A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods

Hsuan-Tien Lin and Chih-Jen Lin

Department of Computer Science and

Information Engineering

National Taiwan University

Taipei 106, Taiwan

cjlin@csie.ntu.edu.tw

Abstract

The sigmoid kernel was quite popular for support vector machines due to its origin from neural networks. Although it is known that the kernel matrix may not be positive semi-definite (PSD), other properties are not fully studied. In this paper, we discuss such non-PSD kernels through the viewpoint of separability. Results help to validate the possible use of non-PSD kernels. One example shows that the sigmoid kernel matrix is conditionally positive definite (CPD) in certain parameters and thus are valid kernels there. However, we also explain that the sigmoid kernel is not better than the RBF kernel in general. Experiments are given to illustrate our analysis. Finally, we discuss how to solve the non-convex dual problems by SMO-type decomposition methods. Suitable modifications for any symmetric non-PSD kernel matrices are proposed with convergence proofs.

Keywords

Sigmoid Kernel, non-Positive Semi-Definite Kernel, Sequential Minimal Optimization, Support Vector Machine

1 Introduction

Given training vectors $x_i \in R^n, i = 1, \ldots, l$ in two classes, labeled by the vector $y \in \{+1, -1\}^l$. The support vector machine (SVM) (Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995) separates the training vectors in a ϕ-mapped (and possibly infinite dimensional) space, with an error cost $C > 0$:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{1}$$

Due to the high dimensionality of the vector variable w, we usually solve (1) through its Lagrangian dual problem:

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

where $Q_{ij} \equiv y_i y_j \phi(x_i)^T \phi(x_j)$ and e is the vector of all ones. Here,

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j) \tag{3}$$

is called the kernel function where some popular ones are, for example, the polynomial kernel $K(x_i, x_j) = (a x_i^T x_j + r)^d$, and the RBF (Gaussian) kernel $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$. By the definition (3), the matrix Q is symmetric and positive semi-definite (PSD). After (2) is solved, $w = \sum_{i=1}^{l} y_i \alpha_i \phi(x_i)$ so the decision function for any test vector x is

$$\text{sgn}(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b), \tag{4}$$

where b is calculated through the primal-dual relationship.

In practice, some non-PSD matrices are used in (2). An important one is the sigmoid kernel $K(x_i, x_j) = \tanh(a x_i^T x_j + r)$, which is related to neural networks. It was first pointed out in (Vapnik 1995) that its kernel matrix might not be PSD for certain

values of the parameters a and r. More discussions are in, for instance, (Burges 1998; Schölkopf and Smola 2002). When K is not PSD, (3) cannot be satisfied and the primal-dual relationship between (1) and (2) does not exist. Thus, it is unclear what kind of classification problems we are solving. Surprisingly, the sigmoid kernel has been used in several practical cases. Some explanations are in (Schölkopf 1997).

Recently, quite a few kernels specific to different applications are proposed. However, similar to the sigmoid kernel, some of them are not PSD either (e.g. kernel jittering in (DeCoste and Schölkopf 2002) and tangent distance kernels in (Haasdonk and Keysers 2002)). Thus, it is essential to analyze such non-PSD kernels. In Section 2, we discuss them by considering the separability of training data. Then in Section 3, we explain the practical viability of the sigmoid kernel by showing that for parameters in certain ranges, it is conditionally positive definite (CPD). We discuss in Section 4 about the similarity between the sigmoid kernel and the RBF kernel, which shows that the sigmoid kernel is less preferable. Section 5 presents experiments showing that the linear constraint $y^T\alpha = 0$ in the dual problem is essential for a CPD kernel matrix to work for SVM.

In addition to unknown behavior, non-PSD kernels also cause difficulties in solving (2). The original decomposition method for solving (2) was designed when Q is PSD and and existing software may have difficulties such as endless loops when using non-PSD kernels. In Section 6, we propose simple modifications for SMO-type decomposition methods which guarantee the convergence to stationary points for non-PSD kernels. Section 7 then discusses some modifications to convex formulas. A comparison between SVM and kernel logistic regression (KLR) is performed. Finally, some discussions are in Section 8.

2 The Separability when Using non-PSD Kernel Matrices

When using non-PSD kernels such as the sigmoid, $K(x_i, x_j)$ cannot be separated as the inner product form in (3). Thus, (1) is not well-defined. After obtaining α from (2), it is not clear how the training data are classified. To analyze what we actually obtained

when using a non-PSD Q, we consider a new problem:

$$\begin{aligned} \min_{\alpha,b,\xi} \quad & \frac{1}{2}\alpha^T Q\alpha + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & Q\alpha + by \geq e - \xi, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{5}$$

It is from substituting $w = \sum_{i=1}^{l} y_i\alpha_i\phi(x_i)$ into (1) so that $w^T w = \alpha^T Q\alpha$ and $y_i w^T\phi(x_i) = (Q\alpha)_i$. Note that in (5), α_i may be negative. This problem was used in (Osuna and Girosi 1998) and some subsequent work. In (Lin and Lin 2003), it shows that if Q is symmetric PSD, the optimal solution α of the dual problem (2) is also optimal for (5). However, the opposite may not be true unless Q is symmetric positive definite (PD).

From now on, we assume that Q (or K) is symmetric but may not be PSD. The next theorem is about the stationary points of (2), that is, the points that satisfy the Karash-Kunh-Tucker (KKT) condition. By this condition, we can get a relation between (2) and (5).

Theorem 1 *Any stationary point $\hat{\alpha}$ of (2) is a feasible point of (5).*

Proof.

Assume that $\hat{\alpha}$ is a stationary point, so it satisfies the KKT condition. For a symmetric Q, the KKT condition of (2) is that there are scalar p, and non-negative vectors λ and μ such that

$$\begin{aligned} Q\hat{\alpha} - e - \mu + \lambda - py = 0, \\ \mu_i \geq 0, \mu_i\hat{\alpha}_i = 0, \\ \lambda_i \geq 0, \lambda_i(C - \hat{\alpha}_i) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

If we consider $\alpha_i = \hat{\alpha}_i$, $b = -p$, and $\xi_i = \lambda_i$, then $\mu_i \geq 0$ implies that $(\hat{\alpha}, -p, \lambda)$ is feasible for (5). □

An immediate implication is that if $\hat{\alpha}$, a stationary point of (2), does not have many nonzero components, the training error would not be large. Thus, even if Q is not PSD, it is still possible that the training error is small. Next, we give a more formal analysis on the separability of training data:

Theorem 2 *Consider the problem (2) without C:*

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ \text{subject to} \quad & 0 \le \alpha_i, i = 1, \ldots, l, \\ & y^T\alpha = 0. \end{aligned} \tag{6}$$

If there exists a attained stationary point $\hat{\alpha}$, then

1. *(5) has a feasible solution with $\xi_i = 0$, for $i = 1, \ldots, l$.*
2. *If C is large enough, then $\hat{\alpha}$ is also a stationary point of (2).*

The proof is directly from Theorem 1, which shows that $\hat{\alpha}$ is feasible for (5) with $\xi_i = 0$. The second property comes from the fact that when $C \ge \max_i \hat{\alpha}_i$, $\hat{\alpha}$ is also stationary for (2).

Thus, if (6) has at least one stationary point, the kernel matrix has the ability to fully separate the training data. This theorem gives an explanation why sometimes non-PSD kernels work. Furthermore, if a global minimum $\hat{\alpha}$ of (6) is attained, it can be the stationary point to have the separability. On the contrary, if $\hat{\alpha}$ is not attained and the optimal objective value goes to $-\infty$, for every C, the global minimum $\hat{\alpha}$ of (2) would have at least one $\hat{\alpha}_i = C$. In this case, the separability of the kernel matrix is not clear.

Next we would like to see if any conditions on a kernel matrix imply an attained global minimum and hence the optimal objective value is not $-\infty$. Several earlier work have given useful results. In particular, it has been shown that a conditionally PSD (CPSD) kernel is good enough for SVM. A matrix K is CPSD (CPD) if for all $v \neq 0$ with $\sum_{i=1}^{l} v_i = 0$, $v^T K v \ge 0$ (> 0). Note that some earlier work use different names: conditionally PD (strictly PD) for the case of ≥ 0 (> 0). More properties can be seen in, for example, (Berg, Christensen, and Ressel 1984). Then, the use of a CPSD kernel is equivalent to the use of a PSD one as $y^T\alpha = 0$ in (2) plays a similar role of $\sum_{i=1}^{l} v_i = 0$ in the definition of CPSD (Schölkopf 2000). For easier analysis here, we will work only on the kernel matrices but not the kernel functions. The following theorem gives properties which imply the existence of optimal solutions of (6).

Theorem 3

1. *A kernel matrix K is CPD if and only if there is Δ such that $K + \Delta ee^T$ is PD.*

2. *If K is CPD, then the solution of (6) is attained and its optimal objective value is greater than $-\infty$.*

Proof.

The "if" part of the first result is very simple by definition. For any $v \neq 0$ with $e^T v = 0$,

$$v^T K v = v^T (K + \Delta e e^T) v > 0,$$

so K is CPD.

On the other hand, if K is CPD but there is no Δ such that $K + \Delta e e^T$ is PD, there are infinite $\{v_i, \Delta_i\}$ with $\|v_i\| = 1, \forall i$ and $\Delta_i \to \infty$ as $i \to \infty$ such that

$$v_i^T (K + \Delta_i e e^T) v_i \leq 0, \forall i. \tag{7}$$

As $\{v_i\}$ is in a compact region, there is a subsequence $\{v_i\}, i \in \mathcal{K}$ which converges to v^*. Since $v_i^T K v_i \to (v^*)^T K v^*$ and $e^T v_i \to e^T v^*$,

$$\lim_{i \to \infty, i \in \mathcal{K}} \frac{v_i^T (K + \Delta_i e e^T) v_i}{\Delta_i} = (e^T v^*)^2 \leq 0.$$

Therefore, $e^T v^* = 0$. By the CPD of K, $(v^*)^T K v^* > 0$ so

$$v_i^T (K + \Delta_i e e^T) v_i > 0 \text{ after } i \text{ is large enough,}$$

a situation which contradicts (7).

For the second result of this theorem, if K is CPD, we have shown that $K + \Delta e e^T$ is PD. Hence, (6) is equivalent to

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T (Q + \Delta y y^T) \alpha - e^T \alpha \\ \text{subject to} \quad & 0 \leq \alpha_i, i = 1, \ldots, l, \\ & y^T \alpha = 0, \end{aligned} \tag{8}$$

which is a strict convex programming problem. Hence, (8) attains a unique global minimum and so does (6). □

Unfortunately, the property that (6) has a finite objective value is not equivalent to the CPD of a matrix. The main reason is that (6) has additional constraints $\alpha_i \geq 0, i =$

$1, \ldots, l$. We illustrate this by a simple example: If

$$K = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & -1 \\ -1 & -1 & 0 \end{bmatrix} \text{ and } y = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix},$$

we can get that

$$\begin{aligned} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha &= \frac{1}{2}[3(\alpha_1 - \frac{2}{3})^2 + 3(\alpha_2 - \frac{2}{3})^2 + 8\alpha_1\alpha_2 - \frac{8}{3}] \\ &\geq -\frac{4}{3} \text{ if } \alpha_1 \geq 0 \text{ and } \alpha_2 \geq 0. \end{aligned}$$

However, K is not CPD as we can easily set $\alpha_1 = -\alpha_2 = 1, \alpha_3 = 0$ which satisfy $e^T\alpha = 0$ but $\alpha^T K\alpha = -2 < 0$.

Moreover, the first result of the above theorem may not hold if K is only CPSD. For example, $K = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ is CPSD as for any $\alpha_1 + \alpha_2 = 0$, $\alpha^T K\alpha = 0$. However, for any $\Delta \neq 0$, $K + \Delta ee^T$ has an eigenvalue $\Delta - \sqrt{\Delta^2 + 1} < 0$. Therefore, there is no Δ such that $K + \Delta ee^T$ is PSD. On the other hand, even though $K + \Delta ee^T$ PSD implies its CPSD, they both may not guarantee the optimal objective value of (6) is finite. For example, if $K = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, it satisfies both properties but the objective value of (6) can be $-\infty$.

Next we use concepts given in this section to analyze the sigmoid kernel.

3 The Behavior of the Sigmoid Kernel

In this section, we consider the sigmoid kernel $K(x_i, x_j) = \tanh(ax_i^T x_j + r)$, which takes two parameters: a and r. For $a > 0$, we can view a as a scaling parameter of the input data, and r as a shifting parameter that controls the threshold of mapping. For $a < 0$, the dot-product of the input data is not only scaled but reversed. In the following table we summarize the behavior in different parameter combinations, which will be discussed in the rest of this section. It concludes that the first case, $a > 0$ and $r < 0$, is more suitable for the sigmoid kernel.

a	r	results
$+$	$-$	K is CPD after r is small; similar to RBF for small a
$+$	$+$	in general not as good as the $(+,-)$ case
$-$	$+$	objective value of (6) $-\infty$ after r large enough
$-$	$-$	easily the objective value of (6) $-\infty$

Case 1: $a > 0$ and $r < 0$

We analyze the limiting case of this region and show that when r is small enough, the matrix K is CPD. We begin with a lemma about the sigmoid function:

Lemma 1 *Given any δ,*

$$\lim_{x \to -\infty} \frac{1 + \tanh(x + \delta)}{1 + \tanh(x)} = e^{2\delta}.$$

The proof is by a direct calculation from the definition of the sigmoid function. With this lemma, we can prove that the sigmoid kernel matrices are CPD when r is small enough:

Theorem 4 *Given any training set, if $x_i \neq x_j$, for $i \neq j$ and $a > 0$, there exists $\hat{r}$ such that for all $r \leq \hat{r}$, $K + ee^T$ is PD.*

Proof.

Let $H^r \equiv (K + ee^T)/(1 + \tanh(r))$, where $K_{ij} = \tanh(ax_i^T x_j + r)$. From Lemma 1,

$$\lim_{r \to -\infty} H^r_{ij} = \lim_{r \to -\infty} \frac{1 + \tanh(ax_i^T x_j + r)}{1 + \tanh(r)} = e^{2ax_i^T x_j}.$$

Let $\bar{H} = \lim_{r \to -\infty} H^r$. Thus, $\bar{H}_{ij} = e^{2ax_i^T x_j} = e^{a\|x_i\|^2} e^{-a\|x_i - x_j\|^2} e^{a\|x_j\|^2}$. If written in matrix products, the first and last terms would form the same diagonal matrices with positive elements. And the middle one is in the form of an RBF kernel matrix. From (Micchelli 1986), if $x_i \neq x_j$, for $i \neq j$, the RBF kernel matrix is PD. Therefore, $\bar{H}$ is PD.

If H^r is not PD after r is small enough, there is an infinite sequence $\{r_i\}$ with $\lim_{i \to \infty} r_i = -\infty$ and $H^{r_i}, \forall i$ are not PD. Thus, for each r_i, there exists $\|v_i\| = 1$ such that $v_i^T H^{r_i} v_i \leq 0$.

Since v_i is an infinite sequence in a compact region, there is a subsequence which converges to $\bar{v} \neq 0$. Therefore, $\bar{v}^T \bar{H} \bar{v} \leq 0$, which contradicts the fact that $\bar{H}$ is PD. Thus, there is $\hat{r}$ such that for all $r \leq \hat{r}$, H^r is PD. By the definition of H^r, $K + ee^T$ is PD as well. □

With Theorems 3 and 4, K is CPD after r is small enough. Theorem 4 also provides a connection between the sigmoid and a special PD kernel related to the RBF kernel when a is fixed and r gets small enough. More discussions are in Section 4.

Case 2: $a > 0$ and $r \geq 0$

It was stated in (Burges 1999) that if $\tanh(ax_i^T x_j + r)$ is PD, then $r \geq 0$ and $a \geq 0$. However, the inverse does not hold so the practical viability is not clear. Here we will discuss this case by checking the separability of training data.

Comparing to Case 1, we show that it is more possible that the objective value of (6) goes to $-\infty$. Therefore, with experiments in Section 4, we conclude that in general using $a > 0$ and $r \geq 0$ is not as good as $a > 0$ and $r < 0$. The following theorem discusses possible situations that (6) has the objective value $-\infty$:

Theorem 5

1. *If there are i and j such that $y_i \neq y_j$ and $K_{ii} + K_{jj} - 2K_{ij} \leq 0$, (6) has the optimal objective value $-\infty$.*

2. *For the sigmoid kernel, if*

$$\max_i (a\|x_i\|^2 + r) \leq 0, \tag{9}$$

then $K_{ii} + K_{jj} - 2K_{ij} > 0$ for any $x_i \neq x_j$.

Proof.

For the first result, let $\alpha_i = \alpha_j = \Delta$ and $\alpha_k = 0$ for $k \neq i, j$. Then, the objective value of (6) is $\Delta^2(K_{ii} - 2K_{ij} + K_{jj}) - 2\Delta$. Thus, $\Delta \to \infty$ leads to a feasible solution of (6) with objective value $-\infty$.

For the second result, now

$$\begin{aligned} & K_{ii} - 2K_{ij} + K_{jj} \\ = \; & \tanh(a\|x_i\|^2 + r) - 2\tanh(a\|x_i^T x_j\| + r) + \tanh(a\|x_j\|^2 + r). \end{aligned} \tag{10}$$

Since $\max_i (a\|x_i\|^2 + r) \leq 0$, by the monotonicity of $\tanh(x)$ and its strict convexity when $x \leq 0$,

$$\begin{aligned} & \frac{\tanh(a\|x_i\|^2 + r) + \tanh(a\|x_j\|^2 + r)}{2} \\ \geq \; & \tanh(\frac{(a\|x_i\|^2 + r) + (a\|x_j\|^2 + r)}{2}) \qquad (11) \\ = \; & \tanh(a\frac{\|x_i\|^2 + \|x_j\|^2}{2} + r) \\ > \; & \tanh(ax_i^T x_j + r). \qquad (12) \end{aligned}$$

Note that the last inequality uses the property that $x_i \neq x_j$.

Then, by (10) and (12), $K_{ii} - 2K_{ij} + K_{jj} > 0$, so the proof is complete. □

The requirement that $x_i \neq x_j$ is in general true if there are no duplicated training instances. Apparently, (9) must happen (for $a > 0$) when r is negative. If (9) is wrong, it is possible that $a\|x_i\|^2 + r \geq 0$ and $a\|x_j\|^2 + r \geq 0$. Then due to the concavity of $\tanh(x)$ at the positive side, "$\geq$" in (11) is changed to "$\leq$." Thus, $K_{ii} - 2K_{ij} + K_{jj}$ may be ≤ 0 and (6) has the optimal objective value $-\infty$.

Case 3: $a < 0$ and $r > 0$

The following theorem tells us that $a < 0$ and large $r > 0$ may not be a good choice.

Theorem 6 *For any given training set, if $a < 0$ and each class has at least one data point, there exists $\bar{r} > 0$ such that for all $r \geq \bar{r}$, (6) has optimal objective value $-\infty$.*

Proof.

Since $K_{ij} = \tanh(ax_i^T x_j + r) = -\tanh(-ax_i^T x_j - r)$, by Theorem 4, there is $-\bar{r} < 0$ such that for all $-r \leq -\bar{r}$, $-K + ee^T$ is PD. That is, there exist $\bar{r} > 0$ such that for all $r \geq \bar{r}$, any α with $y^T\alpha = 0$ and $\alpha \neq 0$ satisfies $\alpha^T Q\alpha < 0$.

Since there is at least one data point in each class, we can find $y_i = +1$ and $y_j = -1$. Let $\alpha_i = \alpha_j = \Delta$, and $\alpha_k = 0$ for $k \neq i, j$ be a feasible solution of (6). The objective value decreases to $-\infty$ as $\Delta \to \infty$. Therefore, for all $r \geq \bar{r}$, (6) has optimal objective value $-\infty$. □

Case 4: $a < 0$ and $r \leq 0$

The following theorem shows that, in this case, the optimal objective value of (6) easily goes to $-\infty$:

Theorem 7 *For any given training set, if $a < 0$, $r \leq 0$, and there are x_i, x_j such that*

$$x_i^T x_j \leq \min(\|x_i\|^2, \|x_j\|^2)$$

and $y_i \neq y_j$, (6) has optimal objective value $-\infty$.

Proof.

By $x_i^T x_j \le \min(\|x_i\|^2, \|x_j\|^2)$, (10), and the monotonicity of $\tanh(x)$, we can get $K_{ii} - 2K_{ij} + K_{jj} \le 0$. Then the proof follows from Theorem 5. □

Note that the situation $x_i^T x_j < \min(\|x_i\|^2, \|x_j\|^2)$ and $y_i \ne y_j$ easily happens if the two classes of training data are not close in the input space. Thus, $a < 0$ and $r \le 0$ are generally not a good choice of parameters.

4 Relation with the RBF Kernel

In this section we extend Case 1 (i.e. $a > 0$, $r < 0$) in Section 3 to show that the sigmoid kernel behaves like the RBF kernel when (a, r) are in a certain range.

Lemma 1 implies that when $r < 0$ is small enough,

$$1 + \tanh(ax_i^T x_j + r) \approx (1 + \tanh(r))(e^{2ax_i^T x_j}). \tag{13}$$

If we further make a close to 0, $e^{a\|x\|^2} \approx 1$ so

$$e^{2ax_i^T x_j} = e^{a\|x_i\|^2} e^{-a\|x_i - x_j\|^2} e^{a\|x_j\|^2} \approx e^{-a\|x_i - x_j\|^2}.$$

Therefore, when $r < 0$ is small enough and a is close to 0,

$$1 + \tanh(ax_i^T x_j + r) \approx (1 + \tanh(r))(e^{-a\|x_i - x_j\|^2}), \tag{14}$$

a form of the RBF kernel.

However, the closeness of kernel elements does not directly imply similar generalization performance. Hence, we need to show that they have nearly the same decision functions. Note that the objective function of (2) is the same as:

$$\begin{aligned} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha &= \frac{1}{2}\alpha^T(Q + yy^T)\alpha - e^T\alpha \qquad (15) \\ &= \frac{1}{1 + \tanh(r)}\left(\frac{1}{2}\tilde{\alpha}^T \frac{Q + yy^T}{1 + \tanh(r)}\tilde{\alpha} - e^T\tilde{\alpha}\right), \end{aligned}$$

where $\tilde{\alpha} \equiv (1+\tanh(r))\alpha$ and (15) follows from the equality constraint in (2). Multiplying the objective function of (2) by $(1 + \tanh(r))$, and setting $\tilde{C} = (1 + \tanh(r))C$, solving

(2) is the same as solving

$$\begin{aligned} \min_{\tilde{\alpha}} \quad & F_r(\tilde{\alpha}) = \frac{1}{2}\tilde{\alpha}^T \frac{Q + yy^T}{1 + \tanh(r)}\tilde{\alpha} - e^T\tilde{\alpha} \\ \text{subject to} \quad & 0 \le \tilde{\alpha}_i \le \tilde{C}, i = 1, \ldots, l, \\ & y^T\tilde{\alpha} = 0. \end{aligned} \tag{16}$$

Given a fixed $\tilde{C}$, as $r \to -\infty$, since $(Q + yy^T)_{ij} = y_i y_j (K_{ij} + 1)$, the problem approaches

$$\begin{aligned} \min_{\tilde{\alpha}} \quad & F_T(\tilde{\alpha}) = \frac{1}{2}\tilde{\alpha}^T \bar{Q}\tilde{\alpha} - e^T\tilde{\alpha} \\ \text{subject to} \quad & 0 \le \tilde{\alpha}_i \le \tilde{C}, i = 1, \ldots, l, \\ & y^T\tilde{\alpha} = 0, \end{aligned} \tag{17}$$

where $\bar{Q}_{ij} = y_i y_j e^{2a x_i^T x_j}$ is a PD kernel matrix when $x_i \neq x_j$ for all $i \neq j$. Then, we can prove the following theorem:

Theorem 8 *Given fixed a and $\tilde{C}$, assume that $x_i \neq x_j$ for all $i \neq j$, and the optimal b of the decision function from (17) is unique. Then for any data point x,*

$$\begin{aligned} & \lim_{r \to -\infty} && \text{decision value at } x \text{ using the sigmoid kernel in (2)} \\ = \quad & && \text{decision value at } x \text{ using (17).} \end{aligned}$$

We leave the proof in Appendix A. Theorem 8 tells us that when $r < 0$ is small enough, the separating hyperplanes of (2) and (17) are almost the same. Similar cross-validation (CV) accuracy will be shown in the later experiments.

(Keerthi and Lin 2003, Theorem 2) shows that when $a \to 0$, for any given $\bar{C}$, the decision value by the SVM using the RBF kernel $e^{-a\|x_i - x_j\|^2}$ with the error cost $\frac{\bar{C}}{2a}$ approaches the decision value of the following linear SVM:

$$\begin{aligned} \min_{\bar{\alpha}} \quad & \frac{1}{2}\sum_i \sum_j \bar{\alpha}_i \bar{\alpha}_j y_i y_j x_i^T x_j - \sum_i \bar{\alpha}_i \\ \text{subject to} \quad & 0 \le \bar{\alpha}_i \le \bar{C}, i = 1, \ldots, l, \\ & y^T\bar{\alpha} = 0. \end{aligned} \tag{18}$$

The same result can be proved for the SVM in (17). Therefore, under the assumption

that the optimal b of the decision function from (18) is unique, for any data point x,

$$\begin{array}{rll} & \lim\limits_{a\to 0} & \text{decision value at } x \text{ using the RBF kernel with } \tilde{C} = \frac{\bar{C}}{2a} \\ = & & \text{decision value at } x \text{ using (18) with } \bar{C} \\ = & \lim\limits_{a\to 0} & \text{decision value at } x \text{ using (17) with } \tilde{C} = \frac{\bar{C}}{2a}. \end{array}$$

Then we can get the similarity between the sigmoid and the RBF kernels as follows:

Theorem 9 *Given a fixed $\bar{C}$, assume that $x_i \neq x_j$ for all $i \neq j$, and each of* (17) *after a is close to* 0 *and* (18) *has a unique b. Then for any data point x,*

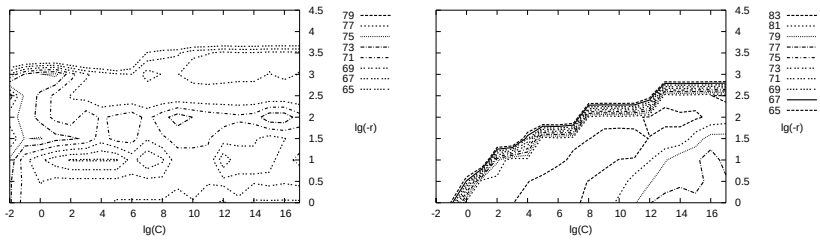
$$\begin{array}{rll} & \lim\limits_{a\to 0} \lim\limits_{r\to-\infty} & \text{decision value at } x \text{ using the sigmoid kernel with } C = \frac{\tilde{C}}{1+\tanh(r)} \\ = & \lim\limits_{a\to 0} & \text{decision value at } x \text{ using (17) with } \tilde{C} = \frac{\bar{C}}{2a} \\ = & \lim\limits_{a\to 0} & \text{decision value at } x \text{ using the RBF kernel with } \tilde{C} = \frac{\bar{C}}{2a} \\ = & & \text{decision value at } x \text{ using the linear kernel with } \bar{C}. \end{array}$$

We can observe the result of Theorems 8 and 9 from Figure 1. The contours show five-fold cross-validation accuracy of the data set heart in different r and C. The contours with $a = 1$ are on the left-hand side, while those with $a = 0.01$ are on the right-hand side. Other parameters considered here are $\log_2 C$ from -2 to 13, with grid space 1, and $\log_2(-r)$ from 0 to 4.5, with grid space 0.5. Detailed description of the data set will be given later in Section 7.

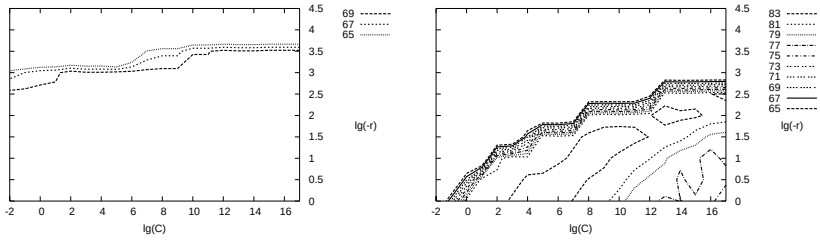
From both sides of Figure 1, we can see that the middle contour (using (17)) is similar to the top one (using tanh) when r gets small. This verifies our approximation in (13) as well as Theorem 8. However, on the left-hand side, since a is not small enough, the data-dependent scaling term $e^{a\|x_i\|^2}$ between (13) and (14) is large and causes a difference between the middle and bottom contours. When a is reduced to 0.01 on the right-hand side, the top, middle, and bottom contours are all similar when r is small. This observation corresponds to Theorem 9.

We observe this on other data sets, too. However, Figure 1 and Theorem 9 can only provide a connection between the sigmoid and the RBF kernels when (a, r) are in a limited range. Thus, in Section 7, we try to compare the two kernels using parameters in other ranges.

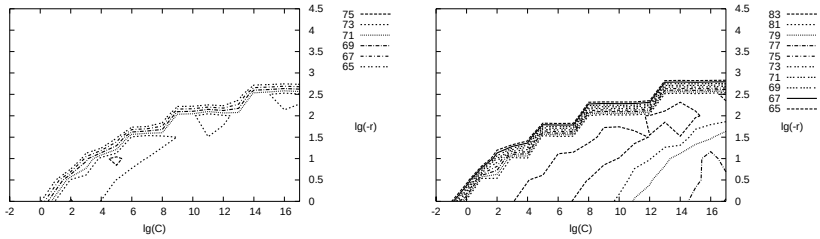
$a = 1$ $a = 0.01$



(a) heart-tanh (b) heart-tanh



(c) heart-(17) (d) heart-(17)



(e) heart-RBF-$\tilde{C}$ (f) heart-RBF-$\tilde{C}$

Figure 1: Performance of different kernels

5 Importance of the Linear Constraint $y^T\alpha = 0$

In Section 3 we showed that for certain parameters, the kernel matrix using the sigmoid kernel is CPD. This is strongly related to the linear constraint $y^T\alpha = 0$ in the dual problem (2). Hence, we can use it to verify the CPD-ness of a given kernel matrix.

Recall that $y^T\alpha = 0$ of (2) is originally derived from the bias term b of (1). It has been known that if the kernel function is PD and $x_i \neq x_j$ for all $i \neq j$, Q will be PD and the problem (6) attains an optimal solution. Therefore, for PD kernels such as the RBF, in many cases, the performance is not affected much if the bias term b is not used. By doing so, the dual problem is

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, i = 1, \ldots, l. \end{aligned} \tag{19}$$

For the sigmoid kernel, we may think that (19) is also acceptable. It turns out that without $y^T\alpha = 0$, in more cases, (19) without the upper bound C, has the objective value $-\infty$. Thus, training data may not be properly separated. The following theorem gives an example on such cases:

Theorem 10 *If there is one $K_{ii} < 0$ and there is no upper bound C of α, (19) has optimal objective value $-\infty$.*

Proof.

Let $\alpha_i = \Delta$ and $\alpha_k = 0$ for $k \neq i$. We can easily see that $\Delta \to \infty$ leads to an optimal objective value $-\infty$. □

Note that for sigmoid kernel matrices, this situation happens when $\min_i(a\|x_i\|^2 + r) < 0$. Thus, when $a > 0$ but r is small, unlike our analysis in Case 1 of Section 3, solving (19) may lead to very different results. This will be shown in the following experiments.

We compare the five-fold cross-validation accuracy with problems (2) and (19). Four data sets, which will be described in Section 7, are considered. We use LIBSVM for solving (2), and a modification of BSVM (Hsu and Lin 2002) for (19). Results of CV accuracy with parameters $a = 1/n$ and $(\log_2 C, r) = [-2, -1, \ldots, 13] \times [-2, -1.8, \ldots, 2]$ are presented in Figure 2. Contours of (2) are on the left column, and those of (19)

are on the right. For each contour, the horizontal axis is $\log_2 C$, while the vertical axis is r. The internal optimization solver of BSVM can handle non-convex problems, so its decomposition procedure guarantees the strict decrease of function values throughout all iterations. However, unlike LIBSVM which always obtains a stationary point of (2) using the analysis in Section 6, for BSVM, we do not know whether its convergent point is a stationary point of (19) or not.

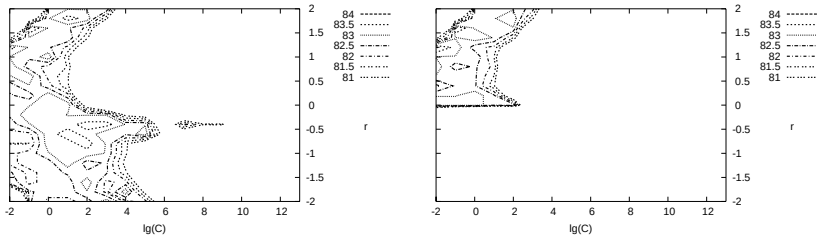
When (2) is solved, from Figure 2, higher accuracy generally happens when $r < 0$ (especially german and diabete). This corresponds to our analysis about the CPD of K when $a > 0$ and r small enough. However, sometimes the CV accuracy is also high when $r > 0$. We have also tried the cases of $a < 0$, results are worse.

The good regions for the right column shift to $r \geq 0$. This confirms our analysis in Theorem 10 as when $r < 0$, (19) without C tends to have the objective value $-\infty$. In other words, without $y^T\alpha = 0$, CPD of K for small r is not useful.

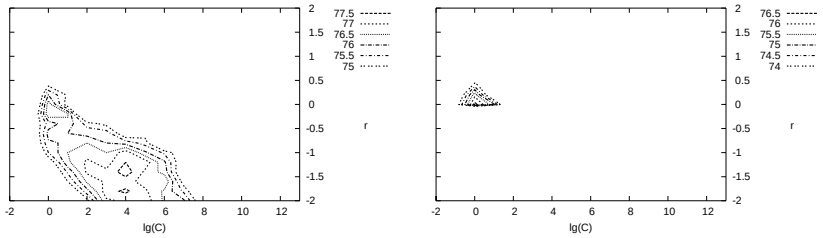
The experiments fully demonstrate the importance of incorporating constraints of the dual problem into the analysis of the kernel. An earlier paper (Sellathurai and Haykin 1999) says that each K_{ij} of the sigmoid kernel matrix is from a hyperbolic inner product. Thus, a special type of maximal margin still exists. However, as shown in Figure 2, without $y^T\alpha = 0$, the performance is very bad. Thus, the separability of non-PSD kernels may not come from their own properties, and a direct analysis may not be useful.

6 SMO-type Implementation for non-PSD Kernel Matrices

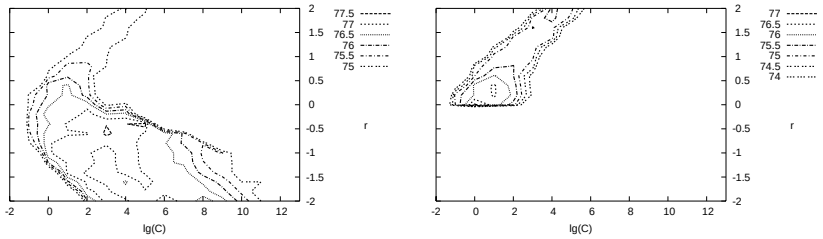
First we discuss how decomposition methods work for PSD kernels and the difficulties for non-PSD cases. In particular, we explain that the algorithm may stay at the same point, so the program never ends. The decomposition method (e.g. (Osuna, Freund, and Girosi 1997; Joachims 1998; Platt 1998; Chang and Lin 2001)) is an iterative process. In each step, the index set of variables is partitioned to two sets B and N, where B is the working set. Then in that iteration variables corresponding to N are fixed while a sub-problem on variables corresponding to B is minimized. Thus, if α^k is the current



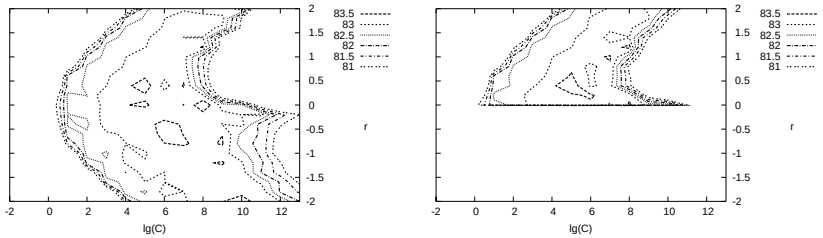
(a) heart (b) heart-nob



(c) german (d) german-nob



(e) diabete (f) diabete-nob



(g) a1a (h) a1a-nob

Figure 2: Comparison of cross validation rates between problems with the linear constraint (left) and without it (right)

solution, the following sub-problem is solved:

$$\begin{aligned} \min_{\alpha_B} \quad & \frac{1}{2}\begin{bmatrix}\alpha_B^T & (\alpha_N^k)^T\end{bmatrix}\begin{bmatrix}Q_{BB} & Q_{BN}\\ Q_{NB} & Q_{NN}\end{bmatrix}\begin{bmatrix}\alpha_B\\ \alpha_N^k\end{bmatrix} - \begin{bmatrix}e_B^T & (e_N^k)^T\end{bmatrix}\begin{bmatrix}\alpha_B\\ \alpha_N^k\end{bmatrix} \\ \text{subject to} \quad & y_B^T\alpha_B = -y_N^T\alpha_N^k, \\ & 0 \le \alpha_i \le C, i \in B. \end{aligned} \tag{20}$$

The objective function of (20) can be simplified to

$$\min_{\alpha_B} \quad \frac{1}{2}\alpha_B^T Q_{BB}\alpha_B + (Q_{BN}\alpha_N^k - e_B)^T\alpha_B$$

after removing constant terms.

The extreme of the decomposition method is the Sequential Minimal Optimization (SMO) algorithm (Platt 1998) whose working sets are restricted to two elements. The advantage of SMO is that (20) can be easily solved without an optimization package. A simple and common way to select the two variables is through the following form of optimal conditions (Keerthi, Shevade, Bhattacharyya, and Murthy 2001; Chang and Lin 2001): α is a stationary point of (2) if and only if α is feasible and

$$\max_{t\in I_{up}(\alpha,C)} -y_t\nabla F(\alpha)_t \le \min_{t\in I_{low}(\alpha,C)} -y_t\nabla F(\alpha)_t, \tag{21}$$

where

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

Thus, when α^k is feasible but not optimal for (2), (21) does not hold so a simple selection of $B = \{i, j\}$ is

$$i \equiv \operatorname*{argmax}_{t\in I_{up}(\alpha^k,C)} -y_t\nabla F(\alpha^k)_t \text{ and } j \equiv \operatorname*{argmin}_{t\in I_{low}(\alpha^k,C)} -y_t\nabla F(\alpha^k)_t. \tag{22}$$

By considering the variable $\alpha_B = \alpha_B^k + d$, and defining

$$\hat{d}_i \equiv y_i d_i \text{ and } \hat{d}_j \equiv y_j d_j,$$

the two-variable sub-problem is

$$\begin{aligned} \min_{\hat{d}_i,\hat{d}_j} \quad & \frac{1}{2}\begin{bmatrix}\hat{d}_i & \hat{d}_j\end{bmatrix}\begin{bmatrix}K_{ii} & K_{ij}\\ K_{ji} & K_{jj}\end{bmatrix}\begin{bmatrix}\hat{d}_i\\ \hat{d}_j\end{bmatrix} + \begin{bmatrix}y_i\nabla F(\alpha^k)_i & y_j\nabla F(\alpha^k)_j\end{bmatrix}\begin{bmatrix}\hat{d}_i\\ \hat{d}_j\end{bmatrix} \\ \text{subject to} \quad & \hat{d}_i + \hat{d}_j = 0, \\ & 0 \le \alpha_i^k + y_i\hat{d}_i, \alpha_j^k + y_j\hat{d}_j \le C. \end{aligned} \tag{23}$$

To solve (23), we can substitute $\hat{d}_i = -\hat{d}_j$ into its objective function:

$$\min_{\hat{d}_j} \quad \frac{1}{2}(K_{ii} - 2K_{ij} + K_{jj})\hat{d}_j^2 + (-y_i \nabla F(\alpha^k)_i + y_j \nabla F(\alpha^k)_j)\hat{d}_j. \tag{24a}$$

$$\text{subject to} \quad L \le \hat{d}_j \le H, \tag{24b}$$

where L and H are upper and lower bounds of $\hat{d}_j$ after including information on $\hat{d}_i$: $\hat{d}_i = -\hat{d}_j$ and $0 \le \alpha_i^k + y_i \hat{d}_i \le C$. For example, if $y_i = y_j = 1$,

$$L = \max(-\alpha_j^k, \alpha_i^k - C) \text{ and } H = \min(C - \alpha_j^k, \alpha_i^k).$$

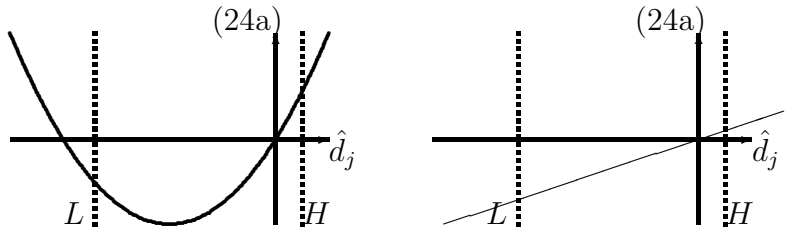
Since $i \in I_{up}(\alpha^k, C)$ and $j \in I_{low}(\alpha^k, C)$, we can clearly see $L < 0$ but H only ≥ 0. If Q is PSD, $K_{ii} + K_{jj} - 2K_{ij} \ge 0$ so (24) is a convex parabola or a straight line. In addition, from the working set selection strategy in (22), $-y_i \nabla F(\alpha^k)_i + y_j \nabla F(\alpha^k)_j > 0$, so (24) is like Figure 3. Thus, there exists $\hat{d}_j < 0$ such that the objective value of (24) is strictly decreasing. In addition, $\hat{d}_j < 0$ also shows the direction toward the minimum of the function.

If $K_{ii} + K_{jj} - 2K_{ij} > 0$, the way to solve (24) is by calculating the minimum of (24a) first:

$$-\frac{-y_i \nabla F(\alpha^k)_i + y_j \nabla F(\alpha^k)_j}{K_{ii} - 2K_{ij} + K_{jj}} < 0. \tag{25}$$

Then, if $\hat{d}_j$ defined by the above is less than L, we reduce $\hat{d}_j$ to the lower bound. If the kernel matrix is only PSD, it is possible that $K_{ii} - 2K_{ij} + K_{jj} = 0$, as shown in Figure 3(b). In this case, using the trick under IEEE floating point standard (Goldberg 1991), we can make sure that (25) to be $-\infty$ which is still defined. Then, a comparison with L still reduces $\hat{d}_j$ to the lower bound. Thus, a direct (but careful) use of (25) does not cause any problem. More details are in (Chang and Lin 2001). The above procedure explains how we solve (24) in an SMO-type software.

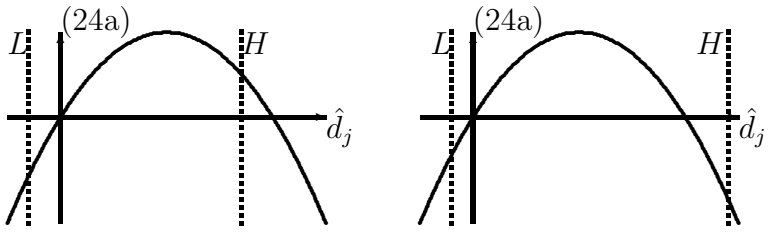
If $K_{ii} - 2K_{ij} + K_{jj} < 0$, which may happen if the kernel is not PSD, (25) is positive. That is, the quadratic function (24a) is concave (see Figure 4) and a direct use of (25) move the solution toward (24a)'s maximum. Therefore, the decomposition method may not have the objective value strictly decreasing, a property usually required for an optimization algorithm. Moreover, it may not be feasible to move along a positive direction $\hat{d}_j$. For example, if $\alpha_i^k = 0, y_i = 1$ and $\alpha_j^k = 0, y_j = -1$, $H = 0$ in (24) so we can neither



(a) $K_{ii} + K_{jj} - 2K_{ij} > 0$ (b) $K_{ii} + K_{jj} - 2K_{ij} = 0$

Figure 3: Solving the convex sub-problem (24)

decrease α_i nor α_j. Thus, under the current setting for PSD kernels, it is possible that the next solution stays at the same point so the program never ends. In the following we propose different approaches to handle this difficulty.



(a) L is the minimum (b) H is the minimum

Figure 4: Solving the concave sub-problem (24)

6.1 Restricting the Range of Parameters

The first approach is to restrict the parameter space. In other words, users are allowed to specify only certain kernel parameters. Then the sub-problem is guaranteed to be convex, so the original procedure for solving sub-problems works without modification.

Lemma 2 *If $a > 0$ and*

$$\max_i (a\|x_i\|^2 + r) \leq 0, \tag{26}$$

any two-variable sub-problem of an SMO algorithm is convex.

We have explained that the sub-problem can be reformulated as (24), so the proof is reduced to show that $K_{ii} - 2K_{ij} + K_{jj} \geq 0$. This, in fact, is nearly the same as the proof of Theorem 5. The only change is that without assuming $x_i \neq x_j$, "> 0" is changed to "≥ 0."

Therefore, if we require that a and r satisfy (26), we will never have an endless loop staying at one α^k.

6.2 An SMO-type Method for General non-PSD Kernels

Results in Section 6.1 depend on properties of the sigmoid kernel. Here we propose an SMO-type method which is able to handle all kernel matrices no matter they are PSD or not. To have such a method, the key is on solving the sub-problem when $K_{ii}-2K_{ij}+K_{jj} < 0$. In this case, (24a) is a concave quadratic function like that in Figure 4. The two sub-figures clearly show that the global optimal solution of (24) can be obtained by checking the objective values at two bounds L and H.

A disadvantage is that this procedure of checking two points is different from the solution procedure of $K_{ii} - 2K_{ij} + K_{jj} \geq 0$. Thus, we propose to consider only the lower bound L which, as $L < 0$, always ensures the strict decrease of the objective function. Therefore, the algorithm is as follows:

$$\begin{aligned} &\text{If } K_{ii} - 2K_{ij} + K_{jj} > 0, \text{then } \hat{d}_j \text{ is the maximum of (25) and } L, \\ &\text{Else } \hat{d}_j = L. \end{aligned} \tag{27}$$

Practically the change of the code may be only from (25) to

$$-\frac{-y_i \nabla F(\alpha^k)_i + y_j \nabla F(\alpha^k)_j}{\max(K_{ii} - 2K_{ij} + K_{jj}, 0)}. \tag{28}$$

When $K_{ii} + K_{jj} - 2K_{ij} < 0$, (28) is $-\infty$. Then the same as the situation of $K_{ii} + K_{jj} - 2K_{ij} = 0$, $\hat{d}_j = L$ is taken.

An advantage of this strategy is that we do not have to exactly solve (24). (28) also shows that a very simple modification from the PSD-kernel version is possible. Moreover, it is easier to prove the asymptotic convergence. The reason will be discussed after

Lemma 3. In the following we prove that any limit point of the decomposition procedure discussed above is a stationary point of (2). In earlier convergence results, Q is PSD so a stationary point is already a global minimum.

If the working set selection is via (22), existing convergence proofs for PSD kernels (Lin 2001; Lin 2002) require the following important lemma which is also needed here:

Lemma 3 *There exists $\sigma > 0$ such that for any k,*

$$F(\alpha^{k+1}) \leq F(\alpha^k) - \frac{\sigma}{2}\|\alpha^{k+1} - \alpha^k\|^2. \tag{29}$$

Proof.

If $K_{ii} + K_{jj} - 2K_{ij} \geq 0$ in the current iteration, (Lin 2002) shows that by selecting σ as the following number

$$\min\{\frac{2}{C}, \min_{t,r}\{\frac{K_{tt} + K_{rr} - 2K_{tr}}{2} \mid K_{tt} + K_{rr} - 2K_{tr} > 0\}\}, \tag{30}$$

(29) holds.

If $K_{ii}+K_{jj}-2K_{ij} < 0$, $\hat{d}_j = L < 0$ is the step chosen so $(-y_i\nabla F(\alpha^k)_i + y_j\nabla F(\alpha^k)_j)\hat{d}_j < 0$. As $\|\alpha^{k+1} - \alpha^k\|^2 = 2\hat{d}_j^2$ from $\hat{d}_i = -\hat{d}_j$, (24a) implies that

$$\begin{aligned} F(\alpha^{k+1}) - F(\alpha^k) &< \frac{1}{2}(K_{ii} + K_{jj} - 2K_{ij})\hat{d}_j^2 \\ &= \frac{1}{4}(K_{ii} + K_{jj} - 2K_{ij})\|\alpha^{k+1} - \alpha^k\|^2 \\ &\leq -\frac{\sigma'}{2}\|\alpha^{k+1} - \alpha^k\|^2, \end{aligned} \tag{31}$$

where

$$\sigma' \equiv -\max_{t,r}\{\frac{K_{tt} + K_{rr} - 2K_{tr}}{2} \mid K_{tt} + K_{rr} - 2K_{tr} < 0\}. \tag{32}$$

Therefore, by defining σ as the minimum of (30) and (32), the proof is complete. □

Next we give the main convergence result:

Theorem 11 *For the decomposition method using (22) for the working set selection and (27) for solving the sub-problem, any limit point of $\{\alpha^k\}$ is a stationary point of (2).*

Proof.

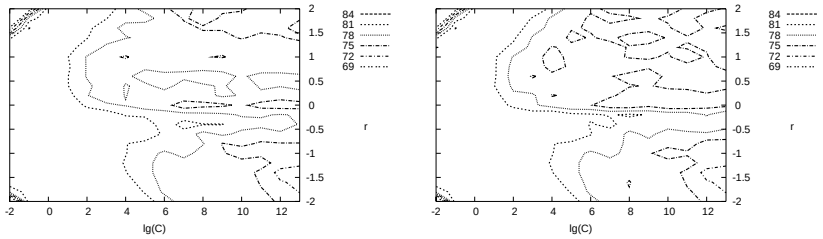
If we carefully check the proof in (Lin 2001; Lin 2002), it can be extended to non-PSD Q if (1) (29) holds and (2) a local minimum of the sub-problem is obtained in each iteration. Now we have (29) from Lemma 3. In addition, $\hat{d}_j = L$ is essentially one of the two local minima of problem (24) as clearly seen from Figure 4. Thus, the same proof follows. □

There is an interesting remark about Lemma 3. If we exactly solve (24), so far we have not been able to establish Lemma 3. The reason is that if $\hat{d}_j = H$ is taken, $(-y_i \nabla F(\alpha^k)_i + y_j \nabla F(\alpha^k)_j)\hat{d}_j > 0$ so (31) may not be true. Therefore, the convergence is not clear. In the whole convergence proof, Lemma 3 is used to obtain $\|\alpha^{k+1} - \alpha^k\| \to 0$ as $k \to \infty$. A different way to have this property is by slightly modifying the sub-problem (20) as shown in (Palagi and Sciandrone 2002). Then the convergence holds when we exactly solve the new sub-problem.

Although Theorem 11 shows only that the improved SMO algorithm converges to a stationary point rather than a global minimum, the algorithm nevertheless shows a way to design a robust SVM software with separability concern. Theorem 1 indicates that a stationary point is feasible for the separability problem (5). Thus, if the number of support vectors of this stationary point is not too large, the training error would not be too large, either. Furthermore, with additional constraints $y^T\alpha = 0$ and $0 \le \alpha_i \le C, i = 1, \ldots, l$, a stationary point may already be a global one. If this happens at parameters with better accuracy, we do not worry about multiple stationary points at others. An example is the sigmoid kernel, where discussion in Section 3 indicates that parameters with better accuracy tends to be with CPD kernel matrices.

It is well known that Neural Networks have similar problems about local minima (Sarle 1997), and a popular way to prevent trapping in a bad one is multiple random initializations. Here we adapt this method and present an empirical study in Figure 5. We use the heart data set, with the same setting as in Figure 2. Figure 5(a) is the contour which uses the zero vector as the initial α^0. Figure 5(b) is the contour by choosing the solution with the smallest of five objective values via different random initial α^0.

The performance of Figures 5(a) and 5(b) is similar, especially in regions with good rates. For example, when $r < -0.5$, the two contours are almost the same, a property



(a) heart-zero (b) heart-random5

Figure 5: Comparison of cross validation rates between approaches without (left) and with (right) five random initializations

which may explain the CPD-ness in that region. In the regions where multiple stationary points may occur (e.g. $C > 2^6$ and $r > +0.5$), two contours are different but the rates are still similar. We observe similar results on other datasets, too. Therefore, the stationary point obtained by zero initialization seems good enough in practice.

7 Modification to Convex Formulations

While (5) is non-convex, it is possible to slightly modify the formulation to be convex. If the objective function is replaced by

$$\frac{1}{2}\alpha^T\alpha + C\sum_{i=1}^{l}\xi_i,$$

then (5) becomes convex. Note that non-PSD kernel K still appears in constraints. The main drawback of this approach is that α_i are in general non-zero, so unlike standard SVM, the sparsity is lost.

There are other formulations which use a non-PSD kernel matrix but remain convex. For example, we can consider the kernel logistic regression (KLR) (e.g., (Wahba 1998)) and use a convex regularization term:

$$\min_{\alpha,b} \quad \frac{1}{2}\sum_{r=1}^{l}\alpha_r^2 + C\sum_{r=1}^{l}\log(1+e^{\xi_r}), \tag{33}$$

where

$$\xi_r \equiv -y_r(\sum_{j=1}^{l} \alpha_j K(x_r, x_j) + b).$$

By defining an $(l+1) \times l$ matrix $\tilde{K}$ with

$$\tilde{K}_{ij} \equiv \begin{cases} K_{ij} & \text{if } 0 \le i, j \le l, \\ 1 & \text{if } i = l+1, \end{cases}$$

the Hessian matrix of (33) is

$$\tilde{I} + C\tilde{K}\text{diag}(\tilde{p})\text{diag}(1-\tilde{p})\tilde{K}^T,$$

where $\tilde{I}$ is an $l+1$ by $l+1$ identity matrix with the last diagonal element replaced by zero. $\tilde{p} \equiv [1/(1+e^{\xi_1}), \ldots, 1/(1+e^{\xi_l})]^T$ and $\text{diag}(\tilde{p})$ is a diagonal matrix generated by $\tilde{p}$. Clearly, the Hessian matrix is always positive semidefinite, so (33) is convex.

In the following we compare SVM (RBF and sigmoid kernels) and KLR (sigmoid kernel). Four data sets are tested: heart, german, diabete, and a1a. They are from (Michie, Spiegelhalter, and Taylor 1994) and (Blake and Merz 1998). The first three data sets are linearly scaled, so values of each attribute are in [-1, 1]. For a1a, its values are binary (0 or 1), so we do not scale it. We train SVM (RBF and sigmoid kernels) by LIBSVM (Chang and Lin 2001), which, an SMO-type decomposition implementation, uses techniques in Section 6 for solving non-convex optimization problems. For KLR, two optimization procedures are compared. The first one, KLR-NT, is a Newton's method implemented by modifying the software TRON (Lin and Moré 1999). The second one, KLR-CG, is conjugate gradient method (see, for example, (Nash and Sofer 1996)). The stopping criteria for the two procedures are set the same to ensure that the solutions are comparable.

For the comparison, we conduct a two-level cross validation. At the first level, data are separated to five folds. Each fold is predicted by training the remaining four folds. For each training set, we perform another five-fold cross validation and choose the best parameter by CV accuracy. We try all $(\log_2 C, \log_2 a, r)$ in the region $[-3, 0, \ldots, 12] \times [-12, -9, \ldots, 3] \times [-2.4, -1.8, \ldots, 2.4]$. Then the average testing accuracy is reported in Table 1. Note that for the parameter selection, the RBF kernel $e^{-a\|x_i - x_j\|^2}$ does not involve with r.

Resulting accuracy is similar for all the three approaches. The sigmoid kernel seems to work well in practice, but it is not better than RBF. As RBF has properties of being PD and having fewer parameters, somehow there is no strong reason to use the sigmoid. KLR with the sigmoid kernel is competitive with SVM, and a nice property is that it solves a convex problem. However, without sparsity, the training and testing time for KLR is much longer. Moreover, CG is worse than NT for KLR. These are clearly shown in Table 2. The experiments are put on Pentium IV 2.8G machines with 1024 MB RAM. Optimized linear algebra subroutines (Whaley, Petitet, and Dongarra 2000) are linked to reduce the computational time for KLR solvers. The time is measured in CPU seconds. Number of support vectors (#SV) and training/testing time are averaged from the results of the first level of five-fold CV. This means that the maximum possible #SV here is 4/5 of the original data size, and we can see that KLR models are dense to this extent.

Table 1: Comparison of test accuracy

			$e^{-a\|x_i-x_j\|^2}$	$\tanh(ax_i^T x_j + r)$		
data set	#data	#attributes	SVM	SVM	KLR-NT	KLR-CG
heart	270	13	83.0%	83.0%	83.7%	83.7%
german	1000	24	76.6%	76.1%	75.6%	75.6%
diabete	768	8	77.6%	77.3%	77.1%	76.7%
a1a	1605	123	83.6%	83.1%	83.7%	83.8%

Table 2: Comparison of time usage

	$\tanh(ax_i^T x_j + r)$					
	#SV			training/testing time		
data set	SVM	KLR-NT	KLR-CG	SVM	KLR-NT	KLR-CG
heart	115.2	216	216	0.02/0.01	0.12/0.02	0.45/0.02
german	430.2	800	800	0.51/0.07	5.76/0.10	73.3/0.11
diabete	338.4	614.4	614.4	0.09/0.03	2.25/0.04	31.7/0.05
a1a	492	1284	1284	0.39/0.08	46.7/0.25	80.3/0.19

8 Discussions

From the results in Sections 3 and 5, we clearly see the importance of the CPD-ness which is directly related to the linear constraint $y^T\alpha = 0$. We suspect that for many non-PSD kernels used so far, their viability is based on it as well as inequality constraints $0 \le \alpha_i \le C, i = 1, \ldots, l$ of the dual problem. It is known that some non-PSD kernels are not CPD. For example, the tangent distance kernel matrix in (Haasdonk and Keysers 2002) may contain more than one negative eigenvalue, a property that indicates the matrix is not CPD. Further investigation on such non-PSD kernels and the effect of inequality constraints $0 \le \alpha_i \le C$ will be interesting research directions.

Even though the CPD-ness of the sigmoid kernel for certain parameters gives an explanation to the practical viability, the quality of the local minimum solution in other parameters may not be guaranteed. This makes it hard to select suitable parameters for the sigmoid kernel. Thus, in general we do not recommend the use of the sigmoid kernel.

In addition, our analysis indicates that for certain parameters the sigmoid kernel behaves like the RBF kernel. Experiments also show that their performance are similar. Therefore, with the result in (Keerthi and Lin 2003) showing that the linear kernel is essentially a special case of the RBF kernel, among existing kernels, RBF should be the first choice for general users.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan via the grant NSC 90-2213-E-002-111. We thank users of LIBSVM (in particular, Carl Staelin), who somewhat forced us to study this issue. We also thank Bernhard Schölkopf and Bernard Haasdonk for some helpful discussions.

References

Berg, C., J. P. R. Christensen, and P. Ressel (1984). *Harmonic Analysis on Semigroups*. New York: Springer-Verlag.

Blake, C. L. and C. J. Merz (1998). UCI repository of machine learning databases. Technical report, University of California, Department of Information and Computer Science, Irvine, CA. Available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Boser, B., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*(2), 121–167.

Burges, C. J. C. (1999). Geometry and invariance in kernel based methods. In B. Schölkopf, C. Burges, and A. Smola (Eds.), *Advances in Kernel Methods: Support Vector Learning*, pp. 89–116. MIT Press.

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

Cortes, C. and V. Vapnik (1995). Support-vector network. *Machine Learning 20*, 273–297.

DeCoste, D. and B. Schölkopf (2002). Training invariant support vector machines. *Machine Learning 46*, 161–190.

Goldberg, D. (1991). What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys 23*(1), 5–48.

Haasdonk, B. and D. Keysers (2002). Tangent distance kernels for support vector machines. In *Proceedings of the 16th ICPR*, pp. 864–868.

Hsu, C.-W. and C.-J. Lin (2002). A simple decomposition method for support vector machines. *Machine Learning 46*, 291–314.

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

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

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

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

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

Lin, C.-J. and J. J. Moré (1999). Newton's method for large-scale bound constrained problems. *SIAM Journal on Optimization 9*, 1100–1127.

Lin, K.-M. and C.-J. Lin (2003). A study on reduced support vector machines. *IEEE Transactions on Neural Networks 14*(6), 1449–1559.

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

Michie, D., D. J. Spiegelhalter, and C. C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Englewood Cliffs, N.J.: Prentice Hall. Data available at `http://www.ncc.up.pt/liacc/ML/statlog/datasets.html`.

Nash, S. G. and A. Sofer (1996). *Linear and Nonlinear Programming*. McGraw-Hill.

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

Osuna, E. and F. Girosi (1998). Reducing the run-time complexity of support vector machines. In *Proceedings of International Conference on Pattern Recognition*.

Palagi, L. and M. Sciandrone (2002). On the convergence of a modified version of SVM^{light} algorithm. Technical Report IASI-CNR 567.

Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola (Eds.), *Advances*

in Kernel Methods - Support Vector Learning, Cambridge, MA. MIT Press.

Sarle, W. S. (1997). Neural Network FAQ. Periodic posting to the Usenet newsgroup comp.ai.neural-nets.

Schölkopf, B. (1997). *Support Vector Learning*. Ph. D. thesis.

Schölkopf, B. (2000). The kernel trick for distances. In *NIPS*, pp. 301–307.

Schölkopf, B. and A. J. Smola (2002). *Learning with kernels*. MIT Press.

Sellathurai, M. and S. Haykin (1999). The separability theory of hyperbolic tangent kernels and support vector machines for pattern classification. In *Proceedings of ICASSP99*.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C. J. C. Burges, and A. J. Smola (Eds.), *Advances in Kernel Methods: Support Vector Learning*, pp. 69–88. MIT Press.

Whaley, R. C., A. Petitet, and J. J. Dongarra (2000). Automatically tuned linear algebra software and the ATLAS project. Technical report, Department of Computer Sciences, University of Tennessee.

A Proof of Theorem 8

The proof of Theorem 8 contains three parts: the convergence of the optimal solution, the convergence of the decision value without the bias term, and the convergence of the bias term. Before entering the proof, we first need to know that (17) has a PD kernel under our assumption $x_i \neq x_j$ for all $i \neq j$. Therefore, the optimal solution $\hat{\alpha}^*$ of (17) is unique. From now on we denote $\hat{\alpha}^r$ as a local optimal solution of (2), and b^r as the associated optimal b value. For (17), b^* denotes its optimal b.

1. The convergence of optimal solution:

$$\lim_{r \to -\infty} \theta_r \hat{\alpha}^r = \hat{\alpha}^*, \text{where } \theta_r \equiv 1 + \tanh(r). \tag{34}$$

Proof.

By the equivalence between (2) and (16), $\theta_r \hat{\alpha}^r$ is the optimal solution of (16). The convergence to $\hat{\alpha}^*$ comes from (Keerthi and Lin 2003, Lemma 2) since $\bar{Q}$ is PD and the kernel of (16) approaches $\bar{Q}$ by Lemma 1. □

2. The convergence of the decision value without the bias term: For any x,

$$\lim_{r \to -\infty} \sum_{i=1}^{l} y_i \hat{\alpha}_i^r \tanh(a x_i^T x + r) = \sum_{i=1}^{l} y_i \hat{\alpha}_i^* e^{2a x_i^T x_j}. \tag{35}$$

Proof.

$$\begin{aligned}
& \lim_{r \to -\infty} \sum_{i=1}^{l} y_i \hat{\alpha}_i^r \tanh(a x_i^T x + r) \\
= & \lim_{r \to -\infty} \sum_{i=1}^{l} y_i \hat{\alpha}_i^r (1 + \tanh(a x_i^T x + r)) \qquad (36) \\
= & \lim_{r \to -\infty} \sum_{i=1}^{l} y_i \theta_r \hat{\alpha}_i^r \frac{1 + \tanh(a x_i^T x + r)}{\theta_r} \\
= & \sum_{i=1}^{l} y_i \lim_{r \to -\infty} \theta_r \hat{\alpha}_i^r \lim_{r \to -\infty} \frac{1 + \tanh(a x_i^T x + r)}{\theta_r} \\
= & \sum_{i=1}^{l} y_i \hat{\alpha}_i^* e^{2a x_i^T x}. \qquad (37)
\end{aligned}$$

(36) comes from the equality constraint in (2) and (37) comes from (34) and Lemma 1. □

3. The convergence of the bias term:

$$\lim_{r \to -\infty} b^r = b^*. \tag{38}$$

Proof.

By the KKT condition that b^r must satisfy,

$$\max_{i \in I_{up}(\hat{\alpha}^r, C)} -y_i \nabla F(\hat{\alpha}^r)_i \leq b^r \leq \min_{i \in I_{low}(\hat{\alpha}^r, C)} -y_i \nabla F(\hat{\alpha}^r)_i,$$

where I_{up} and I_{low} are defined in (21). In addition, because b^* is unique,

$$\max_{i\in I_{up}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_T(\hat{\alpha}^*)_i = b^* = \min_{i\in I_{low}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_T(\hat{\alpha}^*)_i.$$

Note that the equivalence between (2) and (16) implies $\nabla F(\hat{\alpha}^r)_i = \nabla F_r(\theta_r\hat{\alpha}^r)_i$. Thus,

$$\max_{i\in I_{up}(\theta_r\hat{\alpha}^r,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i \le b^r \le \min_{i\in I_{low}(\theta_r\hat{\alpha}^r,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i.$$

By the convergence of $\theta_r\hat{\alpha}^r$ when $r \to -\infty$, after r is small enough, all index i's satisfying $\hat{\alpha}_i^* < \tilde{C}$ would have $\theta_r\hat{\alpha}_i^r < \tilde{C}$. That is, $I_{up}(\hat{\alpha}^*, \tilde{C}) \subseteq I_{up}(\theta_r\hat{\alpha}^r, \tilde{C})$. Therefore, when r is small enough,

$$\max_{i\in I_{up}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i \le \max_{i\in I_{up}(\theta_r\hat{\alpha}^r,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i.$$

Similarly,

$$\min_{i\in I_{low}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i \ge \min_{i\in I_{low}(\theta_r\hat{\alpha}^r,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i.$$

Thus, for $r < 0$ small enough,

$$\max_{i\in I_{up}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i \le b^r \le \min_{i\in I_{low}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_r(\theta_r\hat{\alpha}^r)_i.$$

Taking $\lim_{r\to-\infty}$ on both sides, using Lemma 1 and (34),

$$\lim_{r\to-\infty} b^r = \max_{i\in I_{up}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_T(\hat{\alpha}^*)_i = \min_{i\in I_{low}(\hat{\alpha}^*,\tilde{C})} -y_i\nabla F_T(\hat{\alpha}^*)_i = b^*. \tag{39}$$

□

Therefore, with (37) and (39), our proof of Theorem 8 is complete.