# Training SVM with Indefinite Kernels

**Jianhui Chen** JIANHUI.CHEN@ASU.EDU
**Jieping Ye** JIEPING.YE@ASU.EDU
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA

## Abstract

Similarity matrices generated from many applications may not be positive semidefinite, and hence can't fit into the kernel machine framework. In this paper, we study the problem of training support vector machines with an indefinite kernel. We consider a regularized SVM formulation, in which the indefinite kernel matrix is treated as a noisy observation of some unknown positive semidefinite one (proxy kernel) and the support vectors and the proxy kernel can be computed simultaneously. We propose a semi-infinite quadratically constrained linear program formulation for the optimization, which can be solved iteratively to find a global optimum solution. We further propose to employ an additional pruning strategy, which significantly improves the efficiency of the algorithm, while retaining the convergence property of the algorithm. In addition, we show the close relationship between the proposed formulation and multiple kernel learning. Experiments on a collection of benchmark data sets demonstrate the efficiency and effectiveness of the proposed algorithm.

## 1. Introduction

Kernel methods work by embedding the data into a high-dimensional (possibly infinite-dimensional) feature space, where the embedding is defined implicitly through a kernel function. Evaluating the kernel function on all pairs of data points produces a symmetric and positive semidefinite (PSD) kernel matrix. Support Vector Machine (SVM) with a positive semidefinite kernel matrix has been applied successfully in numerous classification tasks including face recognition, image retrieval, and micro-array gene expression data analysis (Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2001; Tong & Chang, 2001). The PSD property of the kernel matrix ensures the existence of a Reproducing Kernel Hilbert Space (RKHS) and results in a convex formulation for SVM. Thus, a global optimal solution exists.

In practice, however, similarity matrices generated from many applications may not be PSD (Qamra et al., 2005; Roth et al., 2003a; Shimodaira et al., 2001). The problem of learning with a non-PSD similarity matrix (indefinite kernel) has been addressed by many researchers (Wu et al., 2005; Haasdonk, 2005; Lin & Lin, 2003). One simple and popular approach is to generate a PSD kernel matrix by transforming the spectrum of the indefinite kernel matrix (Wu et al., 2005). Several representative transformation methods include *denoise* which neglects the negative eigenvalues (Graepel et al., 1998; Pekalska et al., 2002), *flip* which flips the sign of the negative eigenvalues (Graepel et al., 1998), *diffusion* which applies matrix diffusion on the indefinite kernel (Kondor & Lafferty, 2002), and *shift* which shifts all the eigenvalues by a positive constant (Roth et al., 2003b). One common limitation of these approaches is that the transformation may lead to the loss of valuable information in the data.

Several other works use the non-PSD similarity matrix as a kernel, but they change the formulation of SVM. In (Lin & Lin, 2003), an SMO-type method is proposed to find stationary points for the non-convex dual formulation of SVM with a non-PSD sigmoid kernel. However, this method is based on the assumption that a corresponding RKHS still exists such that SVM formulations are valid. Haasdonk (2005) interprets learning with an indefinite kernel as the minimization of distance between two convex hulls in some pseudo-Euclidean (pE) space. However, it assumes that the representer theorem holds in such a pE space. Ong et al. (2004) associate the indefinite kernels with a Reproducing Kernel Kreĭn Space (RKKS), in which a general representer theorem exists and a regularized risk functional can be defined.

Recently, Luss and d'Aspremont (2007) propose a regularized SVM formulation, in which the indefinite kernel matrix is considered as a noisy observation of some unknown PSD one (proxy kernel). One attractive property

Appearing in *Proceedings of the $25^{th}$ International Conference on Machine Learning*, Helsinki, Finland, 2008.

of this formulation is that the support vectors as well as the proxy kernel can be found simultaneously. However, the convex reformulation in (Luss & d'Aspremont, 2007) involves a nondifferentiable objective function. To facilitate the calculation of the gradient, Luss and d'Aspremont (2007) quadratically smoothed the objective function, resulting in two algorithms including the projected gradient method and the analytic center cutting plan method.

In this paper, we study the problem of training SVM with an indefinite kernel matrix following the formulation in (Luss & d'Aspremont, 2007). We show that this problem can be reformulated as a semi-infinite quadratically constrained linear program (SIQCLP), which includes a finite number of optimization variables with an infinite number of constraints. We then propose an iterative algorithm to solve this SIQCLP problem, which consists of two key steps: computing an intermediate SVM solution by solving a quadratically constraint linear program with a restricted subset of constraints, and updating the subset of constraints based on the obtained intermediate SVM solution. We further show the convergence property of the proposed iterative algorithm.

One limitation of the proposed algorithm is that the computational cost for solving the quadratically constraint linear program depends on the number of constraints, which gradually increases during the iteration. We propose to improve the efficiency of the iterative algorithm by pruning inactive constraints at each iteration. We show that such pruning will not affect the convergence property of the algorithm. In addition, we show the close relationship between the proposed SIQCLP formulation and multiple kernel learning (MKL). More specifically, the intermediate quadratically constraint linear program with a restricted subset of constraints is shown to be equivalent to a regularized version of the multiple kernel learning formulation in (Lanckriet et al., 2004). Thus, efficient algorithms for MKL (Lanckriet et al., 2004; Rakotomamonjy et al., 2007; Sonnenburg et al., 2006) can be applied to solve the SIQCLP problem. We have performed experiments on a collection of benchmark data sets. The presented experimental results demonstrate the efficiency and effectiveness of the proposed algorithm.

## 2. Background

Assume $K \in \mathbb{R}^{n\times n}$ is a valid kernel matrix, that is, $K$ is positive semidefinite (PSD). Let $y = [y_1, \cdots, y_n] \in \mathbb{R}^n$ be the vector of class labels, where $y_i \in \{-1, +1\}$. The dual formulation of 1-norm soft margin SVM classification is given by (Schölkopf & Smola, 2001):

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^d} \quad & \alpha^T e - \tfrac{1}{2}\alpha^T YKY\alpha \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \le \alpha \le C, \end{aligned} \qquad (1)$$

where $\alpha$ is the vector of Lagrange dual variables, $Y = \text{diag}(y)$, $C$ is a pre-specified parameter, and $e$ is a vector of all ones of length $n$.

Since $K$ is PSD, the optimization problem in Eq. (1) is a convex Quadratic Program (QP) (Boyd & Vandenberghe, 2004); hence a global optimal solution can be found via standard optimization techniques such as primal-dual interior point methods (Nocedal & Wright, 1999). In practice, however, many similarity matrices may be non-PSD (indefinite kernels), including sigmoid kernels (Vapnik, 1995) for various values of its parameters and hyperbolic tangent kernels (Smola et al., 2000). Additional examples include the protein sequence similarity measures based on Smith-Waterman and BLAST scores.

In (Luss & d'Aspremont, 2007), the indefinite kernel is considered as a noisy observation of some unknown PSD kernel (proxy kernel), and the following max-min optimization problem is proposed for simultaneous proxy kernel learning and SVM classification:

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^d} \min_{K\in\mathbb{R}^{d\times d}} \quad & \alpha^T e - \tfrac{1}{2}\alpha^T YKY\alpha + \rho\|K - K_0\|_F^2 \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \le \alpha \le C,\ K \succeq 0, \end{aligned} \qquad (2)$$

where $K_0$ is a pre-specified indefinite kernel matrix, $K$ is the unknown proxy kernel matrix, and $\rho > 0$ is the pre-specified parameter, and $||\cdot||_F$ denotes the Frobenius norm of a matrix (Golub & Van Loan, 1996).

The objective function in Eq. (2) is convex in $K$ and concave in $\alpha$, thus a global optimal solution exists. However, direct optimization of Eq. (2) in terms of both $\alpha$ and $K$ leads to a complex optimization problem involving nondifferentiable objective function (Luss & d'Aspremont, 2007). To facilitate the calculation of the gradient, Luss and d'Aspremont (2007) quadratically smoothed the objective function. Two algorithms including the projected gradient method and the analytic center cutting plan method are proposed for the proposed formulation.

## 3. Problem Formulation

We propose to solve the optimization problem in Eq. (2) by first reformulating it as a semi-infinite program (SIP) (Hettich & Kortanek, 1993a). The SIP problem refers to optimization problems that maximizes the functional $F(z)$ subject to a system of constraints on $z$, expressed as $g(z,t) \le 0$ for all $t$ in some set $B$. When the objective is linear and the constraints are quadratic, the optimization problem is known as semi-infinite quadratically constrained linear program (SIQCLP).

For notational simplicity, we denote the objective function in Eq. (2) as:

$$S(\alpha, K) = \alpha^T e - \frac{1}{2}\alpha^T YKY\alpha + \rho\|K - K_0\|_F^2. \quad (3)$$

The optimal solution to the max-min problem in Eq. (2) is a saddle-point for the function $S(\alpha, K)$ subject to the constraints in Eq. (2). Let $(\alpha^*, K^*)$ be optimal to Eq. (2). For any feasible $\alpha$ and $K$ in Eq. (2), we have

$$S(\alpha, K^*) \leq S(\alpha^*, K^*) \leq S(\alpha^*, K); \quad (4)$$

moreover, it can be verified that

$$S(\alpha, K^*) = \min_{\tilde{K} \succeq 0} S(\alpha, \tilde{K}) \leq S(\alpha, K). \quad (5)$$

By adding an additional variable $t \in \mathbb{R}$, the max-min optimization problem in Eq. (2) can be reformulated into a SIQCLP problem as follows:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^d,\, t \in \mathbb{R}} \quad & t \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \leq \alpha \leq C \\ & t \leq S(\alpha, K),\ \forall\ K \succeq 0. \end{aligned} \quad (6)$$

The optimization problem in Eq. (6) has two optimization variables ($t$ and $\alpha$) with an infinite number of (quadratic) constraints, i.e., one quadratic constraint $t \leq S(\alpha, K)$ for each kernel matrix $K$. When there is only one (fixed) kernel matrix $K$ involved in Eq. (6), this optimization problem reduces to a standard SVM problem.

## 4. Algorithm

We propose an iterative algorithm to solve Eq. (6), which is guaranteed to converge to a global optimum. The algorithm is closely related to the *bundle method* (Hiriart-Urruty & Lemarechal, 1993; Teo et al., 2007).

The optimization problem in Eq. (6) maximizes its objective function with respect to two variables $t$ and $\alpha$ with an infinite number of (quadratic) constraints. We approach the optimum by optimizing the variables $t$ and $\alpha$ with a restricted subset of the infinite number of constraints, and then updating the constraint subset based on the obtained suboptimal $t$ and $\alpha$ in an iterative manner. It is similar to the strategy presented in (Sonnenburg et al., 2006). The algorithm belongs to a family of algorithms for solving general SIP problems called the *exchange methods*, in which the constraints are exchanged at each iteration. The global optimality property of the final solution after convergence is guaranteed (Hettich & Kortanek, 1993b).

For a restricted subset of constraints, called a *localization* set of kernel matrices $\mathbf{K} = \{K_i\}_{i=1}^p$, the intermediate suboptimal $t$ and $\alpha$ can be computed by solving the following optimization problem:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^d,\, t \in \mathbb{R}} \quad & t \\ \text{subject to} \quad & \alpha^T y = 0,\ 0 \leq \alpha \leq C \\ & t \leq S(\alpha, K_i),\ i = 1, \cdots, p. \end{aligned} \quad (7)$$

This corresponds to a quadratically constrained linear program (QCLP) with $p$ quadratic constraints. The optimization problem is often called the *restricted master problem*, and the obtained suboptimal solution pair $(\alpha, t)$ is called *intermediate solution*. Note that this QCLP problem can be solved efficiently using general optimization solvers.

To approach the optimum of the SIQCLP problem from a given intermediate solution pair $(t, \alpha)$, we find the next constraint with the maximum violation, i.e., the kernel matrix $K$ that minimizes $S(\alpha, K)$. The optimal $K$ can be computed by solving the following minimization problem:

$$\min_K S(\alpha, K) = \min_K \rho\|K - K_0\|_F^2 - \frac{1}{2}\alpha^T YKY\alpha. \quad (8)$$

If the optimal $K^*$ to Eq. (8) satisfies $t \leq S(\alpha, K^*)$, then the current intermediate solution pair $(t, \alpha)$ is optimal for the optimization problem in Eq. (6). Otherwise, $K^*$ is added into the localization set $\mathbf{K}$. The intermediate solution pair $(t, \alpha)$ is updated by solving the restricted master problem based on the updated $\mathbf{K}$. We repeat this iterative process until convergence. The final solution is guaranteed to be globally optimal (Sonnenburg et al., 2006).

It can be shown (Luss & d'Aspremont, 2007) that the optimal $K^*$ to the optimization problem in Eq. (8) for a fixed $\alpha$ is given by:

$$K^* = \left(K_0 + Y\alpha\alpha^T Y/(4\rho)\right)_+. \quad (9)$$

Here $X_+$ refers to the positive part of a symmetric matrix $X$, i.e., $X_+ = \sum_i \max(0, \lambda_i) x_i x_i^T$, where $\lambda_i$ and $x_i$ are the $i$-th eigenvalue and eigenvector of $X$.

Based on the discussions above, we propose an iterative algorithm to solve the optimization in Eq. (6). The pseudocode is presented in Algorithm 1. Note that the algorithm searches for a quadratic constraint (specified by $K^*$) with the maximum violation in step 1, then updates the intermediate solution $(t, \alpha)$, which is repeated iteratively until convergence. When no more constraints with violation can be found, i.e., all constraints are satisfied, Algorithm 1 converges. In practice, we determine the convergence by comparing an upper and lower bound of the objective as described in the next section.

### 4.1. Convergence Analysis

We analyze the convergence property of Algorithm 1. Recall that Algorithm 1 alternates between the updating of the

**Algorithm 1** Proposed Algorithm

```
Input: Indefinite kernel $K_0$ and class label vector $y$;
Output: $\alpha$ and $\mathbf{K}$;
Initialization: $(t_0, \alpha_0) \leftarrow \max_\alpha S(\alpha, (K_0)_+)$;
Initialization: $i \leftarrow 1, \mathbf{K} = \varnothing$;
Do
    Step 1: compute $K^*$ from Eq. (9) and $K_i \leftarrow K^*$;
    if $S(\alpha_{i-1}, K_i) \geq t_{i-1}$ then exit the loop;
        else update localization set $\mathbf{K} \leftarrow \mathbf{K} \cup \{K_i\}$;
    end if
    Step 2: compute $(t_i, \alpha_i)$ by solving Eq. (7);
    $i \leftarrow i + 1$;
until convergence
```

localization set $\mathbf{K}$ (step 1) and the updating of the intermediate solution pair $(t, \alpha)$ (step 2). At the $i$-th iteration, a new kernel matrix $K_i$ is computed from Eq. (9) based on $\alpha_{i-1}$. That is,

$$S(\alpha_{i-1}, K_i) = \min_{K \succeq 0} S(\alpha_{i-1}, K). \tag{10}$$

With the addition of $K_i$, the localization set is updated as $\mathbf{K} = \{K_j\}_{j=1}^{i}$. We denote

$$l_i^- = \max_{j \leq i} S(\alpha_{j-1}, K_j). \tag{11}$$

Let $(t_i, \alpha_i)$ be the solution pair to the optimization problem in Eq. (7) after the $i$-th iteration. Denote $u_i^+ = t_i$. That is,

$$\begin{aligned} \alpha_i &= \arg\max_\alpha \left( \min_{K \in \mathbf{K}} S(\alpha, K) \right), & (12) \\ u_i^+ \equiv t_i &= \min_{K \in \mathbf{K}} S(\alpha_i, K) = \max_\alpha \min_{K \in \mathbf{K}} S(\alpha, K), & (13) \end{aligned}$$

where $\mathbf{K}$ is the updated restricted localization set.

The following theorem shows that Algorithm 1 makes continuous progress towards the optimal solution:

**Theorem 4.1.** *Let $l_i^-$ and $u_i^+$ be defined in Eq. (11) and Eq. (13), respectively. Let $(\alpha^*, t^*)$ be the optimal solution pair to the optimization problem in Eq. (6). Then*

$$u_i^+ \;\geq\; t^* \;\geq\; l_i^-. \tag{14}$$

*Moreover, the sequence $\{u_i^+\}$ is monotonically decreasing, and the sequences $\{l_i^-\}$ is monotonically increasing.*

*Proof.* For any feasible $\alpha$ in Eq. (6), we have

$$\min_{K \in \mathbf{K}} S(\alpha, K) \geq \min_{K \succeq 0} S(\alpha, K). \tag{15}$$

It follows that the inequality above also holds for their corresponding pointwise maximum with respect to $\alpha$. From Eq. (13) and the equality below

$$t^* = \max_{\alpha \in \mathbb{R}^d} \min_{K \succeq 0} S(\alpha, K), \tag{16}$$

we have $u_i^+ \geq t^*$. On the other hand, it follows from Eq. (10) that

$$K_j = \arg\min_{K \succeq 0} S(\alpha_{j-1}, K), \;\; j = 1, \cdots, i. \tag{17}$$

Thus $\{(\alpha_{j-1}, S(\alpha_{j-1}, K_j)), j = 1, \cdots, i\}$ is a set of feasible solution pairs to the optimization problem in Eq. (6). Since $(\alpha^*, t^*)$ is the optimal solution pair to Eq. (6), we have $t^* \geq S(\alpha_{j-1}, K_j)$, for $j = 1, \cdots, i$. If follows from Eq. (11) that $t^* \geq l_i^-$.

From Eq. (13), we have

$$u_i^+ = \max_\alpha \min_{K \in \mathbf{K}} S(\alpha, K). \tag{18}$$

Thus, the sequence $\{u_i^+\}$ is monotonically decreasing, as the size of localization set $\mathbf{K}$ monotonically increases. It follows from Eq. (11) that $\{l_i^-\}$ is monotonically increasing. This completes the proof of the theorem. □

Based on the result in Theorem 4.1, we can use the gap between $u_i^+$ and $l_i^-$ to trace the convergence of Algorithm 1. When this gap is smaller that a pre-specified tolerance, we stop the algorithm.

### 4.2. Pruning Inactive Constraints

In Algorithm 1, a quadratically constraint linear program (QCLP) is involved at each iteration (step 2). The computational cost for solving QCLP grows with the number of quadratic constraints, which increases by one after each iteration. We show that at each iteration, many (inactive) quadratic constraints can be pruned, while retaining the convergence property of the algorithm.

Assume that $(\alpha_i, t_i)$ is the optimal solution pair at the $i$-th iteration with $\mathbf{K}^i = \{K_j\}_{j=1}^p$ as the localization set. We further partition $\mathbf{K}^i$ into two subsets as $\mathbf{K}^i = \mathbf{K}_{act}^i \cup \mathbf{K}_{ina}^i$ such that

$$t_i = S(\alpha_i, K), \;\; \forall \; K \in \mathbf{K}_{act}^i, \tag{19}$$

and

$$t_i < S(\alpha_i, K), \;\; \forall \; K \in \mathbf{K}_{ina}^i, \tag{20}$$

where the equalities in Eq. (19) and inequalities in Eq. (20) are called *active* and *inactive* constraints (Nocedal & Wright, 1999), respectively. Let $K^*$ be the optimal matrix given in Eq. (9) with $\alpha = \alpha_i$. In Algorithm 1, we use $\mathbf{K}^i \cup K^*$ as the new localization set. We propose to improve the efficiency by removing the inactive constraints from the optimization and updating the new localization set $\mathbf{K}^{i+1}$ as $\mathbf{K}^{i+1} = \mathbf{K}_{act}^i \cup K^*$. Let $(\alpha_{i+1}, t_{i+1})$ be the optimal solution pair at the $(i+1)$-th iteration with the updated $\mathbf{K}^{i+1}$ as the localization set. To show the convergence, we need to prove $t_{i+1} \leq t_i$, as summarized below:

**Lemma 4.1.** *Let $t_i$ and $t_{i+1}$ be defined as above. Then $t_{i+1} \leq t_i$.*

*Proof.* Prove by contradiction. Assume that $t_{i+1} \leq t_i$ doesn't hold, i.e., $t_{i+1} > t_i$.

Let $(\tilde{\alpha}, \tilde{t})$ be the optimal solution pair to the optimization problem in Eq. (7) with $\mathbf{K}^i_{act}$ as the localization set. It is clear that $t_{i+1} \leq \tilde{t}$, as $\mathbf{K}^i_{act} \subset \mathbf{K}^{i+1} = \mathbf{K}^i_{act} \cup K^*$. Thus $\tilde{t} \geq t_{i+1} > t_i$.

For any $K \in \mathbf{K}^i_{act}$, we have $S(\tilde{\alpha}, K) \geq \tilde{t}$. Since $\tilde{t} > t_i$, the following holds for any $K \in \mathbf{K}^i_{act}$:

$$S(\tilde{\alpha}, K) \geq \tilde{t} > t_i = S(\alpha_i, K). \tag{21}$$

For any $\eta \in (0, 1)$, let $\beta = \eta\tilde{\alpha} + (1-\eta)\alpha_i$. Since $S(\alpha, K)$ is concave on $\alpha$ and $t_i = S(\alpha_i, K)$ for any $K \in \mathbf{K}^i_{act}$ from Eq. (19), the following holds for any $K \in \mathbf{K}^i_{act}$:

$$S(\beta, K) \geq \eta S(\tilde{\alpha}, K) + (1-\eta)S(\alpha_i, K) > t_i. \tag{22}$$

Recall that for any $K \in \mathbf{K}^i_{ina}$, we have $S(\alpha_i, K) > t_i$. Since $S(\alpha, K)$ is continuous on $\alpha$, and $\mathbf{K}^i_{ina}$ is a finite set, there exists an $\epsilon \in (0, 1)$ sufficiently close to zero such that

$$S(\beta_\epsilon, K) > t_i, \ \ \forall\ K \in \mathbf{K}^i_{ina}, \tag{23}$$

where $\beta_\epsilon = \epsilon\tilde{\alpha} + (1-\epsilon)\alpha_i$.

It follows from Eqs. (22) and (23) that

$$\hat{t} = \min_{K \in \mathbf{K}^i} S(\beta_\epsilon, K) > t_i. \tag{24}$$

Since $(\beta_\epsilon, \hat{t})$ is a feasible solution pair to Eq. (7) with $\mathbf{K}^i$ as the localization set, Eq. (24) contradicts with our assumption that $(\alpha_i, t_i)$ is the optimal solution pair to Eq. (7). This completes the proof of the lemma. □

Lemma 4.1 shows that the upper bound defined in Eq. (13) with the inactive constraints pruned as above decreases monotonically. As the lower bound defined in Eq. (11) always increases monotonically, the proposed pruning strategy retains the convergence property in Theorem 4.1.

## 5. Relationship with Multiple Kernel Learning

We show the close relationship between the proposed SIQCLP formulation in Eq. (6) and the multiple kernel learning formulation in (Lanckriet et al., 2004).

For a given set of kernel matrices $\{K_i\}_{i=1}^p$ and a class label vector $y$, Lanckriet et al. (2004) propose to learn an optimal convex combination of the $p$ pre-specified kernel matrices by solving the following optimization problem:

$$\begin{aligned} \min_{\{\theta_i\}} \max_{\alpha \in \mathbb{R}^d} \quad & \alpha^T e - \frac{1}{2}\alpha^T Y \left(\sum_{i=1}^p \theta_i K_i\right) Y\alpha \\ \text{subject to} \quad & \sum_{i=1}^p \theta_i \ \text{tr}(K_i) = 1, \\ & \alpha^T y = 0, \ \ 0 \leq \alpha \leq C, \end{aligned} \tag{25}$$

where $Y = \text{diag}(y)$, and $C$ is the pre-specified parameter.

Recall that $K_0$ is a indefinite kernel matrix. For each of the given PSD kernel matrix $K_i$, we denote

$$\mu_i = \|K_i - K_0\|_F^2, \ \ i = 1, \cdots, p, \tag{26}$$

where $\mu_i$ measures the distance between $K_i$ and $K_0$ in terms of Frobenius norm. Consider a regularized version of the optimization problem in Eq. (25) given by:

$$\begin{aligned} \min_{\{\theta_i\}} \max_{\alpha \in \mathbb{R}^d} \quad & \alpha^T e - \frac{1}{2}\alpha^T Y \left(\sum_{i=1}^p \theta_i K_i\right) Y\alpha + \rho\sum_{i=1}^p \theta_i\mu_i \\ \text{subject to} \quad & \sum_{i=1}^p \theta_i = 1, \ \ \alpha^T y = 0, \ \ 0 \leq \alpha \leq C, \end{aligned} \tag{27}$$

where $\rho$ is the pre-specified parameter as in Eq. (6). The optimization problem in Eq. (27) computes an optimal linear combination of the $p$ pre-specified kernel matrices by maximizing the margin for SVM classification, while penalizing kernels with a large deviation from $K_0$.

The following theorem shows the equivalence relationship between the regularized MKL problem in Eq. (27) and the SIQCLP formulation in Eq. (7).

**Theorem 5.1.** *Let $\{K_i\}_{i=1}^p$ be a set of pre-specified PSD kernel matrices. Then the optimization problem in Eq. (7) is equivalent to the one in Eq. (27).*

*Proof.* Since all constraints in Eq. (27) are linear and the objective is convex on $\{\theta_i\}$ and concave on $\alpha$, the minimization and the maximization in Eq. (27) can be exchanges. This leads to the following optimization problem:

$$\begin{aligned} & \max_{\alpha \in \mathbb{R}^d} \min_{K \succeq 0} \alpha^T e - \frac{1}{2}\alpha^T YKY\alpha + \rho\sum_{i=1}^p \theta_i\mu_i \\ = \ & \max_{\alpha \in \mathbb{R}^d} \min_{\sum \theta_i = 1} \alpha^T e - \frac{1}{2}\alpha^T Y \sum_{i=1}^p \theta_i K_i Y\alpha + \rho\sum_{i=1}^p \theta_i\mu_i \\ = \ & \max_{\alpha \in \mathbb{R}^d} \min_{\sum \theta_i = 1} \left(\sum_{i=1}^p \theta_i t_i\right), \end{aligned} \tag{28}$$

where $t_i$ is defined as

$$t_i = \left(\alpha^T e - \frac{1}{2}\alpha^T Y K_i Y\alpha + \rho\mu_i\right). \tag{29}$$

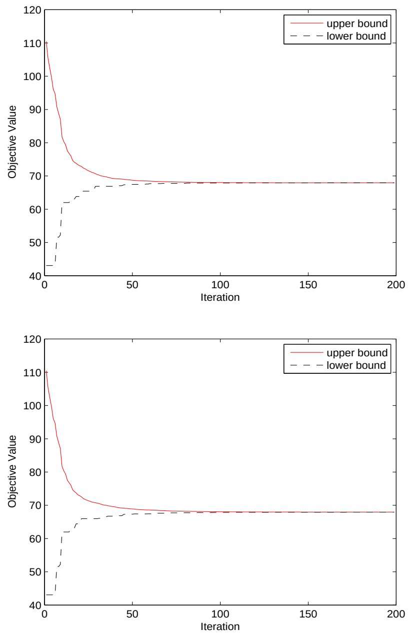

*Figure 1.* Convergence of the proposed algorithms without the pruning strategy applied (top graph) and with the pruning strategy applied (bottom graph).

From Eq. (28) and Eq. (29), the optimization problem in Eq. (27) can be reformulated as:

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^d,\ t\in\mathbb{R}} \quad & t \\ \text{subject to} \quad & \alpha^T y = 0,\ \ 0 \leq \alpha \leq C, \\ & t \leq \alpha^T e - \frac{1}{2}\alpha^T Y K_i Y \alpha + \rho\mu_i, \\ & i = 1, \cdots, p, \end{aligned}$$

which is equivalent to the SIQCLP formulation given in Eq. (7). We complete the proof of this theorem. □

The equivalent result in Theorem 5.1 implies that our proposed SIQCLP formulation in Eq. (6) can be solved by recycling existing efficient MKL implementations (Rakotomamonjy et al., 2007; Sonnenburg et al., 2006).

## 6. Experiments

We experimentally evaluate the convergence property of the proposed algorithms. We also compare the proposed algorithms with other representative ones using a collection of benchmark data sets.

### 6.1. Experimental Setup

We use several benchmark data sets from the UCI repository (Newman et al., 1998) including Sonar, Ionosphere, Breast Cancer, and Diabetes, as well as USPS (Hull, 1994) and Heart[1]. For USPS, we select two classes corresponding to two digits $3$ and $5$, and randomly select $600$ samples for each digit.

In our simulation study, we first generate Gaussian kernels from the data with the parameter value estimated via cross-validation and then construct indefinite kernels through perturbation. More specifically, we randomly generate a matrix $E$ with zero mean and identity covariance matrix, and then apply $\xi\hat{E}$ as the perturbation, where $\hat{E} = (E + E^T)/2$ and $\xi > 0$ is small constant. We set $C = 1$ in SVM. The value of $\rho$ is estimated via cross-validation.

### 6.2. Convergence

In this experiment, we empirically evaluate the convergence property of the proposed algorithms with and without pruning. We also investigate the number of kernel matrices involved when the pruning strategy is employed. We use the sonar data set for this study, and the perturbation matrix is set to be $0.1\hat{E}$.

The results are presented in Figure 1. The top graph in Figure 1 shows the convergence of the upper bound as well as the lower bound of the objective value when the algorithm without the pruning strategy is applied. The bottom graph shows the convergence result for the case when the pruning strategy is applied. We can observe from the figure that the upper bound and lower bound curves approach each other gradually during the iteration. More specifically, the upper bound monotonically decreases, while the lower bound monotonically increases, both approaching the optimal objective value. This is consistent with our convergence results in Section 4.1. Interestingly, our results show that the proposed algorithms with or without the pruning strategy applied result in a similar convergent rate. We further observe that the gap between the upper and lower bound is less that $10^{-2}$ after about $150$ iterations, and it takes about $600$ iterations to attain a gap smaller than $10^{-5}$.

Figure 2 shows the number of kernels (the size of the localization set) involved at each iteration. We can observe from the figure that with the pruning strategy, the number of kernels involved in the algorithm stabilizes around a small constant. In contrast, this number increases gradually when the pruning strategy is not applied. These results demonstrate the advantage of the proposed pruning strategy.

We show in Figure 3 the generalization performance (measured by classification accuracy) of the proposed algorithm with pruning at each of the first 70 iterations. We observe a large variation at the first few iterations, while the accuracy becomes more stable after about 40 iterations. We further run the algorithm until convergence and the resulting accu-

[1] http://www.is.umk.pl/projects/datasets-stat.html#heart

| Data Set | Size | $\lambda^-$ num. | $\lambda^+$ num. | $\lambda_{min}$ | $\lambda_{max}$ | $\lvert\lambda_{max}/\lambda_{min}\rvert$ | Denoise | Flip | Shift | SVM | Indefinite SVM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sonar | 208 | 57.41 | 150.62 | $-1.36$ | 18.42 | 13.55 | 78.57 | 79.52 | 78.10 | 72.86 | 80.95 |
| Ionosphere | 351 | 169.62 | 181.45 | $-25.50$ | 94.49 | 3.71 | 75.57 | 71.43 | 71.41 | 68.00 | 77.43 |
| Breast Cancer | 683 | 323.21 | 359.82 | $-3.51$ | 390.52 | 111.26 | 95.38 | 95.62 | 95.38 | 89.54 | 95.36 |
| Heart | 270 | 125.57 | 144.71 | $-10.96$ | 42.93 | 3.92 | 71.02 | 67.28 | 65.42 | 65.43 | 72.22 |
| USPS-3-5 | 1200 | 520.12 | 680.31 | $-3.54$ | 81.99 | 23.16 | 96.25 | 96.88 | 95.63 | 96.11 | 96.81 |
| Diabetes | 768 | 381.22 | 385.18 | $-3.93$ | 8.13 | 2.06 | 68.83 | 64.28 | 62.98 | 66.23 | 70.08 |

*Table 1.* Comparison of the proposed algorithm with other representative algorithms in terms of classification accuracy (in percentage). All values shown in the table are the averaged ones over 10 partitions of the data into training and test sets with a ratio $4 : 1$. $\lambda^-$ num ($\lambda^+$ num) denotes the number of negative (positive) eigenvalues; $\lambda_{min}$ ($\lambda_{max}$) denotes the minimum (maximum) eigenvalue; $|\lambda_{max}/\lambda_{min}|$ denotes the ratio of the absolute values of $\lambda_{max}$ and $\lambda_{min}$. SVM refers to applying indefinite kernel in the SVM formulation directly, while indefinite SVM refers to our proposed algorithm with the pruning strategy applied.

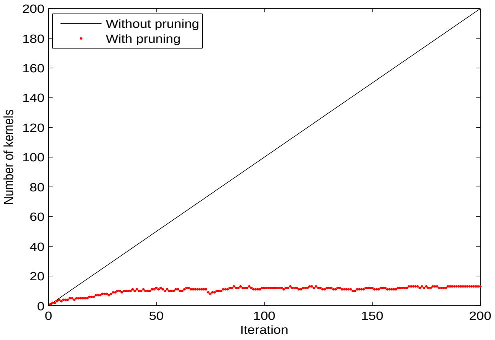

*Figure 2.* The number of kernel matrices involved for the proposed algorithms with and without the pruning.

racy is about 76%. We obtain a similar observation from other data sets. This implies that an early-stopping strategy could be employed for the proposed algorithm.

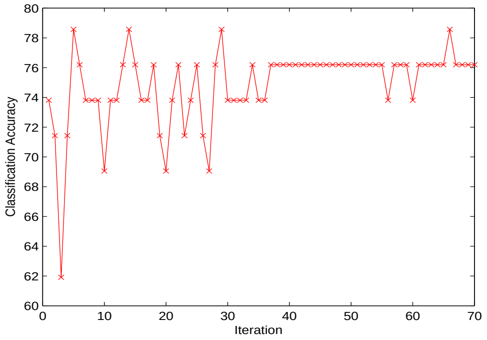

*Figure 3.* Generalization performance (measured by classification accuracy in percentage) of the proposed algorithm with pruning.

### 6.3. Classification Performance

In this experiment, we compare our proposed algorithms (Indefinite SVM) with other representative ones including Denoise, Flip, Shift, and SVM using indefinite kernels in terms of classification accuracy. The presented experimental results are averaged over 10 random partitions of the data into a training and a test set with a ratio $4 : 1$.

The experimental results are summarized in Table 1. We also report the maximum and minimum eigenvalues of the indefinite kernel matrix in the table. We can observe from the table that Indefinite SVM is competitive with all other algorithms in most cases. It outperforms all other algorithms on the Sonar, Ionosphere, Heart, and Diabetes data sets, where the perturbed kernel matrix has a relatively small ratio $|\lambda_{max}/\lambda_{min}|$. For the other two data sets including Breast Cancer and USPS-3-5, where the perturbed kernel matrix has a relatively large ratio $|\lambda_{max}/\lambda_{min}|$, Indefinite SVM is comparable to the best among all other algorithms. These results demonstrate the effectiveness of the proposed learning algorithm, especially when the indefinite kernel matrix is highly non-PSD. A similar trend has been observed in (Luss & d'Aspremont, 2007).

## 7. Conclusion

In this paper, we study the problem of training SVM with an indefinite kernel matrix following the formulation in (Luss & d'Aspremont, 2007). We propose a semi-infinite quadratically constrained linear program formulation, which can be solved iteratively. The algorithm alternates between the computation of an intermediate SVM solution by solving a quadratically constraint linear program with a subset of constraints, and the computation of the new constraint set based on the obtained intermediate SVM solution. We further propose to improve the efficiency of the iterative algorithm by pruning inactive constraints at each iteration. We show that such pruning will not affect the convergence property of the algorithm. In addition, we show the close relationship between the proposed SIQCLP formulation and multiple kernel learning. The presented analysis provides new insights into the nature of this learning formulation.

We have performed a simulation study using a collection of benchmark data sets. Our results verify the convergence property of the proposed algorithms. Our empirical results show that the proposed algorithms with or without the pruning strategy applied result in a similar convergent rate, while a much smaller number of kernel matrices are involved when the pruning strategy is applied. Our results also demonstrate the favorable performance of the proposed algorithms in terms of classification accuracy in comparison with several other representative algorithms. Our future works include the analysis of the convergence rate of the proposed algorithms similar to the analysis conducted in (Teo et al., 2007), the estimation of the regularization parameter $\rho$, and the application of the proposed algorithms to real-world applications involving indefinite kernels such as protein sequence and structure analysis based on various sequence/structure alignment measures.

## Acknowledgments

This research is sponsored in part by funds from the Arizona State University and the National Science Foundation under Grant No. IIS-0612069.

## References

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations*. Johns Hopkins University Press. 3 edition.

Graepel, T., Herbrich, R., Bollmann-Sdorra, P., & Obermayer, K. (1998). Classification on pairwise proximity data. *NIPS* (pp. 438–444).

Haasdonk, B. (2005). Feature space interpretation of svms with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 482–492.

Hettich, R., & Kortanek, K. O. (1993a). Semi-infinite programming: theory, methods, and applications. *SIAM Review*, *35*, 380–429.

Hettich, R., & Kortanek, K. O. (1993b). Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, *35*, 380–429.

Hiriart-Urruty, J.-B., & Lemarechal, C. (1993). *Convex analysis and minimization algorithms II*. Springer.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*, 550–554.

Kondor, R. I., & Lafferty, J. D. (2002). Diffusion kernels on graphs and other discrete input spaces. *ICML* (pp. 315–322).

Lanckriet, G. R. G., Cristianini, N., Bartlett, P. L., Ghaoui, L. E., & Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, *5*, 27–72.

Lin, H.-T., & Lin, C.-J. (2003). A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. *Technical Report, National Taiwan University*.

Luss, R., & d'Aspremont, A. (2007). Support vector machine classification with indefinite kernels. *NIPS*.

Newman, D., Hettich, S., Blake, C., & Merz, C. (1998). UCI repository of machine learning databases.

Nocedal, J., & Wright, S. J. (1999). *Numerical optimization springer series in operations research*. Springer.

Ong, C. S., Mary, X., Canu, S., & Smola, A. J. (2004). Learning with non-positive kernels. *ICML*.

Pekalska, E., Paclik, P., & Duin, R. P. W. (2002). A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, *2*, 175–211.

Qamra, A., Meng, Y., & Chang, E. Y. (2005). Enhanced perceptual distance functions and indexing for image replica recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 379–391.

Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2007). More efficiency in multiple kernel learning. *ICML*.

Roth, V., Laub, J., Kawanabe, M., & Buhmann, J. M. (2003a). Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 1540–1551.

Roth, V., Laub, J., Kawanabe, M., & Buhmann, J. M. (2003b). Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 1540–1551.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. The MIT Press.

Shimodaira, H., Noma, K.-I., Nakai, M., & Sagayama, S. (2001). Dynamic time-alignment kernel in support vector machine. *NIPS* (pp. 921–928).

Smola, A. J., Óvári, Z. L., & Williamson, R. C. (2000). Regularization with dot-product kernels. *NIPS* (pp. 308–314).

Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, *7*, 1531–1565.

Teo, C. H., Smola, A., Vishwanathan, S. V., & Le, Q. V. (2007). A scalable modular convex solver for regularized risk minimization. *KDD* (pp. 727–736).

Tong, S., & Chang, E. (2001). Support vector machine active learning for image retrieval. *ACM Multimedia* (pp. 107–118).

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Wu, G., Zhang, Z., & Chang, E. Y. (2005). An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. *Technical Report, UCSB*.