SHARP QUADRATIC MAJORIZATION IN ONE DIMENSION

JAN DE LEEUW AND KENNETH LANGE

ABSTRACT. Quadratic majorizations for real-valued functions of a real variable are analyzed, and the concept of sharp majorization is introduced and studied. Applications to logistic , probit and robust loss functions are discussed. The univariate quadratic majorizations can be combined with regression, principal component analysis, and multidimensional scaling models to create simple iterative algorithms for complicated multivariate techniques.

1. INTRODUCTION

Majorization algorithms, including the EM algorithm, are used for more and more computational tasks in statistics [De Leeuw, 1994; Heiser, 1995; Lange et al., 2000; Hunter and Lange, 2004]. The basic idea is simple. A function g majorizes a function f at a point y if $g \geq f$ and $g(y) = f(y)$. If we are minimizing a complicated objective function f iteratively, then we construct a majorizing function at the current best solution $x^{(k)}$. We then find a new solution $x^{(k+1)}$ by minimizing the majorization function. Then we construct a new majorizing function at $x^{(k+1)}$, and so on.

Majorization algorithms are worth considering if the majorizing functions can be chosen to be much easier to minimize than the

Date: December 4, 2006.

2000 *Mathematics Subject Classification.* 29M20.

Key words and phrases. Successive Approximation, Iterative Majorization, Convexity.

This research was supported in part by NIH grants GM53275 and MH59490 to KL..

original objective function, for instance linear or quadratic. In this paper we will look in more detail at majorization with quadratic functions. We restrict ourselves to functions of a single real variable. This is not as restrictive as it seems, because many functions $F(x_1, \cdots, x_n)$ in optimization and statistics are *separable* in the sense that

$$F(x_1, \cdots, x_n) = \sum_{i=1}^{n} f_i(x_i),$$

and majorization of the univariate functions f_i automatically gives a majorization of F.

Many of our results generalize without much trouble to real-valued functions on R^n and to constrained minimization over subsets of R^n. The univariate context suffices to explain most of the basic ideas.

2. Majorization

2.1. **Definitions.** We formalize the definition of majorization at a point.

Definition 2.1. Suppose f and g are real-valued functions on R^n. We say that g *majorizes* f *at* y if

- $g(x) \geq f(x)$ for all x,
- $g(y) = f(y)$.

If the first condition can be replaced by

- $g(x) > f(x)$ for all $x \neq y$,

we say that majorization is *strict*.

Thus g majorizes f at y if $d = g - f$ has a minimum, equal to zero, at y. And majorization is strict if this minimum is unique. If g majorizes f at y, then f *minorizes* g at y. Alternatively we also say that f *supports* g at y.

It is also useful to have a global definition, which says that f can be majorized at all y.

Definition 2.2. Suppose f is a real-valued functions on $\mathbb{R}^n$ and g is a real-valued function on $\mathbb{R}^n \otimes \mathbb{R}^n$. We say that g *majorizes* f if

- $g(x,y) \geq f(x)$ for all x and all y,
- $g(x,x) = f(x)$ for all x.

Majorization is $strict$ if the first condition is

- $g(x,y) > f(x)$ for all $x \neq y$.

2.2. **Majorization Algorithms.** The basic idea of majorization algorithms is simple. Suppose our current best approximation to the minimum of f is $x^{(k)}$, and we have a g that majorizes f in $x^{(k)}$. If $x^{(k)}$ already minimizes g we stop, otherwise we update $x^{(k)}$ to $x^{(k+1)}$ by minimizing g. If we do not stop, we have the *sandwich inequality*

$$f(x^{(k+1)}) \leq g(x^{(k+1)}) < g(x^{(k)}) = f(x^{(k)}),$$

and in the case of strict majorization

$$f(x^{(k+1)}) < g(x^{(k+1)}) < g(x^{(k)}) = f(x^{(k)}).$$

Repeating these steps produces a decreasing sequence of function values, and appropriate additional compactness and continuity conditions guarantee convergence of the algorithm. In fact, it is not necessary to actually minimize the majorization function; it is sufficient to have a continuous update function h such that $g[h(y)] < g(y)$ for all y. In that case the sandwich inequality still applies with $x^{(k+1)} = h(x^{(k)})$.

2.3. **Majorizing Differentiable Functions.** We first show that majorization functions must have certain properties at the point where they touch the target.

Theorem 2.1. *Suppose f and g are differentiable at y. If g majorizes f at y, then*

- $g(y) = f(y)$,
- $g'(y) = f'(y)$.

If f and g are twice differentiable at y, then in addition

- $g''(y) \geq f''(y)$.

Proof. If g majorizes f at y then $d = g - f$ has a minimum at y. Now use the familiar necessary conditions for the minimum of a differentiable function, which say the derivative at the minimum is zero and, for a twice-differentiable function, the second derivative is non-negative. □

Theorem 2.1 can be generalized in many directions if differentiability fails. If f has a left and right derivatives in y, for instance, and g is differentiable, then

$$f'_R(y) \leq g'(y) \leq f'_L(y).$$

If f is convex, then $f'_L(y) \leq f'_R(y)$, and $f'(y)$ must exist in order for a differentiable g to majorize f at y. In this case $g'(y) = f'(y)$. For nonconvex f more general differential inclusions are possible using the four Dini derivatives of f at y.

3. Quadratic Majorizers

As we said, it is desirable that the subproblems in which we minimize the majorization function are simple. One way to guarantee this is to try to find a *quadratic majorizer*. This will generally be *convex quadratic majorizers*, because concave ones have no minima and are useless for algorithmic purposes (at least in the case of unconstrained minimization).

The first result, which has been widely applied, applies to functions with a continuous and uniformly bounded second derivative [Böhning and Lindsay, 1988].

Theorem 3.1. *If f is twice differentiable and there is an $B > 0$ such that $f''(x) \leq B$ for all x, then for each y the convex quadratic function*

$$g(x) = f(y) + f'(y)(x - y) + \frac{1}{2}B(x - y)^2.$$

majorizes f at y.

Proof. Use Taylor's theorem in the form

$$f(x) = f(y) + f'(y)(x - y) + \frac{1}{2}f''(\xi)(x - y)^2,$$

with ξ on the line connecting x and y. Because $f''(\xi) \leq B$, this implies $f(x) \leq g(x)$, where g is defined above. □

This result is very useful, but it has some limitations. In the first place we would like a similar result for functions that are not everywhere twice differentiable, or even those that are not everywhere differentiable. Second, the bound does take into account that we only need to bound the second derivative on the interval between x and y, and not on the whole line. This may result in a bound which is not sharp.

Why do we want the bounds on the second derivative to be sharp? The majorization algorithm corresponding to this result is

$$x^{(k+1)} = x^{(k)} - \frac{1}{B}f'(x^{(k)}),$$

By Ostrowski's Theorem [Ortega and Rheinboldt, 1970, Theorem 10.1.3] this sequence converges linearly, say to x_∞, with rate

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - x_\infty\|}{\|x^{(k)} - x_\infty\|} = 1 - \frac{1}{B}f''(x_\infty).$$

The smaller we choose B, the faster our convergence. The result remains true in the more general situation we study in this paper,

where B continuously depends on the majorization point y, provided the algorithm converges to a point with $f'(x_\infty) = 0$. In that case the linear convergence rate is $1 - f''(x_\infty)/B(x_\infty)$.

Example 3.1. If a quadratic g majorizes a twice-differentiable convex function f at y, then g is convex. This follows from $g''(y) \geq f''(y) \geq 0$.

Example 3.2. If a concave quadratic g majorizes a twice-differentiable function f at y, then f is concave at y. This follows from $0 \geq g''(y) \geq f''(y)$.

Example 3.3. Quadratic majorizers can be concave. Take $f(x) = -x^2$ and $g(x) = -x^2 + \frac{1}{2}(x - y)^2$.

Example 3.4. Quadratic majorizers may not exist anywhere. Suppose, for example, that f is a cubic. If g is quadratic, then $d = g - f$ is a cubic, and $d(x)$ is negative for at least one value of x.

Example 3.5. Quadratic majorizers may exist almost everywhere, but not everywhere. Suppose, for example, that $f(x) = |x|$. Then f has a quadratic majorizer at each y except $y = 0$. If $y \neq 0$ we can use, following Heiser [1986], the arithmetic mean-geometric mean inequality in the form

$$\sqrt{x^2 y^2} \leq \frac{1}{2}(x^2 + y^2),$$

and find

$$|x| \leq \frac{1}{2|y|}x^2 + \frac{1}{2}|y|.$$

If g majorizes $|x|$ at 0, then we must have $ax^2 + bx \geq |x|$ for all $x \neq 0$, and thus $a|x| + b\,\mathbf{sign}(x) \geq 1$ for all $x \neq 0$. But for $|x| < \frac{1+|b|}{a}$ and $\mathbf{sign}(x) = -\mathbf{sign}(b)$, we have $a|x| + b\,\mathbf{sign}(x) < 1$.

Example 3.6. For a nice regular example we use the celebrated functions

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \\ \Phi(x) &= \int_{-\infty}^{x} \phi(z)\, dz. \end{aligned}$$

Then

$$\begin{aligned}
\Phi'(x) &= \phi(x), \\
\Phi''(x) &= \phi'(x) = -x\phi(x), \\
\Phi'''(x) &= \phi''(x) = -(1-x^2)\phi(x), \\
\Phi''''(x) &= \phi'''(x) = -x(x^2-3)\phi(x).
\end{aligned}$$

It follows, by setting various derivatives to zero and checking for maxima and minima, that

$$\begin{aligned}
0 \leq \Phi'(x) &= \phi(x) \leq \phi(0), \\
-\phi(1) \leq \Phi''(x) &= \phi'(x) \leq \phi(1), \\
-\phi(0) \leq \Phi'''(x) &= \phi''(x) \leq 2\phi(\sqrt{3}).
\end{aligned}$$

Thus we have the quadratic majorizers

$$\Phi(x) \leq \Phi(y) + \phi(y)(x-y) + \frac{1}{2}\phi(1)(x-y)^2,$$

and

$$\phi(x) \leq \phi(y) - y\phi(y)(x-y) + \phi(\sqrt{3})(x-y)^2.$$

This is illustrated for both Φ and ϕ at the points $y = 0$ and $y = -3$ in Figures 1 and 2.

[Figure 1 about here.]

[Figure 2 about here.]

4. Sharp Quadratic Majorization

We now drop the assumption that the objective function is twice differentiable, even locally, and we try to improve our bound estimates at the same time.

4.1. **Differentiable Case.** Let us first deal with the case in which f is differentiable in y. Consider all $a > 0$ for which

$$f(x) \leq f(y) + f'(y)(x-y) + \frac{1}{2}a(x-y)^2$$

for a fixed y and for all x. Equivalently, we must have

$$a \geq \frac{f(x) - f(y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2}. \tag{1}$$

Define the function

$$\delta(x,y) = \frac{f(x) - f(y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2}$$

for all $x \neq y$. The inequalities (1) have a solution if and only if

$$A(y) = \sup_x \delta(x,y) < \infty.$$

If this is the case, then any $a \geq A(y)$ will satisfy (1). Because we want a to be as small as possible, we will usually prefer to choose $a = A(y)$. This is what we mean by the *sharp quadratic majorization*. If the second derivative is uniformly bounded by B, we have $A(y) \leq B$, and thus our bound improves on the uniform bound considered before.

The function δ has some interesting properties. If f is convex we have $\delta(x,y) \geq 0$ and for a concave f we have $\delta(x,y) \leq 0$. For strictly convex and concave f these inequalities are strict. If $\delta(x,y) \leq 0$ for all x and y, then f must be concave. Consequently $A(y) \leq 0$ only if f is concave, and without loss of generality we can exclude this case from consideration.

Clearly $\delta(x,y)$ is closely related to the second derivative at or near y. If f is twice differentiable at y, then

$$\lim_{x \to y} \delta(x,y) = f''(y). \tag{2}$$

If f is three times differentiable, this can be sharpened to

$$\lim_{x \to y} \frac{\delta(x,y) - f''(y)}{x-y} = \frac{1}{6}f'''(y).$$

Moreover, in the twice differentiable case, the mean value theorem implies there is a ξ in the interval extending from x to y with $\delta(x,y) = f''(\xi)$. We can also derive an integral representation of $\delta(x,y)$ and its first derivative with respect to x [Tom Ferguson, Personal Communication, 03/12/04].

Lemma 4.1. *$\delta(x,y)$ can written as the expectation*

$$\delta(x,y) = \mathbf{E}\{f''[Vy + (1-V)x]\},$$

where the random variable V follows a $\beta(2,1)$ distribution. Likewise

$$\delta'(x,y) = \frac{1}{3}\mathbf{E}\{f'''[Wy + (1-W)x]\},$$

where the random variable W follows a $\beta(2,2)$ distribution. Thus $\delta(x,y)$ and $\delta'(x,y)$ can be interpreted as smoothed versions of f'' and f'''.

Proof. The first representation follows from the second-order Taylor's expansion

$$f(x) = f(y) + f'(y)(x-y) + (x-y)^2 \int_0^1 f''[vy + (1-v)x]v\,dv$$

with integral remainder [Lange, 2004]. This form of the remainder can be deduced by integration by parts. Differentiation under the integral sign yields the second representation. □

In view of Lemma 4.1, $\delta(x,y)$ is jointly continuous in x and y when $f''(x)$ is continuous. Furthermore, if $f''(x)$ tends to ∞ as x tends to $-\infty$ or $+\infty$, then $\delta(x,y)$ is unbounded in x for each fixed y. Thus, quadratic majorizations do not exist for any y if the second derivative grows unboundedly. It also follows from Lemma 4.1 that the best quadratic majorization does not exist if the third derivative f''' is always positive (or always negative). This happens, for instance, if the first derivative f' is strictly convex or strictly concave. Thus as mentioned earlier, cubics do not have quadratic majorizations.

Example 4.1. Majorization may be possible at all points y without the function $A(y)$ being bounded. Suppose the graph of $f''(x)$ is 0 except for an isosceles triangle centered at each integer $n \geq 2$. If we let the base of the triangle be $2n^{-3}$ and the height of the triangle be n, then the area under the triangle is n^{-2}. The formulas

$$f'(x) = \int_0^x f''(y)\,dy, \quad f(x) = \int_0^x f'(y)\,dy$$

define a nonnegative convex function $f(x)$ satisfying

$$f'(x) \leq \sum_{n=2}^{\infty} \frac{1}{n^2} < \infty.$$

To prove the $A(y)$ is finite for every y, recall the limit (2) and observe that

$$\delta(x,y) = \frac{f'(w)(x-y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2} = \frac{f'(w) - f'(y)}{\frac{1}{2}(x-y)}$$

for some w between x and y. It follows that $\delta(x,y)$ tends to 0 as $|x|$ tends to ∞. Because $A(n) \geq f''(n) = n$, it is clear that $A(y)$ is unbounded.

4.2. **Computing the Sharp Quadratic Majorization.** Let us study the case in which the supremum of $\delta(x,y)$ over $x \neq y$ is attained at, say, $z \neq y$. In our earlier notation $A(y) = \delta(z,y)$. Differentiating $\delta(x,y)$ with respect to x gives

$$\delta'(x,y) = \frac{\frac{1}{2}(x-y)^2[f'(x) + f'(y)] - (x-y)[f(x) - f(y)]}{\frac{1}{4}(x-y)^4},$$

and

$$\frac{f(z) - f(y)}{z - y} = \frac{1}{2}[f'(z) + f'(y)] \tag{3}$$

is a necessary and sufficient condition for $\delta'(z,y)$ to vanish. At the optimal z we have

$$A(y) = \delta(z,y) = \frac{f'(z) - f'(y)}{z - y}. \tag{4}$$

It is interesting that the fundamental theorem of calculus allows us to recast equations (3) and (4) as

$$\begin{aligned}\frac{1}{2}[f'(z)+f'(y)] &= \int_0^1 f'[z+t(y-z)]\,dt \\ A(y) &= \int_0^1 f''[z+t(y-z)]\,dt.\end{aligned}$$

When f is convex, $A(y) \geq 0$. For the second derivative at z, we have

$$\delta''(z,y) = \frac{(z-y)^2 f''(z) - [f'(z)-f'(y)](z-y)}{\frac{1}{2}(z-y)^4}.$$

At a maximum we must have $\delta''(z,y) \leq 0$, which is equivalent to

$$f''(z) \leq \frac{f'(z)-f'(y)}{z-y} = A(y). \tag{5}$$

We can achieve more clarity by viewing these questions from a different angle. If the quadratic g majorizes f at y, then it satisfies

$$g(x) = f(y) + f'(y)(x-y) + \frac{1}{2}a(x-y)^2$$

for some a. If z is a second support point, then g not only intersects f at z, but it also majorizes f at z. The condition $g'(z) = f'(z)$ yields

$$a = \frac{f'(z)-f'(y)}{z-y}.$$

If we match this value with the requirement $\delta(z,y) = a$, then we recover the second equality in (4). Conversely, if a point z satisfies the second equality in (4), then it is a second support point. In this case, one can easily check condition (3) guaranteeing that z is a stationary point of $\delta(x,y)$.

4.3. **Optimality with Two Support Points.** Quadratic functions with two support points have occurred in various instances in specific majorization algorithms. For simple symmetric examples involving the absolute value function and other symmetric robust loss functions we refer, for example, to Heiser [1987] and Verboon and

Heiser [1994]. For quadratic approximations to penalized likelihood functions used for variable selection in linear regression we refer to Fan and Li [2001] and Hunter and Li [2005]. In both cases the existence and location of the second support point naturally follows from the symmetry around zero of the robust loss functions and the various penalty functions.

Extensions to the general case, and explicit introduction of the notion of sharp quadratic approximation, started with work by Groenen et al. [2003]. Van Ruitenburg [2005] proves that a quadratic function g majorizing a differentiable function f at two points must be a sharp majorizer. We now summarize in our language Van Ruitenburg's lovely proof of this fact.

Lemma 4.2. *Suppose two quadratic functions $g_1 \neq g_2$ both majorize the differentiable function f at y. Then either g_1 strictly majorizes g_2 at y or g_1 strictly majorizes g_2 at y.*

Proof. We have

$$g_1(x) = f(y) + f'(y)(x-y) + \frac{1}{2}a_1(x-y)^2, \tag{6}$$

$$g_2(x) = f(y) + f'(y)(x-y) + \frac{1}{2}a_2(x-y)^2, \tag{7}$$

with $a_1 \neq a_2$. Subtracting (6) and(7) proves the theorem. □

Lemma 4.3. *Suppose the quadratic function g_1 majorizes a differentiable function f at y and $z_1 \neq y$ and that the quadratic function g_2 majorizes f at y and $z_2 \neq y$. Then $g_1 = g_2$.*

Proof. Suppose $g_1 \neq g_2$. Since both g_1 and g_2 majorize f at y, Lemma 4.2 applies. If g_2 strictly majorizes g_1 at y, then $g_1(z_2) < g_2(z_2) = f(z_2)$, and g_1 does not majorize f. If g_1 strictly majorizes g_2 at y, then similarly $g_2(z_1) < g_1(z_1) = f(z_1)$, and g_2 does not majorize f. Unless $g_1 = g_2$, we reach a contradiction. □

We now come to Van Ruitenburg's main result.

Theorem 4.4. *Suppose a quadratic function g_1 majorizes a differentiable function f at y and at $z \neq y$, and suppose $g_2 \neq g_1$ majorizes f at y. Then g_2 strictly majorizes g_1 at y.*

Proof. Suppose g_1 strictly majorizes g_2. Then $g_2(z) < g_1(z) = f(z)$ and thus g_2 does not majorize f. The result now follows from Lemma 4.2. □

[Figure 3 about here.]

Example 4.2. It is not true, by the way, that a quadratic majorizer can have at most two support points. There can even be an infinite number of them. Consider the function $h(x) = c\sin^2(x)$ for some $c > 0$. Clearly $h(x) \geq 0$ and $h(x) = 0$ for all integer multiples of π. Now define $f(x) = x^2 - h(x)$ and $g(x) = x^2$. Then g is a quadratic majorizer of f at all integer multiples of π. This is plotted in Figure 3 for $c = 10$.

Example 4.3. There is no guarantee that a second support point $z \neq y$ exists. Consider the continuously differentiable convex function

$$f(x) = \begin{cases} x^2 & \text{if } \leq 1 \\ 2x - 1 & \text{if } x > 1, \end{cases}$$

and fix $y > 1$. For $x > 1$

$$\delta(x,y) = \frac{2x - 1 - 2y + 1 - 2(x-y)}{\frac{1}{2}(x-y)^2} = 0.$$

For $x \leq 1$

$$\delta(x,y) = \frac{x^2 - 2y + 1 - 2(x-y)}{\frac{1}{2}(x-y)^2} = \frac{(x-1)^2}{\frac{1}{2}(x-y)^2}.$$

It follows that $\lim_{x\to-\infty}\delta(x,y) = 2$. On the other hand, one can easily demonstrate that $\delta(x,y) < 2$ whenever $x \leq 1$. Hence, $A(y) = 2$, but $\delta(x,y) < 2$ for all $x \neq y$.

4.4. **Even Functions .** Assuming that $f(x)$ is even simplifies the construction of quadratic majorizers. If an even quadratic g satisfies $g(y) = f(y)$ and $g'(y) = f'(y)$, then it also satisfies $g(-y) = f(-y)$ and $g'(-y) = f'(-y)$. If in addition g majorizes f at either y or $-y$, then it majorizes $f$at both y and $-y$, and Theorem 4.4 implies that it is the best possible majorization at both points. This means we only need an extra condition to guarantee that g majorizes f. The next theorem, essentially proved in the references Jaakkola and Jordan [2000]; Groenen et al. [2003]; Hunter and Li [2005] by other techniques, highlights an important sufficient condition.

Theorem 4.5. *Suppose $f(x)$ is an even, differentiable function on R such that the ratio $f'(x)/x$ is decreasing on $(0,\infty)$. Then the even quadratic*

$$g(x) = \frac{f'(y)}{2y}(x^2 - y^2) + f(y)$$

is the best majorizer of $f(x)$ at the point y.

Proof. It is obvious that $g(x)$ is even and satisfies the tangency conditions $g(y) = f(y)$ and $g'(y) = f'(y)$. For the case $0 \leq x \leq y$, we have

$$\begin{aligned} f(y) - f(x) &= \int_x^y f'(z)\,dz \\ &= \int_x^y \frac{f'(z)}{z} z\,dz \\ &\geq \frac{f'(y)}{y}\int_x^y z\,dz \\ &= \frac{f'(y)}{y}\frac{1}{2}(y^2 - x^2) \\ &= f(y) - g(x). \end{aligned}$$

It follows that $g(x) \geq f(x)$. The case $0 \leq y \leq x$ is proved in similar fashion, and all other cases reduce to these two cases given that $f(x)$ and $g(x)$ are even. □

There is an condition equivalent to the sufficient condition of Theorem 4.5 that is sometimes easier to check.

Theorem 4.6. *The ratio $f'(x)/x$ is decreasing on $(0,\infty)$ if and only $f(\sqrt{x})$ is concave. The set of functions satisfying this condition is a closed under the formation of (a) positive multiples, (b) convex combinations, (c) limits, and (d) composition with a concave increasing function $g(x)$.*

Proof. Suppose $f(\sqrt{x})$ is concave and $x > y$. Then the two inequalities

$$\begin{aligned} f(\sqrt{x}) &\leq f(\sqrt{y}) + \frac{f'(\sqrt{y})}{2\sqrt{y}}(x-y) \\ f(\sqrt{y}) &\leq f(\sqrt{x}) + \frac{f'(\sqrt{x})}{2\sqrt{x}}(y-x) \end{aligned}$$

are valid. Adding these, subtracting the common sum $f(\sqrt{x}) + f(\sqrt{y})$ from both sides, and rearranging give

$$\frac{f'(\sqrt{x})}{2\sqrt{x}}(x-y) \leq \frac{f'(\sqrt{y})}{2\sqrt{y}}(x-y).$$

Dividing by $(x-y)/2$ yields the desired result

$$\frac{f'(\sqrt{x})}{\sqrt{x}} \leq \frac{f'(\sqrt{y})}{\sqrt{y}}.$$

Conversely, suppose the ratio is decreasing and $x > y$. Then the mean value expansion

$$f(\sqrt{x}) = f(\sqrt{y}) + \frac{f'(\sqrt{z})}{2\sqrt{z}}(x-y)$$

for $z \in (y,x)$ leads to the concavity inequality.

$$f(\sqrt{x}) \leq f(\sqrt{y}) + \frac{f'(\sqrt{y})}{2\sqrt{y}}(x-y).$$

The asserted closure properties are all easy to check. □

As examples of property (d) of Theorem 4.6, note that the functions $g(x) = \ln x$ and $g(x) = \sqrt{x}$ are concave and increasing.

Hence, if $f(\sqrt{x})$ is concave, then $\ln f(\sqrt{x})$ and $f(\sqrt{x})^{1/2}$ are concave as well.

The above discussion suggests that we look at more general transformations of the argument of f. If we define $\tilde{f}(x) = f(\alpha + \beta x)$ for an arbitrary function $f(x)$, then a brief calculation shows that

$$\begin{aligned}\tilde{A}(y) &= \beta^2 A(\alpha + \beta y)\\ \tilde{z}(y) &= \frac{z(\alpha + \beta y) - \alpha}{\beta}\end{aligned}$$

using the identity $\tilde{\delta}(x, y) = \beta^2 \delta(\alpha + \beta x, \alpha + \beta y)$. An even function $f(x)$ satisfies $\tilde{f}(x) = f(x)$ for $\alpha = 0$ and $\beta = -1$.

4.5. **Non-Differentiable Functions.** If f is not differentiable at y, then we must find a and b such that

$$f(x) \leq f(y) + b(x - y) + \frac{1}{2}a(x - y)^2.$$

for all x. This is an infinite system of linear inequalities in a and b, which means that the solution set is a closed convex subset of the plane.

Analogous to the differentiable case we define

$$\delta(x, y, b) = \frac{f(x) - f(y) - b(x - y)}{\frac{1}{2}(x - y)^2},$$

as well as

$$A(y, b) = \sup_x \delta(x, y, b).$$

If $A(y, b) < +\infty$, we have the sharpest quadratic majorization for given y and b. The sharpest quadratic majorization at y is given by

$$A(y) = \inf_b A(y, b).$$

5. Examples

5.1. **Logistic.** Our first example is the negative logarithm of the logistic cdf

$$\Psi(x) = \frac{1}{1+e^{-x}}.$$

Thus

$$f(x) = \log(1+e^{-x}).$$

Clearly

$$f'(x) = -\frac{e^{-x}}{1+e^{-x}} = \Psi(x) - 1,$$

and

$$f''(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \Psi(x)[1-\Psi(x)].$$

This shows that $f(x)$ is strictly convex. Since $f''(x) \leq 1/4$, a uniform bound is readily available.

The symmetry relations

$$\begin{aligned} f(-x) &= x + f(x), \\ f'(-x) &= -[1+f'(x)] = -\Psi(x), \\ f''(-x) &= f''(x). \end{aligned}$$

demonstrate that $z = -y$ satisfies equation (3) and hence maximizes $\delta(x,y)$. The optimum value is determined by (4) as

$$A(y) = \delta(z,y) = \frac{2\Psi(y)-1}{2y}.$$

The same result was derived, using quite different methods, by Jaakkola and Jordan [2000] and Groenen et al. [2003].

[Figure 4 about here.]

We plot the function $\delta(x,y)$ for $y = 1$ and $y = 8$ in Figure 4. Observe that the uniform bound $1/4$ is not improved much for y close to 0, but for large values of y the improvement is huge. This is because $A(y) \approx (2|y|)^{-1}$ for large $|y|$.

Alternatively, we can majorize $f(x) = \log(1 + e^{-x})$ by writing

$$\log(1 + e^{-x}) = -\frac{1}{2}x + \log(e^{x/2} + e^{-x/2})$$

and majorizing the even function $h(x) = \log(e^{x/2}+e^{-x/2})$. Straightforward but tedious differentiation shows that

$$\begin{aligned}\left[\frac{h'(x)}{x}\right]' &= \frac{1 - e^{2x} + 2xe^x}{2x^2(1 + e^x)^2}\\ &= \frac{1}{2x^2(1 + e^x)^2}\sum_{k=2}^{\infty}\left[2x\frac{x^k}{k!} - \frac{(2x)^{k+1}}{(k+1)!}\right]\\ &= \frac{2}{2x^2(1 + e^x)^2}\sum_{k=2}^{\infty}\frac{x^{k+1}}{k!}\left[1 - \frac{2^k}{k+1}\right]\\ &\leq 0.\end{aligned}$$

Hence, $h'(x)/x$ is decreasing on $(0, \infty)$, and Theorem 4.5 applies.

Of course minimizing a single function $f(x) = \log(1 + e^{-x})$ is not interesting. But a weighted sum of such functions defines logistic regression, and our non-uniform quadratic majorization provides an algorithm for logistic regression with a rapid convergence rate. In De Leeuw [2006a,b] quadratic logistic majorization is used to define majorization algorithms for principal component analysis of binary data. This is extended to multi-category data and to various distance-based association models for cross-tables and indicator matrices in De Leeuw [2006c,d]. The corresponding algorithms are all based on sharp quadratic majorization of the logistic function.

5.2. **Probit.** In Probit and Tobit analysis the negative log-likelihood is a weighted sum of terms involving $f(x) = -\log\Phi(x)$, with Φ the cumulative standard normal. It is well-known that f is strictly convex, implying that $f''(x) > 0$ for all x. It was shown by Böhning [1999] that in addition $f(x) < 1$ for all x as well, which provides a uniform bound for quadratic majorization.

De Leeuw [2006b] gave an alternative proof, based on simple inequalities for Mill's Ratio derived by Sampford [1953]. He also

stated, but did not prove, that uniform quadratic majorization is sharp. In fact, that last result follows easily from the asymptotic expansion of Mill's Ratio. We can use

$$1 - \Phi(x) \sim \frac{\phi(x)}{x}$$

as $x \to \pm\infty$ to show that $\lim_{x\to\pm\infty} \delta(x,y) = 1$ for all y.

Consequently, from the computational point of view, there is an important difference between logit and probit techniques. In the logistic case, which includes applications to logistic regression, Rasch models, and multivariate logistic item response models, we can accelerate convergence by optimal quadratic majorization. For solutions with logistic probabilities close to zero and one, the acceleration will be very substantial. But the corresponding probit techniques cannot be optimized in the same way, because uniform quadratic majorization is already sharp. This makes a large difference in optimizing logistic or probit regression, and the more complicated techniques for principal component analysis and multidimensional scaling discussed before.

5.3. **The Absolute Value Function.** Because $|x|$ is even, Theorem 4.5 yields the majorization

$$g(x) = \frac{1}{2|y|}(x^2 - y^2) + |y| = \frac{1}{2|y|}x^2 + \frac{1}{2}|y|,$$

which is just the result given by the arithmetic/geometric mean inequality in Example 3.5. When $y = 0$, recall that no quadratic majorization exists.

If we approach majorization of $|x|$ directly, we need to find $a > 0$ and b such that

$$a(x-y)^2 + b(x-y) + |y| \geq |x|$$

for all x. Let us compute $A(y,b)$. If $y < 0$ then $b = -1$, and thus

$$A(y,-1) = \sup_{x\neq y} \frac{|x| + x}{\frac{1}{2}(x-y)^2} = \frac{1}{|y|}.$$

If $y > 0$ then $b = +1$, and again

$$A(y, +1) = \sup_{x \neq y} \frac{|x| - x}{\frac{1}{2}(x-y)^2} = \frac{1}{|y|}.$$

In both cases, the best quadratic majorizer can be expressed as

$$\begin{aligned} g(x) &= \frac{1}{2}\frac{1}{|y|}(x-y)^2 + \mathbf{sign}(y)(x-y) + |y| \\ &= \frac{1}{2|y|}x^2 + \frac{1}{2}|y|. \end{aligned}$$

5.4. **The Huber Function.** Majorization for the Huber function, specifically quadratic majorization, has been studied earlier by Heiser [1987] and Verboon and Heiser [1994]. In those papers quadratic majorization functions appear more or less out of the blue, and it is then verified that they are indeed majorization functions. This is not completely satisfactory. Here we attack the problem by applying Theorem 4.5. This automatically leads to the sharpest quadratic majorization.

The Huber function is defined by

$$f(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < c, \\ c|x| - \frac{1}{2}c^2 & \text{otherwise.} \end{cases}$$

Thus we really deal with a family of even functions, one for each $c > 0$. The Huber functions are differentiable with derivative

$$f'(x) = \begin{cases} x & \text{if } |x| < c, \\ c & \text{if } x \geq c, \\ -c & \text{if } x \leq -c. \end{cases}$$

Since it is obvious that $f'(x)/x$ is decreasing $(0, \infty)$, Theorem 4.5 immediately gives the sharpest majorizer

$$g(x) = \begin{cases} \frac{1}{2}\frac{c}{|y|}(x-y)^2 - cx - \frac{1}{2}c^2 & \text{if } y \leq -c, \\ \frac{1}{2}x^2 & \text{if } |y| < c, \\ \frac{1}{2}\frac{c}{|y|}(x-y)^2 + cx - \frac{1}{2}c^2 & \text{if } y \geq +c. \end{cases}$$

6. Iterative Computation of $A(y)$

In general, one must find $A(y)$ numerically. Observe that in previous papers heuristics were used to find a second support point. In some cases, however, thee are no heuristics, and we need to actually optimize $\delta(x,y)$ over x for given y

For a convex function f, two similar iterative algorithms are available. They both depend on minorizing f by the linear function $f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)})$ at the current point $x^{(k)}$ in the search for the maximum z of $\delta(x,y)$. This minorization propels the further minorization

$$\begin{aligned}\delta(x,y) &\geq \frac{f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) - f(y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2} \\ &= \frac{[f'(x^{(k)}) - f'(y)](x-y) + f(x^{(k)}) + f'(x^{(k)})(y - x^{(k)}) - f(y)}{\frac{1}{2}(x-y)^2}.\end{aligned}$$

Maximizing the displayed minorizer drives $\delta(x,y)$ uphill. Fortunately, the minorizer is a function of the form

$$h(w) = \frac{cw + d}{w^2} = \frac{c}{w} + \frac{d}{w^2}$$

with $w = x - y$. The stationary point $w = -2d/c$ furnishes the maximum of $h(w)$ provided

$$h''\left(-\frac{2d}{d}\right) = \frac{2c}{w^3}\Big|_{w=-2d/c} + \frac{6d}{w^4}\Big|_{w=-2d/c} = \frac{c^4}{8d^3}$$

is negative. If $f(x)$ is strictly convex, then

$$d = 2\left[f(x^{(k)}) + f'(x^{(k)})(y - x^{(k)}) - f(y)\right],$$

is negative, and the test for a maximum succeeds. The update can be phrased as

$$x^{(k+1)} = y - 2\,\frac{f(x^{(k)}) + f'(x^{(k)})(y - x^{(k)}) - f(y)}{f'(x^{(k)}) - f'(y)}.$$

A brief calculation based on equations (3) and (4) shows that the iteration map $x^{(k+1)} = g(x^{(k)})$ has derivative

$$g'(z) = \frac{f''(z)(z-y)}{f'(z) - f'(y)} = \frac{f''(z)}{A(y)}$$

at the optimal point z.

On the other hand, the Dinkelbach [1967] maneuver for increasing $h(w)$ considers the function $e(w) = cw + d - h(w^{(k)})w^2$ with value $e(w^{(k)}) = 0$. If we choose

$$w^{(k+1)} = \frac{c}{2h(w^{(k)})}$$

to maximize $e(w)$, then it is obvious that $h(w^{(k+1)}) \geq h(w^{(k)})$. This gives the iteration map

$$x_{n+1} = y + \frac{\frac{1}{2}[f'(x^{(k)}) - f'(y)](x^{(k)} - y)^2}{f(x^{(k)}) - f(y) - f'(y)(x^{(k)} - y)} = y + \frac{f'(x^{(k)}) - f'(y)}{\delta(x^{(k)}, y)}$$

with derivative at z equal to $f''(z)/A(y)$ by virtue of equations (3) and (4). Hence, the two algorithms have the same local rate of convergence. We recommend starting both algorithms near y. In the case of the Dinkelbach algorithm, this entails

$$h(w) \approx \delta(x, y) \approx f''(y) > 0$$

for $f(x)$ strictly convex. Positivity of $h(w^{(0)})$ is required for proper functioning of the algorithm.

In view of the convexity of $f(x)$, it is clear that $f''(z)/A(y) \geq 0$. The inequality $f''(z) \leq A(y)$ follows from the condition $A(y) = A(z)$ determined by Theorem 4.4 and inequality (5). Ordinarily, strict inequality $f''(z) < A(y)$ prevails, and the two iteration maps just defined are locally contractive. Globally, the standard convergence theory for iterative majorization (MM algorithms) suggests that $\lim_{n\to\infty} |x^{(k+1)} - x^{(k)}| = 0$ and that the limit of every convergent subsequence must be a stationary point of $\delta(x, y)$ [Lange, 2004].

REFERENCES

D. Böhning. The Lower Bound Method in Probit Regression. *Computational Statistics and Data Analysis*, 30:13–17, 1999.

D. Böhning and B.G. Lindsay. Monotonicity of Quadratic-approximation Algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988.

J. De Leeuw. Nonlinear Principal Component Analysis and Related Techniques. In M. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, 2006a.

J. De Leeuw. Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition. *Computational Statistics and Data Analysis*, 50(1):21–39, 2006b.

J. De Leeuw. Majorization Methods for Logit, Probit, and Tobit Models. Preprint 489, UCLA Department of Statistics, 2006c. URL `http://preprints.stat.ucla.edu/489/econ50Course.pdf`.

J. De Leeuw. Block Relaxation Methods in Statistics. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data Analysis*, Berlin, 1994. Springer Verlag.

J. De Leeuw. Gifi Goes Logistic. Preprint 492, UCLA Department of Statistics, 2006d. URL `http://preprints.stat.ucla.edu/492/statistDag.pdf`.

W. Dinkelbach. On Nonlinear Fractional Programming. *Management Science*, 13:492–498, 1967.

J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of American Statistical Association*, 96:1348–1360, 2001.

P.J.F. Groenen, P. Giaquinto, and H.L Kiers. Weighted Majorization Algorithms for Weighted Least Squares Decomposition Models. Technical Report EI 2003-09, Econometric Institute, Erasmus University, Rotterdam, Netherlands, 2003.

W.J. Heiser. Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In W.J.

Krzanowski, editor, *Recent Advanmtages in Descriptive Multivariate Analysis*, pages 157–189. Oxford: Clarendon Press, 1995.

W.J. Heiser. Correspondence Analysis with Least Absolute Residuals. *Computational Statistica and Data Analysis*, 5:357–356, 1987.

W.J. Heiser. A Majorization Algorithm for the Reciprocal Location Problem. Technical Report RR-86-12, Department of Data Theory, University of Leiden, 1986.

D. R. Hunter and R. Li. Variable Selection Using MM Algorithms. *The Annals of Statistics*, 33:1617–1642, 2005.

D.R. Hunter and K. Lange. A Tutorial on MM Algorithms. *American Statistician*, 58(30–37), 2004.

T.S. Jaakkola and M. I. Jordan. Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, 10:25–37, 2000.

K. Lange, D.R. Hunter, and I. Yang. Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, 9:1–20, 2000.

J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, N.Y., 1970.

M.R. Sampford. Some Inequalities on Mill's ratio and Related Functions. *Annals of Mathematical Statistics*, 24:130–132, 1953.

J. Van Ruitenburg. Algorithms for Parameter Estimation in the Rasch Model. Measurement and Research Department Reports 2005-04, CITO, Arnhem, Netherlands, 2005.

P. Verboon and W.J. Heiser. Resistant Lower Rank Approximation of Matrices by Iterative Majorization. *Computational Statistics and Data Analysis*, 18:457–467, 1994.

(Jan de Leeuw) Department of Statistics, University of California, Los Angeles, CA 90095

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu

(Kenneth Lange) Departments of Biomathematics,Human Genetics, and Statistics, University of California, Los Angeles, CA 90095,

E-mail address, Kenneth Lange: klange@ucla.edu

URL, Kenneth Lange: http://www.biomath.ucla.edu/faculty/klange/klange.htm

List of Figures

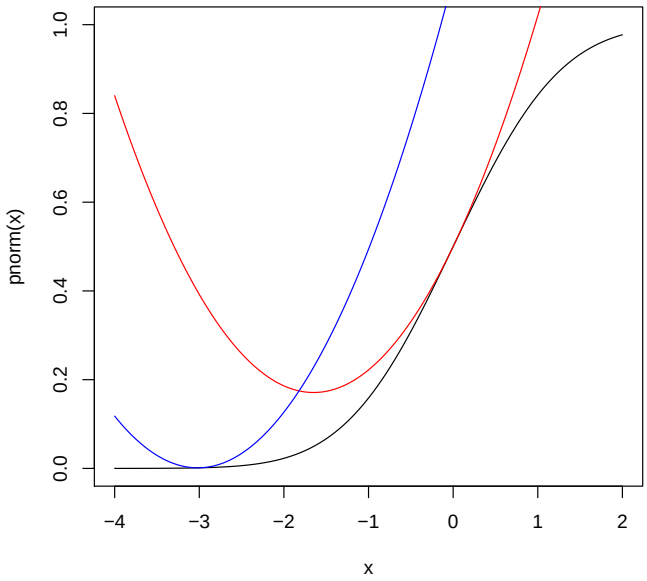

FIGURE 1. Quadratic majorization of cumulative normal

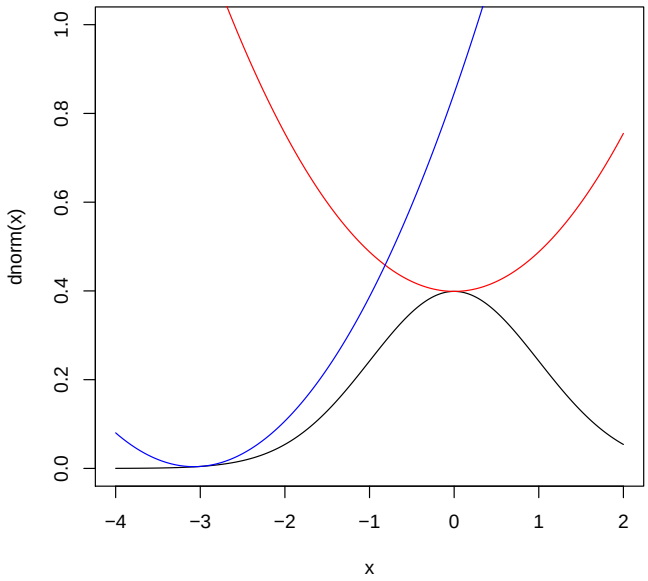

FIGURE 2. Quadratic majorization of normal density

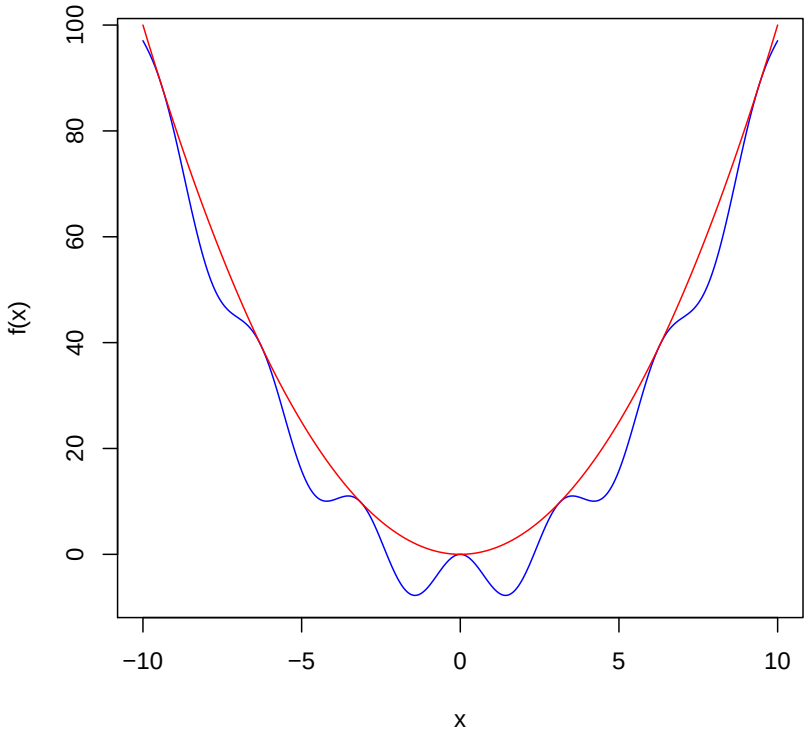

FIGURE 3. Many support points.

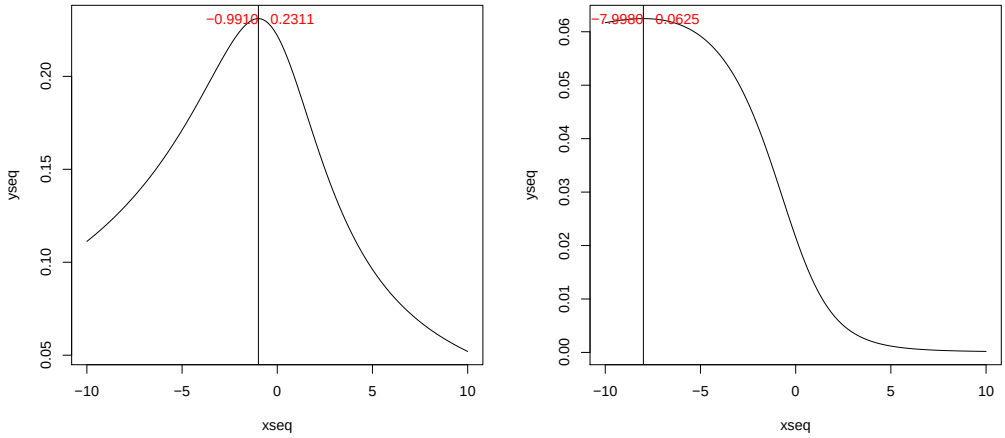

FIGURE 4. δ for logistic at $y = 1$ (left) and $y = 8$ (right).