SHARP LOCAL QUADRATIC MAJORIZATION

JAN DE LEEUW

ABSTRACT. The majorization principle provides us with a way to find global quadratic approximations to functions, and globally convergent algorithms. The main disadvantage iof majorization algorithms is their slow local convergence. Moreover majorizers may not exist. In this note we develop some theory which can be used to improve the local properties of majorization methods, while retaining global and monotone convergence. Some example, mostly of functions on the real line, are discussed in detail.

1. INTRODUCTION

In this paper we study iterative minimization algorithms with a common structure. Say we are minimizing a real valued function f on a set $\mathscr{S}$. In each iteration we have a current best approximation $y \in \mathscr{S}$ to the minimizer. We now use an auxiliary function g that approximates f at y in some sense, and we find the update y^+ of y by minimizing g over x. If y itself is a minimizer of g, we stop. If we do not stop we make a new auxiliary function that approximates f in y^+, and so on.

To simplify matters we shall only study one-dimensional unconstrained minimization in this paper, i.e. we want to find the minimum of f on the real line. To further simplify, we assume in addition that f is everywhere differentiable.

Functions of a single variable are written simply as f, or as $f(\bullet)$. A function of two real variables is $g(\bullet,\bullet)$. For a function of two variables we write

Date: Friday 24th June, 2011.

2000 *Mathematics Subject Classification.* 29M20.

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

$g(\bullet, y)$ for the "horizontal" cross-section where the second variable is fixed at y, and $g(x, \bullet)$ for the corresponding "vertical" cross section. Thus $g(x,y)$ is a single real number, because both x and y are fixed.

We assume our approximations have the following three properties.

$$g(\bullet, y) \text{ is a strictly convex quadratic for all } y, \tag{1}$$

$$g(y,y) = f(y) \text{ for all } y, \tag{2}$$

$$g'(y,y) = f'(y) \text{ for all } y. \tag{3}$$

Thus $g(\bullet, y)$ is a quadratic that coincides with f in y and has the same tangent as f in y. It follows that

$$g(x,y) = f(y) + f'(y)(x-y) + \frac{1}{2}a(x-y)^2 \tag{4}$$

for some $a > 0$. It also follows that, if $a \neq 0$,

$$y^+ = y - \frac{f'(y)}{a}. \tag{5}$$

1.1. **Newton.** The most familiar example of an algorithm of this form is *Newton's method*, which requires in addition to (1), (2), and (3) that

$$g''(y,y) = f''(y). \tag{6}$$

This means $a = f''(y)$ and thus $g(\bullet, y)$ is uniquely defined. Of course Newton's method, in its unmodified form, can only be applied for twice-differentiable functions at locations where $f''(y) > 0$. If we want it to work more generally, we need to modify (6). Another major problem with (6) is that the approximation is strictly local and may be very poor if we move away from y. That is another reason why we need to add various safeguards in actual implementations of Newton's method.

1.2. **Quadratic Majorization.** A second example of our algorithm scheme is *quadratic majorization*, where we require, in addition to (1), (2) and (3) that

$$g(x,y) \geq f(x) \text{ for all } x. \tag{7}$$

If f is twice-differentiable case (7) implies $a = g''(y|y) \geq f''(y)$ [De Leeuw and Lange, 2009]. But (7) is actually much stronger, because it is a global condition, in the sense that it restricts the approximation g for all x. Using a global approximation condition on the one hand is good, because it regularizes the approximations, and forces convergence. On the other hand it is bad, because it constrains the approximation in areas which are not really relevant for the computations, which results in a slow rate of convergence.

If we define

$$\delta(x,y,a) \stackrel{\Delta}{=} f(x) - f(y) - f'(y)(x-y) - \frac{1}{2}a(x-y)^2 \tag{8}$$

then (7) requires $\delta(x,y,a) \leq 0$ for all x. Define

$$\overline{\delta}(y,a) \stackrel{\Delta}{=} \sup_x \delta(x,y,a),$$

and suppose $\mathscr{A}(y)$ is the set of all $a > 0$ such that $\overline{\delta}(y,a) \leq 0$. If $a \in \mathscr{A}(y)$ and $a' \geq a$, then $a' \in \mathscr{A}(y)$. Thus if $\mathscr{A}(y)$ is non-empty, there is an $\overline{a}(y)$ such that either $\mathscr{A}(y) = [\overline{a}(y), +\infty)$ or $\mathscr{A}(y) = (\overline{a}(y), +\infty)$. The set $\mathscr{A}(y)$may be empty, because there may not exist any quadratic majorizations. If f is cubic, for instance, quadratic majorizations do not exist for any y [De Leeuw and Lange, 2009].

The number $\overline{a}(y)$, which is the greatest lower bound of $\mathscr{A}(y)$, defines the *sharpest quadratic majorization* [Van Ruitenburg, 2005; De Leeuw and Lange, 2009]. Observe that under our conditions the sharpest quadratic majorization could be linear.

2. IMPROVEMENTS

Newton's method has fast local convergence, while the majorization method has slow global convergence. There has been quite a lot of work modifying the majorization method to give it faster local convergence, while maintaining the global convergence. Updates using cubic and quartic approximations are discussed in De Leeuw [2006]. Updates that do not minimize $g(\bullet, y)$ but merely decrease it in other systematic ways have been discussed in many places. In this paper, however, we continue to impose (1), (2) and (3), but we relax (7).

The proof of global convergence of the majorization method (i.e. convergence from any initial point) uses the *sandwich inequality* [De Leeuw, 1994]. We introduce the additional convention that we stop the algorithm if $g(\bullet, y)$ is minimized at $x = y$, i.e. if $y^+ = y$. Thus if we do not stop

$$f(y^+) \leq g(y^+, y) < g(y, y) = f(y), \tag{9}$$

and we decrease the value of f.

The sandwich inequality (9) continues to apply under weaker conditions than (7). Three possible alternatives are

$$f(x) \leq g(x,y) \text{ for all } x \text{ such that } g(x,y) \leq g(y,y), \tag{10a}$$

$$f(x) \leq f(y) \text{ for all } x \text{ such that } g(x,y) \leq g(y,y), \tag{10b}$$

$$f(y^+) \leq f(y). \tag{10c}$$

Define $\mathscr{L}(y)$ to be the *level set* $\{x \mid g(x,y) \leq g(y,y)\}$. Then (10a) says that $g(\bullet, y)$ majorizes f on the level set $\mathscr{L}(y)$, while (10b) says that f attains its maximum on the level set $\mathscr{L}(y)$ at y. Condition (10c) merely says that the update y^+ that minimizes the quadratic majorizer $g(\bullet, y)$ is at least as good as y.

Under all three conditions it follows that $f(y^+) \leq f(y)$, i.e. we have monotone convergence. Each of the conditions (10) defines a set $\mathscr{A}(y)$ of a such that g defined by (4) satisfies the condition. Thus each of the conditions can be used to define a notion of sharp quadratic majorization, which chooses $\overline{a}$ to be the glb of $\mathscr{A}$.

For ease of reference, the three conditions are referred to, respectively, as *L-majorization, M-majorization,* and *U-majorization*, where the L stand for level, the M stand for maximum, and the U stands for update. In this paper we shall limit ourselves to studying L-majorization.

3. CUBICS

3.1. **Majorization.** As we mentioned before, there is no quadratic majorization for (non-trivial) cubics. For a cubic we have

$$f(x) = f(y) + f'(y)(x-y) + \frac{1}{2}f''(y)(x-y)^2 + \frac{1}{6}f'''(y)(x-y)^3,$$

with $f'''(y) \neq 0$. Thus we must have

$$f(x) - g(x,y) = \frac{1}{2}(f''(y)-a)(x-y)^2 + \frac{1}{6}f'''(y)(x-y)^3 \leq 0, \tag{11}$$

for all x, which is impossible. We can always make $f(x) - g(x,y)$ arbitrarily large by letting either $x \rightarrow +\infty$ or $x \rightarrow -\infty$.

3.2. **L-majorization.** If $g(x,y)$ has the form (4), then the level set $\mathscr{L}$ is a closed interval with endpoints y and $y - 2f'(y)/a$. For each x in this interval we must have (11), or

$$(x-y)^2[\frac{1}{2}(f''(y)-a) + \frac{1}{6}f'''(y)(x-y)] \leq 0.$$

Since the linear term in square brackets takes its extreme values at the endpoints of the interval, we must have both $a \geq f''(y)$ and

$$\frac{1}{2}(f''(y)-a) - \frac{1}{3}f'''(y)f'(y)/a \leq 0,$$

which we can rewrite as

$$a^2 - af''(y) + \frac{2}{3}f''(y)f'(y) \geq 0. \tag{12}$$

If the quadratic equation corresponding to (12) has no real roots or one real root, then the inequality (12) is satisfied for all a, and thus $\overline{a}(y) = \max(f''(y), 0)$. If $f''(y) > 0$ the sharp quadratic local majorization step is a Newton step.

If the equation has two real roots, they are written as $p(y) \leq q(y)$. We have $p(y) + q(y) = f''(y)$. Thus if $p(y)$ and $q(y)$ are non-negative, then $0 \leq p \leq q \leq f''(y)$, and consequently $\overline{a}(y) = f''(y) \geq 0$. If $p \leq 0$ and $q \geq 0$ then $q = f''(y) - p \geq f''(y)$ and thus $\overline{a}(y) = q$. If both $p \leq 0$ and $q \leq 0$

then $f''(y) \leq p \leq q \leq 0$, and thus $\overline{a}(y) = 0$ and the best local majorization is linear.

Near a local minimum, where $f''(y) \geq 0$ and $f'(y)$ is close to zero, the two roots of the quadratic are approximately $q = f''(y) - \frac{2}{3}f'(y)$ and $p = \frac{2}{3}f'(y)$.

3.3. **Example.** Our numerical example is $f(x) = x^3 - 2x + 1$. The function and its first three derivatives at various points y are given in Table 1. We also give the two roots of the quadratic (**??**), and the bounds $\alpha(y)$.

TABLE 1. Cubic Function

y	-2.000	-1.000	-0.500	0.000	0.500	1.000	2.000
f	-3.000	2.000	1.875	1.000	0.125	0.000	5.000
f'	10.000	1.000	-1.250	-2.000	-1.250	1.000	10.000
f''	-12.000	-6.000	-3.000	0.000	3.000	6.000	12.000
f'''	6.000	6.000	6.000	6.000	6.000	6.000	6.000
$p(y)$	–	-5.236	-4.193	-2.828	-1.193	0.764	–
$q(y)$	–	-0.764	1.193	2.828	4.193	5.236	–
$\overline{a}(y)$	0.000	0.000	1.193	2.828	4.193	6.000	12.000

We show two actual "majorizations" satisfying (10a) in Figure 1. The cubic is in red, the quadratic majorizer in blue, and the interval $\mathscr{I}(y,a)$ in green. Thus, throughout the green interval, the red function must be below the blue function.

4. QUARTICS

For a quartic $f^{iv}(y)$ is independent of y. For non-trivial quartics $f^{iv}(y) \neq 0$.

4.1. **Majorization.** We know that cubics can never be globally majorized by quadratics. We show now that only some quartics can be globally majorized.

$$f(x) - g(x,y) = \tag{13}$$
$$= (x-y)^2\{\frac{1}{2}(f''(y) - a) + \frac{1}{6}f'''(y)(x-y) + \frac{1}{24}f^{iv}(y)(x-y)^2\}.$$

Thus for $x \neq y$ the sign of $f(x) - g(x,y)$ is the sign of the quadratic $h(\bullet, y)$ defined by

$$h(x,y) \stackrel{\Delta}{=} \frac{1}{2}(f''(y) - a) + \frac{1}{6} f'''(y)(x-y) + \frac{1}{24} f^{iv}(y)(x-y)^2. \tag{14}$$

For majorization we require that $h(\bullet, y)$ is less than or equal to zero everywhere. This is only possible if $f^{iv}(y) < 0$. We then have majorization if and only if the discriminant of $h(\bullet, y)$ is non-positive, i.e. if and only if

$$\frac{1}{36}(f'''(y))^2 - \frac{1}{12}(f''(y) - a) f^{iv}(y) \leq 0,$$

or

$$a \geq f''(y) - 3\frac{(f'''(y))^2}{f^{iv}(y)}. \tag{15}$$

Thus sharp global quadratic majorizers exist if and only if majorizers exist if and only if $f^{iv}(y) < 0$. Define $\overline{a}_G(y)$ to be the right hand-side of (15) If we use sharp global quadratic majorization, then the rate of convergence of the majorization method to a local minimum is

$$\lambda(y) = \frac{3(f'''(y))^2}{3(f'''(y))^2 - f''(y) f^{iv}(y)}.$$

4.2. **L-Majorization.** We assume first that $f^{iv}(y) > 0$. Because $h(\bullet, y)$ is a convex quadratic $f(x) - g(x,y) \leq 0$ for all x in an interval if and only if $h(x,y) \leq 0$ at both endpoints of the interval. This means that we must have $a \geq f''(y)$ and

$$a^3 - a^2 \frac{1}{2} f''(y) + a \frac{1}{6} f'''(y) f'(y) - \frac{1}{24} f^{iv}(y)(y) f'(y)^2 \geq 0.$$

If the corresponding cubic has only one real root $p(y)$, then $\overline{a}(y) = \max(0, f''(y), p(y))$. If there are three real roots $p_1(y) \leq p_2(y) \leq p_3(y)$ then define

$$\tilde{a}(y) = \begin{cases} p_1(y) & \text{if } -\infty \leq f''(y) \leq p_1(y), \\ f''(y) & \text{if } p_1(y) \leq f''(y) \leq p_2(y), \\ p_3(y) & \text{if } p_2(y) \leq f''(y) \leq p_3(y), \\ f''(y) & \text{if } p_3(y) \leq f''(y) \leq +\infty. \end{cases}$$

Then the sharp local bound is $\overline{a}_L^+(y) = \max(0, \tilde{a}(y))$.

If $f^{iv}(y) < 0$ matters are more complicated. Observe that in this case f is unbounded below. We define sharp local bound $\overline{a}_L^-(y)$ for this case. Also observe there is a sharp global quadratic majorization. Clearly the $\overline{a}_G(y) \leq \overline{a}_L^-(y)$. Since it is still necessary that $h(x,y) \leq 0$ at the endpoints of the interval we also have $\overline{a}_L^-(y) \leq \overline{a}_L^+(y)$.

Now $h(\bullet, y)$ is a concave quadratic. If it is maximized at a location outside the interval, then $\overline{a}_L^-(y) = \overline{a}_L^+(y)$.

5. Logit

6. Probit

TABLE 2. Iterations

Iteration	Newton	A	B
0	.5	.5	.5
1	.916667	.75	.7981456
2	.8219697	.8125	.8164283
3	.8165148	.8164804	.8164966
4	.8164966	.8164966	

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.

J. De Leeuw. Quadratic and Cubic Majorization. *Medium for Econometric Applications*, 14:44–49, 2006.

J. De Leeuw and K. Lange. Sharp Quadratic Majorization in One Dimension. *Computational Statistics and Data Analysis*, 53:2471–2484, 2009.

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

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

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

FIGURE 1. First Condition, Majorizations