QUADRATIC MAJORIZATION

JAN DE LEEUW

1. INTRODUCTION

Majorization methods are used extensively to solve complicated multivariate optimizaton problems. We refer to de Leeuw [1994]; Heiser [1995]; Lange et al. [2000] for overviews.

The general idea of majorization methods is easy to explain. Suppose we are minimizing a proper lower-semicontinuous $f : \mathbb{R}^n \to (-\infty, +\infty]$,and suppose our current best guess of the solution is y. We now construct an auxiliary function g which is at least as large as f on all of $\mathbb{R}^n$, and which is equal to f at y. Thus g *majorizes* f and *touches* f in y.

One step of our algorithm tries to minimize g over x. Let us ignore problems of existence and uniqueness of the minimum at the moment, and suppose the minimum is attained at z. Then $f(z) \leq g(z)$, because g majorizes f, and $g(z) \leq g(y)$, because z minimizes g. Finally $g(y) = f(y)$, because g touches f in y. Thus $f(z) \leq f(y)$ and we have decreased the function we are trying to minimize. Repeat the same process, i.e. construct a new majorization, now at z, and minimize it. And so on.

It only makes sense to consider this algorithm if the minimization of the majorizing functions g in the substeps is considerably simpler than the original minimization of the target f. Fortunately it turns out that in many practical examples we can routinely construct majorization functions which are easy to minimize. Many of these examples are reviewed in the publications cited above, some particularly attractive additional ones are in Böhning and Lindsay [1988]; Böhning [1992]; Kiers [1995, 1990].

We limit ourselves in this paper to unconstrained minimization. In many cases we can incorporate constraints $x \in S$, by minimizing $f(x) + \delta(x, S)$, where

$$\delta(x, S) = \begin{cases} 0 & \text{if } x \in S, \\ +\infty & \text{if } x \notin S. \end{cases}$$

Date: June 27, 2003.

In other cases slightly more subtle tools, such as penalty functions, Lagrangians and augmented Lagrangians, can be used to turn constrained problems into unconstrained ones. And in yet other cases, we take the constraints into account explicitly, and only minimize the majorization functions g over the subset defined by the constraints.

2. MAJORIZATION

Here we formalize what was said in the introduction.

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

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

Alternatively, g majorizes f at y if $d = g - f$ has a global minimum, equal to zero, at y.

We say that majorization is $strict$ if

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

Alternatively, g majorizes f strictly at y if $d = g - f$ has a unique global minimum, equal to zero, at y.

Example 2.1. Let

$$g(x) = \begin{cases} f(x) + 1 & \text{for all } x \neq y, \\ f(x) & \text{for } x = y. \end{cases}$$

Then g majorizes f strictly at y.

We also give a non-local version of this definition, which deals with the situation in which majorizations exist at all y.

Definition 2.2. Suppose f is real-valued on $\mathbb{R}^n$, and h is real-valued on $\mathbb{R}^n \times \mathbb{R}^n$. Then h is a *majorization scheme* for f if

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

Alternatively, h is a majorization scheme if $h(\bullet, y)$ majorizes f at y for each y.

Or, alternatively, h is a majorization scheme if $f(y) = h(y, y) = \min_x h(x, y)$ for all y.

Strict majorization schemes are defined in the obvious way.

Example 2.2. Observe that if we define $h(x, y) = f(x)$, then h is a (trivial) majorization scheme for f. We can use Example 2.1 to define a strict majorization scheme as

$$h(x, y) = \begin{cases} f(x) + 1 & \text{for all } x \neq y, \\ f(x) & \text{for } x = y. \end{cases}$$

Example 2.3. To show that non-trivial majorizing schemes always exist, simply use

$$h(x, y) = f(x) + (x - y)'A(x - y)$$

with[1] $A \succeq 0$. We have strict majorization if $A \succ 0$.

Example 2.4. Suppose f is Lipschitz, i.e. there is a $K > 0$ such that

$$\|f(x) - f(y)\| \leq K\|x - y\|.$$

Then

$$h(x, y) = f(y) + K\|x - y\|$$

is a majorization scheme for f. Remember that for f to be Lipschitz it is sufficient that f is differentiable and $\|f'\| \leq K$.

Example 2.5. Define the celebrated functions

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

Then

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

[1] We use curly comparison symbols for the Loewner ordering of square symmetric matrices. Thus $A \succeq B$ means that $A - B$ is positive semidefinite, and $A \succ B$ means that $A - B$ is positive definite.

It follows that

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

Thus both ϕ and Φ are Lipschitz and have a bounded second derivative. has the quadratic majorization scheme

$$h(x,y) = f(y) + \frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2})(x-y) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2})(x-y)^2.$$

This is illustrated for $y = 0$ and $y = -3$ in Figure 2.

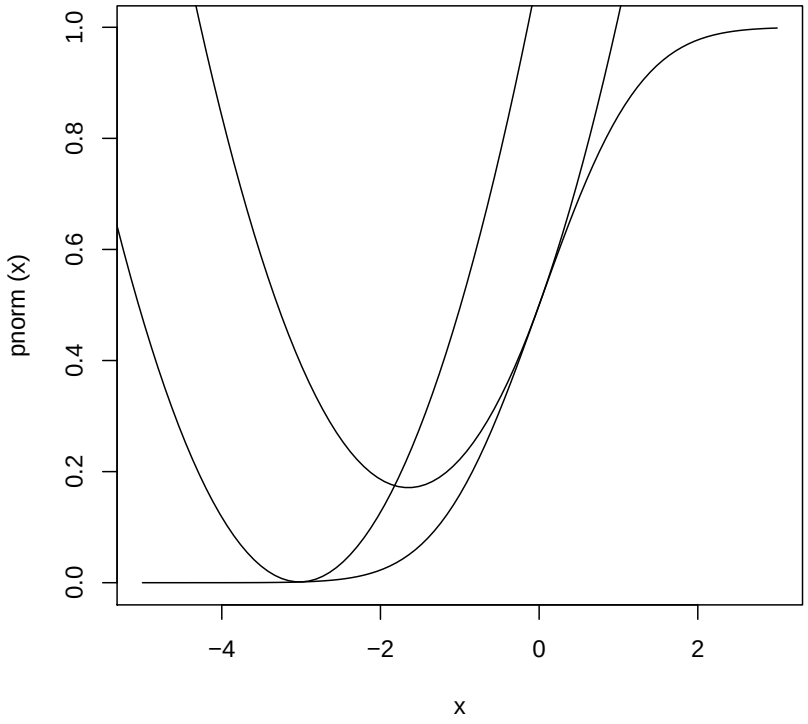

FIGURE 1. Quadratic majorization of cumulative normal

2.1. **Majorization sequences.** If the minimum does not exist, and g is unbounded below on $\mathcal{S}$, we stop. Because g majorizes f on $\mathcal{S}$, it follows that

f is unbounded below on $\mathcal{S}$, and our original problem does not have a solution. If the minimum does exist, it is not necessarily unique. Suppose $\mathcal{R} \subseteq \mathcal{S}$ is the set of minimizers. If $y \in \mathcal{R}$, we stop. Otherwise, we select z from $\mathcal{R}$.

If we don't stop, then $f(z) \leq g(z)$, because g majorizes f, and $g(z) < g(y)$, because z minimizes g over $\mathcal{S}$ (and y does not). Finally $g(y) = f(y)$, because g touches f in y. Thus $f(z) < f(y)$ and we have decreased the function we are trying to minimize. Repeat the same process, i.e. construct a new majorization, now at z, and minimize it.

2.2. If g majorizes f, and g is continuous, then

$$f(y) = g(y) = \lim_{x \to y} g(x) \geq \lim_{x \to y} f(x)$$

$$g(x) - g(y) \geq f(x) - f(y)$$

2.3. **Necessary Conditions.** We first show that majorization functions must have certain properties at the point where they touch the target. We look at the case in which we are minimizing over the whole space, and in which we have some differentiability.

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 [Cartan, 1971, Chapter 8]. □

In the non-differentiable case, life becomes more complicated. There are many extensions of the necessary conditions for a local minimum in the literature, and consequently there are many possible extensions of Theorem 2.1.

Theorem 2.2. *Suppose f is convex and g is differentiable at y. If g majorizes f at y then*

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

- $g'(y) \in \partial f(y)$.

2.4. **Composition.**

Theorem 2.3 (Sum of functions)**.** *Suppose $f = \int v(\bullet, u)dF(u)$, and suppose $w(\bullet, u)$ majorizes $v(\bullet, u)$ at y for all u. Then $g = \int w(\bullet, u)dF(u)$ majorizes f at y.*

Proof. $w(x, u) \geq v(x, u)$ for all x and all u, and thus $g(x) \geq f(x)$. Moreover $w(y, u) = v(y, u)$ for all u and thus $g(y) = f(y)$. □

Theorem 2.4 (Inf of functions)**.** *Suppose $f = \inf_u v(\bullet, u)$ and let $X(u) = \{x | f(x) = v(x, u)\}$. Suppose $y \in X(u)$ and g majorizes $v(\bullet, u)$ at y. Then g majorizes f at y.*

Proof. $g(x) \geq v(x, u) \geq \inf_u v(x, u) = f(x)$, and because $y \in X(u)$ also $g(y) = v(y, u) = f(y)$. □

Observe the theorem is not true for sup, and also we cannot say that if $w(\bullet, u)$ majorizes $v(\bullet, u)$ for all u at y, then $g = \inf_u w(\bullet, u)$ majorizes f at y.

Theorem 2.5 (Composition of functions)**.** *If g majorizes f at y and $\gamma : \mathbb{R} \to \mathbb{R}$ is non-decreasing, then $\gamma \circ g$ majorizes $\gamma \circ f$ at y. If, in addition, γ majorizes the non-decreasing $\eta : \mathbb{R} \to \mathbb{R}$ at $g(y)$, then $\gamma \circ g$ majorizes $\eta \circ f$.*

Proof. $g(x) \geq f(x)$ and thus $\gamma(g(x)) \geq \gamma(f(x))$. Also $g(y) = f(y)$ and thus $\gamma(g(y)) = \gamma(f(y))$. For the second part we have $\gamma(g(x)) \geq \eta(g(x)) \geq \eta(f(x))$ and $\gamma(g(y)) = \eta(g(y)) = \eta(f(y))$. □

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 convex quadratic majorizer. This leads to the algorithm

$$x^{(k+1)} = x^{(k)} - [A(x^{(k)})]^{-1} f'(x^{(k)}).$$

If the sequence converges to $\hat{x}$, then it does so with linear convergence rate equal to the largest eigenvalue of $I - [A(\hat{x})]^{-1} f''(\hat{x})$ [**?**]. We also see that choosing a larger A leads to faster linear convergence.

Thus the question we look at in the rest of this paper is to compute the quadratic majorizers g of f at y. The key result applies to all functions with a bounded and continuous second derivative.

Theorem 3.1. *If $f \in \mathcal{C}^2$. and there is an $A \succeq 0$ such that $f''(x) \preceq A$ for all x, then for each y there is a convex quadratic function g majorizing f at y.*

Proof. Use Taylor's theorem in the form

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

with ξ on the line connecting x and y. Because $f''(\xi) \preceq A$ then this implies $f(x) \leq g(x)$, with

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

□

Theorem 3.2. *If $f = f_1 - f_2$ with f_1 convex quadratic and f_2 convex then for each y there is a convex quadratic function g majorizing f at y.*

Proof. We have $f_2(x) \geq f_2(y) + z'(x-y)$ for all $z \in \partial f_2(y)$. Thus we can use $g(x) = f_1(x) - f_2(y) - z'(x-y)$. □

It might be tempting to assume that global majorization by a convex quadratic implies some kind of smoothness. The following, quite convenient, result shows that this is not the case.

Theorem 3.3. *Suppose $f = \min_k f_k$ and let S_i be the set where $f = f_i$. If $y \in S_i$ and g majorizes f_i at y, then g majorizes f at y.*

Proof. First $g(x) \geq f_i(x) \geq \min_k f_k(x) = f(x)$. Because $y \in S_i$ also $g(y) = f_i(y) = f(y)$. □

This implies that if $f = \min_k f_k$ has a quadratic majorizer at each y, if each of the f_k has a quadratic majorizer at each y.

Example 3.1. 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, at thus d is negative for at least one value of x.

Example 3.2. 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 at $y = 0$. If $y \neq 0$ we can use the AM/GM inequality[2] in the form

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

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 $ax + b\,\mathbf{sign}(x) \geq 1$ for all $x \neq 0$. But for $x < \min(0, \frac{1+b}{a})$ we have $ax + b\,\mathbf{sign}(x) < 1$.

3.1. **One dimension.** Let us look at the one dimensional case first. We must find a solution to the infinite set of linear inequalities

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

Obviously the solutions for $a > 0$ and b, if they exist, form a closed convex set in the plane. Moreover, if (a, b) is a solution, then $(\tilde{a}, b)$ with $\tilde{a} > a$ is also a solution.

Let

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

If f is differentiable at y, we know that $b = f'(y)$, and thus we also define

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

In the differentiable case the system is solvable if and only if

$$\underline{a}(y) \triangleq \sup_{x \neq y} \delta(x, y)$$

is finite and positive, in which case the solution consists of all $a \geq \underline{a}(y)$.

In the non-differentiable case we have to look for those b for which

$$\underline{a}(b, y) \triangleq \sup_{x \neq y} \delta(x, y, b)$$

is finite and positive.

[2]That is, the arithmetic mean-geometric mean inequality: the geometric mean of two different non-negative numbers is smaller than their arithmetic mean. The inequality has been introduced in majorization theory by Heiser in the early eighties.

3.1.1. *The absolute value function.* This extends the analysis in Example 3.2. In our previous result we used the AM/GM inequality, as a trick we pulled out of our hat. We happened to have an inequality handy that covered this case. But it is not clear, for instance, how optimal this approach is. And we do not get information about $y = 0$. So let's compute the optimal quadratic majorization instead.

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 $\overline{a}(b)$. If $y < 0$ then $b = -1$ and thus

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

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

$$\overline{a} = \sup_{x \neq y} \frac{|x| - x}{(x-y)^2} = \frac{1}{2}\frac{1}{|y|}.$$

If $y = 0$ then we must look at

$$\overline{a}(b) = \sup_{x \neq 0} \frac{|x| - bx}{x^2} = \sup_{x \neq 0} \frac{\mathbf{sign}(x) - b}{x},$$

which is clearly $+\infty$. For $y \neq 0$ we see that

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

Thus for $y \neq 0$ the best quadratic majorization is given by the AM/GM inequality, while for $y = 0$ no quadratic majorization exists.

It follows from Theorem 2.2, by the way that $-1 \leq b \leq +1$, because the interval $[-1, +1]$ is the subgradient of $|x|$ at zero. This result does not help here, because for none of these values of b can we find a matching a.

What can we do in this case ? Not much. The most commonly used trick is to consider the function $f(x) = \sqrt{x^2 + \epsilon}$ for some small and fixed $\epsilon > 0$. This is a smoothed version of $|x|$. Now apply the AM/GM inequality to obtain

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

3.1.2. *The Huber function.* Majorization for the Huber function, in particular quadratic majorization, has been studied earlier by Heiser [1987]; Verboon and Heiser [1994]. In those papers quadratic majorization functions appear 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 with our technique, which is tedious but straightforward, and 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 functions, one for each $c>0$. The Huber functions are differentiable, with derivative

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

We can find all quadratic majorizers by making a table of $\frac{f(x)-f(y)-f'(y)(x-y)}{(x-y)^2}$.

	$x\leq -c$	$\lvert x\rvert<c$	$x\geq +c$
$y\leq -c$	0	$\frac{1}{2}\frac{(x+c)^2}{(x-y)^2}$	$\frac{2cx}{(x-y)^2}$
$\lvert y\rvert<c$	$\frac{1}{2}(1-\frac{(x+c)^2}{(x-y)^2})$	$\frac{1}{2}$	$\frac{1}{2}(1-\frac{(x-c)^2}{(x-y)^2})$
$y\geq +c$	$-\frac{2cx}{(x-y)^2}$	$\frac{1}{2}\frac{(x-c)^2}{(x-y)^2}$	0

The sup over x in each cell is

	$x\leq -c$	$\lvert x\rvert<c$	$x\geq +c$
$y\leq -c$	0	$\frac{2c^2}{(c-y)^2}$	$\frac{1}{2}\frac{c}{\lvert y\rvert}$
$\lvert y\rvert<c$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
$y\geq +c$	$\frac{1}{2}\frac{c}{\lvert y\rvert}$	$\frac{2c^2}{(c+y)^2}$	0

Finally, taking the sup of the rows shows that

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

3.2. **Double concave functions.** We know

$$f(x) \leq f(y) + f'(y)(x-y) + \sup_{0\leq\lambda\leq 1} f''(\lambda x + (1-\lambda)y)(x-y)^2$$

3.3. **The scalar case.** To start the multivariate case we will look at majorizing f with quadratics of the form

$$g(x) = \omega^2(x-y)'A(x-y) + b'(x-y) + f(y),$$

where A is now a known matrix. Often, but not aways, A is the identity matrix, corresponding with the algorithm

with linear convergence rate.

For majorization we must have

$$\omega^2 = \sup_{x\neq y} \frac{f(x) - f(y) - b'(x-y)}{(x-y)'A(x-y)}$$

3.4. **The general case.**

Theorem 3.4. *A differentiable f can be majorized by a quadratic g at y if and only if there is an M such that*

$$\frac{f(x) - f(y) - (x-y)'f'(y)}{(x-y)'(x-y)} \leq M$$

for all $x \neq y$. The majorizing quadratics are given by

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

with $\|A\|_\infty \geq M$.

Proof. We have $g(x) \geq f(x)$ for all x if and only if

$$(x-y)'A(x-y) \geq f(x) - f(y) - (x-y)'f'(y) =$$
$$= (x-y)'[\int_0^1 (1-t)f''(y+t(x-y))dt](x-y),$$

i.e. if and only if

$$A \geq \int_0^1 (1-t)f''(y+t(x-y))dt$$

for all x. □

APPENDIX A. BOUNDING THE MULTINORMAL

Theorem A.1. *Suppose* $f(x) = \exp(-\frac{1}{2}x'Ax)$. *Then*

$$h(x,y) = f(y) - f(y)y'A(x-y) + \|A\|_\infty \exp(-\frac{3}{2})(x-y)'(x-y)$$

is a quadratic majorization scheme for f.

Proof. We have

$$\begin{aligned} f'(x) &= -f(x)Ax, \\ f''(x) &= -f(x)[A - Axx'A]. \end{aligned}$$

Suppose z is any vector with $z'Az = 1$. Then

$$z'f''(x)z = -f(x)(1-(z'Ax)^2),$$

and

$$\frac{\partial z'f''(x)z}{\partial x} = f(x)[(2(z'Ax)Az + (1-(z'Ax)^2)Ax],$$

which is zero if and only if $2(z'Ax)z + (1-(z'Ax)^2)x = 0$. Thus $x = \lambda z$, where $2\lambda + \lambda(1-\lambda^2) = 0$, which means $\lambda = 0$ or $\lambda = \pm\sqrt{3}$, and consequently

$$-f(0) \leq z'f''(x)z \leq 2f(\sqrt{3}z) = 2\exp(-\frac{3}{2}).$$

Because $z'Az = 1$, we also have $z'z \geq \|A\|_\infty^{-1}$, and thus finally

$$\|f''(x)\|_\infty \leq 2\|A\|_\infty \exp(-\frac{3}{2})$$

. □

APPENDIX A. INTRODUCTION

I am collecting some material about limits and continuity here, mostly for my own reference.

APPENDIX B. EXTENDED REAL FUNCTIONS

A function is proper if it is not everywhere equal to $+\infty$.

$$\liminf_{x \to y}$$

APPENDIX C. BOUNDS

APPENDIX D. LIMITS OF A FUNCTION AT A POINT

APPENDIX E. SEMICONTINUOUS FUNCTIONS

APPENDIX F. DERIVATIVES OF A FUNCTION AT A POINT

REFERENCES

D. Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.

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

H. Cartan. *Differential Calculus*. Hermann, Paris, France, 1971.

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.

W.J. Heiser. Correspondence analysis with least absolute residuals. *Computational Statistica and Data Analysis*, 5:357–356, 1987.

W.J. Heiser. Convergent computing by iterative majorization: theory and applications in multidimensional data analysis. In W.J. Krzanowski, editor, *Recent Advantages in Descriptive Multivariate Analysis*. Oxford: Clarendon Press, 1995.

H. Kiers. Majorization as a tool for optimizing a class of matrix functions. *Psychometrika*, 55:417–428, 1990.

H. Kiers. Maximization of sums of quotients of quadratic forms and some generalizations. *Psychometrika*, 60:221–245, 1995.

K. Lange, D. Hunter, and I. Yang. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, 9:1–59, 2000.

P. Verboon and W.J. Heiser. Resistant lower rank approximation of matrices by iterative majorization. *Computational Statistics and Data Analysis*, 18:457–467, 1994.

UCLA STATISTICS

E-mail address: deleeuw@stat.ucla.edu