# A Finite Newton Method for Classification Problems

O. L. Mangasarian
Computer Sciences Department
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706
*olvi@cs.wisc.edu*

**Abstract**

A fundamental classification problem of data mining and machine learning is that of minimizing a strongly convex, piecewise quadratic function on the $n$-dimensional real space $R^n$. We show finite termination of a Newton method to the unique global solution starting from any point in $R^n$. If the function is well conditioned, then no stepsize is required from the start, and if not, an Armijo stepsize is used. In either case finite termination is guaranteed to the unique global minimum solution.

## 1 Introduction

This paper establishes finite termination of a Newton method for minimizing a strongly convex, piecewise quadratic function on the $n$-dimensional real space $R^n$. Such a problem is a fundamental one in generating a linear or nonlinear kernel classifier for data mining and machine learning [23, 2, 16, 17, 18, 10, 11, 19, 4]. This work is motivated by [11], where a smoothed version of the present algorithm was shown to converge globally and quadratically but was observed to terminate in a few steps even when the smoothing and stepsize were removed. Another motivating work is [8], where finite termination of a Newton method was established for the least 2-norm solution of a linear program. In fact our algorithm is similar to that of [8], but has a global finite termination property without a stepsize under a well conditioning property. This well conditioning property, (26) below, which is sufficient for finite termination without a stepsize, does not appear to be necessary for a wide range of classification problems tested in [11].

Other related work using a finite Newton method to solve quadratic programs with bound constraints appears in [12, 13, 14, 15]. Our approach differs from this previous work in its finite termination property with or without an Armijo stepsize, in the analysis we present, and in the application to support vector machine classification.

We outline now the contents of the paper. In Section 2 we describe the linear and nonlinear classification problems leading to a piecewise quadratic strongly convex minimization problem. In Section 3 we establish finite global termination of a Newton algorithm without a stepsize but under a well conditioning property. In Section 4 we remove the well conditioning assumption but add the Armijo stepsize and obtain again finite global termination at the unique solution. In Section 5 we briefly discuss some previous numerical results with the proposed method. Section 6 concludes the paper.

A word about our notation. All vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. For a vector $x$ in the $n$-dimensional real space $R^n$, the *plus function* $x_+$ is defined as $(x_+)_i = \max\ \{0, x_i\}$, $i = 1, \ldots, n$, while $x_*$ denotes the subgradient of $x_+$ which is the step function defined as $(x_*)_i = 1$ if $x_i > 0$, $(x_*)_i = 0$ if $x_i < 0$, and $(x_*)_i \in [0, 1]$ if $x_i = 0$, $i = 1, \ldots, n$. The scalar (inner) product of two vectors $x$ and $y$ in the $n$-dimensional real space $R^n$ will be denoted by $x'y$ and the 2-norm of $x$ will be denoted by $\|x\|$. For a matrix $A \in R^{m\times n}$, $A_i$ is the $i$th row of $A$ which is a row vector in $R^n$ and $\|A\|$ is 2-norm of $A$: $\max\limits_{\|x\|=1} \|Ax\|$. A column vector of ones of arbitrary dimension will be denoted by $e$. For $A \in R^{m\times n}$ and $B \in R^{n\times l}$, the kernel $K(A, B)$ [23, 2, 16] is an arbitrary function which maps $R^{m\times n} \times R^{n\times l}$ into $R^{m\times l}$. In particular, if $x$ and $y$ are column vectors in $R^n$ then, $K(x', y)$ is a real number, $K(x', A')$ is a row vector in $R^m$ and $K(A, A')$ is an $m \times m$ matrix. 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 column 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 convex hull of a set $S$ is denoted by $co\{S\}$. The identity matrix of arbitrary order will be denoted by $I$.

## 2 Linear and Nonlinear Kernel Classification

We describe in this section the fundamental classification problems that lead to minimizing a piecewise quadratic strongly convex function. We consider the problem of classifying $m$ points in the $n$-dimensional real space $R^n$, represented by the $m \times n$ matrix $A$, according to membership of each point $A_i$ in the classes +1 or -1 as specified by a given $m \times m$ diagonal matrix $D$ with ones or minus ones along its diagonal. For this problem, the standard support vector machine with a linear kernel $AA'$ [23, 2] is given by the following quadratic program for some $\nu > 0$:

$$\begin{array}{rrcl} \displaystyle\min_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \frac{1}{2}w'w & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \tag{1}$$

As depicted in Figure 1, $w$ is the normal to the bounding planes:

$$\begin{array}{rcrcr} x'w & - & \gamma & = & +1 \\ x'w & - & \gamma & = & -1, \end{array} \tag{2}$$

and $\gamma$ determines their location relative to the origin. The first plane above bounds the class +1 points and the second plane bounds the class -1 points when the two classes are strictly linearly separable, that is when the slack variable $y = 0$. The linear separating surface is the plane

$$x'w = \gamma, \tag{3}$$

midway between the bounding planes (2). If the classes are linearly inseparable then the two planes bound the two classes with a "soft margin" determined by a nonnegative slack variable $y$, that is:

$$\begin{array}{rcrcrcrl} x'w & - & \gamma & + & y_i & \geq & +1, & \text{for } x' = A_i \text{ and } D_{ii} = +1, \\ x'w & - & \gamma & - & y_i & \leq & -1, & \text{for } x' = A_i \text{ and } D_{ii} = -1. \end{array} \tag{4}$$

The 1-norm of the slack variable $y$ is minimized with weight $\nu$ in (1). The quadratic term in (1), which is twice the reciprocal of the square of the 2-norm distance $\frac{2}{\|w\|}$ between the two bounding planes of (2) in the $n$-dimensional space of $w \in R^n$ for a *fixed* $\gamma$, maximizes that distance, often called the "margin". Figure 1 depicts the points represented by $A$, the bounding planes (2) with margin $\frac{2}{\|w\|}$, and the separating plane (3) which separates $A+$, the points represented by rows of $A$ with $D_{ii} = +1$, from $A-$, the points represented by rows of $A$ with $D_{ii} = -1$.

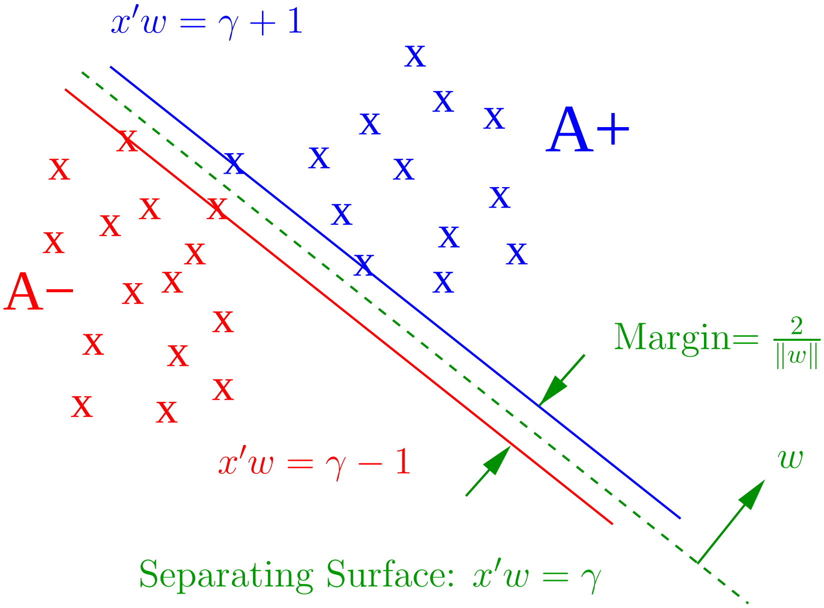

Figure 1: **The bounding planes (2) with margin $\frac{2}{\|w\|}$, and the plane (3) separating $A+$, the points represented by rows of $A$ with $D_{ii} = +1$, from $A-$, the points represented by rows of $A$ with $D_{ii} = -1$.**

In many essentially equivalent formulations of the classification problem [11, 10, 4, 5], 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 (1). In addition the distance between the planes (2) is measured in the $(n+1)$-dimensional space of $(w, \gamma) \in R^{n+1}$, that is $\frac{2}{\|(w,\gamma)\|}$. 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 [17]. Thus using twice the reciprocal squared of the margin instead, yields our modified SVM problem as follows:

$$
\begin{array}{rrcl}
\displaystyle\min_{(w,\gamma,y)\in R^{n+1+m}} & \frac{\nu}{2}y'y+\frac{1}{2}(w'w+\gamma^2) & & \\
\text{s.t.} & D(Aw-e\gamma)+y & \geq & e \\
 & y & \geq & 0.
\end{array}
\tag{5}
$$

It has been shown computationally [19] that this reformulation (5) of the conventional support vector machine formulation (1) yields similar results to (1). At a solution of problem (5), $y$ is given by

$$
y=(e-D(Aw-e\gamma))_+,
\tag{6}
$$

where, as defined in the Introduction, $(\cdot)_+$ replaces negative components of a vector by zeros. Thus, we can replace $y$ in (5) by $(e-D(Aw-e\gamma))_+$ and convert the SVM problem (5) into an equivalent SVM which is an unconstrained optimization problem as follows:

$$
\min_{(w,\gamma)\in R^{n+1}} \quad \frac{\nu}{2}\|(e-D(Aw-e\gamma))_+\|^2+\frac{1}{2}(w'w+\gamma^2).
\tag{7}
$$

This problem is the strongly convex piecewise quadratic minimization problem. Note however that its objective function is not twice differentiable which precludes the use of a regular Newton method. In [11] we smoothed this problem and applied a fast Newton method to solve it. Problem (7) is one of the nonsmooth problems that we shall provide a direct finite Newton method for. The other nonsmooth problem that we will treat is the nonlinear kernel problem, (12) below, which generates a *nonlinear* classifier as described below.

We now describe how the generalized support vector machine (GSVM) [16] generates a nonlinear separating surface by using a completely arbitrary kernel. The GSVM solves the following mathematical program for a general kernel $K(A,A')$, defined in the Introduction:

$$
\begin{array}{rrcl}
\displaystyle\min_{(u,\gamma,y)\in R^{2m+1}} & \nu e'y+f(u) & & \\
\text{s.t.} & D(K(A,A')Du-e\gamma)+y & \geq & e \\
 & y & \geq & 0.
\end{array}
\tag{8}
$$

Here $f(u)$ is some convex function on $R^m$ which suppresses the parameter $u$ and $\nu$ is some positive number that weights the classification error $e'y$ versus the suppression of $u$. A solution of this mathematical program for $u$ and $\gamma$ leads to the nonlinear separating surface

$$
K(x',A')Du=\gamma.
\tag{9}
$$

The linear formulation (1) of Section 2 is obtained if we let $K(A,A') = AA'$, $w = A'Du$ and $f(u) = \frac{1}{2}u'DAA'Du$. We now use a different classification objective which not only suppresses the parameter $u$ but also suppresses $\gamma$ in our nonlinear formulation:

$$\begin{array}{rrcl} \displaystyle\min_{(u,\gamma,y)\in R^{2m+1}} & \frac{\nu}{2}y'y + \frac{1}{2}(u'u+\gamma^2) & & \\ \text{s.t.} & D(K(A,A')Du - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \tag{10}$$

At a solution of (10), $y$ is given by

$$y = (e - D(K(A,A')Du - e\gamma))_+, \tag{11}$$

where, as defined earlier, $(\cdot)_+$ replaces negative components of a vector by zeros. Thus, we can replace $y$ in (10) by $(e - D(K(A,A')Du - e\gamma))_+$ and convert the SVM problem (10) into an equivalent SVM which is an unconstrained optimization problem as follows:

$$\min_{(u,\gamma)\in R^{m+1}} \frac{\nu}{2}\|(e - D(K(A,A')Du - e\gamma))_+\|^2 + \frac{1}{2}(u'u+\gamma^2). \tag{12}$$

Again, as in (7), this problem is a strongly convex unconstrained minimization problem which has a unique solution but its objective function is not twice differentiable. It is for this problem, that generates the nonlinear separating surface (9), and for problem (7) that generates the linear separating surface (3), that we shall develop the finite Newton methods that we describe in the next two sections of the paper.

## 3 Finite Stepless Newton Method

We shall consider in the rest of the paper the following piecewise quadratic strongly convex problem which subsumes both problems (7) and (12):

$$\min_{z\in R^p} f(z) := \frac{\nu}{2}\|(Cz-h))_+\|^2 + \frac{1}{2}(z'z), \tag{13}$$

where $C \in R^{m\times p}$, $h \in R^m$ and $\nu$ is a fixed positive parameter. We note immediately that, since $\|(r-t)_+\| \leq \|(r-t)\|$ for $r,t \in R^m$, the gradient of $f$:

$$\nabla f(z) = \nu C'(Cz-h)_+ + z, \tag{14}$$

is globally Lipschitz continuous with constant $K$ as follows:

$$\|\nabla f(s) - \nabla f(z)\| \leq K\|s - z\|, \forall s, z \in R^p, \text{ where } K = \nu\|C'\|\|C\| + 1. \tag{15}$$

The ordinary Hessian of $f$ does not exist everywhere. However, since $\nabla f(z)$ is Lipschitzian, $f$ belongs to the class $LC^1$ of functions with locally Lipschitzian gradients, and a generalized Hessian exists everywhere [7]. $LC^1$ functions were introduced in [7] and used in subsequent papers such as [9, 22, 21, 3]. The generalized Hessian is defined as follows [7, 3]:

**Definition 3.1 Generalized Hessian** *Let $f : R^p \longrightarrow R$ have a Lipschitz continuous gradient on $R^p$. The generalized Hessian of $f$ at $x$ is the set of $\partial^2 f(z)$ of $p \times p$ matrices defined as [7, 3]:*

$$\begin{array}{l} \partial^2 f(z) := co\{H \in R^{p\times p} \mid \exists x^k \longrightarrow \ x \textit{ such that}: \\ \nabla f \textit{is differentiable at } x^k \textit{ and } \nabla^2 f(x^k) \longrightarrow H\}. \end{array} \tag{16}$$

Furthermore, a generalization of the mean value theorem for the Lipschitzian gradient $\nabla f(z)$ holds [6, 3]:

**Proposition 3.2 Generalized Mean Value Theorem** *Let $f : R^p \longrightarrow R$ have a Lipschitz continuous gradient on $R^p$. Then for $z, s \in R^p$:*

$$\nabla f(z) = \nabla f(s) + \sum_{j=1}^{p} t_j H^j (z - s) \tag{17}$$

*where $H^j \in \partial^2 f(y^j)$, for some $y^j \in (s, z),\ t_j \geq 0,\ j = 1, \ldots, p$ and $\sum_{j=1}^{p} t_j = 1$.*

We shall also need the following lemma [20, 8.1.5] that gives a quadratic bound on a linear Taylor expansion:

**Lemma 3.3 Quadratic Bound Lemma** *Let $f : R^n \longrightarrow R$ have a Lipschitz continuous gradient on $R^p$ with constant $K$. Then for $z, s \in R^p$:*

$$|f(z) - f(s) - \nabla f(s)'(z - s)| \leq \frac{K}{2}\|z - s\|^2. \tag{18}$$

We summarize now properties of the function $f(z)$ in the following lemma and omit the straightforward proof of these properties.

**Lemma 3.4 Properties of** $f(z)$ *The function $f(z)$ defined by (13) has the following properties for all $s, z \in R^p$:*

*(i)* *$f(z)$ is strongly convex with constant $k = 1$:*

$$(\nabla f(s) - \nabla f(z))'(s - z) \geq k \cdot \|s - z\|^2 = \|s - z\|^2. \qquad (19)$$

*(ii)* *$\nabla f(z)$ is Lipschitz continuous with constant $K$:*

$$\|\nabla f(s) - \nabla f(z)\| \leq K\|s - z\|, \, \forall s, z \in R^p, \, \mathit{where}\, K = \nu\|C'\|\|C\| + 1. \qquad (20)$$

*(iii)* *The generalized Hessian of $f(z)$ [7] is:*

$$\partial^2 f(z) = \nu C' diag(Cz - h)_* C + I, \qquad (21)$$

*where $diag(Cz - h)_*$ denotes the $p \times p$ diagonal matrix whose $j^{th}$ diagonal entry is the subgradient of the step function $(\cdot)_+$ as follows:*

$$(diag(Cz - h)_*)_{jj} \begin{cases} = 1, & \mathit{if}\, C_j z - h_j > 0, \\ = [0, 1], & \mathit{if}\, C_j z - h_j = 0, \\ = 0, & \mathit{if}\, C_j z - h_j < 0, \end{cases} \qquad j = 1, \dots, p. \qquad (22)$$

*(iv)*

$$K\|s\|^2 \geq s'(\nu C'C + I)s \geq s'\partial^2 f(z)s \geq \|s\|^2, \, \mathit{where}\, K = \nu\|C'\|\|C\| + 1. \qquad (23)$$

*(v)*

$$\frac{1}{K}\|s\|^2 \leq s'\partial^2 f(z)^{-1}s \leq \|s\|^2. \qquad (24)$$

*(vi)*

$$\|\partial^2 f(s) - \partial^2 f(z)\| \leq \nu\|C'\| \cdot \|C\|. \qquad (25)$$

We note that the inverse $\partial^2 f(z)^{-1}$ appearing in (24) and elsewhere refers to the inverse of $\partial^2 f(z)$ defined by (21)-(22) for an arbitrary but specific value of $(diag(Cz - h)_*)_{jj}, \ j = 1, \dots, p$ in the interval $[0, 1]$ when $C_j z - h_j = 0$.

We are ready now to state and establish global finite termination of a Newton algorithm without a stepsize starting from any point.

**Algorithm 3.5 Stepless Newton Algorithm for (13)** *Start with any $z^0 \in R^p$. For $i = 0, 1, \dots$:*

*(i)* $z^{i+1} = z^i - \partial^2 f(z^i)^{-1}\nabla f(z^i)$.

*(ii)* *Stop if* $\nabla f(z^{i+1}) = 0$.

*(iii)* $i = i + 1$. *Go to (i).*

**Theorem 3.6 Finite Termination of Stepless Newton** *Let $f(z)$ be well conditioned, that is:*

$$\frac{K}{k} = \frac{\nu\|C'\|\|C\| + 1}{1} < 2, \text{ i.e. } \nu\|C'\|\|C\| < 1. \tag{26}$$

*(i) The sequence $\{z^i\}$ of Algorithm 3.5 terminates at the global minimum solution $\bar{z}$ of (13).*

*(ii) The error decreases linearly at each step as follows:*

$$\|z^{i+1} - \bar{z}\| \leq \nu\|C'\|\|C\|\|z^i - \bar{z}\|. \tag{27}$$

**Proof**

(i) We first establish convergence and then finite termination of the sequence $\{z^i\}$. By the Quadratic Bound Lemma (18):

$$\begin{array}{rcl} f(z^i) - f(z^{i+1}) & \geq & \nabla f(z^i)'\partial^2 f(z^i)^{-1}\nabla f(z^i) - \frac{K}{2}\nabla f(z^i)'\partial^2 f(z^i)^{-2}\nabla f(z^i) \\ & = & \nabla f(z^i)'(\partial^2 f(z^i)^{-1} - \frac{K}{2}\partial^2 f(z^i)^{-2})\nabla f(z^i) \\ & = & \frac{K}{2}\nabla f(z^i)'\partial^2 f(z^i)^{-\frac{1}{2}}(\frac{2I}{K} - \partial^2 f(z^i)^{-1})\partial^2 f(z^i)^{-\frac{1}{2}}\nabla f(z^i) \\ & \geq & \frac{K}{2}\frac{1-\nu\|C'\|\|C\|}{1+\nu\|C'\|\|C\|}\nabla f(z^i)'\partial^2 f(z^i)^{-1}\nabla f(z^i) \\ & \geq & \frac{1}{2}\frac{1-\nu\|C'\|\|C\|}{1+\nu\|C'\|\|C\|}\|\nabla f(z^i)\|^2 \geq 0, \end{array} \tag{28}$$

where the first inequality above follows from Quadratic Bound Lemma (18), the next to the last inequality follows from the well conditioned assumption (26) and (24), and the last inequality from (24).

By (28) and the strong convexity of $f(z)$ we have that:

$$\begin{array}{rcl} 0 & \geq & f(z^i) - f(z^0) \\ & \geq & \nabla f(z^0)'(z^i - z^0) + \frac{1}{2}\|z^i - z^0\|^2 \\ & \geq & -\|\nabla f(z^0)\|\|z^i - z^0\| + \frac{1}{2}\|z^i - z^0\|^2. \end{array} \tag{29}$$

Hence,

$$\|z^i - z^0\| \leq \frac{2}{\|\nabla f(z^0)\|},\ i = 1, 2, \ldots\ , \tag{30}$$

and consequently the bounded sequence $\{z^i\}$ has an accumulation point $\bar{z}$ such that $\lim\limits_{j \longrightarrow \infty} z^{i_j} = \bar{z}$. Since $\{f(z^i)\}$ is nonincreasing by (28) and bounded below by $\min\limits_{z \in R^p} f(z)$, it converges and $\lim\limits_{i \longrightarrow \infty} f(z^i) = f(\bar{z})$. Thus:

$$0 = \lim_{j \longrightarrow \infty} (f(z^{i_j}) - f(z^{i_j+1})) \geq \frac{1}{2} \frac{1 - \nu\|C'\|\|C\|}{1 + \nu\|C'\|\|C\|} \lim_{j \longrightarrow \infty} \|\nabla f(z^{i_j})\|^2 \geq 0. \tag{31}$$

Hence $\lim\limits_{j \longrightarrow \infty} \|\nabla f(z^{i_j})\| = \|\nabla f(\bar{z})\| = 0$. Since all accumulation points are stationary, they must all equal the unique minimizer of $f(z)$ on $R^p$. It follows that the whole sequence $\{z^i\}$ must converge to the unique $\bar{z}$ such that $\nabla f(\bar{z}) = 0$.

We now show finite termination by using an argument similar to that of [8].

Our Newton iteration is:

$$\nu C'(Cz^i - h)_+ + z^i + (\nu C' diag(Cz^i - h)_* C + I)(z^{i+1} - z^i) = 0, \tag{32}$$

which we rewrite by subtracting from it the equality:

$$\nu C'(C\bar{z} - h)_+ + \bar{z} = \nabla f(\bar{z}) = 0. \tag{33}$$

This results in the equivalent iteration:

$$\begin{array}{c} \nu C'[(Cz^i - h)_+ - (C\bar{z} - h)_+] + [z^i - \bar{z}] + \\ (\nu C' diag(Cz^i - h)_* C + I)(z^{i+1} - z^i) = 0. \end{array} \tag{34}$$

We establish now that this Newton iteration is satisfied uniquely (since $\partial^2 f(z^i)$ is nonsingular) by $z^{i+1} = \bar{z}$ when $z^i$ is sufficiently close to $\bar{z}$ and hence the Newton iteration terminates at $\bar{z}$ at step (ii) of Algorithm 3.5. Setting $z^{i+1} = \bar{z}$ in (34) and canceling terms gives:

$$\nu C'[(Cz^i - h)_+ - (C\bar{z} - h)_+ + diag(Cz^i - h)_*((C\bar{z} - h) - (Cz^i - h))] = 0. \tag{35}$$

We verify now that the term in the square bracket in (35) is zero when $z^i$ is sufficiently close to $\bar{z}$ by looking at each component $j$, $j = 1, \dots, p$, of the vector enclosed in the square brackets. We consider the nine possible combinations:

(i) $C_j\bar{z} - h_j > 0,\ C_j z^i - h_j > 0$:

$$C_j z^i - h_j - C_j\bar{z} + h_j + 1 \cdot (C_j\bar{z} - h_j) - C_j z^i + h_j = 0.$$

(ii) $C_j\bar{z} - h_j > 0,\ C_j z^i - h_j = 0$:
Cannot occur when $z^i$ is sufficiently close to $\bar{z}$.

(iii) $C_j\bar{z} - h_j > 0,\ C_j z^i - h_j < 0$:
Cannot occur when $z^i$ is sufficiently close to $\bar{z}$.

(iv) $C_j\bar{z} - h_j = 0,\ C_j z^i - h_j > 0$:

$$C_j z^i - h_j - 0 + 1 \cdot (0 - C_j z^i + h_j) = 0.$$

(v) $C_j\bar{z} - h_j = 0,\ C_j z^i - h_j = 0$:

$$0 + 0 + [0, 1](0 + 0) = 0.$$

(vi) $C_j\bar{z} - h_j = 0,\ C_j z^i - h_j < 0$:

$$0 + 0 + 0 \cdot (0 - C_j z^i + h_j) = 0.$$

(vii) $C_j\bar{z} - h_j < 0,\ C_j z^i - h_j > 0$:
Cannot occur when $z^i$ is sufficiently close to $\bar{z}$.

(viii) $C_j\bar{z} - h_j < 0,\ C_j z^i - h_j = 0$:
Cannot occur when $z^i$ is sufficiently close to $\bar{z}$.

(ix) $C_j\bar{z} - h_j < 0,\ C_j z^i - h_j < 0$:

$$0 - 0 + 0 \cdot (C_j\bar{z} - h_j - C_j z^i + h_j) = 0.$$

Hence for $z^i$ is sufficiently close to $\bar{z}$, the Newton iteration is uniquely satisfied by $\bar{z}$ and terminates.

(ii) We establish now the linear termination rate by using the generalization of the mean value theorem Proposition 3.2 as follows:

$$\begin{array}{rcl}(z^{i+1}-\bar{z}) & = & z^i-\bar{z}-\partial^2 f(z^i)^{-1}(\nabla f(z^i)-\nabla f(\bar{z}))\\ & = & z^i-\bar{z}-\partial^2 f(z^i)^{-1}\sum_{j=1}^{p} t^j H^j(z^i-\bar{z})\\ & = & (z^i-\bar{z})-\partial^2 f(z^i)^{-1}[\sum_{j=1}^{p} t^j H^j+\partial^2 f(z^i)-\partial^2 f(z^i)](z^i-\bar{z})\\ & = & -\partial^2 f(z^i)^{-1}[\sum_{j=1}^{p} t^j(H^j-\partial^2 f(z^i))](z^i-\bar{z}),\end{array} \tag{36}$$

where the second equality above follows from the generalization of the mean value theorem Proposition 3.2. Taking norms of the first and last terms above gives:

$$\begin{array}{rcl}\|z^{i+1}-\bar{z}\| & \leq & \|\partial^2 f(z^i)^{-1}\|\|\sum_{j=1}^{p} t^j(H^j-\partial^2 f(z^i))\|\|z^i-\bar{z}\|\\ & \leq & \|\partial^2 f(z^i)^{-1}\|\nu\|C'\|\|C\|\|z^i-\bar{z}\|\\ & \leq & \nu\|C'\|\|C\|\|z^i-\bar{z}\|,\end{array} \tag{37}$$

where use has been made of (24) and (25) of Lemma 3.4 in the last two inequalities above.□

We turn now to a Newton algorithm with an Armijo stepsize.

# 4 Finite Armijo Newton Method

We now drop the well conditioning assumption (26) but add an Armijo stepsize [1] in order guarantee global finite termination of the following algorithm.

**Algorithm 4.1 Armijo Newton Algorithm for (13)** *Start with any* $z^0 \in R^p$*. For* $i = 0, 1, \ldots$ *:*

*(i) Stop if* $\nabla f(z^i - \partial^2 f(z^i)^{-1}\nabla f(z^i)) = 0$.

*(ii)* $z^{i+1} = z^i - \lambda_i \partial^2 f(z^i)^{-1}\nabla f(z^i) = z^i + \lambda_i d^i$,
*where* $\lambda_i = \max\{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ *such that:*

$$f(z^i) - f(z^i + \lambda_i d^i) \geq -\delta\lambda_i \nabla f(z^i)' d^i, \tag{38}$$

*for some $\delta \in (0, \frac{1}{2})$, and $d^i$ is the Newton direction:*

$$d^i = -\partial^2 f(z^i)^{-1}\nabla f(z^i). \tag{39}$$

*(iii) $i = i + 1$. Go to (i).*

**Theorem 4.2 Finite Termination of Armijo Newton** *The sequence $\{z^i\}$ of Algorithm 4.1 terminates at the global minimum solution $\bar{z}$ of (13).*

**Proof** As in the proof of the Stepless Newton Algorithm 3.5, we first establish global convergence of $\{z^i\}$ to the global solution $\bar{z}$, but without using the well conditioned assumption (26). By the Quadratic Bound Lemma 3.3 we have the first inequality below:

$$\begin{array}{rcl} f(z^i) - f(z^i + \lambda_i d^i) + \delta\lambda_i \nabla f(z^i)' d^i & \geq & -\frac{K}{2}\lambda_i^2 \|d^i\|^2 - (1-\delta)\lambda_i \nabla f(z^i)' d^i \\ & = & -\frac{K}{2}\lambda_i^2 \nabla f(z^i)'\partial^2 f(z^i)^{-2}\nabla f(z^i) \\ & & +(1-\delta)\lambda_i \nabla f(z^i)'\partial^2 f(z^i)^{-1}\nabla f(z^i) \\ & \geq & \lambda_i(-\frac{K}{2}\lambda_i + \frac{1-\delta}{K})\|\nabla f(z^i)\|^2, \end{array} \tag{40}$$

where the equality above makes use of the definition of $d^i$ and the last inequality makes use of (24) of Lemma 3.4. Hence if $\lambda_i$ is small enough, that is $\lambda_i \leq \frac{2(1-\delta)}{K^2}$ then the Armijo inequality (38) is satisfied. However by the definition of the Armijo stepsize, $2\lambda_i$ violates the Armijo inequality (38) and hence by (40):

$$(-\frac{K}{2}2\lambda_i + \frac{1-\delta}{K}) < 0, \text{ or equivalently } \lambda_i > \frac{1-\delta}{K^2}. \tag{41}$$

It follows that:

$$\lambda_i \geq \min\{1, \frac{1-\delta}{K^2}\} =: \tau > 0, \tag{42}$$

and by the Armijo inequality (38) and (24) of Lemma 3.4 that:

$$f(z^i) - f(z^{i+1}) \geq -\delta\lambda_i \nabla f(z^i)' d^i \geq \delta\tau \nabla f(z^i)'\partial^2 f(z^i)^{-1}\nabla f(z^i) \geq \frac{\delta\tau}{K}\|\nabla f(z^i)\|^2. \tag{43}$$

By exactly the same arguments following inequalities (28), we have that the sequence $\{z^i\}$ converges to unique minimum solution $\bar{z}$ of (13).

Having established convergence of $\{z^i\}$ to $\bar{z}$, finite termination of the Armijo Newton to $\bar{z}$ again follows exactly the finiteness proof of Theorem 3.6 because of step (i) of the current algorithm which checks whether the next iterate obtained by a stepless Newton is stationary.□

## 5 Numerical Experience

Our numerical experience is based on 16 test problems of [11] for which a smoothed version of (7) and (12) was solved using a regular Newton method with an Armijo step size. However all of these test problems were also solved in [11] using the proposed Stepless Newton Algorithm 3.5 here and gave the same results as the smoothed method and with the number of Newton steps varying between 5 and 8 for all problems. The test problems were all of type of (7) or (12), and varied in size with $m$ between 110 and 32,562 and $n$ between 2 and 123. Running times varied between 1 and 85 seconds on a 200 MHz PentiumPro with 64 megabytes of RAM. Linear classifiers were used for the largest problems, $m = 32,562$, so that the Newton method solved the minimization problem (7) in $R^{n+1}$ with $n = 123$ for these large problems.

## 6 Conclusion

We have presented fast, finitely terminating Newton methods for solving a fundamental classification problem of data mining and machine learning. The methods are simple and fast and can be applied to other problems such as linear programming [8]. An interesting open question is whether these finite methods can be shown to have polynomial run time and whether they can be applied to an even broader class of problems.

## Acknowledgments

The research described in this Data Mining Institute Report 01-11, December 2001, was supported by National Science Foundation Grants CCR-9729842 and CDA-9623632, by Air Force Office of Scientific Research Grant F49620-00-1-0085 and by the Microsoft Corporation.

## References

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

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

[3] F. Facchinei. Minimization of $SC^1$ functions and the Maratos effect. *Operations Research Letters*, 17:131–137, 1995.

[4] G. Fung and O. L. Mangasarian. Proximal support vector machine classifiers. In F. Provost and R. Srikant, editors, *Proceedings KDD-2001: Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, CA*, pages 77–86, New York, 2001. Asscociation for Computing Machinery. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-02.ps.

[5] G. Fung and O. L. Mangasarian. Incremental support vector machine classification. In H. Mannila R. Grossman and R. Motwani, editors, *Proceedings of the Second SIAM International Conference on Data Mining, Arlington, Virginia, April 11-13,2002*, pages 247–260, Philadelphia, 2002. SIAM. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-08.ps.

[6] J.-B. Hiriart-Urruty. Refinements of necessary optimality conditions in nondifferentiable programming II. *Mathematical Programming Study*, 9:120–139, 1982.

[7] J.-B. Hiriart-Urruty, J. J. Strodiot, and V. H. Nguyen. Generalized hessian matrix and second-order optimality conditions for problems with $C^{L1}$ data. *Applied Mathematics and Optimization*, 11:43–56, 1984.

[8] C. Kanzow, H. Qi, and L. Qi. On the minimum norm solution of linear programs. Preprint, University of Hamburg, Hamburg, 2001. http://www.math.uni-hamburg.de/home/kanzow/paper.html. Journal of Optimization Theory and Applications, to appear.

[9] D. Klatte and K. Tammer. On second-order sufficient optimality conditions for $C^{1,1}$-optimization problems. *Optimization*, 19:169–179, 1988.

[10] Y.-J. Lee and O. L. Mangasarian. RSVM: Reduced support vector machines. Technical Report 00-07, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, July 2000. Proceedings of the First SIAM International Conference on Data Mining, Chicago, April 5-7, 2001, CD-ROM Proceedings. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-07.ps.

[11] Yuh-Jye Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. *Computational Optimization and Applications*, 20:5–22, 2001. Data Mining Institute, University of Wisconsin, Technical Report 99-03. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-03.ps.

[12] W. Li and J. J. Swetits. A Newton method for convex regression, data smoothing and quadratic programming with bounded constraints. *SIAM Journal on Optimization*, 3:466–488, 1993.

[13] W. Li and J. J. Swetits. A new algorithm for solving strictly convex quadratic programs. *SIAM Journal on Optimization*, 7:595–619, 1997.

[14] K. Madsen, H. B. Nielsen, and M. C. Pinar. A finite continuation algortihm for bound constrained quadratic programming. *SIAM Journal on Optimization*, 9:62–83, 1998.

[15] K. Madsen, H. B. Nielsen, and M. C. Pinar. Bound constrained quadratic programming via piecewise quadratic functions. *Mathematical Programming*, 85:135–156, 1999.

[16] 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.

[17] 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.

[18] O. L. Mangasarian and D. R. Musicant. Active support vector machine classification. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 577–583, Cambridge, MA, 2001. MIT Press. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-04.ps.

[19] O. L. Mangasarian and D. R. Musicant. Lagrangian support vector machines. *Journal of Machine Learning Research*, 1:161–177, 2001. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-06.ps.

[20] J. M. Ortega. *Numerical Analysis, A Second Course.* Academic Press, New York, 1972.

[21] L. Qi. Superlinearly convergent approximate Newton methods for LC$^1$ optimization problems. *Mathematical Programming*, 64:277–294, 1994.

[22] L. Qi and J. Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58:353–368, 1993.

[23] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, second edition, 2000.