Smooth Support Vector Machines

Yuh-Jye Lee

March 6, 2000

Preliminary Thesis Proposal
Computer Sciences Department
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706
Committee: O.L. Mangasarian (chair), M.C. Ferris, J.W. Shavlik

Abstract

Smoothing methods, extensively used for solving important mathematical programming problems and applications, are proposed here to generate and solve an unconstrained smooth reformulation of support vector machines for pattern classification using completely arbitrary kernels. We term such reformulations smooth support vector machines (SSVMs). A fast Newton-Armijo algorithm for solving the SSVMs converges globally and quadratically. Numerical results and comparisons are given to demonstrate the effectiveness and speed of the algorithm. On six publicly available datasets, tenfold cross validation correctness of SSVM was the highest compared with four other methods as well as the fastest on five of these datasets. Using nonlinear kernels, SSVMs also obtained very distinct nonlinear separations for the checkerboard and the two-spiral datasets. A medical application is also proposed here. A linear support vector machine (SVM) is used to extract 6 features from a total of 31 features in a dataset of 253 breast cancer patients. Five features are nuclear features obtained during a non-invasive diagnostic procedure while one feature, tumor size, is obtained during surgery. The linear SVM selected the 6 features in the process of classifying the patients into node-positive (patients with some metastasized lymph nodes) and node-negative (patients with no metastasized lymph nodes). The 6 features were then used in a Gaussian kernel nonlinear SVM to classify the patients into three prognostic groups: good (node-negative), intermediate (1 to 4 metastasized nodes) and poor (more than 4 metastasized nodes). Very well separated Kaplan-Meier survival curves were constructed for the three groups with pairwise p-value of less than 0.009 based on the logrank statistic. New patients can be assigned to one of these three prognostic groups with its associated survival curve, based only on 6 features obtained before and during surgery, but without the potentially risky procedure of removing lymph nodes to determine how many of them have metastasized.

Contents

1 Introduction **2**

1.1 **Smooth Support Vector Machines** 2

1.2 **A Medical Application to Breast Cancer Prognosis** . . 3

1.3 **Notation** 4

2 Smooth Support Vector Machines **5**

2.1 **Support Vector Machines** 5

2.2 **Smooth Support Vector Machines** 8

2.3 **A Newton-Armijo Algorithm for Smooth Support Vector Machines** 11

2.4 **SSVM with a Nonlinear Kernel** 13

2.5 **Numerical Results and Comparisons** 13

2.6 **Summary and Future Work** 15

3 An Application: Breast Cancer Prognosis **22**

3.1 **Kaplan-Meier Survival Curve and Logrank Test** 22

3.2 **Breast Cancer Prognosis** 23

3.3 **Summary and Future Work** 26

4 Appendix **32**

1 Introduction

Support vector machines [40, 14] have been extensively used in machine learning and data mining [27, 7, 5, 6, 4] as well as in medical applications [34, 44, 43, 42, 45, 46, 22]. In the classification problem, we attempt to classify points belonging to two given sets in R^n by a linear or nonlinear separating surface. The learning process utilizes the training input data to generate a separating surface which assigns each training input data point to the appropriate category. The separating surface is then tested on the unseen data. In order to have a good generalization ability a support vector machine not only minimizes the training error but also attempts to improve the generalization ability of the separating surface. For the linearly separable case, this is achieved by a separating hyperplane mid-way between the furthest parallel planes apart that bound the two sets. From a statistical learning theory point of view, such a support vector machine will minimize the empirical error of the training data as well as the error bound for the unseen data [40].

1.1 Smooth Support Vector Machines

Smoothing methods have been extensively used for solving important mathematical programming problems [10, 11, 9, 12, 13, 18, 39, 15]. In this work, we propose a new formulation of support vector machines with linear (13) and nonlinear kernels (30) [40, 6, 8, 29] for pattern classification. We begin with the linear case which can be converted to an unconstrained optimization problem (15). Since the objective function of this unconstrained optimization problem is not twice differentiable, we employ a smoothing technique to introduce the smooth support vector machine (SSVM) (17). The smooth support vector machine has important mathematical properties such as strong convexity and infinitely often differentiability. Based on these properties, we can prove that when the smoothing parameter α in the SSVM approaches infinity, the unique solution of the SSVM converges to the unique solution of the original optimization problem (13). Practically speaking α never exceeds 10. We also prescribe a Newton-Armijo algorithm (*i.e.* a Newton method with a stepsize determined by the simple Armijo rule which eventually becomes redundant) to solve the SSVM and show that this algorithm globally and quadratically converges to the unique solution of the SSVM. To construct a nonlinear classifier, a nonlinear kernel [40, 29] is used to obtain the nonlinear support vector machine (30) and its corresponding smooth version (31) with a completely arbitrary kernel. The smooth formulation (31) with a nonlinear kernel retains the strong convexity and twice differentiability and thus we can apply Newton-Armijo algorithm to solve it. Numerical tests and comparisons are given in Section 2.5. It turns out that SSVM gave the highest tenfold cross validation correctness [37] on six publicly available datasets as well as the fastest times for five of the datasets when compared with four other methods [2, 5]. Using nonlinear kernels, SSVM also obtained very clear separation for the checkerboard pattern of [20, 32] as depicted in Figure 3 and Figure 4 as well as for the two-spiral as depicted in Figure 6.

1.2 A Medical Application to Breast Cancer Prognosis

In this part of the work we attempt to classify breast cancer patients using a criterion that is closely related to the decision whether a patient is prescribed to have chemotherapy treatment or not. This criterion is the presence of metastasized lymph nodes under the patient's armpits (node-positive) or their absence (node-negative). Lymph nodes are removed during surgery in conjunction with the removal of the malignant tumor from the breast. This potentially risky procedure which can cause arm swelling and increased susceptibility to infection can possibly be eliminated by using the classification procedures proposed here. By using a linear support vector machine (8) we first select 6 out of 31 available features to classify patients into node-positive and node-negative patients. We then use these six features in a nonlinear support vector machine (11) to classify 253 breast cancer patients into three prognosis groups: a good prognosis group (GPG) which is node-negative, an intermediate prognosis group (IPG) which has 1-4 metastasized lymph nodes, and a poor prognosis group (PPG) which has more than 4 metastasized lymph nodes. It turns out that these three groups have well separated Kaplan-Meier survival curves [19, 21]. We use the logrank statistic to test the difference between any two groups. The p-values between each pair of groups are less than 0.009 (Figure 8). That means these curves are statistically distinct. In addition, 72.8% of the patients in GPG did not take chemotherapy, while 87.5% of the patients in PPG took chemotherapy. Only 56.7% of patients in IPG took chemotherapy. Thus our classification procedure can be utilized to assign new patients to one of three prognostic groups with an associated survival curve and a possible indication of the utilization of chemotherapy or not.

1.3 Notation

- All vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. The scalar (inner) product of two vectors x and y in the n-dimensional real space R^n will be denoted by $x'y$.

- For a vector $x \in R^n$, the plus function x_+ is defined as $(x_+)_i = \max\{0, x_i\}$, $i = 1, \ldots, n$.

- For a real value $z \in R$, the step function z_* is defined as $z_* = 1$ if $z > 0$ and $z_* = 0$ otherwise.

- For a matrix $A \in R^{m\times n}$, A_i is the ith row of A which is a row vector in R^n. A column vector of one of arbitrary dimension will be denoted by e.

- We shall employ the MATLAB "dot" notation [35] to signify application of a function to all components of a matrix or a vector. For example if $A \in R^{m\times n}$, then $A^2_{\bullet} \in R^{m\times n}$ will denote the matrix obtained by squaring each element of A.

- For $x \in R^n$ and $1 \le p < \infty$, the norm $\|x\|_p$ will denote the $p-$norm:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \tag{1}$$

  and

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|. \tag{2}$$

- For a general norm $\|\cdot\|$ on R^n, the dual norm $\|\cdot\|'$ on R^n is defined as

$$\|x\|' := \max_{\|y\|=1} x'y. \tag{3}$$

  For $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$, the p-norm $\|\cdot\|_p$ and q-norm $\|\cdot\|_q$ are dual norms. Thus the 1-norm $\|\cdot\|_1$ and ∞-norm $\|\cdot\|_\infty$ are dual norms while the 2-norm is self-dual.

- If f is a real valued function defined on the n-dimensional real space R^n, the gradient of f at x is denoted by $\nabla f(x)$ which is a row vector in R^n and the $n \times n$ Hessian matrix of second partial derivatives of f at x is denoted by $\nabla^2 f(x)$.

- The level set of f is defined as $L_\mu(f) = \{x | f(x) \leq \mu\}$ for a given real number μ.

- The base of the natural logarithm will be denoted by ε.

2 Smooth Support Vector Machines

2.1 Support Vector Machines

We begin with the simple linear support vector machine formulation as follows:
(SVM)

$$\begin{array}{rlcl} \displaystyle\min_{w,\gamma,y} & \nu e'y + \|w\| & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \tag{4}$$

Here, ν is a positive weight, $\|\cdot\|$ is an arbitrary norm and the $m \times n$ matrix A represents m given points in R^n which belong to class 1 or -1 depending on whether the corresponding elements of the given $m \times m$ diagonal matrix D are 1 or -1 respectively. The plane in R^n, with normal w and distance from the origin $\frac{|\gamma|}{\|w\|}$ from the origin, generated by a solution (w, γ) of the above mathematical program:

$$x'w = \gamma, \tag{5}$$

strictly separates the set $A+ := \{A_i \mid D_{ii} = 1\}$ from $A- := \{A_i \mid D_{ii} = -1\}$, if and only if the slack variable $y \in R^m$ is zero when ν is sufficiently large as shown in Figure 1. The margin or distance (measured by a norm $\|\cdot\|'$ that is dual to the norm $\|\cdot\|$ of (4)) between the bounding planes $x'w = \gamma \pm 1$ is given by $\frac{2}{\|w\|}$ [28]. These two planes are parallel to the plane $x'w = \gamma$ and bound the sets $A+$ and $A-$. For $\nu \geq \bar{\nu}$ for some $\bar{\nu} > 0$ [30, Theorem 4], the margin is maximized by a solution of (4) that suppresses the slack variable y first. When the sets $A+$ and $A-$ are not strictly linearly separable, some components y_i of the slack variable y will be positive. The points A_i in either $A+$ or $A-$ corresponding to positive y_i, as well as those corresponding to zero y_i but with positive Lagrange multipliers, are termed *support vectors*. Points A_i with positive y_i lie on the wrong sides of the bounding planes $x'w = \gamma \pm 1$, while points with positive multipliers and zero slacks y_i lie on the bounding planes. Moreover $e'y$ provides an upper bound of the number of misclassified

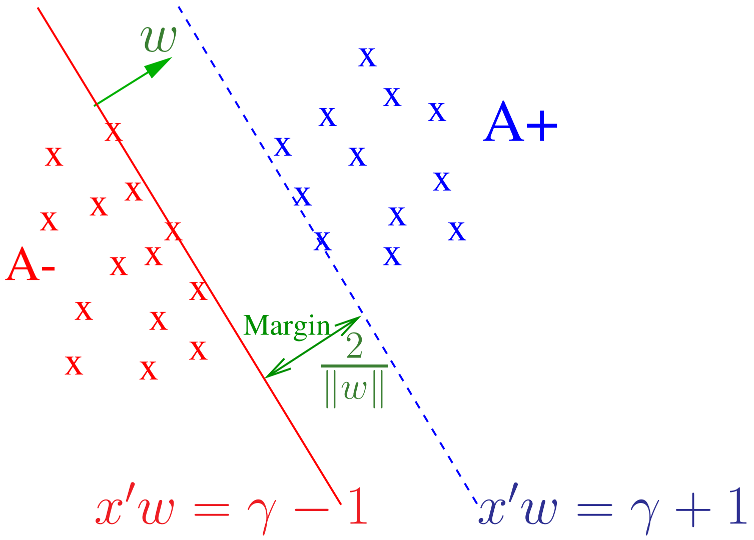

Figure 1: The separating plane $x'w = \gamma$, midway and parallel the two bounding parallel planes $x'w = \gamma \pm 1$ with maximal separation margin $\frac{2}{\|w\|}$, is obtained by solving the mathematical program (4).

points of the training data [8]. A new point $x \in R^n$ is classified in class 1 or -1 according to whether the *decision function*

$$(x'w - \gamma)_*, \tag{6}$$

yields 1 or 0 respectively.

Using different objective functions in (4) leads to different SVMs. For example, replacing $\|w\|$ by $\frac{1}{2}\|w\|_2^2$ we have the conventional support vector machine [40, 14]

$(SVM_{\|\cdot\|_2^2})$

$$\begin{array}{lrcl} \min\limits_{w,\gamma,y} & \nu e'y + \frac{1}{2}\|w\|_2^2 & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & y & \geq & 0 \end{array} \tag{7}$$

and using $\|w\|_1$ instead of $\|w\|$ in SVM we have the following linear programming formulation:

$(SVM_{\|\cdot\|_1})$

$$\begin{array}{llcll} \min\limits_{w,\gamma,s,y} & & \nu e'y + e's & & \\ \text{s.t.} & & D(Aw - e\gamma) + y & \geq & e \\ & -s \leq & w & \leq & s \\ & & y & \geq & 0. \end{array} \tag{8}$$

The linear programming formulation $SVM_{\|\cdot\|_1}$ (8) has been shown [5] to have one of the best feature selection properties among all norms including the conventional support vector machine (7). We will use $SVM_{\|\cdot\|_1}$ for feature selection for breast cancer prognosis in the next section.

In many classification problems, the training data can not be linearly separated in the original space but may be linearly separated in a higher dimensional space after applying some nonlinear map. Using the kernel technique we can achieve this goal without knowing the nonlinear map. We now define a general kernel map as follows:

Definition 2.1 Let $A \in R^{m\times n}$ and $B \in R^{n\times l}$. The **kernel** $K(A, B)$ maps $R^{m\times n} \times R^{n\times l}$ into $R^{m\times l}$.
Typical kernel functions are the following:

- **Polynomial Kernel** $(AA' + \mu aa')^d_\bullet$,
  where $[\cdot]^d_\bullet$ denotes component-wise exponentiation as in MATLAB [35].
- **Radial Basis (Gaussian) Kernel** $\varepsilon^{-\mu\|A_i - A_j\|^2}$, $i, j = 1, \ldots, m$.
- **Neural Network Kernel** $(AA' + \mu aa')_{\bullet *}$,
  where $[\cdot]_{\bullet *}$ denotes the step function: $R \longrightarrow \{0, 1\}$ component-wise.

In order to obtain a nonlinear support vector machine with a nonlinear separating surface, we make use of the following transformation based on the dual formulation for the conventional support vector machine $SVM_{\|\cdot\|_2^2}$ [29]:

$$w = A'Du \tag{9}$$

Substitution in (4) gives:

$$\begin{array}{lrcl} \min\limits_{u,\gamma,y} & \nu e'y + \|A'Du\| & & \\ \text{s.t.} & D(AA'Du - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \tag{10}$$

Replacing AA' by a nonlinear kernel $K(A, A') : R^{m\times n} \times R^{n\times m} \longrightarrow R^{m\times m}$, and $\|A'Du\|$ by a convex function of u on R^m, typically a norm, results in the following nonlinear generalized support vector machine [29]:
(GSVM)

$$\begin{array}{lrcl} \min\limits_{u,\gamma,y} & \nu e'y + \|u\|_p & & \\ \text{s.t.} & D(K(A, A')Du - e\gamma) + y & \geq & e \\ & y & \geq & 0, \end{array} \tag{11}$$

with the following nonlinear separating surface instead of a plane:

$$K(x', A')Du = \gamma. \tag{12}$$

Note that the linear SVM is the special case of (11) where $K(A, A') = AA'$. This nonlinear GSVM, which is capable of discriminating between elements of fairly complex sets, such as the black and white squares of a checkerboard [33, 23], will be used in the following sections. We note that even though the nonlinear separating surface (12) is capable of achieving fairly complex separation, it is not suited for feature selection, because the kernel $K(A, A')$ creates interactions between all the features through the variable u which is not present in linear support vector machine $SVM_{\|\cdot\|_1}$ (8). That is the reason we used to apply the linear $SVM_{\|\cdot\|_1}$ for feature selection and the nonlinear GSVM for the final separation.

2.2 Smooth Support Vector Machines

In this section we will apply the smoothing techniques which have been extensively used for solving many important problems [10, 11, 9, 12, 13, 18, 39, 15] to support vector machine. In our smooth approach, the square of 2-norm of the slack variable y is minimized with weight $\frac{\nu}{2}$ instead of the 1-norm of y as in (4). In addition the distance between the two bounding planes $x'w = \gamma \pm 1$ is measured in the $(n+1)$-dimensional space of $(w, \gamma) \in R^{n+1}$, that is $\frac{2}{\|(w,\gamma)\|_2}$. Measuring the margin in this $(n+1)$-dimensional space instead of R^n induces strong convexity and has little or no effect on the problem as was shown in [32]. Thus using twice the reciprocal squared of the margin instead, yields our modified SVM problem as follows:

$$\begin{array}{lll} \min\limits_{w,\gamma,y} & \frac{\nu}{2}y'y + \frac{1}{2}(w'w + \gamma^2) & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq \; e \\ & \quad y & \geq \; 0. \end{array} \tag{13}$$

Appling the **KKT** condition of the above quadratic programming problem [25], at a solution of problem (13), y is given by

$$y = (e - D(Aw - e\gamma))_+, \tag{14}$$

where, as defined earlier, $(\cdot)_+$ replaces negative components of a vector by zeros. Thus, we can replace y in (13) by $(e - D(Aw - e\gamma))_+$ and convert the SVM problem (13) into an equivalent SVM which is an unconstrained optimization problem as follows:

$$\min_{w,\gamma} \quad \tfrac{\nu}{2}\|(e - D(Aw - e\gamma))_+\|_2^2 + \tfrac{1}{2}(w'w + \gamma^2). \tag{15}$$

This problem is a strongly convex minimization problem without any constraints. It is easy to show that it has a unique solution. However, the objective function in (15) is not twice differentiable which precludes the use of a fast Newton method. We thus apply the smoothing techniques of [10, 11] and replace x_+ by a very accurate smooth approximation (see Lemma 2.2 below) that is given by $p(x, \alpha)$, the integral of the sigmoid function $\frac{1}{1+\varepsilon^{-\alpha x}}$ of neural networks [24], that is

$$p(x, \alpha) = x + \frac{1}{\alpha}\log(1 + \varepsilon^{-\alpha x}), \ \alpha > 0. \tag{16}$$

This p function with a smoothing parameter α is used here to replace the plus function of (15) to obtain a smooth support vector machine **(SSVM)** :

$$\min_{(w,\gamma)\in R^{n+1}} \Phi_\alpha(w, \gamma) \ := \ \min_{(w,\gamma)\in R^{n+1}} \frac{\nu}{2}\|p(e - D(Aw - e\gamma), \alpha)\|_2^2 + \frac{1}{2}(w'w + \gamma^2). \tag{17}$$

We will now show that the solution of problem (13) is obtained by solving problem (17) with α approaching infinity. We take advantage of the twice differentiable property of the objective function of (17) to utilize a quadratically convergent algorithm for solving the smooth support vector machine (17).

We begin with a simple lemma that bounds the square difference between the plus function $(x)_+$ and its smooth approximation $p(x, \alpha)$.

Lemma 2.2 For $x \in R$ and $|x| < \rho$: $p(x, \alpha)^2 - (x_+)^2 \leq (\frac{\log 2}{\alpha})^2 + \frac{2\rho}{\alpha}\log 2$, where $p(x, \alpha)$ is the p function of (16) with smoothing parameter $\alpha > 0$.
Proof We consider two cases. For $0 < x < \rho$,

$$\begin{aligned} p(x, \alpha)^2 - (x_+)^2 &= \frac{1}{\alpha^2}\log^2(1 + \varepsilon^{-\alpha x}) + \frac{2x}{\alpha}\log(1 + \varepsilon^{-\alpha x}) \\ &\leq (\frac{\log 2}{\alpha})^2 + \frac{2\rho}{\alpha}\log 2. \end{aligned}$$

For $-\rho < x \leq 0,\ p(x,\alpha)^2$ is a monotonically increasing function, so we have

$$p(x,\alpha)^2 - (x_+)^2 = p(x,\alpha)^2 \leq p(0,\alpha)^2 = (\frac{\log 2}{\alpha})^2.$$

Hence, $p(x,\alpha)^2 - (x_+)^2 \leq (\frac{\log 2}{\alpha})^2 + \frac{2\rho}{\alpha}\log 2.$ □

We now show that as the smoothing parameter α approaches infinity the unique solution of our smooth problem (17) approaches the unique solution of the equivalent SVM problem (15). We shall do this for a function $f(x)$ given in (18) below that subsumes the equivalent SVM function of (15) and for a function $g(x,\alpha)$ given in (19) below which subsumes the SSVM function of (17).

Theorem 2.3 Let $A \in R^{m\times n}$ and $b \in R^{m\times 1}$. Define the real valued functions $f(x)$ and $g(x,\alpha)$ in the n-dimensional real space R^n:

$$f(x) = \frac{1}{2}\|(Ax-b)_+\|_2^2 + \frac{1}{2}\|x\|_2^2 \tag{18}$$

and

$$g(x,\alpha) = \frac{1}{2}\|p((Ax-b),\alpha)\|_2^2 + \frac{1}{2}\|x\|_2^2, \tag{19}$$

with $\alpha > 0$.

(i) There exists a unique solution $\bar{x}$ of $\min\limits_{x\in R^n} f(x)$ and a unique solution $\bar{x}_\alpha$ of $\min\limits_{x\in R^n} g(x,\alpha)$.

(ii) For all $\alpha > 0$, we have the following inequality:

$$\|\bar{x}_\alpha - \bar{x}\|_2^2 \leq \frac{m}{2}((\frac{\log 2}{\alpha})^2 + 2\xi\frac{\log 2}{\alpha}), \tag{20}$$

where ξ is defined as follows:

$$\xi \quad = \quad \max_{1\leq i\leq m} |(A\bar{x}-b)_i|. \tag{21}$$

Thus, $\bar{x}_\alpha$ converges to $\bar{x}$ as α goes to infinity with an upper bound given by (20).

Proof (i) To show the existence of unique solutions, we know that since $x_+ \leq p(x,\alpha)$, the level sets $L_\nu(g(x,\alpha))$ and $L_\nu(f(x))$ satisfy

$$L_\nu(g(x,\alpha)) \subseteq L_\nu(f(x)) \subseteq \{x \mid \|x\|_2^2 \leq 2\nu\}, \tag{22}$$

for $\nu \geq 0$. Hence $L_\nu(g(x,\alpha))$ and $L_\nu(f(x))$ are compact subsets in R^n and the problems $\min\limits_{x\in R^n} f(x)$ and $\min\limits_{x\in R^n} g(x,\alpha)$ have solutions. By the strong convexity of $f(x)$ and $g(x,\alpha)$ for all $\alpha > 0$, these solutions are unique.

(ii) To establish convergence, we note that by the first order optimality condition and strong convexity of $f(x)$ and $g(x,\alpha)$ we have that

$$f(\bar{x}_\alpha) - f(\bar{x}) \geq \nabla f(\bar{x})(\bar{x}_\alpha - \bar{x}) + \frac{1}{2}\|\bar{x}_\alpha - \bar{x}\|_2^2 = \frac{1}{2}\|\bar{x}_\alpha - \bar{x}\|_2^2, \tag{23}$$

$$g(\bar{x},\alpha) - g(\bar{x}_\alpha,\alpha) \geq \nabla g(\bar{x}_\alpha,\alpha)(\bar{x} - \bar{x}_\alpha) + \frac{1}{2}\|\bar{x} - \bar{x}_\alpha\|_2^2 = \frac{1}{2}\|\bar{x} - \bar{x}_\alpha\|_2^2. \tag{24}$$

Since the p function dominates the plus function we have that $g(x,\alpha) - f(x) \geq 0$ for all $\alpha > 0$. Adding (18) and (19) and using this fact gives:

$$\begin{array}{rcl} \|\bar{x}_\alpha - \bar{x}\|_2^2 & \leq & (g(\bar{x},\alpha) - f(\bar{x})) - (g(\bar{x}_\alpha,\alpha) - f(\bar{x}_\alpha)) \\ & \leq & g(\bar{x},\alpha) - f(\bar{x}) \\ & = & \frac{1}{2}\|p((A\bar{x} - b),\alpha)\|_2^2 - \frac{1}{2}\|(A\bar{x} - b)_+\|_2^2. \end{array} \tag{25}$$

Application of Lemma 2.2 gives :

$$\|\bar{x}_\alpha - \bar{x}\|_2^2 \;\leq\; \tfrac{m}{2}((\tfrac{\log 2}{\alpha})^2 + 2\xi\tfrac{\log 2}{\alpha}), \tag{26}$$

where ξ is a fixed positive number defined in (21). The last term in (26) will converge to zero as α goes to infinity. Thus $\bar{x}_\alpha$ converges to $\bar{x}$ as α goes to infinity with an upper bound given by (20). $\square$

We will now describe a Newton-Armijo algorithm for solving the smooth problem (17).

2.3 A Newton-Armijo Algorithm for Smooth Support Vector Machines

By making use of the results of the previous section and taking advantage of the twice differentiability of the objective function of problem (17), we prescribe a quadratically convergent Newton algorithm with an Armijo stepsize

[1, 17, 3] that makes the algorithm globally convergent.

Algorithm 2.4 Newton-Armijo Algorithm for SSVM (17)
Start with any $(w^0, \gamma^0) \in R^{n+1}$. Having (w^i, γ^i), stop if the gradient of the objective function of (17) is zero, that is $\nabla\Phi_\alpha(w^i, \gamma^i) = 0$. Else compute (w^{i+1}, γ^{i+1}) as follows:

(i) **Newton Direction**: Determine direction $d^i \in R^{n+1}$ by setting equal to zero the linearization of $\nabla\Phi_\alpha(w, \gamma)$ around (w^i, γ^i) which gives $n+1$ linear equations in $n+1$ variables:

$$\nabla^2\Phi_\alpha(w^i, \gamma^i)d^i = -\nabla\Phi_\alpha(w^i, \gamma^i)'. \tag{27}$$

(ii) **Armijo Stepsize** [1]: Choose a stepsize $\lambda_i \in R$ such that:

$$(w^{i+1}, \gamma^{i+1}) = (w^i, \gamma^i) + \lambda_i d^i \tag{28}$$

where $\lambda_i = \max\{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ such that :

$$\Phi_\alpha(w^i, \gamma^i) - \Phi_\alpha((w^i, \gamma^i) + \lambda_i d^i) \geq -\delta\lambda_i \nabla\Phi_\alpha(w^i, \gamma^i)d^i \tag{29}$$

where $\delta \in (0, \frac{1}{2})$.

Note that a key difference between our smoothing approach and that of the classical SVM [40, 14] is that we are solving here a linear system of equations (27) instead of solving a quadratic program as is the case with the classical SVM. Furthermore, we can show that our smoothing Algorithm 2.4 converges globally to the unique solution and takes a pure Newton step after a finite number of iterations automatically. This leads to the following quadratic convergence result:

Theorem 2.5 Let $\{(w^i, \gamma^i)\}$ be a sequence generated by Algorithm 2.4 and $(\bar{w}, \bar{\gamma})$ be the unique solution of problem (17).

(i) The sequence $\{(w^i, \gamma^i)\}$ converges to the unique solution $(\bar{w}, \bar{\gamma})$ from any initial point (w^0, γ^0) in R^{n+1}.

(ii) For any initial point (w^0, γ^0), there exists an integer $\bar{i}$ such that the stepsize λ_i of Algorithm 2.4 equals 1 for $i \geq \bar{i}$ and the sequence $\{(w^i, \gamma^i)\}$ converges to $(\bar{w}, \bar{\gamma})$ quadratically.

Although this theorem can be inferred from [17, Therem 6.3.4, pp 123-125], we give its proof in the Appendix for the sake of completeness. We note here that even though the Armijo stepsize is needed to guarantee that Algorithm 2.4 is globally convergent, in most of our numerical tests Algorithm 2.4 converged from any starting point without the need for an Armijo stepsize.

2.4 SSVM with a Nonlinear Kernel

In Section 2.2 the smooth support vector machine formulation constructed a linear separating surface (5) for our classification problem. We now describe how to construct a nonlinear separating surface which is implicitly defined by a kernel function. Like we have done in Section 2.2, we simply change the objective function in $GSVM$ (11) by replacing $\nu e'y$ by $\frac{\nu}{2}y'y$ and $\|u\|_p$ by $\|(u,\gamma)\|_2^2$ in our nonlinear formulation:

$$\begin{array}{lrcl} \min\limits_{u,\gamma,y} & \frac{\nu}{2}y'y + \frac{1}{2}(u'u+\gamma^2) & & \\ \text{s.t.} & D(K(A,A')Du - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \tag{30}$$

We repeat the same arguments as in Section 2.2 to obtain the SSVM with a nonlinear kernel $K(A, A')$:

$$\min_{u,\gamma} \ \frac{\nu}{2}\|p(e - D(K(A,A')Du - e\gamma), \alpha)\|_2^2 + \frac{1}{2}(u'u + \gamma^2), \tag{31}$$

where $K(A, A')$ is a kernel map from $R^{m\times n} \times R^{n\times m}$ to $R^{m\times m}$. We note that this problem, which is capable of generating highly nonlinear separating surfaces, still retains the strong convexity and differentiability properties for any arbitrary kernel. All of the results of the previous sections still hold. Hence we can apply the Newton-Armijo Algorithm 2.4 directly to solve (31).

We turn our attention now to numerical testing.

2.5 Numerical Results and Comparisons

We demonstrate the effectiveness and speed of the smooth support vector machine (SSVM) approach by comparing it numerically with the robust linear program (RLP) algorithm [2], the feature selection concave minimization (FSV) algorithm, the support vector machine using the 1-norm approach

($SVM_{\|\cdot\|_1}$) and the classical support vector machine ($SVM_{\|\cdot\|_2^2}$) [40, 5, 14]. All our experiments were run on the University of Wisconsin Computer Sciences Department Ironsides cluster. This cluster of four Sun Enterprise E6000 machines consists of 16 UltraSPARC II 248 MHz processors and 2 gigabytes of RAM on each node, resulting in a total of 64 processors and 8 gigabytes of RAM. All linear programming formulations were solved using the CPLEX package [16] called from within MATLAB [35].

In order to evaluate how well each algorithm generalizes to future data, we performed tenfold cross-validation on each data set [37]. To evaluate the efficacy of SSVM, we compared computational times of SSVM with those of RLP and $SVM_{\|\cdot\|_1}$. We ran all tests on six publicly available datasets: the Wisconsin Prognostic Breast Cancer Database [34] and five datasets from the Irvine Machine Learning Database Repository [36]. It turned out that tenfold testing correctness of the SSVM is the highest for these five methods on all datasets tested as indicated in Table 1. For performance speed, the SSVM was always faster than RLP and $SVM_{\|\cdot\|_1}$, except on the WPBC(60 months) dataset where it is slower than the RLP. Note however that the number of examples of WPBC(60 months) is the smallest one for these six datasets. Times were not available for $SVM_{\|\cdot\|_2^2}$ and FSV from [5]. The numerical results of FSV and $SVM_{\|\cdot\|_2^2}$ are quoted from [5]. We summarize all these results in Table 1.

To test the effectiveness of the SSVM in generating a highly nonlinear separating surface, we tested it on the checkerboard dataset of [20] and two-spiral dataset. They are depicted in Figure 2 and Figure 5 respectively. We used the following symmetric sixth degree polynomial kernel and a symmetric sinusoidal kernel introduced in [32] for the checkerboard and Gaussian kernel for the two-spiral dataset. Utilized they within a smooth support vector machine formulation (31).

Polynomial Kernel $((\frac{A}{\lambda}-\rho)(\frac{A}{\lambda}-\rho)'-\mu)^d_\bullet$

Sinusoidal Kernel $(sin(\frac{A}{\lambda}-\rho)sin(\frac{A}{\lambda}-\rho)'-\mu)^d_\bullet$

The results and values of the parameters λ, ρ, μ used are given in Figure 3 and Figure 4 as well as that of the parameter ν of the nonlinear SSVM (31) for the checkerboard problem. We note that the boundaries of the checkerboard are as sharp as those of [32], obtained by a linear programming solution, and considerably sharper than those of [20], obtained by a Newton approach to a quadratic programming formulation. The result of two-spiral problem is shown in Figure 6.

2.6 Summary and Future Work

We have proposed a new formulation, SSVM, which is a smooth unconstrained optimization reformulation of the traditional quadratic program associated with a SVM. It can be shown that the unique solution of SSVM will converge to the solution of modified support vector machine as the smoothing parameter goes to infinity. We have prescribed a very fast Newton-Armijo algorithm to solve the SSVM. We also extended SSVM to nonlinear separation surfaces by using kernel techniques. The numerical results show that the SSVM is faster than other methods and has better generalization ability. Future work includes other smooth formulations, feature selection via SSVM and smooth support vector regression. We also plan to use the row and column generation chunking algorithms of [6, 31] for solving the SSVM for extremely large classification problems.

Ten-Fold Training Correctness, % Ten-Fold Testing Correctness, % Ten-Fold Computational Time, *sec.*					
Data Set	**Method**				
$m \times n$	SSVM	RLP	$SVM_{\|\cdot\|_1}$	$SVM_{\|\cdot\|_2^2}$	FSV
WPBC(24 months) 155×32	86.16 **83.47** **4.05**	85.23 67.12 6.47	74.40 71.08 6.25	82.86 75.46 -	73.97 66.42 -
WPBC(60 months) 110×32	80.20 **68.18** 3.15	87.58 63.50 **2.81**	71.21 66.23 3.72	75.54 66.21 -	70.70 67.05 -
Ionosphere 351×34	94.12 **89.63** **8.84**	94.78 86.04 10.38	88.92 86.10 42.41	94.56* 85.75* -	90.47 84.07 -
Cleveland 297×13	87.32 **86.13** **4.61**	86.31 83.87 12.74	85.30 84.55 18.71	84.70* 75.86* -	83.57 80.94 -
Pima Indians 768×8	78.11 **78.12** **9.93**	76.48 76.16 195.21	75.52 74.47 286.59	* * -	75.22 74.60 -
BUPA Liver 345×6	70.37 **70.33** **4.19**	68.98 64.34 17.26	67.83 64.03 19.94	60.22* 60.95* -	68.18 65.20 -

Table 1: **Ten-fold cross-validation correctness results on six datasets by five different methods. Bold type indicates the best result and asterisk (*) indicates that the full experiment had not been carried out because of excessive computational time [5]. Dash (-) indicates computational times were not given in the cited reference [5].**

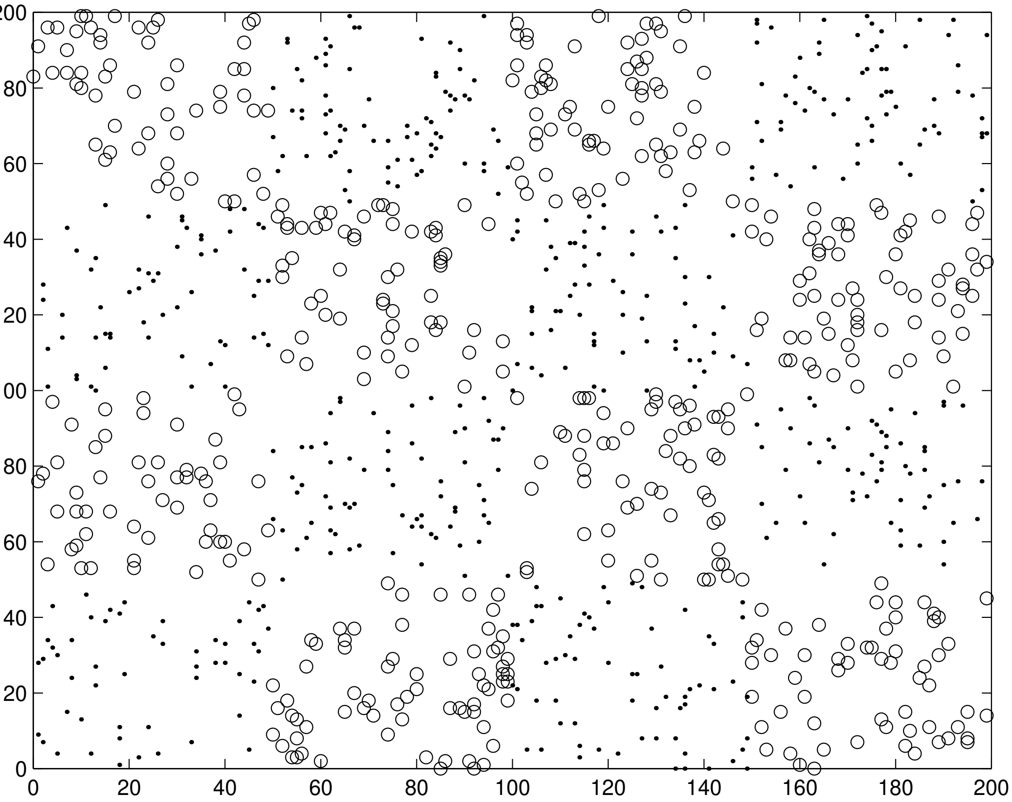

Figure 2: **Checkerboard training dataset**

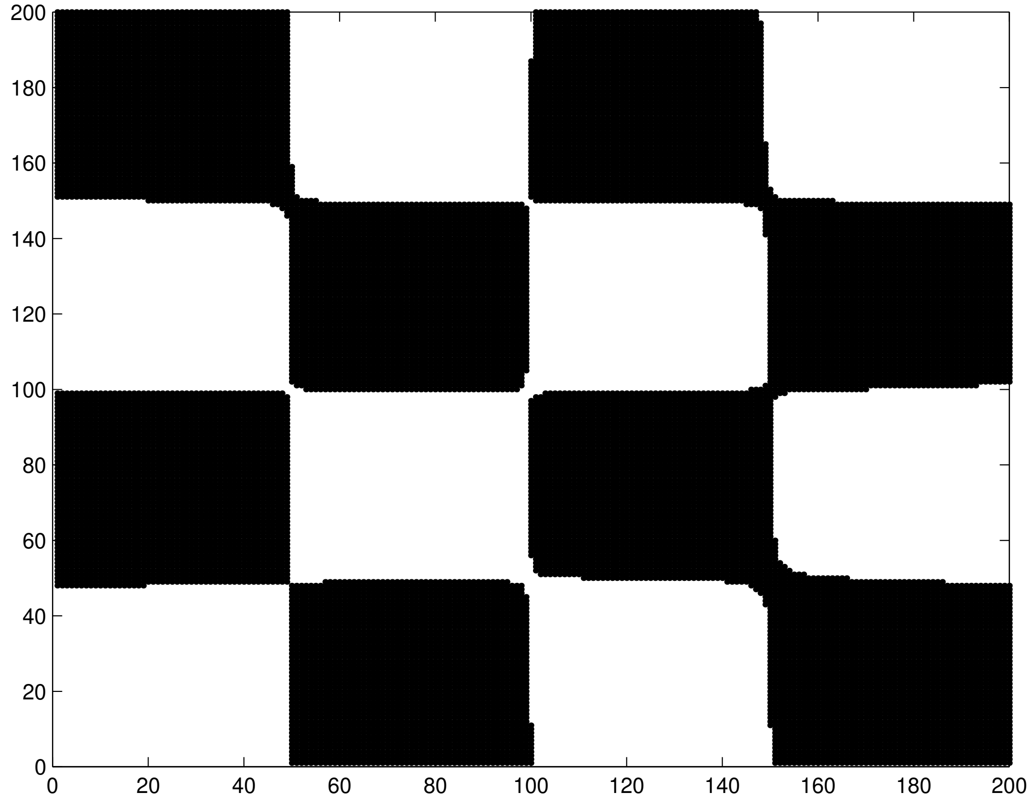

Figure 3: **Indefinite sixth degree polynomial kernel separation of the checkerboard dataset** $(\nu = 320,000, \lambda = 100, \rho = 1, d = 6, \mu = 0.5)$

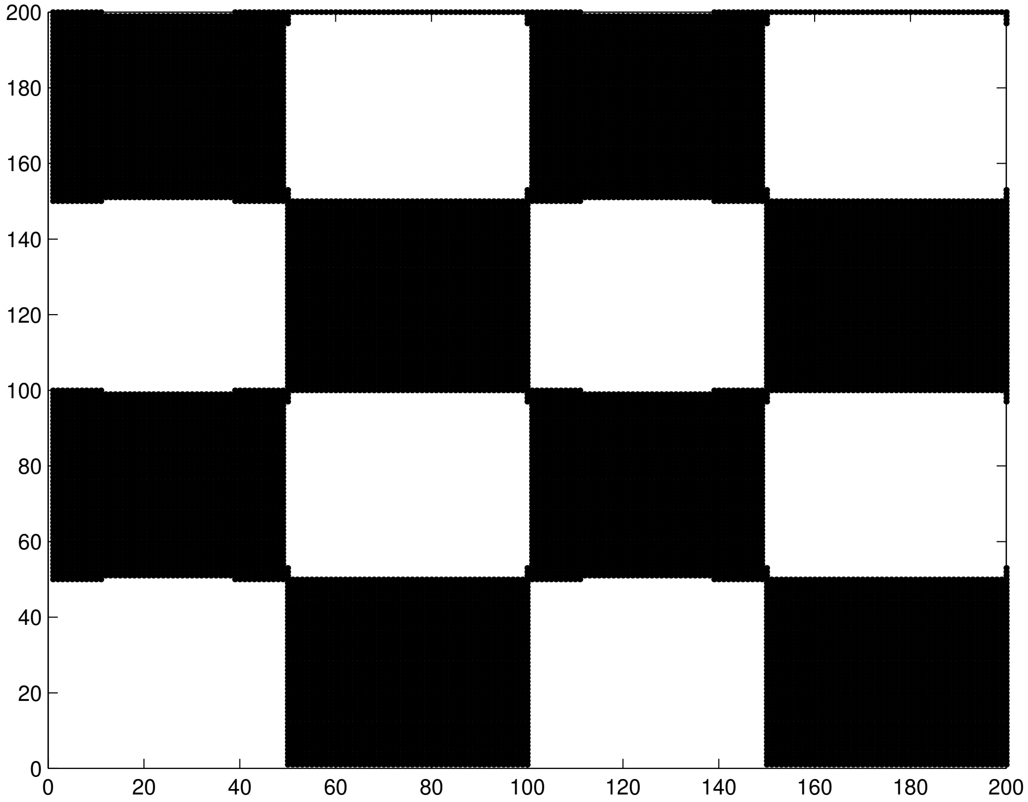

Figure 4: **Indefinite sinusoidal kernel separation of the checkerboard dataset** $\nu = 2, \lambda = \frac{50}{\pi}, \rho = 2\pi, d = 2, \mu = 1)$

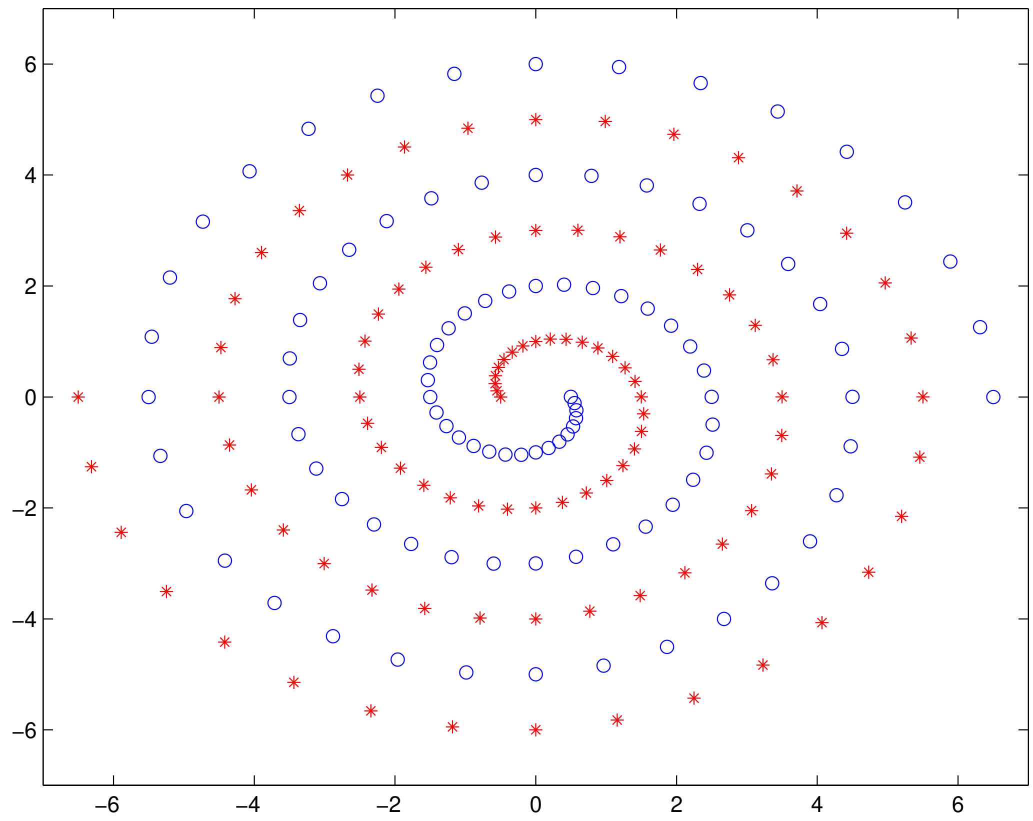

Figure 5: **Two spirals dataset**

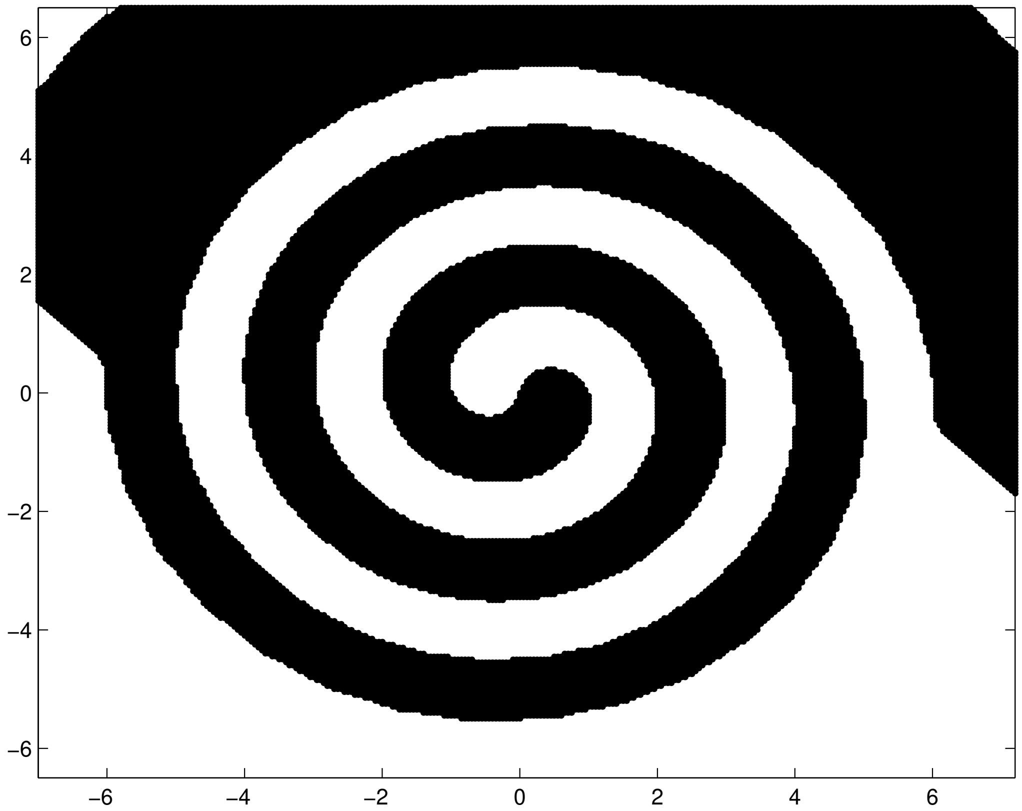

Figure 6: **Gaussian kernel separation of two spirals dataset** $(\nu = 100, \mu = 4)$

3 An Application: Breast Cancer Prognosis

In this section of our work we shall use the support vector machines described in the previous section to discriminate between 253 breast cancer patients. The dataset for these patients is publicly available [41]. Our criterion for discrimination will be the absence or presence of metastasized lymph nodes (1 to 30 nodes) removed from the patient's armpit during surgery. Lymph node metastasis is a strong indicator for chemotherapy. The total number of features used to constitute the n-dimensional space in which the separation is accomplished is made up of the mean, standard error and maximum (worst) value of ten cytological nuclear measurements of size, shape and texture taken from the patients breast by a non-surgical fine needle aspirate procedure [38, 34] together with the tumor size excised from the patient's breast during surgery. Taken together these constitute the 31-dimensional feature space over which classification according to lymph node metastasis will be achieved. We shall use the linear support vector machine, $SVM_{\|\cdot\|_1}$ to select six features out of 31 features and then apply to the GSVM with a Gaussian kernel to classify patients into three groups.

We introduce some statistics tools of survival analysis which will be used in our study.

3.1 Kaplan-Meier Survival Curve and Logrank Test

Kaplan-Meier survival probability represents expected percent of surviving patients as a function of time which can be expressed as follows:

$$S_{t_j} = \begin{cases} 1, & j = 0 \\ S_{t_{j-1}}(1 - \text{risk}_{t_j}), & j = 1, 2, \ldots, \end{cases} \tag{32}$$

where risk_{t_j} is the instantaneous risk of death at jth observed time which is estimated by the ratio between the number of new patients died at time t_j and the number of patients remaining in study group at time t_j. The Kaplan-Meier survival curve is typically plotted as a step function that stars with a horizontal line at the initial survival probability, namely $S_{t_0} = 1$, and then steps down to the other survival probabilities as we move from one observed time to the next.

In our study, we would like to compare the Kaplan-Meier curves of two groups, say **G**ood and **P**oor. Intuitively, if group **G**ood's survival curve dominates group **P**oor, then group **G**ood has a better survival rate than group

Poor. Is this difference significant? We employ the logrank test [21] to answer this question. The logrank statistic is a comparative measure of the survival times of two groups which calculates a weighted sum of the total difference through all observed times which provides an overall comparison of different Kaplan-Meier survival curves. We evaluate the difference of the mortality ratio between the two groups: $\frac{m_{1j}}{m_{2j}}$, and the population ratio between the two groups: $\frac{n_{1j}}{n_{2j}}$ at the jth observed time. Here, m_{ij} is the number of deaths in group i and n_{ij} is the number of people in group i at the jth observed time. The logrank statistic for two groups has a Chi-squared distribution with degree of freedom one [21]. Larger logrank statistic indicates significantly different survival curves. That corresponds to a small p-value. The p-value is the probability that the two survival curves are statistically the same. In our experience in comparing two Kaplan-Meier survival curves, the curves are well separated when the p-value is less than 0.05 but converse is not true (*see* Figure 11). In the logrank test, we tend to reject the null-hypothesis (survival curves are the same) if the p-value is small enough. (Conventionally, p-value < 0.05. It is equivalent to the logrank statistic > 7.89.)

3.2 Breast Cancer Prognosis

Before proceeding with our classification process we exhibit Kaplan-Meier survival curves [21] for 140 of the 253 patients who had chemotherapy and for the remaining 113 patients that did not have chemotherapy in Figure 7. The solid curve gives the percent of surviving patients as a function of time for the group that received chemotherapy, while the dashed curve gives the corresponding survival curve for the group that did not receive chemotherapy. One possible interpretation of this curve is that the patients not receiving chemotherapy and who have a better survival curve than those who are receiving chemotherapy, are not being compromised by not receiving chemotherapy. Another interpretation is that most patients who had chemotherapy were preselected for having lymph node metastasis and hence had inherently worse cancers. The p-value based on the logrank statistic [21] relating these two curves is 0.0018. Hence these two curves are statistically distinct.

Our next result is the classification of the 253 patients into three prognostic groups as follows:

1. **Good Prognosis** "Node-negative patients": Patients with no metas-

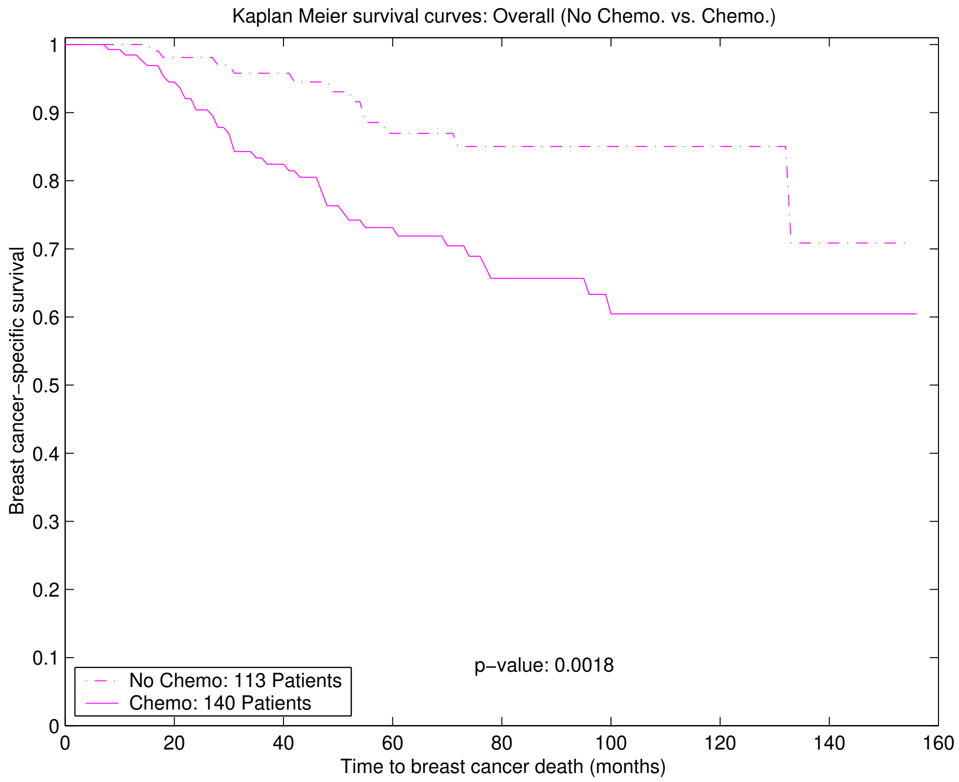

Figure 7: Kaplan-Meier survival curve (solid line) for 140 of the 253 patients who received chemitherapy and the corresponding survival curve (dashed line) for the remaining 113 patients that did not receive chemotherapy. p-value=0.0018.

tasized lymph nodes.

2. **Intermediate Prognosis** "Node-positive patients" with 1 to 4 metastasized lymph nodes.

3. **Poor Prognosis** "Node-positive patients" with more than 4 metastasized lymph nodes.

To achieve this classification we first used $SVM_{\|\cdot\|_1}$ (8) to separate node-negative patients from node-positive patients. By using a sufficiently small value of ν in (8), $\nu = 0.02$, the formulation $SVM_{\|\cdot\|_1}$ (8) suppressed 25 of the 31 features and selected a total of 6 features as follows:

1. **Mean Area**

2. **Standard Error of Area**

3. **Worst Area**

4. **Worst Texture**

5. **Worst Perimeter**

6. **Tumor Size** (from surgery)

Note that the first five are cytological features obtained from the fine needle aspirate taken during diagnosis.

We next used the nonlinear support vector machine GSVM (11) with a Gaussian kernel with $\nu = 0.096$ to achieve our three-class separation as follows:

1. **GPG**: Good prognosis group (0 lymph node metastasis)
   was separated from all the rest for which:

2. **IPG**: Intermediate prognosis group (1-4 lymph node metastasis)
   was separated from:

   PPG: Poor prognosis group (lymph node metastasis > 4).

Note that the separation in Step 2 above included some node-negative patients in IPG because the separation of Step 1 is not perfect and GPG does not include *all* node-negative patients. Tenfold cross validation correctness [37] for the above separation was 72.8%.

Based on the above separation we constructed three Kaplan-Meier survival curves for each of the three prognostic groups: GPG, IPG and PPG as shown in Figure 8. Note that the three curves are very well separated and the p-values between the three groups are as follows:

1. GPG-IPG p-value=0.0090
2. GPG-PPG p-value=9.5769e-06
3. IPG-PPG p-value=0.0037

Note that these p-values are substantially smaller than the conventional 0.05 cutoff used to reject the null-hypothesis (survival curves are the same). Hence these curves are indeed statistically significantly different from each other.

It is worth noting within each of the three prognosis groups how many patients received and did not receive chemotherapy:

1. **GPG**: 19 patients received chemotherapy; 51 did not (27.2%;72.8%)
2. **IPG**: 72 patients received chemotherapy; 55 did not (56.7%;43.3%)
3. **PPG**: 49 patients received chemotherapy; 7 did not (87.5%;12.5%)

Within each prognosis group, there was no statistically significant difference based on p-values, between the survival curves of patients who took and did not take chemotherapy.

In contrast to the three survival curves based on the GPG, IPG and PPG groups, we constructed three other survival curves based strictly on the number of metastasized lymph nodes, the conventional indicator for chemotherapy prescription, as shown in Figure 12. It is interesting to observe the large p-values between the node-negative survival curve and the 1-4 lymph node survival curve (p-value=0.1137), and between the latter curve and the 5-30 lymph node survival curve (p-value=0.028). This indicates that our criteria for GPG-IPG-PPG classification based on the six features selected by $SVM_{\|\cdot\|_1}$, which do not include lymph node status, yield statistically more distinct survival curves than those based on lymph node status by itself.

3.3 Summary and Future Work

Based on the results presented in this work, the following conclusions can be made:

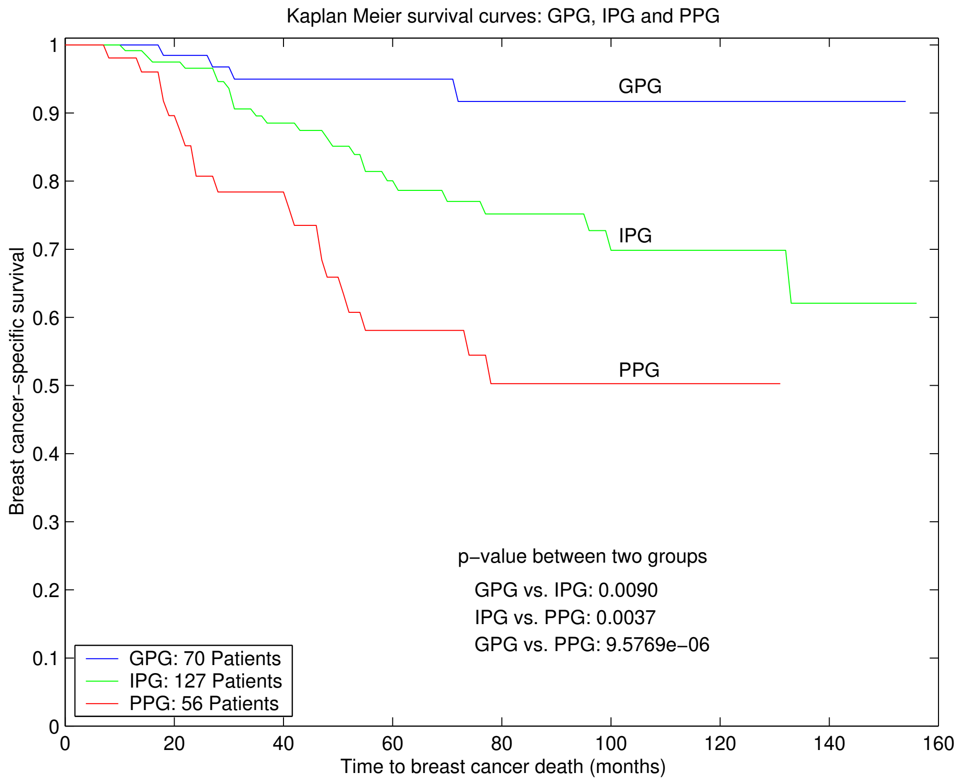

Figure 8: Kaplan-Meier survival curves for GPG, IPG and PPG: the good, intermediate and poor prognosis groups. The classification criterion is based on the number of metastasized lymph nodes. The features used for the classification are: mean area, standard error of area, worst texture, worst perimeter and tumor size. The pairwise p-values are: GPG-IPG= 0.0090, GPG-PPG= 9.5769e-06, IPG-PPG= 0.0037.

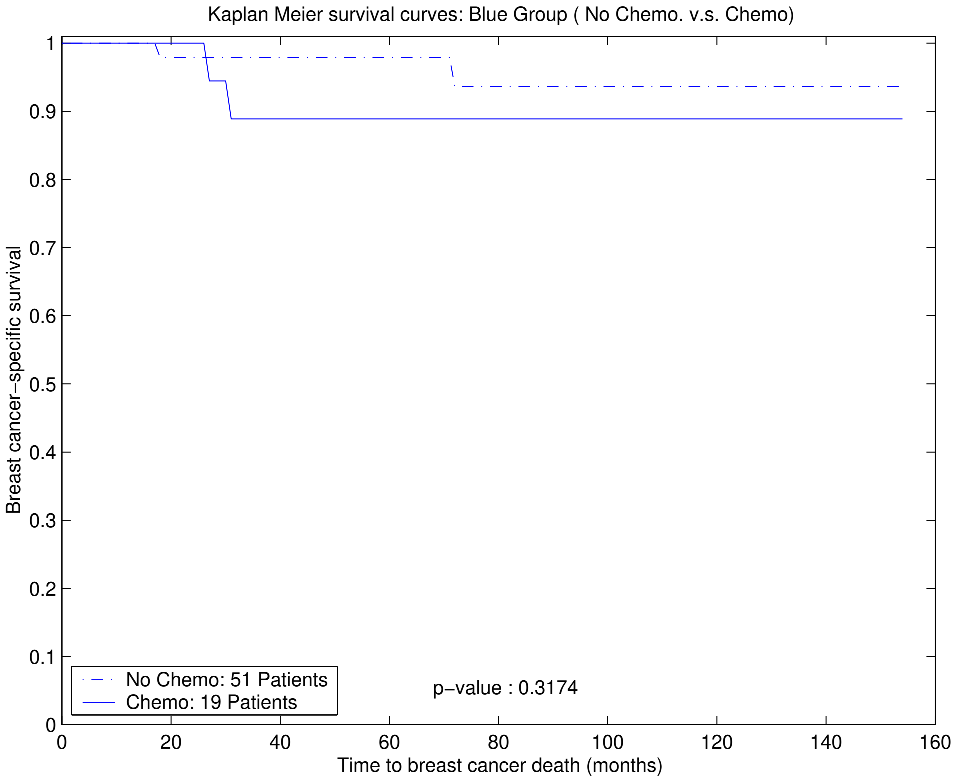

Figure 9: Kaplan-Meier survival curve (solid line) for 51 of the 70 patients in GPG who received chemotherapy and the corresponding survival curve (dashed line) for the remaining 19 patients that did not receive chemotherapy. $p-$value $= 0.3174$.

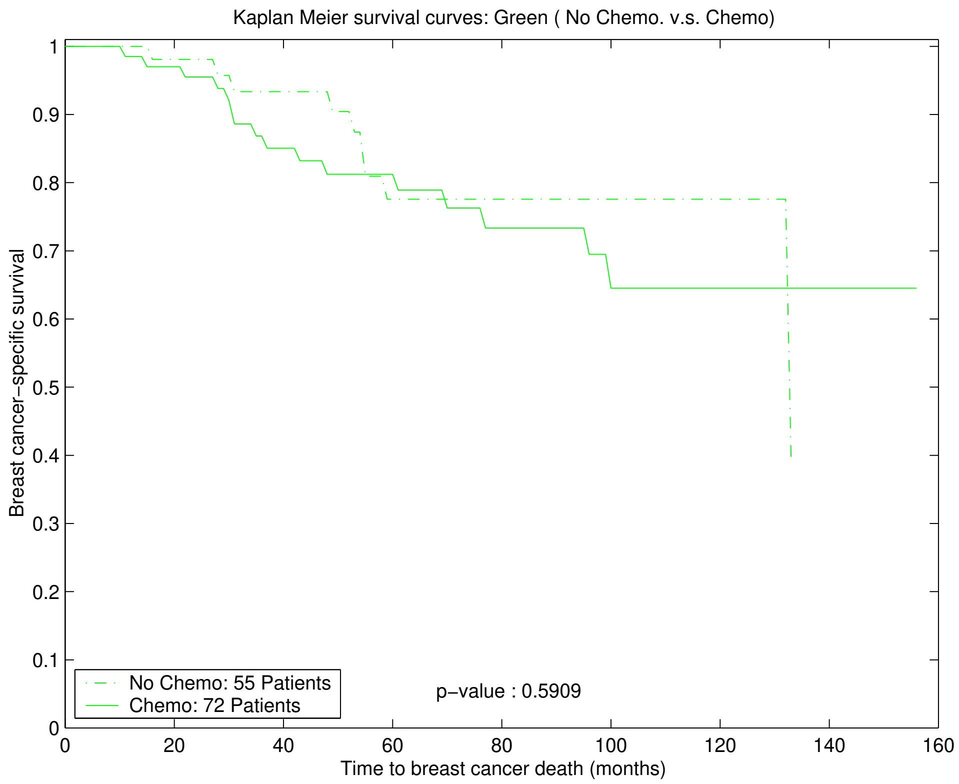

Figure 10: Kaplan-Meier survival curve (solid line) for 72 of the 127 patients in IPG who received chemotherapy and the corresponding survival curve (dashed line) for the remaining 55 patients that did not receive chemotherapy. $p-$value $= 0.5909$.

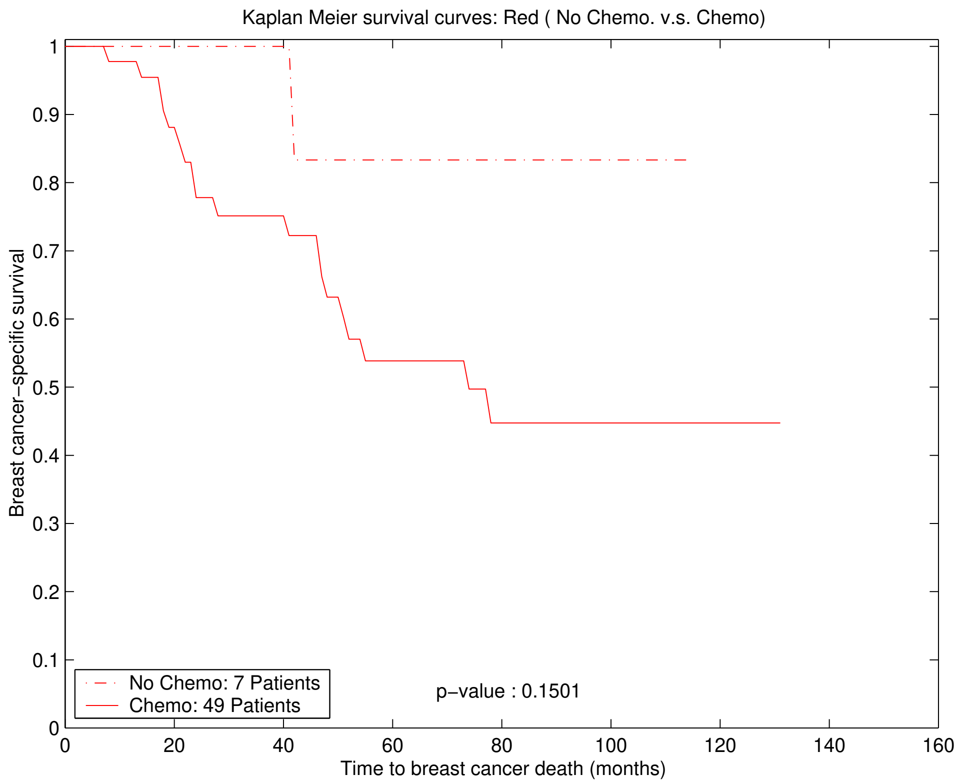

Figure 11: Kaplan-Meier survival curve (solid line) for 49 of the 56 patients in PPG who received chemotherapy and the corresponding survival curve (dashed line) for the remaining 7 patients that did not receive chemotherapy. $p-$value $= 0.1501$.

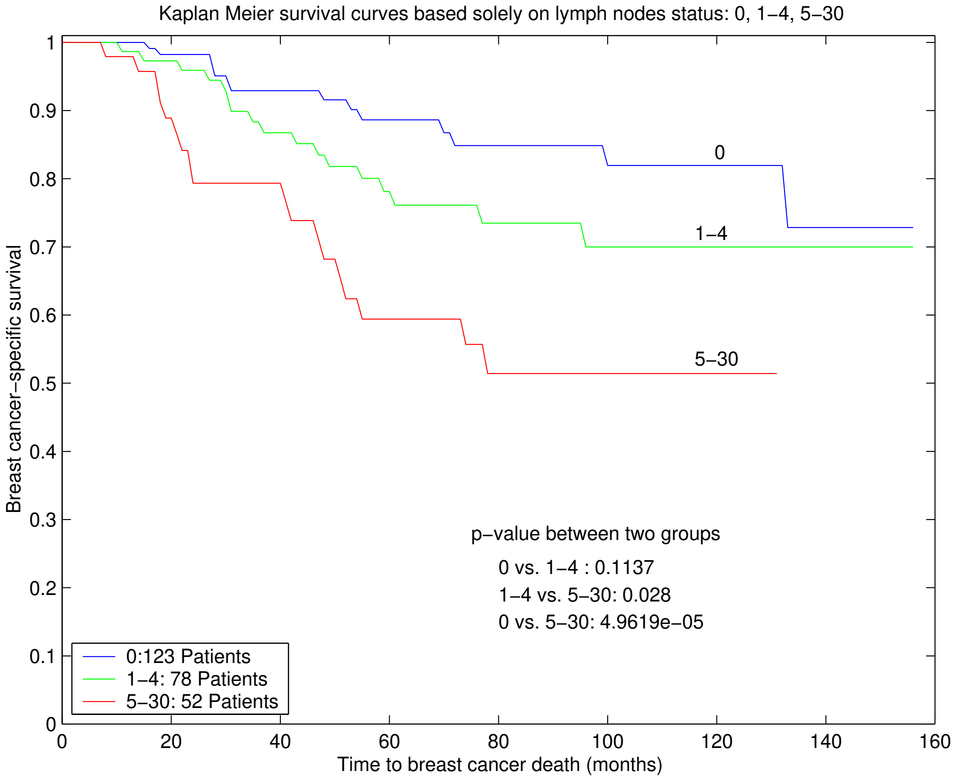

Figure 12: Kaplan-Meier survival curves for 3 groups based strictly on lymph node status: 0, [1-4] and [5-30]. The pairwise p-values are: [0]-[1-4]:0.1137, [1-4]-[5-30]:0.028, [0]-[5-30]:4.9619e-05.

- Support vector machines can be used to:
  - Select features that are important for prognosis.
  - Classify breast cancer patients into three groups with well separated survival curves.
- Most (72.8%) patients in good group received no chemotherapy.
- Most (87.5%) patients in poor group received chemotherapy.
- Just over half the patients (56.7%) in intermediate group received chemotherapy.
- Survival curves and possible chemotherapy decision for new patients can be determined by assigning them to one of the three groups based on five nuclear features and tumor size without the need for lymph node status.

Future work includes prognosis depending on other classification criteria such as whether patients had chemotherapy or not and the type of chemotherapy they had, as well as other features such as age.

4 Appendix

We establish here global quadratic convergence for a Newton-Armijo Algorithm for solving the unconstrained minimization problem:

$$\min_{x \in R^n} f(x), \tag{33}$$

where f is a twice differentiable real valued function on R^n satisfying conditions (i) and (ii) below. This result, which can also be deduced from [17, Theorem 6.3.4, pp 123-125] is given here for completeness.

Newton-Armijo Algorithm for Unconstrained Minimization Let f be a twice differentiable real valued function on R^n. Start with any $x^0 \in R^n$. Having x^i, stop if the gradient of f at x^i is zero, that is $\nabla f(x^i) = 0$. Else compute x^{i+1} as follows:

(i) **Newton Direction**: Determine direction $d^i \in R^n$ by setting equal to zero, the linearization of $\nabla f(x)$ around x^i which gives n linear equations in n variables:

$$\nabla^2 f(x^i)d^i = -\nabla f(x^i)'. \tag{34}$$

(ii) **Armijo Stepsize** [1]: Choose a stepsize $\lambda_i \in R$ such that:

$$x^{i+1} = x^i + \lambda_i d^i \tag{35}$$

where $\lambda_i = \max\{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ such that :

$$f(x^i) - f(x^i + \lambda_i d^i) \geq -\delta\lambda_i \nabla f(x^i)d^i \tag{36}$$

where $\delta \in (0, \frac{1}{2})$.

Global Quadratic Convergence Theorem Let f be a twice differentiable real valued function on R^n and such that :

(i) $\|\nabla^2 f(y) - \nabla^2 f(x)\|_2 \leq \kappa\|y - x\|_2$, for all $x, y \in R^n$ and some $\kappa > 0$,

(ii) $y\nabla^2 f(x)y \geq \mu\|y\|_2^2$, for all $x, y \in R^n$ and some $\mu > 0$.

Let $\{x^i\}$ be a sequence generated by the Newton-Armijo Algorithm. Then:

(a) The sequence $\{x^i\}$ converges to $\bar{x}$ such that $\nabla f(\bar{x}) = 0$, where $\bar{x}$ is the unique global minimizer of f on R^n.

(b) $\|x^{i+1} - \bar{x}\|_2 \leq \frac{\kappa}{2\mu}\|x^i - \bar{x}\|_2^2$, for $i \geq \bar{i}$ for some $\bar{i}$.

Proof (a) Since

$$f(x^i) - f(x^{i+1}) \geq \delta\lambda_i \nabla f(x^i)\nabla^2 f(x^i)^{-1}\nabla f(x^i)' \geq 0, \tag{37}$$

the sequence $\{x^i\}$ is contained in the compact level set:

$$L_{f(x^0)}(f(x)) \subseteq \{x| \ \|x - x^0\|_2 \leq \frac{2}{\mu}\|\nabla f(x^0)\|_2\}. \tag{38}$$

Hence

$$y\nabla^2 f(x^i)y \leq \nu\|y\|_2^2, \ \text{for all } i = 0, 1, 2, \ldots, \ y \in R^n, \ \text{for some } \nu > 0. \tag{39}$$

It follows that

$$-\nabla f(x^i)d^i = \nabla f(x^i)\nabla^2 f(x^i)^{-1}\nabla f(x^i)' \geq \frac{1}{\nu}\|\nabla f(x^i)\|^2, \tag{40}$$

and by Theorem 2.1, Example 2.2 (iv) of [26] we have that each accumulation point of $\{x^i\}$ is stationary. Since all accumulation points of $\{x^i\}$ are identical to the global unique minimizer of the strongly convex function f, it follows that $\{x^i\}$ converges to $\bar{x}$ such that $\nabla f(\bar{x}) = 0$ and $\bar{x}$ is the unique global minimizer of f on R^n.

(b) We first show that the Armijo inequality holds from a certain i_0 onward with $\lambda_i = 1$. Suppose not, then there exists a subsequence $\{x^{i_j}\}$ such that

$$f(x^{i_j}) - f(x^{i_j} + d^{i_j}) < -\delta\nabla f(x^{i_j})d^{i_j},\ d^{i_j} = -\nabla^2 f(x^{i_j})^{-1}\nabla f(x^{i_j})'. \tag{41}$$

By the twice differentiability of f:

$$f(x^{i_j} + d^{i_j}) - f(x^{i_j}) = \nabla f(x^{i_j})d^{i_j} + \frac{1}{2}d^{i_j\prime}\nabla^2 f(x^{i_j} + t^{i_j}d^{i_j})d^{i_j}, \tag{42}$$

for some $t^{i_j} \in (0, 1)$. Hence

$$-\nabla f(x^{i_j})d^{i_j} - \frac{1}{2}d^{i_j\prime}\nabla^2 f(x^{i_j} + t^{i_j}d^{i_j})d^{i_j} < -\delta\nabla f(x^{i_j})d^{i_j}. \tag{43}$$

Collecting terms and substitution for d^{i_j} gives:

$$\begin{aligned} -\tfrac{1}{2}\nabla f(x^{i_j})\nabla^2 f(x^{i_j})^{-1}\nabla^2 f(x^{i_j} + t^{i_j}d^{i_j})\nabla^2 f(x^{i_j})^{-1}\nabla f(x^{i_j})' \\ < -(1-\delta)\nabla f(x^{i_j})\nabla^2 f(x^{i_j})^{-1}\nabla f(x^{i_j})'. \end{aligned} \tag{44}$$

Dividing by $\|\nabla^2 f(x^{i_j})\|_2^2$ and letting j go to infinity gives:

$$-\frac{1}{2}\bar{q}\nabla^2 f(\bar{x})^{-1}\bar{q}' \leq -(1-\delta)\bar{q}\nabla^2 f(\bar{x})^{-1}\bar{q}',$$

or

$$(\frac{1}{2} - \delta)\bar{q}\nabla^2 f(\bar{x})^{-1}\bar{q}' \leq 0, \tag{45}$$

where $\bar{q} = \lim_{j\to\infty}\frac{\nabla f(x^{i_j})}{\|\nabla f(x^{i_j})\|_2}$.

This contradicts the positive definiteness of $\nabla^2 f(\bar{x})$ for $\delta \in (0, \frac{1}{2})$. Hence for $i \geq \bar{i}$ for some $\bar{i}$, we have a pure Newton iteration, that is

$$x^{i+1} = x^i - \nabla^2 f(x^i)^{-1} \nabla f(x^i)', \; i \geq \bar{i}. \tag{46}$$

By part (a) above $\{x^i\}$ converges to $\bar{x}$, such that $\nabla f(\bar{x}) = 0$. We now establish the quadratic rate given by the inequality in (b). Since for $i \geq \bar{i}$:

$$\begin{aligned}
&x^{i+1} - \bar{x} = x^i - \bar{x} - \nabla^2 f(x^i)^{-1} \left[\nabla f(\bar{x} + t(x^i - \bar{x}))\right]_{t=0}^{t=1} \\
&= (x^i - \bar{x}) - \nabla^2 f(x^i)^{-1} \int_0^1 (\nabla^2 f(\bar{x} + t(x^i - \bar{x}) - \nabla^2 f(x^i) + \nabla^2 f(x^i))(x^i - \bar{x}) dt,
\end{aligned} \tag{47}$$

we have, upon cancelling the terms $(x^i - \bar{x})$ inside and outside the integral sign and taking norms, that :

$$\begin{aligned}
\|x^{i+1} - \bar{x}\|_2 \leq \;& \|\nabla^2 f(x^i)^{-1}\|_2 \int_0^1 \|\nabla^2 f(\bar{x} + t(x^i - \bar{x}) - \nabla^2 f(x^i)\|_2 \|x^i - \bar{x}\|_2 dt \\
\leq \;& \|\nabla^2 f(x^i)^{-1}\|_2 \|x^i - \bar{x}\|_2 \int_0^1 \kappa(1-t) \|x^i - \bar{x}\|_2 dt \\
= \;& \frac{\kappa}{2\mu} \|x^i - \bar{x}\|_2^2. \qquad \square
\end{aligned} \tag{48}$$

References

[1] L. Armijo. Minimization of functions having Lipschitz-continuous first partial derivatives. *Pacific Journal of Mathematics*, 16:1–3, 1966.

[2] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[3] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, second edition, 1999.

[4] P. S. Bradley, Usama M. Fayyad, and O. L. Mangasarian. Data mining: Overview and optimization opportunities. *INFORMS Journal on Computing*, 11:217–238, 1999. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-01.ps.

[5] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, pages 82–90, San Francisco, California, 1998. Morgan Kaufmann. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

[6] P. S. Bradley and O. L. Mangasarian. Massive data discrimination via linear support vector machines. *Optimization Methods and Software*, 13:1–10, 2000. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

[7] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Feature selection via mathematical programming. *INFORMS Journal on Computing*, 10(2):209–217, 1998. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/95-21.ps.

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

[9] B. Chen and P. T. Harker. Smooth approximations to nonlinear complementarity problems. *SIAM Journal of Optimization*, 7:403–420, 1997.

[10] Chunhui Chen and O. L. Mangasarian. Smoothing methods for convex inequalities and linear complementarity problems. *Mathematical Programming*, 71(1):51–69, 1995.

[11] Chunhui Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5(2):97–138, 1996.

[12] X. Chen, L. Qi, and D. Sun. Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities. *Mathematics of Computation*, 67:519–540, 1998.

[13] X. Chen and Y. Ye. On homotopy-smoothing methods for variational inequalities. *SIAM Journal on Control and Optimization*, 37:589–616, 1999.

[14] V. Cherkassky and F. Mulier. *Learning from Data - Concepts, Theory and Methods*. John Wiley & Sons, New York, 1998.

[15] P. W. Christensen and J.-S. Pang. Frictional contact algorithms based on semismooth newton methods. In *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, (editors)*, pages 81–116, Dordrecht, Netherlands, 1999. Kluwer Academic Publishers.

[16] CPLEX Optimization Inc., Incline Village, Nevada. *Using the CPLEX(TM) Linear Optimizer and CPLEX(TM) Mixed Integer Optimizer (Version 2.0)*, 1992.

[17] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, N.J., 1983.

[18] M. Fukushima and L. Qi. *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[19] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[20] L. Kaufman. Solving the quadratic programming problem arising in support vector classification. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 147–167. MIT Press, 1999.

[21] David G. Kleinbaum. *Survival Analysis*. Springer-Verlag, New York, 1996.

[22] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg. Breast cancer survival and chemotherapy: a support vector machine analysis. Technical Report 99-10, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, December 1999. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Volume 55, 2000, 1-10. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-10.ps.

[23] Yuh-Jye Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. Technical Report 99-03, Data Mining Institute, Com-

puter Sciences Department, University of Wisconsin, Madison, Wisconsin, September 1999. *Computational Optimization and Applications* 20(1), October 2001, to appear. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-03.ps.

[24] O. L. Mangasarian. Mathematical programming in neural networks. *ORSA Journal on Computing*, 5(4):349–360, 1993.

[25] O. L. Mangasarian. *Nonlinear Programming*. SIAM, Philadelphia, PA, 1994.

[26] O. L. Mangasarian. Parallel gradient distribution in unconstrained optimization. *SIAM Journal on Control and Optimization*, 33(6):1916–1925, 1995. ftp://ftp.cs.wisc.edu/tech-reports/reports/1993/tr1145.ps.

[27] O. L. Mangasarian. Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, 1(2):183–201, 1997. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/96-05.ps.

[28] O. L. Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24:15–23, 1999. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps.

[29] O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.

[30] O. L. Mangasarian and R. R. Meyer. Nonlinear perturbation of linear programs. *SIAM Journal on Control and Optimization*, 17(6):745–752, November 1979.

[31] O. L. Mangasarian and D. R. Musicant. Massive support vector regression. Technical Report 99-02, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, August 1999. *Machine Learning*, to appear. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-02.ps.

[32] O. L. Mangasarian and D. R. Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*,

10:1032–1037, 1999. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-18.ps.

[33] O. L. Mangasarian and D. R. Musicant. Data discrimination via nonlinear generalized support vector machines. In M. C. Ferris, O. L. Mangasarian, and J.-S. Pang, editors, *Complementarity: Applications, Algorithms and Extensions*, pages 233–251, Dordrecht, January 2001. Kluwer Academic Publishers. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/99-03.ps.

[34] O. L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, July-August 1995.

[35] MATLAB. *User's Guide*. The MathWorks, Inc., Natick, MA 01760, 1994-2001. http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml.

[36] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1992. www.ics.uci.edu/∼mlearn/MLRepository.html.

[37] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36:111–147, 1974.

[38] W. N. Street. Cancer diagnosis and prognosis via linear-programming-based machine learning. Computer Sciences Department, Mathematical Programming Technical Report and Ph.D. Dissertation 94-14, University of Wisconsin, Madison, Wisconsin, August 1994. http://www.cs.wisc.edu/tech-reports/.

[39] P. Tseng. Analysis of a non-interior continuation method based on chen-mangasarian smoothing functions for complementarity problems. In *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, (editors)*, pages 381–404, Dordrecht, Netherlands, 1999. Kluwer Academic Publishers.

[40] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[41] W. H. Wolberg, Y.-J. Lee, and O. L. Mangasarian. WPBCC: Wisconsin Prognostic Breast Cancer Chemotherapy Database.

Computer Sciences Department, University of Wisconsin, Madison, ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/WPBCC/, 1999.

[42] W. H. Wolberg, W. N. Street, D. H. Heisey, and O. L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, 26:792–796, 1995.

[43] W. H. Wolberg, W. N. Street, D. N. Heisey, and O. L. Mangasarian. Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. *Archives of Surgery*, 130:511–516, 1995.

[44] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Breast cytology diagnosis via digital image analysis. *Analytical and Quantitative Cytology and Histology*, 15(6):396–404, December 1993.

[45] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Computer-derived nuclear features compared with axillary lymph node status for breast carcinoma prognosis. *Cancer Cytopathology*, 81:172–179, 1997.

[46] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Importance of nuclear morphology in breast cancer prognosis. *Clinical Cancer Research*, 5:3542–3548, 1999.