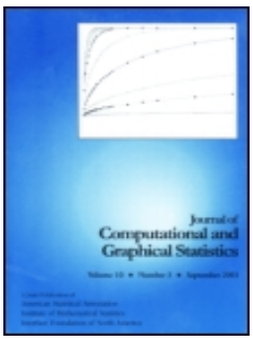

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Optimization Transfer Using Surrogate Objective Functions

Kenneth Lange , David R. Hunter & Ilsoon Yang

To cite this article: Kenneth Lange , David R. Hunter & Ilsoon Yang (2000) Optimization Transfer Using Surrogate Objective Functions, Journal of Computational and Graphical Statistics, 9:1, 1-20

To link to this article: https://doi.org/10.1080/10618600.2000.10474858

DISCUSSION ARTICLE

Optimization Transfer Using Surrogate Objective Functions

Kenneth LANGE, David R. HUNTER, and Ilsoon YANG

The well-known EM algorithm is an optimization transfer algorithm that depends on the notion of incomplete or missing data. By invoking convexity arguments, one can construct a variety of other optimization transfer algorithms that do not involve missing data. These algorithms all rely on a majorizing or minorizing function that serves as a surrogate for the objective function. Optimizing the surrogate function drives the objective function in the correct direction. This article illustrates this general principle by a number of specific examples drawn from the statistical literature. Because optimization transfer algorithms often exhibit the slow convergence of EM algorithms, two methods of accelerating optimization transfer are discussed and evaluated in the context of specific problems.

Key Words: Convexity; EM algorithm; Majorization; Maximum likelihood; Newton's method.

1. INTRODUCTION

Although the repeated successes of the EM algorithm in computational statistics have prompted a veritable alphabet soup of generalizations (Dempster, Laird, and Rubin 1977; Little and Rubin 1987; McLachlan and Krishnan 1997), all of these generalizations retain the overall missing data perspective. This article surveys a different extension that features optimization transfer rather than missing data. The EM algorithm transfers maximization from the log-likelihood $L(\theta)$ of the observed data to a surrogate function $Q(\theta \mid \theta^n)$ depending on the current iterate θ^n through the complete data. The key ingredient in making this transfer successful is the fact that $L(\theta) - Q(\theta \mid \theta^n)$ attains its minimum at $\theta = \theta^n$. Thus, if we determine the next iterate θ^{n+1} to maximize $Q(\theta \mid \theta^n)$, then the

Kenneth Lange is Professor, Departments of Biomathematics and Human Genetics, University of California School of Medicine, Los Angeles, CA 90095-1766 (E-mail: klange@ucla.edu). David R. Hunter is Assistant Professor, Department of Statistics, Penn State University, University Park, PA 16802-2111 (E-mail: dhunter@stat.psu.edu). Ilsoon Yang is Senior Statistician, Schering-Plough Research Institute, 2015 Galloping Hill Road, Kenilworth, NJ 07033 (E-mail: ilsoon.yang@spcorp.com).

Journal of Computational and Graphical Statistics, Volume 9, Number 1, Pages 1–20

well-known inequality

$$\begin{aligned} L(\theta^{n+1}) &= Q(\theta^{n+1} \mid \theta^n) + L(\theta^{n+1}) - Q(\theta^{n+1} \mid \theta^n) \\ &\geq Q(\theta^n \mid \theta^n) + L(\theta^n) - Q(\theta^n \mid \theta^n) \\ &= L(\theta^n) \end{aligned}$$

shows that we increase $L(\theta)$ in the process. The EM derives its numerical stability from this ascent property.

The ascent property of the EM algorithm ultimately depends on the entropy inequality

$$\mathrm{E}_a\left[\ln \mathrm{b}(\mathrm{Z})\right] \leq \mathrm{E}_a\left[\ln \mathrm{a}(\mathrm{Z})\right] \tag{1.1}$$

for probability densities $a(z)$ and $b(z)$. Inequality (1.1) is an immediate consequence of Jensen's inequality and the convexity of $-\ln(z)$. In the EM setting, we denote the complete data by X with likelihood $f(X \mid \theta)$ and the observed data by Y with likelihood $g(Y \mid \theta)$. In inequality (1.1) we replace Z by X given Y, $b(Z)$ by the conditional density $f(X \mid \theta)/g(Y \mid \theta)$, and $a(Z)$ by the conditional density $f(X \mid \theta^n)/g(Y \mid \theta^n)$. Setting $Q(\theta \mid \theta_n) = \mathrm{E}\left[\ln \mathrm{f}(\mathbf{X} \mid \theta) \mid \mathrm{Y} = \mathrm{y}, \theta^n\right]$ and $L(\theta) = g(Y \mid \theta)$ then gives

$$\begin{aligned} Q(\theta \mid \theta^n) - L(\theta) &= \mathrm{E}\left\{\ln\left[\frac{\mathrm{f}(\mathbf{X} \mid \theta)}{\mathrm{g}(\mathrm{Y} \mid \theta)}\right] \mid \mathrm{Y}, \theta^n\right\} \\ &\leq \mathrm{E}\left\{\ln\left[\frac{\mathrm{f}(\mathbf{X} \mid \theta^n)}{\mathrm{g}(\mathrm{Y} \mid \theta^n)}\right] \mid \mathrm{Y}, \theta^n\right\} \\ &= Q(\theta^n \mid \theta^n) - L(\theta^n). \end{aligned}$$

In other words, if we redefine $Q(\theta \mid \theta^n)$ by adding the constant $L(\theta^n) - Q(\theta^n \mid \theta^n)$ to it, then

$$L(\theta) \geq Q(\theta \mid \theta^n) \tag{1.2}$$

for all θ, with equality for $\theta = \theta^n$. The EM algorithm proceeds by alternately forming the minorizing function $Q(\theta \mid \theta^n)$ in the E step and then maximizing it with respect to θ in the M step.

If we want to minimize an arbitrary objective function $L(\theta)$, then we can transfer optimization to a majorizing function $Q(\theta \mid \theta^n)$, defined as in inequality (1.2) but with the inequality sign reversed. Minimizing $Q(\theta \mid \theta^n)$ then drives $L(\theta)$ downhill. "Optimization transfer" seems to us to be a good descriptive term for this process. The alternative term, "iterative majorization," is less desirable in our opinion. First, it suffers from the fact that "majorization" also refers to an entirely different topic in mathematics (Marshall and Olkin 1979). Second, as often as not, we seek to minorize rather than majorize. Regardless of nomenclature, optimization transfer shares with the EM algorithm the exploitation of convexity in constructing surrogate optimization functions.

In those cases where it is impossible to optimize $Q(\theta \mid \theta^n)$ exactly, the one-step Newton update

$$\theta^{n+1} = \theta^n - d^2Q(\theta^n \mid \theta^n)^{-1}dL(\theta^n)^t \tag{1.3}$$

can be employed. Here d denotes the first differential with respect to θ, and d^2 denotes the second differential. In differentiating $Q(\theta \mid \theta^n)$, we always differentiate with respect to the left argument θ, holding the right argument θ^n fixed. Note that the first differential of $Q(\theta \mid \theta^n)$ satisfies $dQ(\theta^n \mid \theta^n) = dL(\theta^n)$ because $L(\theta) - Q(\theta \mid \theta^n)$ has a stationary point at $\theta = \theta^n$. Also observe that in most practical problems, $Q(\theta \mid \theta^n)$ is either strictly concave or strictly convex or can be rendered so by an appropriate change of variables. This fact ensures that the inverse $d^2Q(\theta^n \mid \theta^n)^{-1}$ exists in the approximate optimization transfer algorithm (1.3). This algorithm generalizes the EM gradient algorithm introduced by Lange (1995b) and enjoys the same local convergence properties as exact optimization transfer.

In common with the EM algorithm, optimization transfer tends to substitute simple optimization problems for difficult optimization problems. Simplification usually relies on one or more of the following devices: (1) separation of parameters; (2) avoidance of large matrix inversions; (3) linearization; (4) substitution of a differentiable surrogate function for a nondifferentiable objective function; and (5) graceful handling of equality and inequality constraints. Optimization transfer also shares with the EM algorithm an agonizingly slow convergence in some problems. Besides bringing to the attention of the statistical community the wide variety of optimization transfer algorithms, this article suggests remedies that accelerate their convergence.

Sorting out the history of optimization transfer is as problematic as sorting out the history of the EM algorithm. The general idea appears in the numerical analysis text of Ortega and Rheinboldt (1970, pp. 253–255) in the context of line search methods. De Leeuw and Heiser (1977) presented an algorithm for multidimensional scaling based on majorizing functions; subsequent work in this area was summarized by Borg and Groenen (1997). Huber and Dutter treated robust regression (Huber 1981). Böhning and Lindsay (1988) enunciated a quadratic lower bound principle. In medical imaging, De Pierro (1995) used optimization transfer in emission tomography, and Lange and Fessler (1995c) used it in transmission tomography. The recent articles of de Leeuw (1994), Heiser (1995), and Becker, Yang, and Lange (1997) took a broader view and dealt with the general principle.

In the remainder of this article, Section 2 reviews some of the methods of constructing majorizing and minorizing functions. Each method is illustrated by one or two known examples taken from the fragmentary literature on optimization transfer. (The material on asymmetric least squares and separation of parameters in multidimensional scaling is new.) We hope that readers come away with the impression that construction of a surrogate function via convexity is no more of an art than the clever specification of a complete data space in an EM algorithm. Section 3 briefly mentions the local and global convergence theory of optimization transfer and the theoretical criterion for judging its rate of convergence. Sections 4 and 5 deal with two different techniques for accelerating convergence, and Section 6 provides examples of the effectiveness of acceleration. Section 7 concludes the article with a discussion of open problems and other applications of optimization transfer.

2. CONSTRUCTING OPTIMIZATION TRANSFER ALGORITHMS

There are several ways of exploiting convexity in constructing majorizing and minorizing functions. Suppose $f(u)$ is convex with differential $df(u)$. The inequality

$$f(v) \geq f(u) + df(u)(v-u) \tag{2.1}$$

provides a linear minorizing function at the heart of many optimization transfer algorithms.

Example 1: Bradley–Terry Model of Ranking. In the sports version of the Bradley and Terry model (Bradley and Terry 1952; Keener 1993), each team i in a league of teams is assigned a rank parameter $\theta_i > 0$. Assuming ties are impossible, team i beats team j with probability $\theta_i/(\theta_i + \theta_j)$. If this outcome occurs y_{ij} times during a season of play, then the log-likelihood of the league satisfies

$$\begin{aligned} L(\theta) &= \sum_{i,j} y_{ij} \{\ln \theta_i - \ln(\theta_i + \theta_j)\} \\ &\geq \sum_{i,j} y_{ij} \left\{ \ln \theta_i - \ln(\theta_i^n + \theta_j^n) - \frac{\theta_i + \theta_j - \theta_i^n - \theta_j^n}{\theta_i^n + \theta_j^n} \right\} \\ &= Q(\theta \,|\, \theta^n) \end{aligned}$$

based on inequality (2.1) with $f(u) = -\ln u$ for $u > 0$. The scheme

$$\theta_i^{n+1} = \frac{\sum_{j \neq i} y_{ij}}{\sum_{j \neq i} (y_{ij} + y_{ji})/(\theta_i^n + \theta_j^n)}$$

obviously maximizes $Q(\theta \,|\, \theta^n)$ at each iteration. Because $L(\theta) = L(c\theta)$ for $c > 0$, we constrain $\theta_1 = 1$ and omit the update θ_1^{n+1}.

Example 2: Least Absolute Deviation Regression. Given observations $y_1, \ldots, y_n$ and regression functions $\mu_1(\theta), \ldots, \mu_n(\theta)$, least absolute deviation regression seeks to minimize $\sum_{i=1}^m |y_i - \mu_i(\theta)|$ with respect to a parameter vector θ. If we let $r_i^2(\theta)$ denote the squared residual $[y_i - \mu_i(\theta)]^2$ and invoke the convexity of the function $f(u) = -\sqrt{u}$, then inequality (2.1) implies

$$\begin{aligned} -\sum_{i=1}^m |y_i - \mu_i(\theta)| &= -\sum_{i=1}^m \sqrt{r_i^2(\theta)} \\ &\geq -\sum_{i=1}^m \sqrt{r_i^2(\theta^n)} - \frac{1}{2} \sum_{i=1}^m \frac{r_i^2(\theta) - r_i^2(\theta^n)}{\sqrt{r_i^2(\theta^n)}}. \end{aligned}$$

Thus, we transfer minimization of $\sum_{i=1}^m |y_i - \mu_i(\theta)|$ to minimization of the surrogate function $\sum_{i=1}^m w_i(\theta^n)\{y_i - \mu_i(\theta)\}^2$, where the weight $w_i(\theta) = 1/|y_i - \mu_i(\theta)|$. Although the resulting iteratively reweighted least squares algorithm (Mosteller and Tukey 1977; Rousseeuw and Leroy 1987; Schlossmacher 1973) is actually an EM algorithm, this subtle

fact is far harder to deduce than our simple derivation of the algorithm from convexity considerations (Lange and Sinsheimer 1993). The preceding arguments generalize in interesting and useful ways to estimation with elliptically symmetric distributions such as the multivariate t (Huber 1981; Lange, Little, and Taylor 1989; Lange and Sinsheimer 1993).

Sometimes it is preferable to majorize or minorize by a quadratic function rather than a linear function (Böhning and Lindsay 1988; de Leeuw 1994). This will often be the case for a convex objective function $f(u)$ with bounded curvature. To be more precise, suppose the Hessian $d^2 f(u)$ satisfies $B \succ d^2 f(u)$ for some matrix $B \succ \mathbf{0}$ in the sense that $B - d^2 f(u)$ and B are both positive definite. Then it is trivial to prove that

$$f(v) \leq f(u) + df(u)(v-u) + \frac{1}{2}(v-u)^t B(v-u). \tag{2.2}$$

Example 3: Logistic Regression. Böhning and Lindsay (1988) considered logistic regression with observation y_i, covariate vector x_i, and success probability

$$\pi_i(\theta) = \frac{e^{x_i^t \theta}}{1 + e^{x_i^t \theta}}$$

at trial i. Straightforward calculations show that over m trials the observed information satisfies

$$-d^2 L(\theta) = \sum_{i=1}^m \pi_i(1-\pi_i) x_i x_i^t \leq \frac{1}{4} \sum_{i=1}^m x_i x_i^t.$$

The log-likelihood $L(\theta)$ is therefore concave, and inequality (2.2) applies with objective function $f(\theta) = -L(\theta)$ and $B = \frac{1}{4}\sum_{i=1}^m x_i x_i^t$. Optimization transfer in this instance is similar to Newton's method for maximizing $L(\theta)$ except that the constant matrix B is substituted for $-d^2 L(\theta)$ at each iteration. The advantage of optimization transfer is that B need be inverted only once, rather than at each iteration.

Example 4: Multidimensional Scaling. Multidimensional scaling attempts to represent q objects as faithfully as possible in p-dimensional space given a weight $w_{ij} > 0$ and a dissimilarity measure y_{ij} for each pair of objects i and j. If $\theta_i \in R^p$ is the position of object i, then the $p \times q$ parameter matrix θ with ith column θ_i is estimated by minimizing the stress

$$\begin{aligned}
\sigma^2(\theta) &= \sum_{1 \leq i < j \leq q} w_{ij}(y_{ij} - \|\theta_i - \theta_j\|)^2 \\
&= \sum_{1 \leq i \leq j \leq q} w_{ij} y_{ij}^2 - 2 \sum_{1 \leq i < j \leq q} w_{ij} y_{ij} \|\theta_i - \theta_j\| + \sum_{1 \leq i < j \leq q} w_{ij} \|\theta_i - \theta_j\|^2,
\end{aligned}$$

where $\|\theta_i - \theta_j\|$ is the Euclidean distance between θ_i and θ_j. The stress function is invariant under translations, rotations, and reflections of R^p. To avoid translation and rotation ambiguities, we take θ_1 to be the origin $\mathbf{0}$ and the first $p-1$ coordinates of θ_2 to be 0. Convergence to one member of a pair of reflected minima immediately determines the other member.

The Cauchy–Schwarz inequality

$$-\|\theta_i - \theta_j\| \cdot \|\theta_i^n - \theta_j^n\| \leq -(\theta_i - \theta_j)^t(\theta_i^n - \theta_j^n)$$

allows us to effect an optimization transfer to the quadratic majorizing function

$$Q(\theta \mid \theta^n) = \sum_{1 \leq i < j \leq q} w_{ij} y_{ij}^2 - 2 \sum_{1 \leq i < j \leq q} \frac{w_{ij} y_{ij}}{\|\theta_i^n - \theta_j^n\|} (\theta_i - \theta_j)^t (\theta_i^n - \theta_j^n) + \sum_{1 \leq i < j \leq q} w_{ij} \|\theta_i - \theta_j\|^2 \quad (2.3)$$

and minimize $Q(\theta \mid \theta^n)$ instead of $\sigma^2(\theta)$ (de Leeuw and Heiser 1977; Groenen 1993).

Example 5: Asymmetric Least Squares. Efron (1991) proposed the method of asymmetric least squares for regression problems in which there is a reason to penalize positive residuals and negative residuals differently. Consider the function

$$\rho(r) = \begin{cases} r^2 & r \leq 0 \\ wr^2 & r > 0 \end{cases},$$

where w is a positive constant. Asymmetric least squares minimizes the quantity $\sum_{i=1}^m \rho\{y_i - \mu_i(\theta)\}$ for observations y_i and corresponding regression functions $\mu_i(\theta)$. Newton's method and the Gauss–Newton algorithm are natural candidates to use in this context. However, the Hessian of the objective function exhibits discontinuities. A way of circumventing this difficulty is to transfer optimization to a quadratic majorizing function. If we define $r_i(\theta) = y_i - \mu_i(\theta)$ and set

$$\zeta[r \mid r_i(\theta^n)] = \begin{cases} wr^2 - 2(w-1)r_i(\theta^n)r + (w-1)r_i(\theta^n)^2 & r_i(\theta^n) \leq 0 \\ wr^2 & r_i(\theta^n) > 0 \end{cases},$$

for $w > 1$ and

$$\zeta[r \mid r_i(\theta^n)] = \begin{cases} r^2 & r_i(\theta^n) \leq 0 \\ r^2 + 2(w-1)r_i(\theta^n)r - (w-1)r_i(\theta^n)^2 & r_i(\theta^n) > 0 \end{cases}$$

for $w < 1$, then the quadratic $\sum_{i=1}^m \zeta[r_i(\theta) \mid r_i(\theta^n)]$ majorizes the objective function.

A third method of constructing a majorizing function depends directly on the inequality $f(\sum_i \alpha_i v_i) \leq \sum_i \alpha_i f(v_i)$ defining a convex function $f(u)$. Here the coefficients α_i are nonnegative and sum to 1. It is helpful to extend this inequality to

$$f(c^t v) \leq \sum_i \frac{c_i w_i}{c^t w} f\Big(\frac{c^t w}{w_i} v_i\Big) \quad (2.4)$$

when all components c_i and w_i of the vectors c and w are positive. One of the virtues of applying inequality (2.4) in defining a surrogate function is that it separates parameters in the surrogate function. This feature is critically important in high-dimensional problems.

Example 6: Transmission Tomography. In transmission tomography, high energy photons are beamed from an external X-ray source and pass through the body to an

external detector. Statistical image reconstruction proceeds by dividing the plane region of an X-ray slice into small rectangular pixels and assigning a nonnegative attenuation coefficient θ_j to each pixel j. A photon sent from the source along projection i (line of flight) has probability $\exp(-l_i^t\theta)$ of avoiding absorption by the body, where l_i is the vector of intersection lengths l_{ij} of the ith projection with the jth pixel. If we assume that a Poisson number of photons with mean d_i depart along projection i, then a Poisson number y_i of photons with mean $d_i \exp(-l_i^t\theta)$ is detected. Because different projections behave independently, the log-likelihood reduces to

$$L(\theta) = \sum_i \left(-d_i e^{-l_i^t\theta} + y_i \ln d_i - y_i l_i^t\theta - \ln y_i!\right). \tag{2.5}$$

We now drop irrelevant constants and abbreviate the log-likelihood in (2.5) as $L(\theta) = -\sum_i f_i(l_i^t\theta)$ using the strictly convex functions $f_i(u) = d_i e^{-u} + y_i u$. Owing to the nonnegativity constraints $\theta_j \geq 0$ and $l_{ij} \geq 0$, inequality (2.4) yields

$$\begin{aligned} L(\theta) &= -\sum_i f_i\left(l_i^t\theta\right) \\ &\geq -\sum_i \sum_j \frac{l_{ij}\theta_j^n}{l_i^t\theta^n} f_i\left(\frac{l_i^t\theta^n}{\theta_j^n}\theta_j\right) \\ &= Q(\theta \mid \theta^n), \end{aligned}$$

with equality when $\theta_j = \theta_j^n$ for all j. By construction, maximization of $Q(\theta \mid \theta^n)$ separates into a sequence of one-dimensional problems, each of which can be solved approximately by one step of Newton's method (Lange 1995b).

In a different medical imaging context, De Pierro (1995) introduced a fourth method of optimization transfer. If $f(u)$ is convex, then he invoked the inequality

$$f(c^t v) \leq \sum_i \alpha_i f\left\{\frac{c_i}{\alpha_i}(v_i - w_i) + c^t w\right\}, \tag{2.6}$$

where $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, and $\alpha_i > 0$ whenever $c_i \neq 0$. In contrast to inequality (2.4), there are no positivity restrictions on the components c_i or w_i. However, we must somehow tailor the α_i to the problem at hand. Among the candidates for the α_i are $|c_i|^p/\|c\|_p^p$ with $\|c\|_p^p = \sum_i |c_i|^p$. When $p = 0$, we interpret α_i as 0 when $c_i = 0$ and as $1/m$ when c_i is one among m nonzero coefficients.

Example 7: Ordinary Linear Regression. Application of inequality (2.6) to the least squares criterion $\sum_{i=1}^m (y_i - x_i^t\theta)^2$ implies

$$\begin{aligned} \sum_{i=1}^m (y_i - x_i^t\theta)^2 &\leq \sum_{i=1}^m \sum_j \alpha_{ij}\left\{y_i - \frac{x_{ij}}{\alpha_{ij}}(\theta_j - \theta_j^n) - x_i^t\theta^n\right\}^2 \\ &= Q(\theta \mid \theta^n). \end{aligned}$$

Minimization of the surrogate function $Q(\theta \mid \theta^n)$ then yields the updates

$$\theta_j^{n+1} = \theta_j^n + \frac{\sum_{i=1}^m x_{ij}(y_i - x_i^t\theta^n)}{\sum_{i=1}^m \frac{x_{ij}^2}{\alpha_{ij}}},$$

which involve no matrix inversion (Becker, Yang, and Lange 1997). It seems intuitively reasonable to put $\alpha_{ij} = |x_{ij}|/(\sum_k |x_{ik}|)$ in this context.

Example 8: Poisson Regression. In a Poisson regression model with observation y_i for case i, it is convenient to write the mean $d_i e^{x_i^t \theta}$ as a function of a fixed offset $d_i > 0$ and a covariate vector x_i. Inequality (2.6) applies to the log-likelihood

$$L(\theta) = \sum_{i=1}^{m} \left(-d_i e^{x_i^t \theta} + y_i \ln d_i + y_i x_i^t \theta - \ln y_i! \right)$$

because the function $f_i(u) = -d_i e^u + y_i u$ is concave. In maximizing the corresponding surrogate function, one step of Newton's method yields the update

$$\theta_j^{n+1} = \theta_j^n + \frac{\sum_{i=1}^m x_{ij}(y_i - d_i e^{x_i^t \theta^n})}{\sum_{i=1}^m d_i e^{x_i^t \theta^n} x_{ij}^2 / \alpha_{ij}}.$$

Readers can consult Becker, Yang, and Lange (1997) for details and other examples of how De Pierro's method operates in generalized linear models. It is noteworthy that minorization by a quadratic function fails for Poisson regression because the functions $f_i(u)$ do not have bounded curvature.

Example 9: Separation of Parameters in Multidimensional Scaling Even after transferring optimization of the stress function to a quadratic majorizing function in Example 4 (p. 5), we face the difficulty of solving a large, nonsparse system of linear equations in minimizing the quadratic. This suggests that we attempt to separate parameters. In view of the convexity of the Euclidean norm $\|\cdot\|$ and the square function x^2, the offending part of the quadratic (2.3) can itself be majorized via the inequalities

$$\begin{aligned}
\|\theta_i - \theta_j\|^2 &= \left\| \frac{1}{2} 2(\theta_i - \theta_i^n) + \frac{1}{2} 2(-\theta_j + \theta_j^n) + \theta_i^n - \theta_j^n \right\|^2 \\
&\leq \left\{ \frac{1}{2} \|2(\theta_i - \theta_i^n) + \theta_i^n - \theta_j^n\| + \frac{1}{2} \|2(-\theta_j + \theta_j^n) + \theta_i^n - \theta_j^n\| \right\}^2 \\
&\leq \frac{1}{2} \|2(\theta_i - \theta_i^n) + \theta_i^n - \theta_j^n\|^2 + \frac{1}{2} \|2(-\theta_j + \theta_j^n) + \theta_i^n - \theta_j^n\|^2 \\
&= 2 \left\| \theta_i - \frac{1}{2}(\theta_i^n + \theta_j^n) \right\|^2 + 2 \left\| \theta_j - \frac{1}{2}(\theta_i^n + \theta_j^n) \right\|^2.
\end{aligned}$$

Once again equality occurs throughout if $\theta_i = \theta_i^n$ and $\theta_j = \theta_j^n$.

3. LOCAL AND GLOBAL CONVERGENCE

The local and global convergence properties of optimization transfer exactly parallel the corresponding properties of the EM and EM gradient algorithms. This is hardly surprising because the relevant theory relies entirely on optimization transfer and never mentions missing data. The current development follows Lange (1995b) closely.

To describe the local rate of convergence in the neighborhood of an optimal point θ^∞, we introduce the map $M(\theta)$ taking the current iterate θ^n into the next iterate $\theta^{n+1} = M(\theta^n)$. A first-order Taylor expansion around the point θ^∞ gives

$$\theta^{n+1} \approx \theta^\infty + dM(\theta^\infty)(\theta^n - \theta^\infty),$$

and correctly suggests that θ^n converges geometrically fast to θ^∞ with rate determined by the dominant eigenvalue of $dM(\theta^\infty)$. If it is impossible to maximize $Q(\theta \mid \theta^n)$ exactly, one can always iterate according to Equation (1.3). In this case, it is easy to see that the iteration map $M(\theta) = \theta - d^2Q(\theta \mid \theta)^{-1}dL(\theta)$ has differential $dM(\theta^\infty) = I - d^2Q(\theta^\infty \mid \theta^\infty)^{-1}d^2L(\theta^\infty)$ at θ^∞. Because Newton's method converges at a quadratic rate and optimization transfer at a linear (geometric) rate, both optimization transfer and its gradient version converge at the geometric rate determined by the dominant eigenvalue of $I - d^2Q(\theta^\infty \mid \theta^\infty)^{-1}d^2L(\theta^\infty)$.

Global convergence depends on several weak assumptions which are usually easy to check for a particular optimization transfer algorithm. In the case of maximization, we assume that the iteration map $M(\theta)$ is continuous and satisfies $L[M(\theta)] \geq L(\theta)$, with equality iff θ is a fixed point of $M(\theta)$. If we assume further that the set of fixed points of $M(\theta)$ coincides with the set of stationary points of $L(\theta)$, then $L(\theta)$ serves as a Lyapunov function for $M(\theta)$ (Luenberger 1984), and classical arguments imply that any limit point of the sequence $\theta^{n+1} = M(\theta^n)$ is a stationary point of $L(\theta)$. As a corollary, if $L(\theta)$ possesses a single stationary point—for example, if $L(\theta)$ is a strictly concave log-likelihood function—then optimization transfer is guaranteed to converge to it provided the iterates θ^n stay within a compact set. The hypotheses of this convergence theorem may be weakened slightly (McLachlan and Krishnan 1997), but this simple version suffices for our purposes.

We now turn to some interesting remarks of de Leeuw (1994) and Heiser (1995) regarding the construction of surrogate functions. Most objective functions $L(\theta)$ can be expressed as the difference

$$L(\theta) = f(\theta) - g(\theta) \tag{3.1}$$

of two concave functions. The class of functions permitting such nonunique decompositions is incredibly rich and furnishes the natural domain for optimization transfer. This class is closed under finite sums, products, maxima, and minima and includes all piecewise affine functions and twice continuously differentiable functions (Konno, Thach, and Tuy 1997). The point of the decomposition (3.1) is that we can transfer maximization of $L(\theta)$ to the concave function

$$Q(\theta \mid \theta^n) = f(\theta) - dg(\theta^n)(\theta - \theta^n)$$

because $-g(\theta) + dg(\theta^n)(\theta - \theta^n) \geq -g(\theta^n)$ holds for all θ, with equality at $\theta = \theta^n$. This transfer works even if $g(\theta)$ fails to be differentiable at θ^n provided we use an appropriately defined subdifferential.

Taking second differentials in Equation (3.1) gives the decomposition

$$d^2L(\theta) = N(\theta) + P(\theta) \tag{3.2}$$

of $d^2L(\theta)$ into a sum of a negative definite matrix $N(\theta) = d^2f(\theta)$ and a positive definite matrix $P(\theta) = -d^2g(\theta)$. The matrices $N(\theta)$ and $P(\theta)$ together determine the local convergence rate of optimization transfer through the dominant eigenvalue of

$$I - N(\theta^\infty)^{-1}\left[N(\theta^\infty) + P(\theta^\infty)\right] = -N(\theta^\infty)^{-1}P(\theta^\infty)$$

at the global maximum point θ^∞ of $L(\theta)$. Away from θ^∞, the decomposition (3.2) also provides the basis for acceleration of the algorithm. This brings up the intriguing question of whether we should highlight the decomposition (3.2) as having priority over optimization transfer. Indeed, the ascent algorithm

$$\theta^{n+1} = \theta^n - N(\theta^n)^{-1}dL(\theta^n)^t \tag{3.3}$$

is well defined regardless of whether $N(\theta^n)$ corresponds to the second differential $d^2Q(\theta^n \mid \theta^n)$ of a surrogate function $Q(\theta \mid \theta^n)$.

Example 1 (p. 4) illustrates our point. If we assume that there are just two teams, and team 1 always beats team 2, then

$$\begin{aligned} d^2L(\theta) &= -\frac{y_{12}}{\theta_1^2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{y_{12}}{(\theta_1+\theta_2)^2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \\ &= -y_{12}\begin{pmatrix} \theta_1^{-2} & 0 \\ 0 & (\theta_1+\theta_2)^{-2} \end{pmatrix} + \frac{y_{12}}{(\theta_1+\theta_2)^2}\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}. \end{aligned}$$

Both of these decompositions take the form (3.2), but only the first arises from the stated optimization transfer. In fact, no optimization transfer can account for the second decomposition. If, on the contrary, we suppose that

$$d^2f(\theta) = -y_{12}\begin{pmatrix} \theta_1^{-2} & 0 \\ 0 & (\theta_1+\theta_2)^{-2} \end{pmatrix},$$

then we immediately deduce

$$\frac{\partial^3}{\partial\theta_1\partial\theta_2\partial\theta_2}f(\theta) = 2y_{12}(\theta_1+\theta_2)^{-3} \neq 0 = \frac{\partial^3}{\partial\theta_2\partial\theta_1\partial\theta_2}f(\theta),$$

contradicting the required equality of mixed partial derivatives.

It is clear, however, that the first decomposition is preferable to the second. First, it is equally simple, and second, it leads to faster convergence when extended to the larger league. According to the theory in Lange (1995b), the local convergence rate λ of optimization transfer is determined by the maximum value of the function

$$1 - \frac{v^t d^2L(\theta^\infty)v}{v^t N(\theta^\infty)v}$$

for $v \neq 0$. Given that $d^2L(\theta^\infty)$ is negative definite, two different decompositions involving negative definite parts $N_1(\theta^\infty) \succ N_2(\theta^\infty)$ lead to convergence rates satisfying the reversed inequality $\lambda_1 \leq \lambda_2$. In the Bradley–Terry model, it is obvious that $N_1(\theta) \succ N_2(\theta)$ for all θ, and this ordering persists when we add more teams.

4. QUASI-NEWTON ACCELERATION

Optimization transfer typically performs well far from the optimum point. However, Newton's method enjoys a quadratic convergence rate in contrast to the linear convergence rate of optimization transfer. These considerations suggest that a hybrid algorithm that begins as pure optimization transfer and gradually makes the transition to Newton's method may hold the best promise of acceleration. We now describe one such algorithm based on quasi-Newton approximation (Jamshidian and Jennrich 1993, 1997; Lange 1995a).

If the symmetric matrix H_n approximates $-d^2L(\theta^n)^{-1}$ in maximum likelihood estimation with loglikelihood $L(\theta)$, then a quasi-Newton scheme employing H_n iterates according to $\theta^{n+1} = \theta^n + H_n dL(\theta^n)^t$. Updating H_n can be based on the inverse secant condition $-H_{n+1}g_n = s_n$, where $g_n = dL(\theta^n) - dL(\theta^{n+1})$ and $s_n = \theta^n - \theta^{n+1}$. The unique symmetric, rank-one update to H_n satisfying the inverse secant condition was furnished by Davidon's (1959) formula

$$H_{n+1} = H_n - c_n v_n v_n^t \tag{4.1}$$

with constant c_n and vector v_n specified by

$$\begin{aligned} c_n &= \frac{1}{(s_n + H_n g_n)^t g_n} \\ v_n &= s_n + H_n g_n. \end{aligned} \tag{4.2}$$

Although several alternative updates have been proposed since 1959, the Davidon update (4.1) has recently enjoyed a revival among numerical analysts (Conn, Gould, and Toint 1991; Khalfan, Byrd, and Schnabel 1993).

Approximating $-d^2L(\theta^n)^{-1}$ rather than $-d^2L(\theta^n)$ has the evident advantage of avoiding the matrix inversions of Newton's method. In fact, if one computes updates to the approximation of $-d^2L(\theta^n)^{-1}$ via the Sherman–Morrison formula (Press, Teukolsky, Vetterling, and Flannery 1992), then large matrix inversions can be avoided altogether.

Since optimization transfer already entails the approximation of $-d^2L(\theta^n)^{-1}$ by $-d^2Q(\theta^n \mid \theta^n)^{-1}$, it is more sensible to use a quasi-Newton scheme to approximate the difference

$$d^2Q(\theta^n \mid \theta^n)^{-1} - d^2L(\theta^n)^{-1}$$

by a symmetric matrix M_n and set

$$H_n = M_n - d^2Q(\theta^n \mid \theta^n)^{-1}$$

for an improved approximation to $-d^2L(\theta^n)^{-1}$. The inverse secant condition for M_{n+1} is

$$-M_{n+1}g_n = s_n - d^2Q(\theta^{n+1} \mid \theta^{n+1})^{-1}g_n. \tag{4.3}$$

Davidon's symmetric rank-one update (4.1) with s_n appropriately redefined in (4.2) can be used to construct M_{n+1} from M_n.

Given M_n, the next iterate in the quasi-Newton search can be expressed as

$$\theta^{n+1} = \theta^n + M_n dL(\theta^n)^t - d^2Q(\theta^n \mid \theta^n)^{-1} dL(\theta^n)^t. \tag{4.4}$$

When the exact optimization transfer increment $\Delta\theta^n$ is known, Equation (4.4) can be simplified by the substitution

$$-d^2Q(\theta^n \mid \theta^n)^{-1} dL(\theta^n)^t \approx \Delta\theta^n.$$

The availability of $\Delta\theta^n$ also simplifies the inverse secant condition (4.3). With the understanding that $d^2Q(\theta^n \mid \theta^n)^{-1} \approx d^2Q(\theta^{n+1} \mid \theta^{n+1})^{-1}$, condition (4.3) becomes

$$-M_{n+1} g_n = s_n + \Delta\theta^n - \Delta\theta^{n+1}. \tag{4.5}$$

Thus, quasi-Newton acceleration can be phrased entirely in terms of the score $dL(\theta^n)^t$ and the exact optimization transfer increments (Jamshidian and Jennrich 1997).

In implementing quasi-Newton acceleration, we must invert $d^2Q(\theta^n \mid \theta^n)$. Finding a surrogate function that separates parameters renders $d^2Q(\theta^n \mid \theta^n)$ diagonal and eases this part of the computational burden. We also need some initial approximation M_1. The choice $M_1 = \mathbf{0}$ works well because it guarantees that the first iterate of the accelerated algorithm is either optimization transfer or its gradient version. Finally, we must often deal with the problem of θ^{n+1} decreasing rather than increasing $L(\theta)$. When this occurs, one can reduce the contribution of $M_n dL(\theta^n)^t$ by step-halving until

$$\theta^{n+1} = \theta^n + \frac{1}{2^k} M_n dL(\theta^n)^t - d^2Q(\theta^n \mid \theta^n)^{-1} dL(\theta^n)^t \tag{4.6}$$

does lead to an increase in $L(\theta)$ (Lange 1995a). Alternatively, Jamshidian and Jennrich (1997) recommended conducting a limited line search along the direction implied by the update (4.4). If this search is unsuccessful, then they suggest resetting $M_n = \mathbf{0}$ and beginning the approximation process anew.

5. SCHULTZ–HOTELLING ACCELERATION

The quasi-Newton acceleration seeks to improve the approximation $-d^2Q(\theta^n \mid \theta^n)^{-1}$ to $-d^2L(\theta^n)^{-1}$. In many high-dimensional problems, the difficulty may be more inversion rather than evaluation of $-d^2L(\theta^n)$. If $d^2L(\theta^n)$ and $d^2Q(\theta^n \mid \theta^n)^{-1}$ are reasonably easy to compute, then we can use the Schultz and Hotelling correction (Householder 1975; Press et al. 1992)

$$C_n = 2B_n - B_n A_n B_n \tag{5.1}$$

to the approximate inverse B_n of a matrix A_n to concoct a second accelerated algorithm. Indeed, all we have to do is iterate according to

$$\theta^{n+1} = \theta^n + C_n dL(\theta^n)^t \tag{5.2}$$

based on inserting $A_n = -d^2L(\theta^n)$ and $B_n = -d^2Q(\theta^n \mid \theta^n)^{-1}$ in formula (5.1). If the Schultz–Hotelling acceleration (5.2) is correctly implemented, it entails only matrix times vector multiplication and not matrix times matrix multiplication.

The Schultz–Hotelling formula (5.1) is nothing more than one step of Newton's method for computing the inverse of a matrix. To prove that the Schultz–Hotelling acceleration (5.2) is indeed faster than optimization transfer, we note that B_n is positive definite and that $B_n^{-1} - A_n = d^2L(\theta^n) - d^2Q(\theta^n \mid \theta^n)$ is nonnegative definite because $L(\theta) - Q(\theta \mid \theta^n)$ attains its minimum at $\theta = \theta^n$. Assuming that $B_n^{-1} - A_n$ is actually positive definite, we have

$$\begin{aligned} C_n &= B_n + B_n(B_n^{-1} - A_n)B_n \\ &\succ B_n \end{aligned}$$

in the positive definite partial order $\succ$. From this inequality and standard properties of $\succ$, we deduce that $B_n^{-1} \succ C_n^{-1}$ and that $-C_n^{-1} \succ -B_n^{-1}$ (Horn and Johnson 1985). Because the matrices $-C_n^{-1}$ and $-B_n^{-1}$ correspond to choices of the negative definite matrix $N(\theta)$ in Equation (3.2), our remarks at the end of Section 3 now indicate that the local rate of convergence of the Schultz–Hotelling acceleration improves that of optimization transfer.

The Schultz–Hotelling correction (5.1) is the first of a hierarchy of corrections. If we put

$$\begin{aligned} H_{nk} &= B_n \sum_{j=0}^{k} (I - A_n B_n)^j \\ &= \sum_{j=0}^{k} (I - B_n A_n)^j B_n \\ &= B_n^{\frac{1}{2}} \sum_{j=0}^{k} \left\{ B_n^{\frac{1}{2}} (B_n^{-1} - A_n) B_n^{\frac{1}{2}} \right\}^j B_n^{\frac{1}{2}}, \end{aligned}$$

then we can show that the H_{nk} are better and better positive definite approximations to $-d^2L(\theta^n)^{-1}$ and that the accelerated algorithms

$$\theta^{n+1} = \theta^n + H_{nk} dL(\theta^n)^t \tag{5.3}$$

exhibit better and better local rates of convergence. These positive findings are offset by the increasing computational complexity as we ascend in the hierarchy.

6. NUMERICAL RESULTS

This section revisits three of the theoretical examples from Section 2 and compares the numerical performance of optimization transfer, both unmodified and accelerated, with Newton's method. Because Newton's method requires the inversion of a $p \times p$ matrix at each step of a p-dimensional problem, the relative performance of the competing algorithms improves as p grows in our numerical examples. We measure the performance of the various algorithms in floating point operations (flops) until convergence. All algorithms are implemented in MATLAB, which automatically counts flops.

Example 10: Bradley–Terry Model. For the 30 teams of the U.S. National Football League, the accelerated optimization transfer method of Equation (4.4) is faster

Table 1. Performance of Four Methods for Maximum Likelihood Estimation in the Bradley–Terry Model Applied to 1997 National Football League Data on 30 Teams

Method	*Iterations*	*Flops*
Newton	6	351,822
Optimization Transfer	1234	2,216,776
Quasi-Newton	30	297,396
Schultz–Hotelling	594	3,273,498

than Newton's method in fitting the Bradley–Terry model of Example 2 (p. 4). Table 1 summarizes the number of iterations and flop counts for the various methods on the win-loss results of the 1997 regular season games. The Schultz–Hotelling acceleration embodied in Equation (5.3) with $k = 1$ converges in fewer iterations than unaccelerated optimization transfer, but it requires more flops due to the extra work of computing $d^2L(\theta)$ and $d^2Q(\theta \mid \theta^n)$.

All computer runs started at $(1, \ldots, 1)^t$ with the first parameter fixed at 1. Convergence was declared whenever the L_2 norm of the current parameter increment fell below 10^{-8}. There are certainly other possible convergence criteria, such as the change in the log-likelihood function $L(\theta)$ or the L_2 norm of the score vector $dL(\theta)$. These criteria tend to be less stringent than the one we employ due to the flatness of the likelihood function in the neighborhood of the maximum. Our limited experience suggests that relative to Newton's method, optimization transfer and its variants suffer more from more stringent convergence criteria.

As the number of teams grows, quasi-Newton acceleration of optimization transfer improves relative to Newton's method. Figure 1 shows the results of tests using

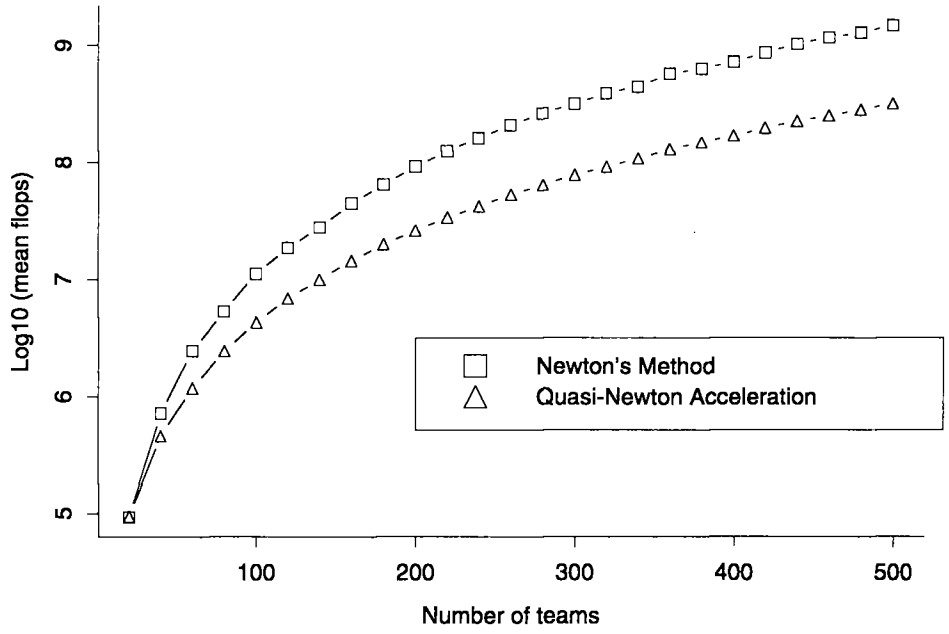

Figure 1. Newton's method compared with quasi-Newton accelerated optimization transfer for Bradley–Terry maximum likelihood. Points indicate $\log_{10}$ of the mean flops until convergence for ten runs.

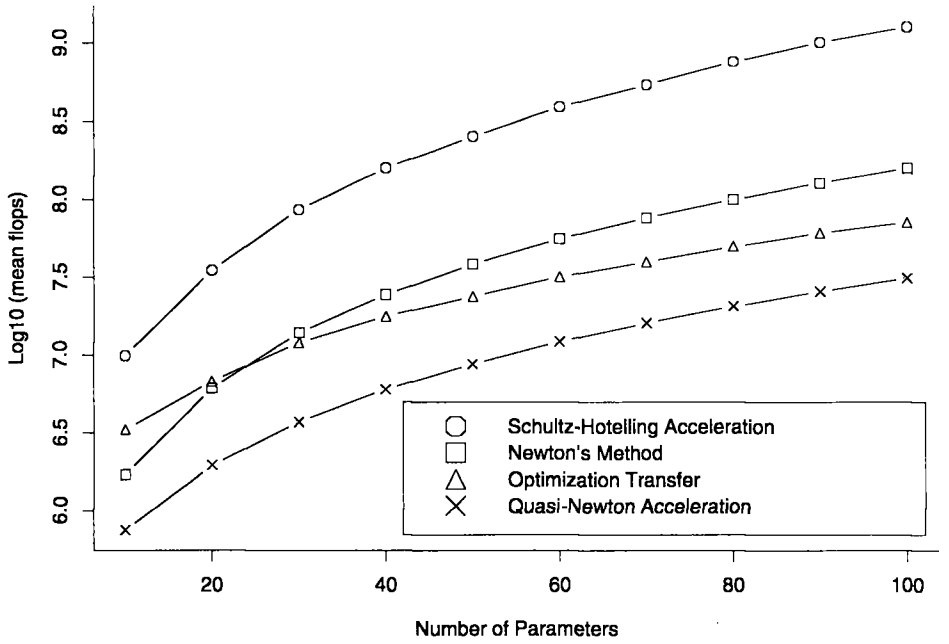

Figure 2. Mean flops until convergence for 100 independent logistic regression datasets of 1,000 simulated observations each.

simulated leagues of various sizes. The win-loss data were constructed by creating 10-team conferences. Each team played exactly two games with every other team in its conference and three games outside of its conference. The Bradley–Terry model determined the outcome of each game, with each team's rank parameter randomly sampled from $[1/2, 1]$. Figure 1 plots average flops until convergence for ten independent seasons at each league size. Newton's method converged in 4 to 7 iterations for each problem, whereas the quasi-Newton acceleration took anywhere from 11 iterations for 10 teams to 49 iterations for 120 teams. The quasi-Newton implementation here omits step-halving by using Equation (4.4) rather than Equation (4.6).

Example 11: Logistic Regression. Böhning and Lindsay (1988) reported that optimization transfer compares favorably with Newton's method in the logistic regression model of Example 3 (p. 5), particularly as the number of parameters increases. They tested both methods on simulated data with all true parameters equal to 0. In this case, the surrogate matrix $B = \frac{1}{4}\sum_{i=1}^{m} x_i x_i^t$ differs little from the observed information $-d^2L(\theta^n)$ for θ^n close to $\mathbf{0} = (0, \ldots, 0)^t$, so optimization transfer capitalizes strongly on its single matrix inversion.

To conduct a more realistic comparison, we generated logistic parameter values and covariates from normal $(0, 4)$ and $(0, 1/p)$ distributions, respectively, where p is the number of parameters. These choices imply that $x_i^t\theta$ has mean 0 and variance 4 for each case i. The results summarized in Figure 2 compare four algorithms starting at $\mathbf{0}$ and stopping according to the stringent convergence criterion of Example 10 (p. 13). The figure emphasizes the superiority of accelerated optimization transfer over Newton's method. Even

unadorned optimization transfer surpasses Newton's method on large enough problems. Once again, the Schultz–Hotelling acceleration of Equation (5.3) with $k = 1$ increases flops considerably despite reducing iterations.

For the runs summarized in Figure 2, Newton's method typically converged in about 7 iterations, regardless of the size of the problem. The iteration count of optimization transfer increases steadily from 70 to 116 as the number of parameters increases from 10 to 100. Schultz–Hotelling acceleration requires about half as many iterations and quasi-Newton acceleration about one-sixth as many iterations as optimization transfer.

Example 12: Multidimensional Scaling. We tested the optimization transfer algorithm of Example 9 (p. 8) on data obtained from a list of latitude and longitude locations for 329 United States cities (Boyer and Savageau 1989). Ignoring the earth's curvature, and taking all weights $w_{ij} = 1$, we treated latitude and longitude as planar coordinates and computed a Euclidean distance matrix (y_{ij}) for the 329 cities. This presumably yields a unique minimum of the two-dimensional scaling problem and facilitates assessment of convergence.

Submatrices of the large 329×329 distance matrix provide ready examples for comparing algorithms on problems of various sizes. As usual, Newton's method was one of the tested algorithms. In this problem, the optimization transfer algorithm of de Leeuw and Heiser (1977), briefly described in Example 4 (p. 5), serves as a substitute for the Schultz–Hotelling acceleration. Figure 3 summarizes the performance of the four algorithms on problems with varying numbers of parameters. The convergence criterion is the same as in Example 10 (p. 13).

The results in Figure 3 for a given number of cities represent averages over ten different runs for the same subset of cities. The runs differ only in their initial points, which were randomly chosen on $[0, 1]$. The parameters in these problems are not completely free to vary. As suggested in Example 4, the first three parameter values (both coordinates of the first city's location and the first coordinate of the second city's location) are held at zero. In the de Leeuw and Heiser method, the center of mass of the solution is held at the origin; this makes solutions unique only up to rotations and reflections about the origin.

Iteration counts for these problems are considerably higher than in our other examples. As the number of cities increases from 10 to 200, Newton's method requires from 39 to 124 iterations and optimization transfer from 2,000 to 24,000 iterations. The other two methods seem to converge in roughly constant numbers of iterations for most problems, about 200 for quasi-Newton and about 500 for de Leeuw–Heiser. Although we see in Figure 3 that both optimization transfer and the de Leeuw–Heiser method surpass Newton's method for large problems, the bottom line is that quasi-Newton accelerated optimization transfer is far superior to the other three methods.

7. DISCUSSION

In this article we have attempted to bring to the attention of the statistical public a potent principle for the construction of optimization algorithms. This optimization transfer

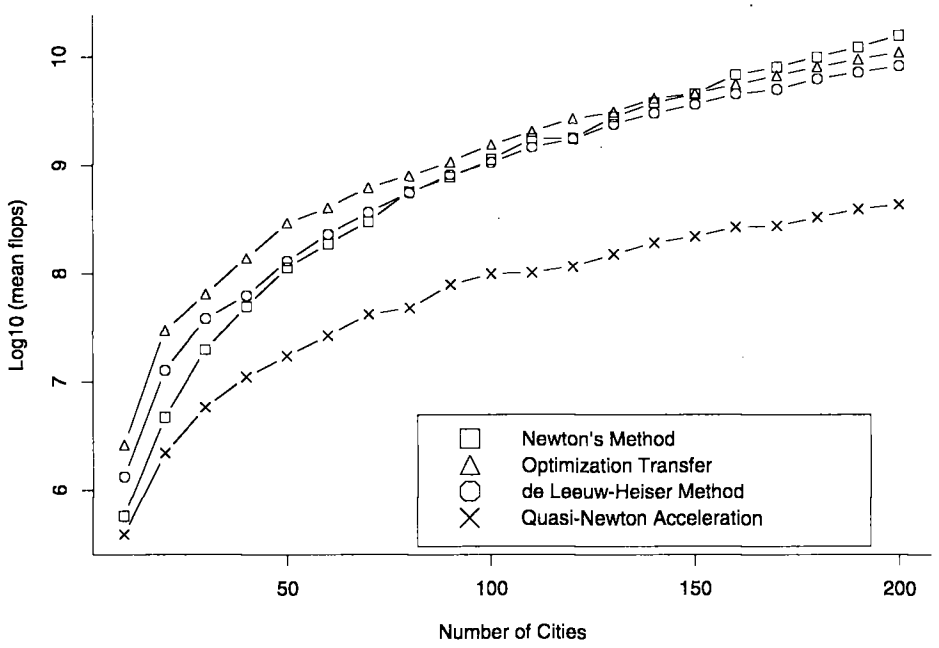

Figure 3. Mean number of flops for ten runs of various multidimensional scaling problems using four iterative algorithms. The number of parameters is twice the number of cities in each case.

principle includes the EM algorithm as a special case. Many specific EM algorithms can even be derived more easily by invoking optimization transfer rather than missing data. Example 2 (p. 4) on least absolute deviation regression is a case in point. Because of the limitations of space, we have omitted deriving other interesting optimization transfer algorithms. Among these algorithms are methods for convex programming (Lange 1994), multinomial logistic regression (Böhning 1992), quantile regression (Hunter and Lange 2000), and estimation in proportional hazards and proportional odds models (Böhning and Lindsay 1988; Hunter and Lange 2000).

We have featured four methods of exploiting convexity in the construction of optimization transfer algorithms. These methods hardly exhaust the possibilities. For instance, generalizations of the arithmetic-geometric mean inequality implicitly applied in Example 2 (p. 4) have proved their worth in geometric programming and should be born in mind (Peressini, Sullivan, and Uhl 1988). The well-studied method of majorization (not to be confused with majorizing functions as we have defined them) opens endless doors in devising inequalities (Marshall and Olkin 1979). Finally, the literature on differences of convex functions suggests useful devices for isolating a concave part of a log-likelihood (Konno, Thach, and Tuy 1997).

As the Bradley–Terry model makes evident, the $N + P$ decomposition (3.2) of the negative observed information can be achieved in more than one way. All such decompositions are not equal. They can be judged by how well the ascent algorithm (3.3) performs and how hard it is to code. In any case, the algorithm (3.3) can be accelerated in exactly the same manner as optimization transfer. It would be helpful to

identify a necessary and sufficient condition guaranteeing that $N(\theta^n)$ equals $d^2Q(\theta^n \mid \theta^n)$ for some surrogate function $Q(\theta \mid \theta^n)$.

Our limited experience suggests that Schultz–Hotelling acceleration leads to smaller gains than quasi-Newton acceleration. In high-dimensional problems, however, it is burdensome to carry along an approximate inverse of the observed information matrix. Schultz–Hotelling acceleration avoids this burden just as the method of conjugate gradients does. Until the Schultz–Hotelling acceleration is thoroughly tested on image reconstruction problems, we reserve final judgment about its effectiveness.

Other means of accelerating optimization transfer are certainly possible. For example, de Leeuw and Heiser (1980) reported that a simple step-doubling scheme (Heiser 1995; Lange 1995b) roughly halves the number of iterations required for convergence without appreciably increasing the computational complexity of each iteration.

We have ignored practical issues such as the existence of multiple modes on a likelihood surface, parameter equality constraints, parameter bounds, and the imposition of Bayesian priors. Our philosophy on these issues was expounded in the discussion of Lange (1995b) and need not be repeated here.

We close by challenging our fellow statisticians to develop their own applications of optimization transfer. This is no more a black art than devising EM algorithms, and the rewards, in our opinion, are equally great. If this article stimulates even a small fraction of the research activity generated by the Dempster, Laird, and Rubin (1977) article on the EM algorithm, we will be well satisfied.

ACKNOWLEDGMENTS

The first author thanks the U.S. Public Health Service for supporting his research through grant GM53275. We also thank Bruce Lindsay and Andreas Buja for their input on the title of the article.

[Received April 1999. Revised October 1999.]

REFERENCES

Becker, M. P., Yang, I., and Lange, K. (1997), "EM Algorithms Without Missing Data," *Statistical Methods in Medical Research*, 6, 38–54.

Böhning, D. (1992), "Multinomial Logistic Regression Algorithm," *Annals Institute of Statistical Mathematics*, 44, 197–200.

Böhning, D., and Lindsay, B. G. (1988), "Monotonicity of Quadratic Approximation Algorithms," *Annals Institute of Statistical Mathematics*, 40, 641–663.

Borg, I., and Groenen, P. (1997), *Modern Multidimensional Scaling*, New York: Springer-Verlag.

Boyer, R., and Savageau, D. (1989), *Places Rated Almanac*, New York: Prentice Hall.

Bradley, R. A., and Terry, M. E. (1952), "Rank Analysis of Incomplete Block Designs," *Biometrika*, 39, 324–345.

Conn, A. R., Gould, N. I. M., and Toint, P. L. (1991), "Convergence of Quasi-Newton Matrices Generated by the Symmetric Rank One Update," *Mathematical Programming*, 50, 177–195.

Davidon, W. C. (1959), "Variable Metric Methods for Minimization," *AEC Research and Development Report ANL–5990*, Argonne National Laboratory.

de Leeuw, J. (1994), "Block Relaxation Algorithms in Statistics," in *Information Systems and Data Analysis*, ed. H. H. Bock, W. Lenski, and M. M. Richter, Berlin: Springer-Verlag, pp. 308–325.

de Leeuw, J., and Heiser, W. J. (1977), "Convergence of Correction Matrix Algorithms for Multidimensional Scaling," in *Geometric Representations of Relational Data*, eds. J. C. Lingoes, E. Roskam, and I. Borg, Ann Arbor: Mathesis Press, pp. 735–752.

——— (1980), "Multidimensional Scaling With Restrictions on the Configuration," in *Multivariate Analysis, Vol. V*, ed. P. R. Krishnaiah, Amsterdam: North-Holland, pp. 501–522.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

De Pierro, A. R. (1995), "A Modified Expectation Maximization Algorithm for Penalized Likelihood Estimation in Emission Tomography," *IEEE Transactions on Medical Imaging*, 14, 132–137.

Efron, B. (1991), "Regression Percentiles Using Asymmetric Squared Error Loss," *Statistica Sinica*, 1, 93–125.

Groenen, P. J. F. (1993), *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*, Leiden, the Netherlands: DSWO Press.

Heiser, W. J. (1995), "Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis," in *Recent Advances in Descriptive Multivariate Analysis*, ed. W. J. Krzanowski, Oxford: Clarendon Press, pp. 157–189.

Horn, R. A., and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge: Cambridge University Press.

Householder, A. S. (1975), *The Theory of Matrices in Numerical Analysis*, New York: Dover.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Hunter, D. R., and Lange, K. (1999), "Computing Estimates in the Proportional Odds Model," unpublished manuscript.

——— (2000), "Quantile Regression van an MM Algorithm," *Journal of Compuational and Graphical Statistics*, 9, 60–77.

Jamshidian, M., and Jennrich, R. I. (1993), "Conjugate Gradient Acceleration of the EM Algorithm," *Journal of the American Statistical Association*, 88, 221–228.

——— (1997), "Quasi-Newton Acceleration of the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 59, 569–587.

Keener, J. P. (1993), "The Perron-Frobenius Theorem and the Ranking of Football Teams," *SIAM Review*, 35, 80–93.

Khalfan, H. F., Byrd, R. H., and Schnabel, R. B. (1993), "A Theoretical and Experimental Study of the Symmetric Rank-One Update," *SIAM Journal on Optimization*, 3, 1–24.

Konno, H., Thach, P. T., and Tuy, H. (1997), *Optimization on Low Rank Nonconvex Structures*, Dordecht, the Netherlands: Kluwer Academic Publishers.

Lange, K. (1994), "An Adaptive Barrier Method for Convex Programming," *Methods Applications Analysis*, 1, 392–402.

——— (1995a), "A Quasi-Newton Acceleration of the EM Algorithm," *Statistica Sinica*, 5, 1–18.

——— (1995b), "A Gradient Algorithm Locally Equivalent to the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 57, 425–437.

Lange, K., and Fessler, J. A. (1995c), "Globally Convergent Algorithms for Maximum A Posteriori Transmission Tomography," *IEEE Transactions on Image Processing*, 4, 1430–1438.

Lange, K., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the t Distribution," *Journal of the American Statistical Association*, 84, 881–896.

Lange, K., and Sinsheimer, J. (1993), "Normal/Independent Distributions and Their Applications in Robust Regression," *Journal of Computational and Statistical Graphics*, 2, 175–198.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Luenberger, D. G. (1984), *Linear and Nonlinear Programming* (2nd ed.), Reading, MA: Addison-Wesley.

Marshall, A. W., and Olkin, I. (1979), *Inequalities: Theory of Majorization and its Applications*, San Diego: Academic Press.

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA: Addison-Wesley.

Ortega, J. M., and Rheinboldt, W. C. (1970), *Iterative Solutions of Nonlinear Equations in Several Variables*, New York: Academic Press.

Peressini, A. L., Sullivan, F. E., and Uhl, J. J. Jr. (1988), *The Mathematics of Nonlinear Programming*, New York: Springer.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in Fortran: The Art of Scientific Computing* (2nd ed.), Cambridge: Cambridge University Press.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Schlossmacher, E. J. (1973), "An Iterative Technique for Absolute Deviations Curve Fitting," *Journal of the American Statistical Association*, 68, 857–859.