ACCELERATING MAJORIZATION ALGORITHMS

JAN DE LEEUW

ABSTRACT. This is a programmatic paper that reviews the construction of majorization algorithms and their rate of convergence. It then discusses some important examples of majorization algorithms, and reviews the literature on how to accelerate the convergence.

1. FIXED POINT ITERATIONS

Suppose $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. We say that x is a *fixed point* of G if $G(x) = x$. *Picard iteration* or *functional iteration* is a simple algorithm to compute fixed points. We start with some $x^{(0)} \in \mathbb{R}^n$ and define subsequent points by

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

Definition 1.1. Let $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$. Then $x^\star$ is a *point of attraction* of the iteration (1) if there is an open neighborhood $\mathcal{N}$ of $x^\star$ such that , for any $x^{(0)} \in \mathcal{N}$, the iterates $x^{(k)}$ converge to $x^\star$.

Remark 1.1. For continuous G points of attraction are fixed points.

Theorem 1.1 (Ostrowski Theorem)**.** *Suppose that $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ has a fixed point $x^\star$ and is F-differentiable at $x^\star$. If $\rho(\mathcal{D}G(x^\star)) < 1$ then $x^\star$ is a point of attraction of the iteration* (1)*.*

Date: September 7, 2008 — 10h 36min — Typeset in LUCIDA BRIGHT.

Key words and phrases. Majorization, Picard Iteration, Fixed Points, Convergence Acceleration, EM, Multidimensional Scaling.

Remark 1.2. Remember that the spectral radius $\rho(A)$ of a square matrix A is its largest singular value.

Remark 1.3. Picard sequences can converge to fixed points which are *not* points of attraction.

Theorem 1.2 (Linear Convergence Theorem)**.** *Under the conditions of Theorem* 1.1 *the sequence has $\mathcal{R}$-convergence rate $\rho(\mathcal{D}G(x^\star))$. If $\rho(\mathcal{D}G(x^\star)) >$ 0 convergence is linear.*

Remark 1.4. For precise definitions we refer to Ortega and Rheinboldt [1970, Chapters 9-10]. They also discuss the case $\rho(\mathcal{D}G(x^\star)) = 1$, in which we may or may not have linear convergence. In fact we may or may not have convergence at all. If $\rho(\mathcal{D}G(x^\star)) = 0$ we have superlinear and, under additional regularity conditions, quadratic convergence. If $\rho(\mathcal{D}G(x^\star)) > 1$ we know from Ostrowski [1966, Theorem 22.2] that $x^\star$ is a *point of repulsion*, i.e. we have divergence from at least some starting points.

Remark 1.5. For a more general theory we can relax F-differentiability and use Clarke Generalized Jacobians $\mathcal{D}G(x^\star)$ or other set-valued generalized derivatives. Some partial results are available.

Remark 1.6. We know that $\rho(A) \geq \|A\|$, where $\|A\|$ is any induced matrix norm. Thus the convergence speed of Picard iteration can be bounded below by any induced norm, using $\|\mathcal{D}G(x_\star)\| \leq \rho(\mathcal{D}G(x_\star)) < 1$.

2. Block Relaxation

We now discuss a specific class of algorithms in which the linear convergence rate can be specified more precisely, because a simpler expression for the spectral radius can be given.

Suppose $g : \mathbb{R}^n \otimes \mathbb{R}^n \longrightarrow \mathbf{R}$. Start with $(x^{(0)}, y^{(0)})$. Then define

$$y^{(k+1)} = \underset{y}{\mathbf{argmin}}\ g(x^{(k)}, y),$$

$$x^{(k+1)} = \underset{x}{\mathbf{argmin}}\ g(x, y^{(k+1)}),$$

assuming the minima exist and are unique. For the map $(x^{(k+1)}, y^{(k+1)}) = G(x^{(k)}, y^{(k)})$ with fixed point $(x^\star, y^\star)$ we find that

$$\mathcal{D}G = \begin{bmatrix} \mathcal{D}_{11}^{-1}\mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21} & 0 \\ \mathcal{D}_{22}^{-1}\mathcal{D}_{21} & 0 \end{bmatrix},$$

where the $\mathcal{D}_{st}$ are the submatrices of the Hessian of g. Of course we assume these second derivatives exist. The diagonal submatrices $\mathcal{D}_{11}$ and $\mathcal{D}_{22}$ are positive definite, and thus non-singular, at a fixed point. Thus $\rho(\mathcal{D}G) = \rho(\mathcal{D}_{11}^{-1}\mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21})$.

Remark 2.1. Note that the eigenvalues of $\mathcal{D}G$ are real and non-negative at a fixed point. Thus the spectral radius of $\mathcal{D}G$ is simply its largest eigenvalue, which is the square of the largest singular value of $\mathcal{D}_{11}^{-\frac{1}{2}}\mathcal{D}_{12}\mathcal{D}_{22}^{-\frac{1}{2}}$. And this is also the convergence rate of the block relaxation algorithm.

Remark 2.2. If the minima exist, but are not unique, we need a differentiable selection from the point-to-set **Argmin**'s. We need a selection anyway in any computer implementation of the algorithm.

Remark 2.3. These results can be extended, and have been extended, to cases in which there are more than two blocks.

Remark 2.4. These results can be extended, and have been extended, to cases in which the minimization in the blocks imposes equality and inequality constraints and we use the Lagrange multiplier method.

3. Majorization

In majorization algorithm we want to minimize a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. To construct an algorithm we assume we have a *majorization function*

$g : \mathbb{R}^n \otimes \mathbb{R}^n \longrightarrow \mathbb{R}$ such that

$$f(x) \leq g(x,y) \qquad \forall x,y,$$
$$f(x) = g(x,x) \qquad \forall x.$$

If we apply block relaxation to the majorization function we obtain

$$y^{(k+1)} = \underset{y}{\textbf{argmin}}\; g(x^{(k)},y) = x^{(k)},$$
$$x^{(k+1)} = \underset{x}{\textbf{argmin}}\; g(x,y^{(k+1)}),$$

and thus

$$x^{(k+1)} = \underset{x}{\textbf{argmin}}\, g(x,x^{(k)}).$$

The general results for block relaxation apply, but we have some additional structure which can be used to simplify the expression for the spectral radius.

If the functions are two times differentiable, then

$$\mathcal{D}f(x) = \mathcal{D}_1 g(x,x),$$
$$\mathcal{D}^2 f(x) = \mathcal{D}_{11} g(x,x) + \mathcal{D}_{12} g(x,x).$$

Also $\mathcal{D}^2 f(x) \lesssim \mathcal{D}_{11} g(x,x)$, in the Loewner sense, i.e. $\mathcal{D}_{11} g(x,x) - \mathcal{D}^2 f(x)$ is positive semi-definite. The off-diagonal matrix $\mathcal{D}_{12} g(x,x)$ is symmetric and negative definite.

As a consequence

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

Thus the convergence speed of the majorization algorithm at the fixed point $x^\star$ is $1 - \lambda_n(x^\star)$, where $\lambda_n(x^\star)$ is the smallest generalized eigenvalue of

$$\mathcal{D}^2 f(x^\star) z = \lambda \mathcal{D}_{11} g(x^\star, x^\star) z$$

Remark 3.1. If the gradient algorithm

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

converges to a fixed point $x^\star$ with $\mathcal{D}f(x^\star) = 0$, then it has the same convergence rate as the majorization algorithm.

Remark 3.2. If we need to maximize f we use a minorization function g.

4. Four Majorization Examples

4.1. **Multdimensional Scaling.** In (Euclidean, Metric) Multidimensional Scaling (MDS) we have M positive semi-definite matrices C_m of order K, and two M-vectors w and δ, both with positive elements, such that $\sum_{m=1}^M w_m C_m = I$ and $\frac{1}{2}\sum_{m=1}^M \delta_m^2 = 1$.

The problem is to minimize

$$f(x) = 1 + \frac{1}{2}\mathbf{tr}\; x'x - \sum_{m=1}^M w_m \delta_m \sqrt{x' C_m x}. \tag{2}$$

It is shown in De Leeuw [1984] that near a local minimum of stress we have $x' C_m x > 0$ for all m. If we define $d_m(x) = \sqrt{x' C_m x}$ and

$$B(x) = \sum_{m=1}^M w_m \frac{\delta_m}{d_m(x)} C_m,$$

then the partials are

$$\mathcal{D}f(x) = x - B(x)x.$$

The Hessian is

$$\mathcal{D}^2 f(x) = I - H(x),$$

with

$$H(x) = \sum_{m=1}^M w_m \frac{\delta_m}{d_m(x)} \left\{ C_m - \frac{C_m x x' C_m}{x' C_m x} \right\}.$$

Observe that $H(x)$ is positive semi-definite, and that its smallest eigenvalue, corresponding to the eigenvector x, is equal to zero. At a strict local minimum $\mathcal{D}^2\sigma(x)$ is positive definite, which means that $H(x)$ has all its eigenvalues less than one. It also means that the Newton-Raphson correction to x is is $(I - H(x))^{-1}B(x)$.

The SMACOF algorithm for MDS [De Leeuw, 1977] is derived from the Cauchy-Schwarz inequality which says

$$d_m(x) \geq \frac{1}{d_m(y)} x' C_m y.$$

Thus

$$f(x) \leq g(x,y) = 1 + \frac{1}{2} x'x - x'B(y)y, \tag{3}$$

and we take $x^{(k+1)} = G(x^{(k)}) = B(x^{(k)})x^{(k)}$. Since $\mathcal{D}G(x) = H(x)$, the derivative of the iteration function at a strict local minimum has all its eigenvalues between zero and one, with the largest one strictly less than one and the smallest one precisely equal to zero [De Leeuw, 1988].

4.2. **The EM Algorithm.**

The EM algorithm [Dempster et al., 1977] was designed to maximize functions of the form

$$f(x) = \log \int_\Omega \pi(x,z)dz,$$

where $\pi(x,z) \geq 0$. By Jensen's Inequality

$$f(x) - f(y) = \log \left\{ \frac{\int_\Omega \pi(y,z) \frac{\pi(x,z)}{\pi(y,z)} dz}{\int_\Omega \pi(y,z)dz} \right\} \geq \frac{\int_\Omega \pi(y,z) \log \frac{\pi(x,z)}{\pi(y,z)} dz}{\int_\Omega \pi(y,z)dz}.$$

Define

$$\pi(x) = \int_\Omega \pi(x,z)dz,$$
$$\pi(z \mid x) = \frac{\pi(x,z)}{\pi(x)},$$

and also $\eta(x,z) = \log \pi(x,z)$. Then we see that

$$g(x,y) = f(y) + \int_\Omega \pi(z \mid y)\eta(x,z)dz - \int_\Omega \pi(z \mid y)\eta(y,z)dz$$

is a minorization function for f. The minorization algorithm is

$$x^{(k+1)} = \mathop{\textbf{argmax}}_{x} \int_{\Omega} \pi(z \mid x^{(k)})\eta(x,z)dz.$$

Now assume we can differentiate under the integral sign and define

$$\begin{aligned} H(x) &= \int_{\Omega} \pi(z|x)\mathcal{D}_{11}\eta(x,z)dz, \\ A(x) &= \int_{\Omega} \pi(z|x)\mathcal{D}_{1}\eta(x,z)\mathcal{D}_{1}\eta(x,z)'dz, \\ D(x) &= \int_{\Omega} \pi(z|x)\mathcal{D}_{1}\eta(x,z)dz. \end{aligned}$$

Then

$$\begin{aligned} \mathcal{D}g_{11}(x,x) &= H(x), \\ \mathcal{D}^2 f(x) &= H(x) - [A(x) - D(x)D(x)'], \end{aligned}$$

and thus

$$\mathcal{D}G(x) = A(x) - D(x)D(x)'.$$

4.3. **Eigenvalues and Pagerank.** The dominant eigenvector of a positive semi-definite matrix A is found by maximizing

$$f(x) = \frac{x'Ax}{x'x}$$

By Cauchy-Schwartz

$$f(x) \geq g(x,y) = \frac{(x'Ay)^2}{x'x.y'Ay},$$

which gives the minorization algorithm

$$x^{(k+1)} = \mathop{\textbf{argmax}}_{x} g(x, x^{(k)}) \propto Ax^{(k)}.$$

Since the update is only defined up to a proportionality factor, we can set

$$x^{(k+1)} = \frac{Ax^{(k)}}{\|Ax^{(k)}\|}.$$

Now

$$\mathcal{D}G(x) = \frac{1}{\|Ax\|}\left\{A - \frac{Axx'A}{x'Ax}\right\},$$

and thus at an eigenvector $x^\star$ corresponding with the largest eigenvalue λ_1 of A we have

$$\rho(\mathcal{D}G(x^\star)) = \frac{\lambda_2}{\lambda_1} \leq 1,$$

where λ_2 is the second largest eigenvalue of A (of course we can have $\lambda_1 = \lambda_2$).

4.4. **Logistic Scaling.** Consider the problem of minimizing

$$f(x) = -\sum_{j=1}^{m} p_j \log \pi_j(x),$$

where the elements of the vector p are non-negative and add up to one, and where

$$\pi_j(x) = \frac{\exp(x_j)}{\sum_{\ell=1}^{m} \exp(x_\ell)}.$$

Now

$$\mathcal{D}f(x) = \pi(x) - p,$$
$$\mathcal{D}^2 f(x) = \Pi(x) - \pi(x)\pi(x)'.$$

To find the majorization function we observe that $\mathcal{D}^2 f(x) \lesssim \frac{1}{2}I$ and consequently $f(x) \leq g(x,y)$ with

$$g(x,y) = f(y) + (\pi(y) - p)'(x - y) + \frac{1}{4}(x - y)'(x - y),$$

which leads to the algorithm

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

In a more general, and more interesting, situation we have

$$\pi_j(x) = \frac{\exp(\eta_j(x))}{\sum_{\ell=1}^{m} \exp(\eta_\ell(x))},$$

with the η_j known functions of x. Then $f(x) \leq g(x,y$ with

$$g(x,y) = f(y) + (\pi(y) - p)'(\eta(x) - \eta(y)) + \frac{1}{4}(\eta(x) - \eta(y))'(\eta(x) - \eta(y)).$$

Letting $\tilde{\eta}^{(k)} = \eta(x^{(k)}) - 2(\pi(x^{(k)}) - p)$ we find the majorization algorithm

$$x^{(k+1)} = \underset{x}{\mathbf{argmin}}(\eta(x) - \tilde{\eta}^k)'(\eta(x) - \tilde{\eta}^k).$$

Thus each step of the algorithm involves solving a generally nonlinear least squares problem. What we have accomplished is to replace a logistic maximum likelihood problem by a sequence of (unweighted) least squares problems. This is quite independent of the nature of the functions η_j, which could be linear functions as in regression, distance functions as in MDS, or inner products as in PCA.

To compute the convergence rate we need some definitons. Let $V(x) = \Pi(x) - \pi(x)\pi(x)'$ and $D(x) = \mathcal{D}\eta(x)$. Finally

$$H(x) = \mathcal{D}^2\eta(x)(\pi(x) - p) = \sum_{j=1}^{m}(\pi_j(x) - p_j)\mathcal{D}^2\eta_j(x).$$

Then

$$\mathcal{D}g_{12}(x^\star, x^\star) = D(x^\star)'(V(x^\star) - \frac{1}{2}I)D(x^\star),$$
$$\mathcal{D}g_{11}(x^\star, x^\star) = H(x^\star) + \frac{1}{2}D(x^\star)'D(x^\star),$$

and thus

$$\mathcal{D}G(x^\star) = I - [H(x^\star) + \frac{1}{2}D(x^\star)'D(x^\star)]^{-1}[H(x^\star) + D(x^\star)'V(x^\star)D(x^\star)].$$

In the linear case with $\eta = Ax$ we have $H(x) = 0$ and $D(x) = A$. Thus

$$\mathcal{D}G(x^\star) = I - 2K'VK,$$

where $K = A(A'A)^{-1/2}$.

References

J. De Leeuw. Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, 5:163–180, 1988.

J. De Leeuw. Differentiability of kruskal's stress at a local minimum. *Psychometrika*, 49:111–113, 1984.

J. De Leeuw. Applications of convex analysis to multidimensional scaling. In J.R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, editors, *Recent developments in statistics*, pages 133–145, Amsterdam, The Netherlands, 1977. North Holland Publishing Company.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

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

A. M. Ostrowski. *Solution of Equations and Systems of Equations.* Academic Press, New York, N.Y., 1966.

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