SOME MAJORIZATION TECHNIQUES

JAN DE LEEUW

Abstract. Majorization algorithms generalize the EM algorithm. In this paper we discuss several distinct, although related, techniques to construct majorization function, and we show the algorithms they imply.

1. Introduction

Majorization algorithms [De Leeuw, 1994; Heiser, 1995; Lange et al., 2000] are used with increasing frequency in statistical computation. They generalize the EM algorithm to a much broader class of problems and they can usually be tailored to handle very high-dimensional problems.

The general idea is simple. If we want to minimize f over $X \subseteq \mathbb{R}^n$, we construct a *majorization function* g on $X \times X$ such that

$$\begin{aligned} f(x) &\leq g(x, y) \qquad \forall x, y \in X, \\ f(x) &= g(x) \qquad \forall x \in X. \end{aligned}$$

Thus g, considered as a function of x is never below f and tocuhes f at y.

The majorization algorithm corresponding with this majorization function g updates x at iteration k by

$$x^{(k+1)} \in \operatorname*{\mathbf{argmin}}_{x \in X} g(x, x^{(k)}),$$

unless we already have

$$x^{(k)} \in \operatorname*{\mathbf{argmin}}_{x \in X} g(x, x^{(k)}),$$

Date: March 24, 2006.

2000 *Mathematics Subject Classification.* 90C30.

Key words and phrases. Mathematical programming, Nonlinear Programming.

in which case we stop. Convergence follows, under some additional simple conditions, from the *sandwich inequality*, which says that if we do not stop at iteration k, then

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

Consider the example $f(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$. The function has local minima at $+1$ and -1, with function value $-\frac{1}{4}$, and a local maximum at 0 with function value 0. A majorization function can be constructed by using $x^2 \geq y^2 + 2y(x - y)$, giving $g(x, y) = \frac{1}{4}x^4 + \frac{1}{2}y^2 - xy$. This leads to the majorization algorithm $x^{(k+1)} = \sqrt[3]{x^{(k)}}$. This converges linearly, with convergence rate $\frac{1}{3}$, to either $+1$ or -1, depending on where we start.

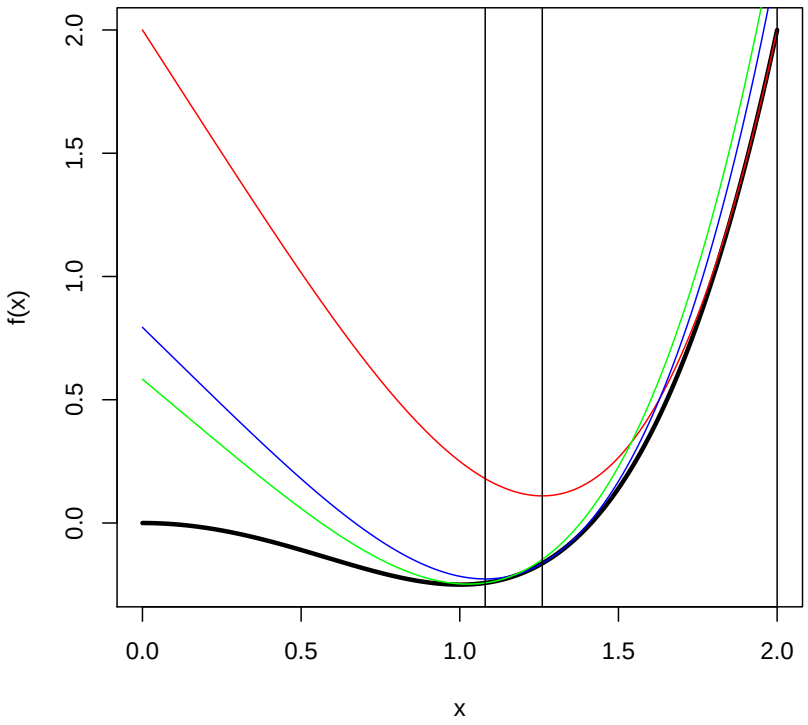

FIGURE 1. Majorization Example

In Figure 1 we start at 2. The majorization function at this point, drawn in red, touches f at 2 and is minimized at $\sqrt[3]{2} \approx 1.2599$. We compute the new majorization function at this point, in blue, and minimize it at $\sqrt[9]{2} \approx 1.0801$. The next step (green) takes us to $\sqrt[27]{2} \approx 1.0260$.

Observe that our algorithm is $x^{(k)} = x^{3^{-k}}$. An equally valid majorization algorithm is $x^{(k)} = (-1)^k x^{1/3^k}$, which also minimizes the majorization function in each step. It produces a decreasing sequence of loss function values converging to $\frac{1}{4}$, but the sequence of solutions is not convergent and has two converging subsequences, one converging to -1 and one to $+1$.

In most cases majorization methods converge at a linear rate, with the rate equal to the largest eigenvalue in modulus of the matrix

$$I - [\mathcal{D}_{11}g(x,x)]^{-1}\mathcal{D}^2 f(x) = -[\mathcal{D}_{11}g(x,x)]^{-1}\mathcal{D}_{12}g(x,x)$$

where the derivatives are evaluated at the fixed point x [Ortega and Rheinboldt, 1970, page 300-301]. In some special cases we can have sub-linear or super-linear convergence, but linear convergence is the rule.

2. Using Elementary Inequalities

The first way to construct majorization functions is the simplest one. There are many inequalities in the literature of the form $F(x,y) \geq 0$ with equality if and only if $x = y$. Such inequalities can often be used to construct majorization functions. Since this is not really a systematic approach, we merely illustrate it by a rather detailed example.

After a suitable choice of coordinates and normalization the Euclidean multidimensional scaling problem can be formulated as minimization of

$$\sigma(x) = 1 + x'x - 2\sum_{i=1}^{n} w_i \delta_i d_i(x). \tag{1}$$

Here the w_i are known positive *weights*, the δ_i are the *dissimilarities*, and the $d_i(x)$ are the *Euclidean distances*, defined by $d_i(x) = \sqrt{x'A_i x}$. The A_i are known positive semi-definite matrices that satisfy $\sum_{i=1}^{n} w_i A_i = I$.

In most cases of interest the dissimilarities will be positive, but we shall cover the more general case in which there can be both positive and negative ones. Decomposing δ_i into its positive and negative parts, i.e. $\delta_i = \delta_i^+ - \delta_i^-$ with both δ_i^+ and δ_i^- non-negative. Now we can write

$$\sigma(x) = 1 + x'x - 2\sum_{i=1}^n w_i\delta_i^+ d_i(x) + 2\sum_{i=1}^n w_i\delta_i^- d_i(x). \tag{2}$$

If $d_i(y) > 0$ then by, respectively, the Cauchy-Schwartz and the Arithmetic-Geometric Mean Inequality

$$\frac{1}{d_i(y)}x'A_iy \leq d_i(x) \leq \frac{1}{d_i(y)}\frac{1}{2}(x'A_ix + y'A_iy).$$

Thus

$$\sum_{i=1}^n w_i\delta_i^+ d_i(x) \geq x'B^+(y)y, \tag{3a}$$

with

$$B^+(y) = \sum_{i=1}^n w_i\frac{\delta_i^+}{d_i(y)}A_i. \tag{3b}$$

And

$$\sum_{i=1}^n w_i\delta_i^- d_i(x) \leq \frac{1}{2}(x'B^-(y)x + y'B_-(y)y), \tag{3c}$$

with

$$B^-(y) = \sum_{i=1}^n w_i\frac{\delta_i^-}{d_i(y)}A_i. \tag{3d}$$

Observe that both B^+ and B^- are positive semi-definbite. Combining these results gives

$$\sigma(x) \leq 1 + x'x - 2x'B^+(y)y + x'B^-(y)x + y'B^-(y)y. \tag{4}$$

The right-hand side of (4) gives a quadratic majorization function, and the corresponding algorithm

$$x^{(k+1)} = [I + B^-(x^{(k)})]^{-1}]B^+(x^{(k)})x^{(k)}.$$

At a stationary point x the derivative of the algorithmic map is

$$[I + B^-(x)]^{-1}[(B^+(x) - H^+(x)) + H^-(x)],$$

where

$$H^+(x) = \sum_{i=1}^n w_i \frac{\delta_i^+}{d_i^3(x)} A_i x x' A_i,$$

and

$$H^-(x) = \sum_{i=1}^n w_i \frac{\delta_i^-}{d_i^3(x)} A_i x x' A_i.$$

The matrices H^+, H^-, and $B^+ - H^+$ are all positive semidefinite.

3. Integrals

Supposed want to maximize

$$f(x) = \log \int_Z \exp\{u(x,z)\} dz.$$

Because we are maximizing we will now construct a minorization function and a minorization algorithm.

Of course the logarithm in the definition of f is really irrelevant here and the exponent merely guarantees that we are integrating a positive function. Write

$$f(x) - f(y) = \log \frac{\int_Z \exp\{u(y,z)\} \frac{\exp\{u(x,z)\}}{\exp\{u(y,z)\}} dz}{\int_Z \exp\{u(y,z)\} dz}.$$

Jensen's inequality, or equivalently the concavity of the logarithm, tells us that

$$f(x) - f(y) \geq \frac{\int_Z \exp\{u(y,z)\} \log \frac{\exp\{u(x,z)\}}{\exp\{u(y,z)\}} dz}{\int_Z \exp\{u(y,z)\} dz}.$$

Define

$$\pi(z \mid y) = \frac{\exp\{u(y,z)\}}{\int_Z \exp\{u(y,z)\} dz}.$$

Then

$$f(x) \geq f(y) + \int_Z \pi(z \mid y) u(x,z) dz - \int_Z \pi(z \mid y) u(y,z) dz,$$

which defines our minorization function.

A step of the minorization algorithm simply maximizes (in the "M" step) the "expectation" $\int_Z \pi(z \mid y) u(x,z) dz$. Computing, and possibly simplifying, this expectation is the "E" step. The algorithm is especially attractive, of course, if the integral defining the expectation can be evaluated in closed

form. This is often the case in exponential family problems in statistics, where we want to compute maximum likelihood estimates.

4. Using Convexity

Suppose we want to minimize $f(x)$ on a convex set X. Under very general conditions we can write f as the difference of two convex functions. It is sufficient to assume, for example, that f is twice continuously differentiable. It is necessary and sufficient that f is the indefinite integral of a function of locally bounded variation [Hartman, 1959].

If $f = u - v$, with u and v both convex, then we use

$$v(x) \geq v(y) + \mathcal{D}v(y)(x - y)$$

to construct the convex majorization function

$$g(x, y) = u(x) - v(y) - \mathcal{D}v(y)(x - y).$$

The majorization method reduces optimization of an arbitrary function to solving a sequence of convex optimization problems. Of course matters simplify if $u(x)$ can be chosen to be quadratic.

5. Using Taylor's Theorem

By Taylor's theorem

$$f(x) \leq f(y) + (x - y)'\mathcal{D}f(y) + \frac{1}{2}\max_{0\leq\xi\leq 1}(x - y)'\mathcal{D}^2 f(\xi x + (1 - \xi)y)(x - y),$$

and the right hand side can be used as a majorization function. Of course this general approach can also be applied if we only use the linear term in the Taylor expansion, or if we use third or higher order terms [De Leeuw, 2006]. And by replacing max by min we can use it to construct minorization functions.

But let us continue with *quadratic majorization*. Suppose there is a matrix B such that $\mathcal{D}^2 f(x) \lesssim B$, in the sense that $B - \mathcal{D}^2 f(x)$ is positive semi-definite

for all x. Then clearly

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

is a majorization function for f. We also write $g_y(x)$ for $g(x,y)$ to emphasize that g_y is a function of x that majorizes f at y. Observe that g_y has both the same function value and the same derivative as f at y.

By defining the current *target*

$$z = y - B^{-1}\mathcal{D}f(y),$$

and by completing the square, we see that

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

Thus step k of the majorization algorithm solves the least squares problem

$$\min_{x\in X}(x - z^{(k)})'B(x - z^{(k)}).$$

We can choose the matrix B to be scalar, for instance by using an upper bound for the largest eigenvalue of $\mathcal{D}^2 f(\xi)$. In that case computing the target simplifies, and all majorization subproblems are unweighted least squares problems.

In the case in which X is all of $\mathbb{R}^n$ the quadratic majorization algorithm simply becomes

$$x^{(k+1)} = x^{(k)} - B^{-1}\mathcal{D}f(x^{(k)}).$$

This algorithm will in general have a linear convergence rate $1-\lambda(x)$, where $\lambda(x)$ is the smallest eigenvalue of $B^{-1}\mathcal{D}^2 f(x)$ and x is the fixed point. A smaller B will give a more rapid convergence rate, but in general we cannot expect to see anything faster than linear convergence. If our bound B is really bad, then we may see very slow linear convergence.

References

J. De Leeuw. Quadratic and Cubic Majorization. Preprint series, UCLA Department of Statistics, 2006.

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.

P. Hartman. On Functions Representable as a Difference of Two Convex Functions. *Pacific Journal of Mathematics*, 9:707–713, 1959.

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.

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.

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