MAJORIZATION METHODS IN STATISTICS

JAN DE LEEUW AND GEORGE MICHAILIDIS

1. INTRODUCTION

It is a pleasure to comment on such a well-written and obviously important paper.

We agree with the basic explicit message of Lange, Hunter, and Yang (LHY). Their so-called "optimization transfer" algorithms form a very interesting and versatile class. One of the main reasons for this is that taylor-made statistical techniques written in interpreted languages are becoming more and more common. For such techniques, and in such computational environments, taylor-made algorithms in the "optimization transfer" class are, at least initially, very convenient, although perhaps ultimately not optimal.

We also agree with what we read as a more implicit message in LHY. The usual derivations of the EM algorithm tend to be somewhat mysterious, because they confound statistics and numerical analysis. The notion of likelihood and of missing data can be used to provide statistical interpretations of the algorithm, but the engine that drives EM is majorization based on Jensen's inequality (or, what amounts to the same thing, on the concavity of the logarithm function).

1.1. **Terminology.**

Although we realize that this a minor point, we are not very happy with the "optimization transfer" terminology. The basic reasoning behind it is that optimization is transferred to a surrogate function that approximates the original function but is simpler to handle. This, however, is much too general for our taste. In the steepest descent method, we minimize the one-dimensional surrogate in the direction of the negative gradient. In Newton's method we minimize a quadratic approximation. Both algorithms consequently use "optimization transfer", but LHY deal with a much more specific class of algorithms.

Consequently we shall continue to use the more specific term "majorization methods", to distinguish them explicitly from other "optimization transfer" methods.

1.2. **Presentation.** For obvious reasons, LHY start with the derivation of the EM algorithm and then strip away the statistical interpretations to arrive at the core of the algorithm, which is based on majorization. In our presentation, we will follow a somewhat different route.

Convergence (both local and global) of majorization algorithms follows easily from the fact that they are block relaxation algorithms. Thus we discuss this larger class first, in particular because it includes many interesting statistical algorithms, and because there are some interesting relationships with Gibbs sampling. Augmentation algorithms form an intermediate class of algorithms, which also deserve some attention, if only because the general idea of augmentation also plays a role in Markov Chain Monte Carlo techniques.

Because the didactic bottom-up approach, which is to start with EM and then generalize it, has been presented so well by LHY, we take the top-down approach. The presentation is similar to de Leeuw [1994]. Many more details, examples, and references can be found in the unpublished, and sadly incomplete, monograph by de Leeuw and Michailidis [1999a].

2. BLOCK RELAXATION

The most general class of algorithms we discuss are *Block Relaxation* (BR) algorithms. The problem is to minimize a function $g(x, y)$ defined on $X \otimes Y$, where X, Y are usually subsets of $\mathbb{R}^n$; that is, the arguments of the function can be partitioned into two blocks. The algorithm starts with some $x^{(0)}$, then computes a corresponding $y^{(0)}$ by minimizing $g(x^{(0)}, y)$ on Y, then computes $x^{(1)}$ by minimizing $g(x, y^{(0)})$ on X, and so on.

This idea can easily be generalized to more than two blocks, although we then face the problem of deciding how we are going to cycle through the blocks.

2.1. **Examples.** The most familiar forms of BR are *Cyclic Coordinate Descent* (CCD), in which each block corresponds to a single coordinate, and *Alternating Least Squares* (ALS), in which the loss function is a sum of squares and each sub-problem corresponds to a linear least squares problem. CCD is used in the Jacobi method for computing eigenvalues and in the Gauss-Seidel, Gauss-Jacobi, and similar relaxation methods for systems of linear equations [Golub and Loan, 1997]. In multivariate analysis, ALS algorithms are a natural alternative to computing singular value decompositions, correspondence analysis solutions, and canonical correlation analysis [Gifi, 1990].

Applications of BR in statistics are discussed in Oberhofer and Kmenta [1974] and Jensen et al. [1991], although in both cases various wheels are reinvented.

2.2. **Global Convergence.** The nice property of BR is that it is convergent from any starting point under quite weak conditions. It is sufficient that f is continuous and $X \otimes Y$ is compact, but even weaker conditions are possible. However, it is not necessary that the minima in both subproblems are unique. The easiest way to prove global convergence is to use the general theory due to Zangwill [1969].

A common variation on BR is not to go all the way. Thus we do not minimize $g(x, y^{(k)})$ over $x \in X$, say, but we use a map $U : X \otimes Y \to X$ so that $g(U(x^{(k)}, y^{(k)}), y^{(k)}) < g(x^{(k)}, y^{(k)})$; and similarly, perhaps, for the other block. As long as the map U is continuous (more precisely closed) Zangwill's general theory applies.

2.3. **Local Convergence.** Under quite general conditions BR methods converge linearly, and the convergence rate corresponds to the largest eigenvalue of the matrix

$$\mathcal{M} = \mathcal{D}_{11}^{-1}\mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21}$$

constructed from the corresponding blocks of second partials of g. Obviously this supposes these partials exist. Also the formulas must be adapted analogously in case the subproblems incorporate either equality or inequality constraints.

3. AUGMENTATION

Suppose the problem is to minimize a function $f(x)$ on $X \subseteq \mathbb{R}^n$. One possible strategy is to find a second function $g(x, y)$ on $X \otimes Y$, where $Y \subseteq \mathbb{R}^m$, such that

$$f(x) = \min_{y \in Y} g(x, y)$$

for all $x \in X$. Such a function $g(\bullet, \bullet)$ is called an *augmentation* of $f(\bullet)$. Augmentation algorithms now minimize g by applying block relaxation.

3.1. **Examples.** The most familiar example for statisticians is perhaps the Yates algorithms for unbalanced factorial designs [Yates, 1934]. The design is made balanced by adding a suitable number of pseudo-observations to each cell. The least squares loss function is then minimized over both parameters and pseudo-observations using ALS.

Other examples include the Thomson refactoring method in factor analysis [Thomson, 1934], where the diagonal elements of the correlation matrix (communalities) are used to augment the least squares loss function, a similar refactoring method in least squares squared distance scaling, and imputation methods for missing values in singular value decomposition and matrix approximation.

3.2. **Local Convergence.** Since the Hessian of f is given by

$$\mathcal{D}^2 f = \mathcal{D}_{11} - \mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21}$$

it can be seen that the local convergence rate for augmentation algorithms is the largest eigenvalue of

$$\mathcal{M} = \mathcal{I} - \mathcal{D}_{11}^{-1}\mathcal{D}^2 f.$$

4. MAJORIZATION

In majorization algorithms the problem is, once again, to minimize $f(x)$ on X. Moreover, suppose that we have a second function $g(x, y)$ on $X \otimes X$ such that

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

Another way of saying this is that

$$\begin{aligned} f(x) &= \min_{y \in X} g(x, y), \\ x &= \operatorname*{argmin}_{y \in X} \ g(x, y). \end{aligned}$$

We see that majorization algorithms are a narrower class than augmentation algorithms, because (i) $X = Y$ and (ii) the presence of the argmin condition, which makes one of the BR subproblems trivial to solve. In majorization, we also have that $\mathcal{D}^2 f = \mathcal{D}_{11} + \mathcal{D}_{12}$, and thus $\mathcal{M} = -\mathcal{D}_{11}^{-1}\mathcal{D}_{12}$.

Global convergence follows from the *Sandwich Inequality*

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

where the first inequality is due to the majorization conditions, and the second one from the fact that the majorization function g is minimized in each step.

4.1. **History.** As LHY point out, majorization algorithms were perhaps first used systematically in statistics in the area of multidimensional scaling [de Leeuw, 1977]. The general EM algorithm [Dempster et al., 1977] was discussed around the same time, and comparing the two clearly showed what they had in common. As soon as the general principle was isolated, it became quite popular in various multivariate analysis procedures [de Leeuw, 1990; Kiers, 1990; Heiser, 1995; Verboon, 1994], mainly in psychometrics.

Recently, various applications of majorization to graph drawing and location analysis are discussed in de Leeuw and Michailidis [1999b]. One of the classical algorithms in that area is the majorization algorithm of Weiszfeld [1937] (see also Vosz and Eckhardt [1980] and Eckhardt [1980] for a more modern exposition).

4.2. **The Classical Inequalities.** In order for majorization algorithms to work in practice, we need to find a majorizing function g that is easy to minimize. In de Leeuw [1994] we distinguish type I majorizations, that employ linear majorizers for convex functions, and type II majorizations, that use quadratic majorizers for functions with bounded second derivatives (some simple examples are provided in Borg and Groenen [1997]). The two approaches are combined in a clever way in Groenen et al. [1997].

Any inequality of the form $F(x,y) \leq G(a(x),b(y))$, with equality if and only if $x = y$, can be used to derive majorization algorithms. We have seen that Jensen's inequality leads to the EM algorithm, while Young's inequality is used in de Leeuw and Michailidis [1999b], and its special case, the Arithmetic Mean-Geometric Mean inequality, features in Heiser [1987].

4.3. **Dinkelbach Majorization.** It is perhaps useful to point out that again, as in EM, the assumptions that drive the majorization algorithm can be relaxed. As in GEM, it suffices to decrease the majorization function with a continuous map. More importantly, the concept of majorization itself can be relaxed. The Sandwich Inequality guarantees that

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

But actually this implication is all we need to construct a convergent majorization algorithm. Consider the (common) problem of minimizing a ratio $f(x) = a(x)/b(x)$ over $x \in X$, where $b(x) > 0$. Define $g(x,y) = a(x) - f(y)b(x)$. Then $g(x^{(k)},x^{(k)}) = 0$, and if $g(x^{(k+1)},x^{(k)}) \leq 0$ then $f(x^{(k+1)}) \leq f(x^{(k)})$. Thus minimizing $a(x) - f(x^{(k)})b(x)$ over $x \in X$ in each step provides a convergent algorithm. In the context of fractional programming this was proposed by Dinkelbach [1967]. An application of this idea in the field of psychometrics can be found in Kiers [1995].

REFERENCES

I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, 1997.

J. de Leeuw. Applications of convex analysis to multidimensional scaling. In B. van Cutsem et al., editor, *Recent advantages in Statistics*, Amsterdam, Netherlands, 1977. North Holland Publishing Company.

J. de Leeuw. Multivariate analysis with optimal scaling. In S. Das Gupta and J. Sethuraman, editors, *Progress in Multivariate Analysis*, Calcutta, India, 1990. Indian Statistical Institute.

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 and G. Michailidis. Block relaxation algorithms in statistics. http://www.stat.uc la.edu/ deleeuw/block.pdf, 1999a.

J. de Leeuw and G. Michailidis. Multivariate data analysis using constrained pulling. http://www.stat.ucl a.edu/ deleeuw/pull.pdf, 1999b.

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.

W. Dinkelbach. On nonlinear fractional programming. *Management Science*, 13:492–498, 1967.

U. Eckhardt. Weber's problem and Weiszfeld's algorithm in general spaces. *Mathematical Programming*, 18:186–196, 1980.

A. Gifi. *Nonlinear multivariate analysis*. Wiley, Chichester, England, 1990.

G.H. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins Univeristy Press (3rd ed), Baltimore, 1997.

P.J.F. Groenen, W.J. Heiser, and J.J. Meulman. Global optimization in least squares multidimensional scaling by distance smoothing. Technical report, Department of Data Theory, University of Leiden, 1997.

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 Advanmtages in Descriptive Multivariate Analysis*. Oxford: Clarendon Press, 1995.

S. T. Jensen, S. Johansen, and S. L. Lauritzen. Globally convergent algorithms for maximizing a likelihood function. *Biometrika*, 78:867–877, 1991.

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.

W. Oberhofer and J. Kmenta. A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica*, 42: 579–590, 1974.

G.H. Thomson. Hotelling's method modfied to give Spearman's *g*. *Journal of Educational Psychology*, 25:366–374, 1934.

P. Verboon. *A Robust Approach to Nonlinear Multivariate Analysis*. PhD thesis, University of Leiden, 1994. Also published by DSWO Press.

H. Vosz and U. Eckhardt. Linear convergence of a generalized Weiszfeld's method. *Computing*, 25:243–251, 1980.

E. Weiszfeld. Sur le point par lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematics Journal*, 43:355–386, 1937.

F. Yates. The analysis of multiple classifications with unequal numbers in different classes. *Journal of the American Statistical Association*, 29: 51–66, 1934.

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

DEPARTMENT OF STATISTICS, UCLA
E-mail address: deleeuw@stat.ucla.edu

DEPARTMENT OF STATISTICS, THE UNIVERSITY OF MICHIGAN
E-mail address: gmichail@umich.edu