MAJORIZATION OF DIFFERENT ORDER

JAN DE LEEUW

1. Introduction

The majorization method [De Leeuw, 1994; Heiser, 1995; Lange et al., 2000] for minimization of real valued loss functions has become very popular in statistics and computer science (under a wide variety of names).

We give a brief introduction. Suppose the problem is to minimize a real valued function $\phi(\bullet)$ over $\Theta \subseteq \mathbb{R}^p$,

We say that a real valued function $\psi(\bullet)$ *majorizes* $\phi(\bullet)$ *over* Θ *in* $\xi \in \Theta$ if

$$\phi(\theta) \leq \psi(\theta) \qquad \forall \theta \in \Theta, \tag{1a}$$

$$\phi(\xi) = \psi(\xi). \tag{1b}$$

In words, $\psi(\bullet)$ must be above $\phi(\bullet)$ in all of Θ, and touches $\phi(\bullet)$ in ξ. We say that $\psi(\bullet)$ *strictly majorizes* $\phi(\bullet)$ *over* Θ *in* $\xi \in \Theta$ if we have (1a) and

$$\phi(\theta) = \psi(\theta) \text{ if and inly if } \theta = \xi. \tag{2}$$

In words, $\psi(\bullet)$ must be above $\phi(\bullet)$ in all of Θ, and touches $\phi(\bullet)$ *only* in ξ.

Now suppose that we have a function $\psi(\bullet, \bullet)$ on $\Theta \otimes \Theta$ such that

$$\phi(\theta) \leq \psi(\theta.\xi) \qquad \forall \theta, \xi \in \Theta, \tag{3a}$$

$$\phi(\xi) = \psi(\xi, \xi) \qquad \forall \xi \in \Theta. \tag{3b}$$

Thus for all $\xi \in \Theta$ the function $\psi(\bullet, \xi)$ majorizes $\phi(\bullet)$ over Θ in ξ. In this case we simply say that $\psi(\bullet, \bullet)$ *majorizes* $\psi(\bullet)$ *over* Θ. Also, $\psi(\bullet, \bullet)$ *strictly*

Date: January 17, 2005.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

majorizes $\psi(\bullet)$ *over* Θ if for all $\xi \in \Theta$ the function $\psi(\bullet, \xi)$ strictly majorizes $\phi(\bullet)$ over Θ in ξ.

Each majorization function can be used to define an algorithm. In each step of such a *majorization algorithm* we find the update $\theta^{(k+1)}$ by minimizing $\psi(\bullet, \theta^{(k)})$ over Θ, i.e. we choose

$$\theta^{(k+1)} \in \underset{\theta\in\Theta}{\mathbf{Argmin}}\, \psi(\theta, \theta^{(k)}).$$

The minimum of $\psi(\bullet, \theta^{(k)})$ over Θ may not be unique, and consequently $\mathbf{Argmin}(\bullet)$ is a set-valued map. If the minimum is unique, we use the single-valued version and set

$$\theta^{(k+1)} = \underset{\theta\in\Theta}{\mathbf{argmin}}\, \psi(\theta, \theta^{(k)}).$$

The algorithm includes a simple stopping rule. If

$$\theta^{(k)} \in \underset{\theta\in\Theta}{\mathbf{Argmin}}\, \psi(\theta, \theta^{(k)})$$

then we stop. If we never stop, then we obviously generate an infinite sequence.

Now suppose the algorithm generates an infinite sequence. For each step of the algorithm the *sandwich inequality*

$$\phi(\theta^{(k+1)}) \leq \psi(\theta^{(k+1)}, \theta^{(k)}) < \psi(\theta^{(k)}, \theta^{(k)}) = \phi(\theta^{(k)}) \tag{4}$$

shows that an iteration decreases the value of the loss function. The strict inequality $\psi(\theta^{(k+1)}, \theta^{(k)}) < \psi(\theta^{(k)}, \theta^{(k)})$ follows from the fact that we do not stop, which implies that $\theta^{(k)}$ is not a minimizer of $\psi(\bullet, \theta^{(k)})$.

This is used to prove convergence of the algorithm, using general results such as those of Zangwill [1969].

Each algorithm we discuss defines an algorithmic map $\alpha(\bullet)$ from Θ into Θ.

From the computational point of view the trick is to find a majorization which is relatively simple to minimize.

2. Example

We use an simple example inspired by the likelihood function for logistic regression. It is only one-dimensional, so it is not reprentative for applications of majorization methods. In general. majorization shines in high-dimensional problems. Nevertheless, the example can be used to illustrate quite a few general techniques and principles.

Suppose the problem is to minimize

$$\phi(\theta) \stackrel{\Delta}{=} \eta(2+\theta) + \eta(1-\theta),$$

where

$$\eta(\theta) \stackrel{\Delta}{=} \log(1+\exp(\theta)).$$

Define

$$\pi(\theta) \stackrel{\Delta}{=} \mathcal{D}\eta(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{1}{1+\exp(-\theta)}.$$

Clearly $\pi(\bullet)$ is increasing, which means that $\eta(\bullet)$ is strictly convex. Thus $\phi(\bullet)$ is strictly convex as well.

By setting the derivative

$$\mathcal{D}\phi(\theta) = \pi(2+\theta) - \pi(1-\theta).$$

equal to zero, we see that the minimum of $\phi(\bullet)$ is attained when $2+\theta = 1-\theta$, or when $\theta = -\frac{1}{2}$. The minimum is equal to $2\eta(1.5) \approx 3.402827$.

3. Quadratic Majorization

From

$$\mathcal{D}^2\sigma(\theta) = \mathcal{D}\pi(\theta) = \pi(\theta)(1-\pi(\theta)).$$

we see that $0 \leq \mathcal{D}^2\sigma(\theta) \leq \frac{1}{4}$. Since

$$\mathcal{D}^2\phi(\theta) = \mathcal{D}^2\sigma(2+\theta) + \mathcal{D}^2\sigma(1-\theta)$$

it follows that $0 \leq \mathcal{D}^2\phi(\theta) \leq \frac{1}{2}$.

By the mean value theorem

$$\phi(\theta) \leq \phi(\xi) + \mathcal{D}\phi(\xi)(\theta-\xi) + \frac{1}{2}\max_{0\leq\lambda\leq 1}\mathcal{D}^2\phi(\lambda\theta+(1-\lambda)\xi)(\theta-\xi)^2,$$

and thus

$$\phi(\theta) \leq \phi(\xi) + \mathcal{D}\phi(\xi)(\theta - \xi) + \frac{1}{4}(\theta - \xi)^2.$$

This implies that

$$\psi(\theta, \xi) = \phi(\xi) + (\pi(2 + \xi) - \pi(1 - \xi))(\theta - \xi) + \frac{1}{4}(\theta - \xi)^2$$

defines a majorization of $\phi(\bullet)$. Minimization of the majorization gives the algorithm

$$\theta^{(k+1)} = \theta^{(k)} - 2(\pi(2 + \theta^{(k)}) - \pi(1 - \theta^{(k)})).$$

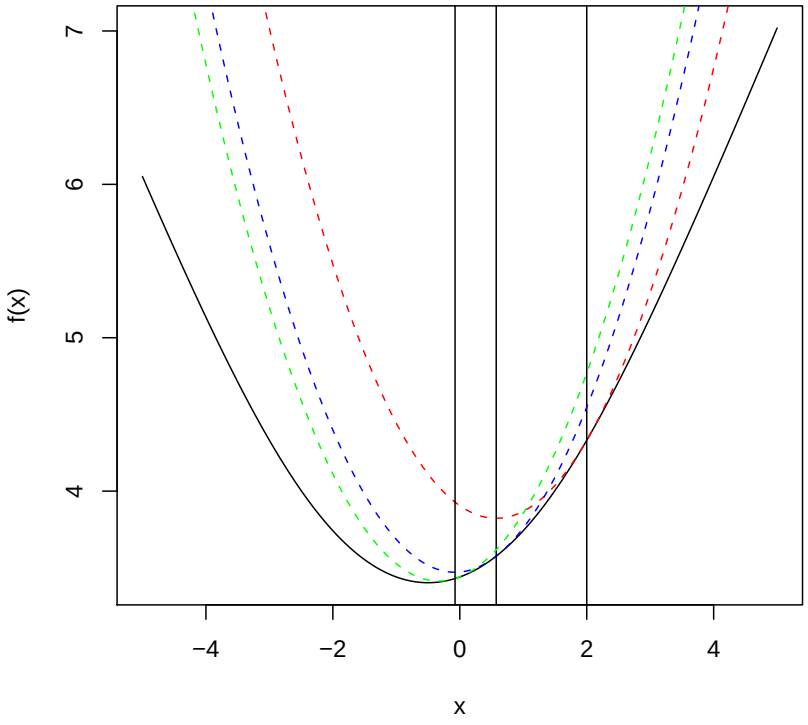

FIGURE 1. Quadratic Majorization

In Figure 1 the quadratic majorization at $\theta = 2$ is in red. It is minimized at $\theta \approx 0.5738553$, where we draw the second majorization, in blue. This majorization is minimized at $\theta \approx -0.0745591$, where the third majorization

function, in green, is drawn. And so on. Starting from $\theta = 2$ it takes 25 iterations to converge to seven decimals precision. We give the successive values of $\theta^{(k)}$ in the "quadratic" column of Table 1.

This algorithm converges linearly with rate

$$1 - \frac{\mathcal{D}^2\phi(-.5)}{\mathcal{D}_{11}\psi(-.5.-.5)} = 1 - \frac{2\sigma(1.5)}{0.5} \approx 0.4034142.$$

The upper bound $\mathcal{D}^2\phi(\theta) \leq \frac{1}{2}$ is actually not sharp. Some numerical computation gives $\mathcal{D}^2\phi(\theta) \leq 0.305142$. The majorization algorithm corresponding to this sharper bound is

$$\theta^{(k+1)} = \theta^{(k)} - 3.277163(\pi(2 + \theta^{(k)}) - \pi(1 - \theta^{(k)}))$$

which has a convergence rate of 0.0224456, about four times as fast as the one using the bound $\frac{1}{2}$. The actual iterations are given in the "sharp" column of Table 1.

Using a sharper upper bound pays off, but in most cases computing the sharpest possible bound requires more computation than the problem we are actually trying to solve. In such cases it is better to use the weaker bounds, that do not require actual iterative computation.

For comparison purposes we also looked at Newton's method for this example. The iterations are

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\pi(2 + \theta^{(k)}) - \pi(1 - \theta^{(k)})}{\sigma''(2 + \theta) + \sigma''(1 - \theta)}$$

and in this case we have global quadratic convergence from any starting point.

4. Relaxation

For the sandwich inequality to apply it is not necessary to actually minimize the majorization function. It suffices to decrease it. In fact, let $F(\bullet)$ be a mapping of Θ into Θ such that

$$\psi(F(\theta^{(k)}), \theta^{(k)}) \leq \psi(\theta^{(k)}, \theta^{(k)})).$$

Then the sandwich inequality still applies, and under strict majorization we still have $\phi(\theta^{(k+1)}) < \phi(\theta^{(k)})$ if we set $\theta^{(k+1)} = F(\theta^{(k)})$.

In our quadratic majorization example we can consider the algorithm

$$\theta^{(k+1)} = \theta^{(k)} - K(\pi(2 + \theta^{(k)}) - \pi(1 - \theta^{(k)})).$$

Now

$$\psi(\theta^{(k+1)}, \theta^{(k)}) = \phi(\theta^{(k)}) + (\frac{1}{4}K^2 - K)(\pi(2 + \theta^{(k)}) - \pi(1 - \theta^{(k)}))^2,$$

and thus $\psi(\theta^{(k+1)}, \theta^{(k)}) \leq \phi(\theta^{(k)})$ for $0 \leq K \leq 4$.

The case $K = 0$ is uninteresting, because any point is a fixed point and nothing changes. The case $K = 4$ is of some interest, however. We move to a point equally far from the minimum as the current solution, or majorization point, but on the other side of the minimum. This is sometimes known as *over-relaxation*. At this over-relaxed point we have $\psi(\theta^{(k+1)}, \theta^{(k)}) = \phi(\theta^{(k)})$, but in the case of strict relaxation this still gives $\phi(\theta^{(k+1)}) < \phi(\theta^{(k)})$.

The linear convergence rate of the algorithm with step-size K is

$$|1 - K\phi''(-.5))| \approx |1 - 0.2982929K|$$

Thus for $K = 4$ we obtain a rate of 0.1931716 and convergence which is about twice as fast as before (see the "overrel" column in Table 1). The first two iterations of the over-relaxed algorithm are in Figure 2.

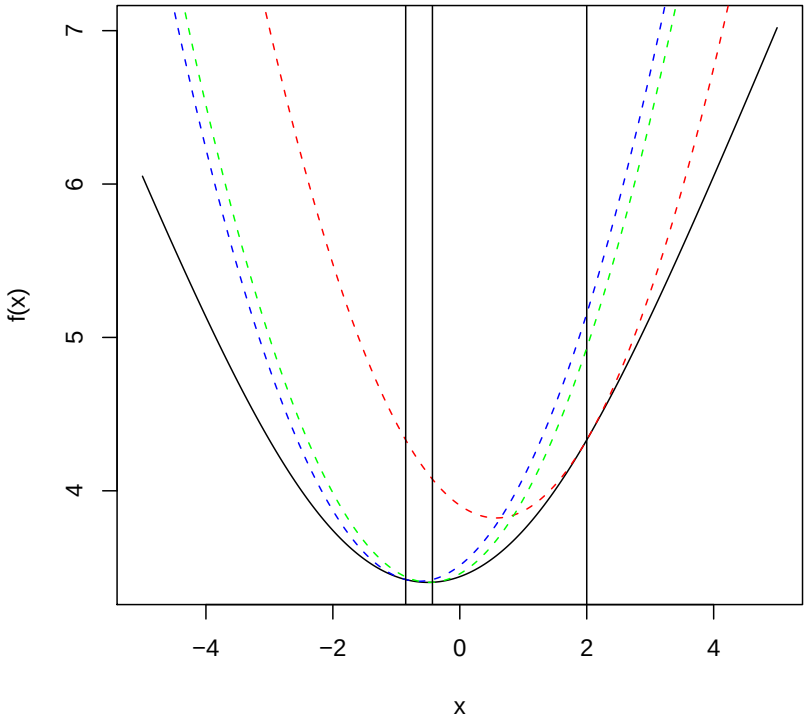

FIGURE 2. Over-relaxation

For the very special choice

$$K = \frac{1}{2 + \sigma(1.5)} \approx 3.352410$$

we have superlinear convergence (but of course we can only use this step-size if we already know the solution). See the "optrel" column in Table 1.

5. CUBIC MAJORIZATION

So far our majorization methods have linear convergence (unless we are very lucky). It is quite straightforward, however, to construct majorization methods with superlinear convergence.

Define

$$\mu(\theta) = \sigma'(\theta) = \pi''(\theta) = \pi(\theta)(1-\pi(\theta))(1-2\pi(\theta)).$$

Some simple computation gives

$$-\frac{1}{18}\sqrt{3} \leq \mu(\theta) \leq +\frac{1}{18}\sqrt{3}.$$

This means that

$$\frac{1}{9}\sqrt{3} \leq \phi'''(\theta) = \mu(2+\theta) - \mu(1-\theta) \leq \frac{1}{9}\sqrt{3},$$

and thus

$$\begin{aligned}\psi(\theta,\xi) = \phi(\xi) + (\pi(2+\xi) - \pi(1-\xi))(\theta-\xi) + \\ + \frac{1}{2}(\sigma(1-\xi)+\sigma(2+\xi))(\theta-\xi)^2 + \frac{\sqrt{3}}{54}|\theta-\xi|^3\end{aligned}$$

is a majorization of $\phi(\bullet)$. The majorization function seems somewhat non-standard, because it involves the absolute value of the cubic term. Nevertheless it is two times continuously differentiable. In fact, it is also strictly convex, because the second derivative is

$$\mathcal{D}_{11}\psi(\theta,\xi) = \begin{cases}(\sigma(1-\xi)+\sigma(2+\xi)) + \frac{\sqrt{3}}{9}(\theta-\xi) & \text{for } \theta \geq \xi,\\ (\sigma(1-\xi)+\sigma(2+\xi)) - \frac{\sqrt{3}}{9}(\theta-\xi) & \text{for } \theta \leq \xi.\end{cases}$$

which is clearly positive.

To find the minimum we set the first derivative equal to zero. The first derivative at ξ is equal to $\pi(2+\xi) - \pi(1-\xi)$. If this is positive then the minimum is attained at a value smaller than ξ. In this case the quadratic

$$\pi(2+\xi) - \pi(1-\xi) + (\sigma(1-\xi)+\sigma(2+\xi))\zeta - \frac{\sqrt{3}}{18}\zeta^2 = 0$$

has two real roots $\zeta_1 < 0 < \zeta_2$ and the minimum we look for is attained at $\xi + \zeta_1$. If the derivative at zero $\pi(2+\xi) - \pi(1-\xi)$ is negative, then

$$(\pi(2+\xi) - \pi(1-\xi)) + (\sigma(1-\xi)+\sigma(2+\xi))\zeta + \frac{\sqrt{3}}{18}\zeta^2 = 0$$

again has two real roots $\zeta_1 < 0 < \zeta_2$ and the minimum of the majorization function is attained at $\xi + \zeta_2$.

For the derivative of the algorithmic map $\xi + \zeta(\xi)$ we find

$$1 - \frac{\phi''(\xi) + \zeta(\xi)\phi'''(\xi)}{\phi''(\xi) + \zeta(\xi)\frac{1}{18}\sqrt{3}}.$$

At a fixed point $\zeta(\xi) = 0$ and thus the derivative is zero, which implies superlinear convergence.

6. Quartic Majorization

Define

$$\lambda(\theta) = \pi'''(\theta) = \pi(\theta)(1 - \pi(\theta))(1 - 6\pi(\theta) + 6\pi^2(\theta)).$$

We find that

$$-\frac{1}{24} \leq \lambda(\theta) \leq \frac{1}{8}.$$

Thus

$$\phi''''(\theta) = \lambda(2 + \theta) + \lambda(1 - \theta) \leq \frac{1}{4}.$$

The majorization function is

$$\begin{aligned}\psi(\theta, \xi) = \phi(\xi) + (\pi(2 + \xi) - \pi(1 - \xi))(\theta - \xi) + \\ + \frac{1}{2}(\sigma(1 - \xi) + \sigma(2 + \xi))(\theta - \xi)^2 + \frac{1}{6}(\mu(2 + \xi) - \mu(1 - \xi))(\theta - \xi)^3 + \\ + \frac{1}{96}(\theta - \xi)^4.\end{aligned}$$

The second partials are

$$\begin{aligned}\mathcal{D}_{11}\psi(\theta, \xi) = (\sigma(1 - \xi) + \sigma(2 + \xi)) + \\ (\mu(2 + \xi) - \mu(1 - \xi))(\theta - \xi) + \frac{1}{8}(\theta - \xi)^2.\end{aligned}$$

This quadratic has no real roots (conjecture so far), and since it is positive for $\theta = \xi$ the quartic majorization function is strictly convex.

Setting the derivative equal to zero means solving a cubic with only one real root. This root gives the minimum of the majorization function.

7. Lipschitz Majorization

We started with quadratic majorization, and consequently we can wonder if linear majorization is also possible in our example. In fact it is, but it does not produce an useful algorithm.

We use $0 \leq \phi'(\theta) \leq 1$. By the mean value theorem

$$\phi(\theta) = \phi(\xi) + \phi'(\xi)(x - y)$$

for some ξ between x and y. This implies that

$$\psi(x, y) = \phi(\xi) + |x - y|$$

is a majorization function, leading to the algorithm

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

We have convergence in one step, because we simply stay in place, no matter where we start.

The majorization is illustrated in the figure below, where we majorize at y equal to $-2, 0$ and $+2$. Clearly majorization is strict, and the majorization functions have a unique minimum.

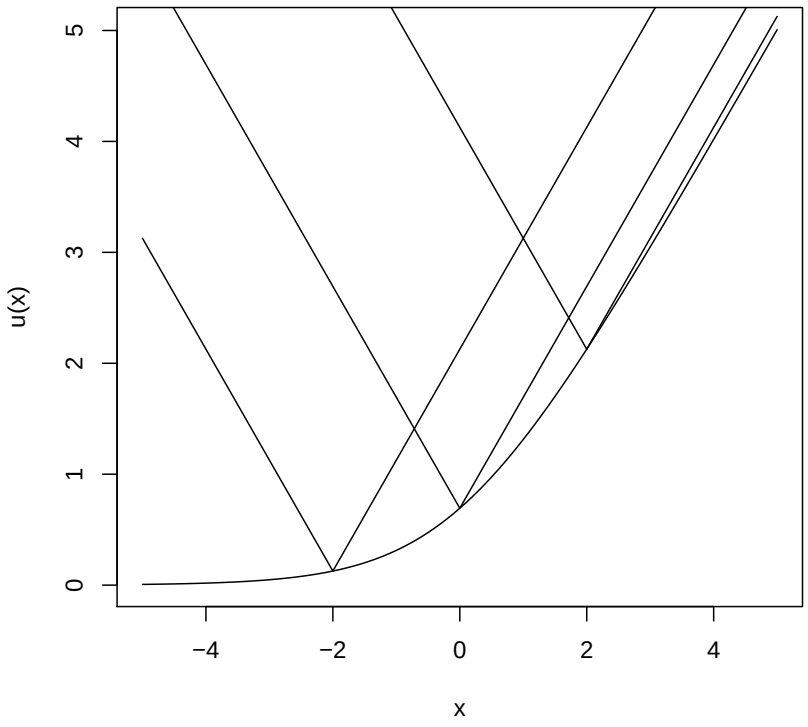

FIGURE 3. Lipschitz Majorization

We can generalize this example by observing that for all multivariate differentiable functions we have

$$f(\theta) = f(\xi) + (x - y)'g(\xi) \leq f(\xi) + \|g(\xi)\|\|x - y\|$$

where the two norms are dual to each other. Thus if $\|g(\xi)\| \leq K$ we have the majorization function

$$f(\theta) \leq f(\xi) + K\|x - y\|$$

If fact this majorization applies to all Lipschitz functions with constant K, even if they are not differentiable.

Appendix A. Example

	quadratic	overrel	optrel	newton	cubic	quartic
1	2.0	2.0	2.0	2.0	2.0	2.0
2	0.5738553	-0.8522895	-.3905107	-1.327843	-0.2206191	0.1497816
3	-0.0745591	-0.4310767	-0.5000229	-0.5101132	-0.4878135	-0.4827805
4	-0.3291210	-0.5133209	-0.5	-0.5	-0.4999761	-0.4999997
5	-0.4311166	-0.4974267			-0.5	-0.5

Table 1. Iterations

Appendix B. Code

```
f<-function(x) log(1+exp(1-x))+log(1+exp(2+x))
pif<-function(x) 1/(1+exp(-x))
nuf<-function(x) pif(x)*(1-pif(x))
vuf<-function(x) pif(x)*(1-pif(x))*(1-2*pif(x))

gqua<-function(x,y) f(y)+
        ((pif(2+x)-pif(1-x))*(x-y))+
        0.25*(x-y)^2
gcub<-function(x,y) f(y)+
        ((pif(2+y)-pif(1-y))*(x-y))+
        (0.5*(nuf(1-y)+nuf(2+y))*((x-y)^2))+
        ((sqrt(3)/108)*abs((x-y)^3))
gqur<-function(x,y) f(y)+
        ((pif(2+y)-pif(1-y))*(x-y))+
        (0.5*(nuf(1-y)+nuf(2+y))*((x-y)^2))+
        ((1/6)*(vuf(2+y)-vuf(1-y))*((x-y)^3))+
        (((x-y)^4)/192)

upd<-function(x) x-2*(pif(2+x)-pif(1-x))
vpd<-function(x,K) x-K*(pif(2+x)-pif(1-x))
new<-function(x) x-(pif(2+x)-pif(1-x))/(nuf(1-x)+nuf
    (2+x))
```

```
relax<-function(x) abs(1-2*x*nuf(1.5))

scub<-function(y) {
a<-pif(2+y)-pif(1-y); b<-nuf(1-y)+nuf(2+y); c<-sqrt(3)
   /36
if (a > 0) min(y+solve.polynomial(c(a,b,-c)))
else max(y+solve.polynomial(c(a,b,c)))
}

squr<-function(y) {
a<-pif(2+y)-pif(1-y); b<-nuf(1-y)+nuf(2+y)
c<-(vuf(2+y)-vuf(1-y))/2; d<-1/48
y+solve.polynomial(c(a,b,c,d))
}

quad.plot<-function(a) {
b<-upd(a); c<-upd(b)
pdf("quad.pdf")
plot(x,f(x),type="l")
points(x,gqua(x,a),type="l",col="red",lty=2)
abline(v=a)
points(x,gqua(x,b),type="l",col="blue",lty=2)
abline(v=b)
points(x,gqua(x,c),type="l",col="green",lty=2)
abline(v=c)
dev.off()
}

relax.plot<-function(a) {
b<-vpd(a); c<-vpd(b)
pdf("relax.pdf")
plot(x,f(x),type="l")
points(x,gqua(x,a),type="l",col="red",lty=2)
```

```
abline(v=a)
points(x,gqua(x,b),type="l",col="blue",lty=2)
abline(v=b)
points(x,gqua(x,c),type="l",col="green",lty=2)
abline(v=c)
dev.off()
}
```

References

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. Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In W.J. Krzanowski, editor, *Recent Advanmtages in Descriptive Multivariate Analysis*. Oxford: Clarendon Press, 1995.

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

W. I. Zangwill. *Nonlinear Programming: a Unified Approach*. Prentice-Hall, Englewood-Cliffs, N.J., 1969.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

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

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