An Adaptive Barrier Method for Convex Programming

Kenneth Lange

Dedicated to the memory of my brother Charles.

ABSTRACT. This paper presents a new barrier method for convex programming. The method involves an optimization transfer principle. Instead of minimizing the objective function $f(x)$ directly, one minimizes the amended function $f(x) - \mu \sum_i x_i^n \ln x_i$ to produce the next iterate x^{n+1} from the current iterate x^n. If the feasible region is contained in the unit simplex, then this strategy forces a decrease in $f(x)$. The barrier parameter μ is kept constant during the process and not sent gradually to 0 as in the classical barrier method. Under mild assumptions on $f(x)$ and the linear constraints, the method converges to the global minimum of $f(x)$. If this minimum occurs in the interior of the feasible region, then the rate of convergence is linear.

1. Introduction

The current revival of interior point methods in convex programming has stimulated a healthy interest in perfecting the classical logarithmic barrier method [3, 5, 18]. This paper discusses an adaptive version of the logarithmic barrier method motivated by the EM algorithm from computational statistics [2, 11]. At the heart of the EM algorithm is a maximization transfer principle. Instead of maximizing the objective function directly, the EM algorithm maximizes at each iteration a related, but simpler function. In the process, the objective function is increased. In convex programming the difficulty is not so much that the objective function is complicated, but rather that optimization is hindered by nonnegativity constraints. Introduction of adaptive, logarithmic barriers effectively removes the nonnegativity constraints at each iteration while permitting parameters to converge to zero over a sequence of iterations.

The pioneering work of Karmarkar [9] contains the germ of the optimization transfer principle. However, in Karmarkar's projective scaling algorithm, the operation of the principle is obscured by repeated re-centering. In the algorithm to be presented here, explicit re-centering is de-emphasized, and the optimization transfer principle emerges more clearly.

The standard convex programming problem is to minimize a convex function $f(x)$ subject to the constraints $Ax = b$ and $x \geq \mathbf{0}$. Here x is a column vector with s components, A is an $r \times s$ matrix of full rank, and $\mathbf{0}$ is a vector of zeros. In the

logarithmic barrier method, the modified function

$$f(x) - \mu^n \sum_{i=1}^{s} \ln x_i \qquad (1)$$

is minimized for a sequence of constants μ^n tending to 0. The new method keeps the barrier parameter μ constant, but changes the barrier function (1) slightly. Suppose that one of the constraints $\sum_{j=1}^{s} a_{ij} x_j = b_i$ has all coefficients a_{ij} positive. If this is the case, then a trivial reparameterization makes it possible to replace this constraint by $\sum_{i=1}^{s} x_i = 1$. If no constraint has all coefficients positive, then Karmarkar [9] has noted for a linear objective function how to add a bounding constraint and a slack variable so that $\sum_{i=1}^{s} x_i = 1$ is part of the constraint structure.

Assuming that the constraint $\sum_{i=1}^{s} x_i = 1$ is present, consider the function

$$f(x) - \mu \sum_{i=1}^{s} x_i^n \ln x_i \qquad (2)$$

defined using the current iterate $x^n > \mathbf{0}$. Our new method operates by taking the next iterate x^{n+1} to be the minimum of (2) subject to $Ax = b$. Because $f(x)$ is convex and $-\mu \sum_{i=1}^{s} x_i^n \ln x_i$ is strictly convex, the minimum is uniquely defined and satisfies $x^{n+1} > \mathbf{0}$. Equally important is the fact that $f(x^{n+1}) \leq f(x^n)$, with strict inequality unless $x^{n+1} = x^n$. This assertion can be proved by observing that $\mu \sum_{i=1}^{s} x_i^n \ln x_i$ has its maximum subject to $\sum_{i=1}^{s} x_i = 1$ and $x > \mathbf{0}$ at $x = x^n$. Hence,

$$\begin{aligned}
f(x^{n+1}) &= f(x^{n+1}) - \mu \sum_{i=1}^{s} x_i^n \ln x_i^{n+1} + \mu \sum_{i=1}^{s} x_i^n \ln x_i^{n+1} \\
&\leq f(x^n) - \mu \sum_{i=1}^{s} x_i^n \ln x_i^n + \mu \sum_{i=1}^{s} x_i^n \ln x_i^{n+1} \\
&\leq f(x^n) - \mu \sum_{i=1}^{s} x_i^n \ln x_i^n + \mu \sum_{i=1}^{s} x_i^n \ln x_i^n \\
&= f(x^n).
\end{aligned}$$

Thus, minimization of (2) translates into a decrease of $f(x)$. The algorithm continues until the generated sequence x^n converges.

Several features of the new algorithm are noteworthy. First, the monotonicity property $f(x^{n+1}) \leq f(x^n)$ promotes numerical stability. Second, in common with ordinary barrier and penalty methods, the algorithm does not prescribe how to find the minimum of (2). In this sense, perhaps the term method should be substituted for algorithm. However, the EM algorithm suffers from the same ambiguity, and we will use the terms algorithm and method interchangeably. We will briefly discuss a one-step Newton's method for numerically computing the next iterate. Third, the algorithm is adaptive. Those components x_i^n that seek to converge to a boundary $x_i = 0$ are allowed to do so because of the presence of the multiplier x_i^n in the barrier term $-\mu x_i^n \ln x_i$. Fourth, the barrier constant μ is truly constant.

The remainder of this paper illustrates the application of the algorithm in one theoretical example and one numerical example. After these examples, convergence properties of the algorithm are investigated. Under fairly mild conditions on $f(x)$ and the linear constraints $Ax = b$, the algorithm can be shown to converge to the

minimum of $f(x)$ over the feasible region. If the minimum occurs at an interior point, the rate of convergence is linear, with faster convergence for smaller μ. The rate of convergence to a boundary point is left unresolved. Also left unresolved are issues of computational complexity. For general convex functions, complexity analysis is apt to be difficult. A reasonable place to start might be the linear or quadratic case [6, 20, 14].

Of course, it remains to be seen whether the algorithm is merely a mathematical curiosity or whether it offers an effective, practical method of optimization in the presence of nonnegativity constraints. It may be that the algorithm must be modified to form a successful stand-alone strategy or part of a hybrid strategy for optimization. For instance, the barrier constant μ in (2) might be gradually reduced to 0 as it is in the ordinary barrier method. One can also contemplate using the algorithm when $f(x)$ fails to be convex. These considerations all suggest that the algorithm warrants a more thorough examination than the modest one undertaken here.

Finally, a reviewer of this paper has kindly pointed out recent related work of Censor and Zenios [1] on proximal point algorithms. Censor and Zenios also provide a convergence analysis for the optimization transfer principle in convex programming. However, their arguments are quite different from those presented here. Our development has the added advantage of lending more insight into the case of nonconvex programming. This issue is addressed in our concluding discussion.

2. Examples and numerical implementation

2.1. A theoretical example. The loglikelihood for a multinomial distribution takes the form $\sum_{i=1}^{s} m_i \ln x_i$, where m_i is the observed number of counts in category i and x_i is the probability attached to category i. Maximizing $\sum_{i=1}^{s} m_i \ln x_i$ subject to the constraints $x_i \geq 0$ and $\sum_{i=1}^{s} x_i = 1$ gives the explicit maximum likelihood estimates $x_i = m_i/m$. Here m denotes the total number of counts $\sum_{i=1}^{s} m_i$. To compute the maximum likelihood estimates iteratively, define the Lagrangian

$$-\sum_{i=1}^{s} m_i \ln x_i - \mu \sum_{i=1}^{s} x_i^n \ln x_i + \lambda (\sum_{i=1}^{s} x_i - 1).$$

Setting the ith partial derivative of the Lagrangian equal to 0 gives

$$-\frac{m_i}{x_i} - \frac{\mu x_i^n}{x_i} + \lambda = 0.$$

Multiplying this equation by x_i produces

$$-m_i - \mu x_i^n + \lambda x_i = 0, \tag{3}$$

which in turn can be summed on i and solved for λ. Substituting the result $\lambda = m + \mu$ in (3) yields $x_i^{n+1} = (m_i + \mu x_i^n)/(m + \mu)$. (This argument for $m = 0$, incidentally, shows that $\mu \sum_{i=1}^{s} x_i^n \ln x_i$ is maximized by $x = x^n$ subject to $\sum_{i=1}^{s} x_i = 1$.) At first glance, it is not obvious that x^n tends to m_i/m, but the algebraic reduction

$$x_i^{n+1} - \frac{m_i}{m} = \frac{m_i + \mu x_i^n}{m + \mu} - \frac{m_i}{m}$$

$$= \frac{\mu}{m + \mu}(x_i^n - \frac{m_i}{m})$$

shows that x_i^n approaches m_i/m at the linear rate $\mu/(m + \mu)$.

Notation. In deriving and displaying various formulas, a compact notation is helpful. Let $R(y \mid x) = f(y) - \mu \sum_{i=1}^{s} x_i \ln y_i$. Because $R(y \mid x)$ has two arguments x and y, there are various first and second differentials depending on which arguments are differentiated. Define

$$d^{10} R(y \mid x) = df(y) - \mu \left(\frac{x_1}{y_1}, \ldots, \frac{x_s}{y_s} \right),$$

$$d^{20} R(y \mid x) = d^2 f(y) + \mu \begin{pmatrix} \frac{x_1}{y_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{x_s}{y_s^2} \end{pmatrix}, \tag{4}$$

$$d^{11} R(y \mid x) = -\mu \begin{pmatrix} \frac{1}{y_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{y_s} \end{pmatrix}. \tag{5}$$

These formulas simplify in an obvious manner when $y = x$.

All vectors except first differentials like $df(y)$ and $d^{10} R(y \mid x)$ are taken to be column vectors. A superscript t indicates vector or matrix transpose. Let I denote an identity matrix and $\mathbf{1}$ and $\mathbf{0}$ denote vectors or matrices of all 1's or 0's. To convert a vector z with entries z_i to a diagonal matrix with diagonal entries z_i, write z as a capital letter Z.

2.2. An approximate version of the algorithm.

For those problems lacking an exact solution x^{n+1}, one step of Newton's method offers a natural approximation [6, 7]. In practice, a crude approximation of x^{n+1} may be acceptable. As long as the surrogate function $R(x \mid x^n)$ is decreased, then the objective function $f(x)$ is decreased as well.

Newton's method can be implemented by minimizing the second-order Taylor's approximation

$$R(x^n + \delta^n \mid x^n) - R(x^n \mid x^n) \approx d^{10} R(x^n \mid x^n) \delta^n$$
$$+ \frac{1}{2} (\delta^n)^t d^{20} R(x^n \mid x^n) \delta^n \tag{6}$$

subject to the constraints $A\delta^n = \mathbf{0}$. The next iterate is then defined by $x^{n+1} = x^n + \delta^n$. Because $d^{10} R(x^n \mid x^n) = df(x^n) - \mu \mathbf{1}^t$ and $\mathbf{1}^t \delta = 0$ is implied by the constraint $\sum_{i=1}^{s} x_i = 1$, $df(x^n)$ can be substituted for $d^{10} R(x^n \mid x^n)$ in (6). The minimum point of (6) then satisfies

$$df(x^n)^t + d^{20} R(x^n \mid x^n) \delta^n + A^t \lambda^n = 0 \tag{7}$$

for some vector λ^n of Lagrange multipliers. Solving (7) for δ^n and then applying $A\delta^n = \mathbf{0}$ yields

$$\delta^n = -d^{20} R(x^n \mid x^n)^{-1} [df(x^n)^t + A^t \lambda^n],$$
$$\lambda^n = -[A \, d^{20} R(x^n \mid x^n)^{-1} A^t]^{-1} A \, d^{20} R(x^n \mid x^n)^{-1} df(x^n)^t.$$

Combining these two results gives

$$\delta^n = -G^n (I - A^t [AG^n A^t]^{-1} AG^n) df(x^n)^t, \tag{8}$$

where $G^n = d^{20} R(x^n \mid x^n)^{-1}$.

The matrix $d^{20}R(x^n \mid x^n)^{-1}$ appearing in (8) is ill-conditioned when any x_i^n is close to 0. The remedy is to re-express

$$d^{20}R(x^n \mid x^n) = (X^n)^{-\frac{1}{2}}\left[(X^n)^{\frac{1}{2}}\,d^2 f(x^n)(X^n)^{\frac{1}{2}} + \mu I\right](X^n)^{-\frac{1}{2}}$$

and

$$d^{20}R(x^n \mid x^n)^{-1} = (X^n)^{\frac{1}{2}}\left[(X^n)^{\frac{1}{2}}\,d^2 f(x^n)(X^n)^{\frac{1}{2}} + \mu I\right]^{-1}(X^n)^{\frac{1}{2}}.$$

To avoid violating the boundary constraint $x^{n+1} > \mathbf{0}$, one can choose a small positive constant ϵ and replace $x^{n+1} = x^n + \delta^n$ by $x^{n+1} = x^n + \alpha\delta^n$, where α is the largest number in $[0,1]$ consistent with $x^{n+1} \geq \epsilon x^n$. If some x_i^n is near 0 and the corresponding δ_i^n is negative and relatively large in magnitude compared to x_i^n, then this test can slow convergence. An ad hoc repair is to set $\delta_i^n = 0$ when δ_i^n is negative but very small in magnitude. This tactic permits more progress in those components at some distance from the boundary and escape from the boundary by a small component x_i when warranted.

2.3. A numerical example.

The linear programming problem of Klee and Minty [10] demonstrates the exponential complexity of the simplex method. This problem amounts to maximizing the mth component of an m-dimensional vector x subject to the inequality constraints $0 \leq x_1 \leq 1$ and $\beta x_{i-1} \leq x_i \leq 1 - \beta x_{i-1}$ for i in $\{2,\dots,m\}$. Here β is a constant satisfying $0 < \beta < 1/2$. The maximum occurs at $x = (0,\dots,0,1)^t$. The Klee-Minty example can be rephrased in standard form by defining $w_1 = x_1$ and $w_i = (x_i - \beta x_{i-1})/\beta^{i-1}$ for $i = 2,\dots,m$ [4, pp. 95, 110]. Setting $\theta = 1/\beta$ and introducing slack variables w_i for $i = m+1,\dots,2m$, the transformed problem consists of minimizing $f(w) = -\sum_{i=1}^m w_i$ subject to $w \geq \mathbf{0}$ and the linear constraints $2\sum_{j=1}^{i-1} w_j + w_i + w_{m+i} = \theta^{i-1}$ for $i = 1,\dots,m$. Adding these m constraints gives a single constraint,

$$\sum_{i=1}^{2m} c_i w_i = \frac{1 - \theta^m}{1 - \theta},$$

where

$$c_i = \begin{cases} 2(m-i)+1 & 1 \leq i \leq m, \\ 1 & m+1 \leq i \leq 2m, \end{cases}$$

with all coefficients positive. Since expanding A to include this constraint creates a new constraint matrix of less than full rank, it is better to minimize $f(w) - \mu\sum_{i=1}^{2m} c_i w_i^n \ln w_i$ subject to the existing constraints rather than the objective function (2). The optimal point w has $w_i = 0$ for $i = 1,\dots,m-1$ and $w_m = \theta^{m-1}$.

Table 1 records the performance of the approximate convex programming algorithm with increment δ^n given by (8) for $m = 8$ and $\beta = 1/4$ in the standard form Klee-Minty problem. The algorithm started at the point $x^1 = (.001,\dots,.001)$ in the original variables. The barrier constant was $\mu = .1$, and the boundary constant was $\epsilon = .01$. Any component w_i^n with proposed increment $\delta_i^n \geq -10^{-6}$ was ignored in calculating the maximum $\alpha \in [0,1]$ consistent with $w^n + \alpha\delta^n \geq \epsilon w^n$. With α computed in this manner, w_i^{n+1} was set to the proposed value $w_i^n + \alpha\delta_i^n$ whenever the proposed value was positive; otherwise, w_i^{n+1} was set to w_i^n.

It is evident from Table 1 that the approximate algorithm converges reasonably fast. When the barrier constant $\mu = 1$, the algorithm converges to the same final value in 51 iterations, and when $\mu = .01$, it converges in 19 iterations. The affine

Iteration	Value of $f(w)$	Iteration	Value of $f(w)$
1	-16.38	10	-16357.26
2	-151.24	11	-16381.00
3	-1538.61	12	-16382.71
4	-15074.90	13	-16383.39
5	-16208.39	14	-16383.69
6	-16269.88	15	-16383.94
7	-16305.08	16	-16383.98
8	-16329.23	17	-16383.99
9	-16345.83	18	-16384.00

TABLE 1. Table 1: Iterations on the Klee-Minty Example

scaling algorithm experiences similar difficulties to the simplex method in solving this problem [13].

3. Convergence of the algorithm

Proof that the algorithm is globally convergent requires precise assumptions. Suppose that $f(x)$ is convex and continuously differentiable on the compact feasible region $\{x : Ax = b, x \geq \mathbf{0}\}$. If the algorithm is to make sense, then the interior $\{x : Ax = b, x > \mathbf{0}\}$ of the feasible region must be nonempty. The matrix A is assumed to be of full rank and to have $\mathbf{1}^t$ as its first row; correspondingly, b has first component $b_1 = 1$.

For any subset $S \subset \{1, \ldots, s\}$, let S^c be the complement of S. If the manifold $M_S = \{x : Ax = b; x_i > 0, i \in S; x_i = 0, i \in S^c\}$ is nonempty, then assume that it contains at most a finite number of stationary or critical points of $f(x)$. Stationary points can be characterized in terms of the projection matrix P_S mapping any row vector v^t onto its entries v_i for $i \in S$. If S has $|S|$ elements, then P_S is an $s \times |S|$ matrix. Now $y \in M_S$ is stationary if and only if

$$df(y)P_S + \nu^t A P_S = 0 \tag{9}$$

for some Lagrange multiplier ν. To solve for ν in (9), it is convenient to assume that AP_S always has full row rank r.

On any manifold M_S, each stationary point coincides with a minimum point of $f(x)$ restricted to M_S. If M_S contains two minimum points, then the convexity of $f(x)$ implies that the line segment between them consists entirely of minimum points. Thus, the number of stationary points on M_S is either 0, 1, or ∞. The third possibility can be ruled out when $f(x)$ is strictly convex. It can also be ruled out for $f(x)$ linear when the linear programming problem is both primal and dual nondegenerate [4, pp. 21, 197]. In this case every stationary point of M_S is an extreme point of the feasible region, and M_S contains an extreme point only when M_S reduces to that extreme point.

Mindful of these definitions, we commence our verification of the global convergence of a sequence x^n generated by the algorithm.

Lemma 1. *The sequence x^n satisfies $\lim_{n \to \infty} \|x^{n+1} - x^n\| = 0$. Furthermore, if a positive limit $\lim_{k \to \infty} x_i^{n_k} > 0$ exists for the ith component of a subsequence x^{n_k}, then $\lim_{k \to \infty} x_i^{n_k}/x_i^{n_k+1} = 1$ as well.*

Proof. The inequalities

$$\mu \sum_{i=1}^{s} x_i^n \ln x_i^{n+1} \leq \mu \sum_{i=1}^{s} x_i^n \ln x_i^n$$

$$R(x^{n+1} \mid x^n) \leq R(x^n \mid x^n)$$

can be rearranged to yield

$$0 \leq \mu \sum_{i=1}^{s} x_i^n \ln(x_i^n / x_i^{n+1})$$

$$\leq f(x^n) - f(x^{n+1}).$$

Since $f(x^n)$ is decreasing and bounded below, this second pair of inequalities implies that $\lim_{n \to \infty} \sum_{i=1}^{s} x_i^n \ln(x_i^n / x_i^{n+1}) = 0$.

Now suppose the first claim of the lemma is false. Then for some subsequence x^{n_k}, $\liminf_{k \to \infty} \|x^{n_k+1} - x^{n_k}\| > 0$. Invoking compactness and passing to a subsequence if necessary, we can further assume that $\lim_{k \to \infty} x^{n_k} = y$ and $\lim_{k \to \infty} x^{n_k+1} = z$ exist with $y \neq z$. Applying the known information inequality [19, p. 58]

$$\sum_{i=1}^{s} x_i^n \ln \frac{x_i^n}{x_i^{n+1}} \geq \frac{1}{2} \sum_{i=1}^{s} x_i^n (x_i^n - x_i^{n+1})^2$$

to the subsequence x^{n_k} and using the fact that $\lim_{n \to \infty} \sum_{i=1}^{s} x_i^n \ln(x_i^n / x_i^{n+1}) = 0$, we can conclude that

$$\sum_{i=1}^{s} y_i (y_i - z_i)^2 = 0.$$

This last equality can only be true if $z_i = y_i$ for all i with $y_i > 0$. Since $\sum_{i=1}^{s} y_i = \sum_{i=1}^{s} z_i = 1$, in fact, all $z_i = y_i$. This contradiction proves the first claim of the lemma. The second assertion of the lemma follows from the first assertion and the fact that the quantity

$$x_i^n (x_i^n - x_i^{n+1})^2 = x_i^n (x_i^{n+1})^2 \left(\frac{x_i^n}{x_i^{n+1}} - 1 \right)^2$$

is being driven to 0.

Lemma 2. *Suppose y is a limit point of the sequence x^n. If S is the set $\{i : y_i > 0\}$, then y coincides with the unique stationary point of the manifold M_S.*

Proof. Let $\lim_{k \to \infty} x^{n_k} = y$ for some subsequence x^{n_k}. Because x^{n_k+1} minimizes $R(y \mid x^n)$ subject to the constraints $Ax = b$, there exists a Lagrange multiplier vector λ^{n_k} such that

$$df(x^{n_k+1}) - \mu(\omega^{n_k})^t + (\lambda^{n_k})^t A = \mathbf{0}^t, \tag{10}$$

where ω^{n_k} is the column vector with ith entry $x_i^{n_k} / x_i^{n_k+1}$. There are a variety of ways of solving for the Lagrange multiplier λ^{n_k} in (10). The most expeditious in the current context is to employ the projection P_S introduced in equation (9). By assumption the matrix AP_S is of full rank. Hence, multiplying (10) on the right by $P_S P_S^t A^t$ and solving yields

$$(\lambda^{n_k})^t = -\left[df(x^{n_k+1}) P_S - \mu(\omega^{n_k})^t P_S \right] P_S^t A^t (A P_S P_S^t A^t)^{-1}.$$

Because P_S annihilates those entries $\omega_i^{n_k}$ with i outside S, Lemma 1 and the continuity of $df(x)$ jointly imply the existence of the limit $\lambda^\infty = \lim_{k\to\infty} \lambda^{n_k}$. This fact now makes it possible to multiply (10) by P_S on the right, take limits, and conclude that

$$df(y)P_S - \mu\mathbf{1}^t P_S + (\lambda^\infty)^t A P_S = \mathbf{0}. \tag{11}$$

Since the first row of A coincides with $\mathbf{1}^t$, equation (11) can be rewritten as

$$df(y)P_S + (\lambda^\infty - \mu e_1)^t A P_S = \mathbf{0}, \tag{12}$$

where e_1 is the vector having first entry 1 and remaining entries 0. But equation (12) is precisely condition (9) characterizing the stationary point of M_S.

Theorem 3. *The sequence x^n converges to the global minimum of $f(x)$ subject to the constraints $Ax = b$ and $x \geq 0$.*

Proof. Since there are only a finite number of manifolds M_S and at most a finite number of stationary points per manifold, Lemma 2 indicates that the set of limit points Γ of x^n is finite. Γ is nonempty because the feasible set is compact. According to a result of Ostrowski [16, p. 173], Γ is also connected because the sequence x^n satisfies $\lim_{n\to\infty} \|x^{n+1} - x^n\| = 0$. The only way a finite set can be connected is for it to reduce to a single point. Thus, $\lim_{n\to\infty} x^n = x^\infty$ exists.

It remains to show that x^∞ satisfies the Karush-Kuhn-Tucker conditions [8, p. 221]. As in the proof of Lemma 2, it is possible to take limits in the equation

$$\mu(\omega^n)^t = df(x^{n+1}) + (\lambda^n)^t A$$

because $\lim_{n\to\infty} \lambda^n = \lambda^\infty$ exists. This shows that $\lim_{n\to\infty} \omega^n = \omega^\infty$ exists and satisfies

$$\mu(\omega^\infty - \mathbf{1})^t = df(x^\infty) + (\lambda^\infty - \mu e_1)^t A,$$

where again e_1 has first entry 1 and remaining entries 0. Owing to Lemma 1, if $x_i^\infty > 0$, then $\omega_i^\infty - 1 = 0$. If $x_i^\infty = 0$, then we must verify that $\omega_i^\infty - 1 \geq 0$. However, $x_i^{n_k+1} \leq x_i^{n_k}$ must hold for some subsequence x^{n_k} in order for $\lim_{n\to\infty} x_i^n = 0$. It follows that

$$\omega_i^\infty - 1 = \lim_{k\to\infty} \frac{x_i^{n_k}}{x_i^{n_k+1}} - 1 \geq 0.$$

This establishes the Karush-Kuhn-Tucker conditions and proves that x^∞ furnishes the global minimum.

Given that every sequence x^n converges to the minimum of $f(x)$, it is natural to ask for the local rate of convergence. If $M(x)$ denotes the algorithm map, this rate of convergence is determined by the dominant eigenvalue of the differential $dM(x)$ at the optimal point x^∞, provided x^∞ is an interior point [15, p. 145]. Making the assumption that x^∞ is an interior point, it is possible to compute $dM(x^\infty)$ by implicit differentiation.

Theorem 4. *Suppose that $f(x)$ is twice continuously differentiable and that the global minimum x^∞ is an interior point. If $d^2 f(x^\infty)$ is positive definite on the null space of A, then the sequence x^n converges to x^∞ at a linear rate. If $d^2 f(x^\infty)$ is positive*

definite on all of R^s, then this rate is no greater than the dominant eigenvalue of the matrix

$$\left[d^2 f(x^\infty) + \mu(X^\infty)^{-1}\right]^{-1} \mu(X^\infty)^{-1}. \tag{13}$$

Proof. Consider an interior point x. The stationarity condition satisfied by $y = M(x)$ and its associated Lagrange multiplier λ can be expressed as

$$
\begin{aligned}
d^{10} R(y \mid x)^t + A^t \lambda &= \mathbf{0}, \\
Ay - b &= \mathbf{0}.
\end{aligned}
\tag{14}
$$

Differentiating the left-hand-side of this system of equations with respect to both y and λ gives the matrix

$$
\begin{pmatrix} d^{20} R(y \mid x) & A^t \\ A & \mathbf{0} \end{pmatrix},
$$

which is invertible because $d^{20} R(y \mid x)$ is positive definite on the null space of A and A has full rank [12, p. 312]. Hence, the implicit function theorem [8, p. 171] implies that $M(x)$ and λ are continuously differentiable functions of x on the interior of the feasible region.

As in the proof of Lemma 2, we can solve for λ in the first equation of (14) by multiplying on the left by $(AA^t)^{-1}A$. This yields

$$Q d^{10} R(M(x) \mid x)^t = \mathbf{0}, \tag{15}$$

where $Q = I - A^t(AA^t)^{-1}A$ is the orthogonal projection onto the null space of A. Differentiating equation (15) with respect to x and rearranging produces

$$Q d^{20} R(M(x) \mid x) dM(x) = -Q d^{11} R(M(x) \mid x). \tag{16}$$

It is possible to deduce from (16) that all eigenvalues of $dM(x^\infty)$ lie on the half-open interval $[0, 1)$. Suppose $\omega \neq 0$ is an eigenvalue of $dM(x^\infty)$ with eigenvector u. Since $dM(x^\infty)$ maps into the null space of A, it follows that $Qu = u$. This holds even if u is complex. Now convert both sides of (16) into quadratic forms by multiplying on the left by the conjugate transpose u^* and on the right by u. Solving for ω and using $Qu = u$ and $dM(x^\infty)u = \omega u$ then yield

$$\omega = -\frac{u^* d^{11} R\big(M(x^\infty) \mid x^\infty\big) u}{u^* d^{20} R\big(M(x^\infty) \mid x^\infty\big) u}. \tag{17}$$

In view of formulas (4) and (5) and the fact that $M(x^\infty) = x^\infty$, equation (17) becomes

$$\omega = \frac{\mu u^*(X^\infty)^{-1} u}{u^* d^2 f(x^\infty) u + \mu u^*(X^\infty)^{-1} u}. \tag{18}$$

Because $d^2 f(x^\infty)$ is positive definite on the null space of A, expression (8) implies that $0 \leq \omega < 1$. Since ω is real, its eigenvector u can be taken real as well.

The last assertion of the theorem follows from the fact [8, p. 85] that the maximum value of the Rayleigh quotient (18) over $\{u \in R^s : u \neq \mathbf{0}\}$ coincides with the dominant eigenvalue of the matrix (13).

3.1. Example. The objective function $f(x) = -\sum_{i=1}^s m_i \ln x_i$ minimized in our multinomial example has a diagonal Hessian matrix with ith diagonal element m_i/x_i^2. At the maximum likelihood values $x_i^\infty = m_i/m$, the matrix (13) consequently reduces to $\mu/(m+\mu)I$. Now not only does the maximum value of the Rayleigh quotient (18) coincide with the dominant eigenvalue of the matrix (13) [8, p. 85], but the whole set of critical values of the Rayleigh quotient coincides with the set of eigenvalues of this matrix. It follows that the Rayleigh quotient (18) has the constant value $\mu/(m+\mu)$, and consequently the dominant eigenvalue, and indeed, all eigenvalues of $dM(x)$ equal $\mu/(m+\mu)$ as well.

4. Discussion

It is worth noting to what extent our theoretical development extends to nonconvex functions $f(x)$. The approximate version of the method, which depends on the quadratic approximation (6), is adversely affected if the Hessian $d^{20}R(x^n \mid x^n)$ fails to be positive definite on the null space of A. Fortunately, subtracting the logarithmic barrier term $\mu\sum_{i=1}^s x_i^n \ln x_i$ from $f(x)$ tends to convexify the problem so that $d^{20}R(x^n \mid x^n)$ can be positive definite even when $d^2 f(x^n)$ is not. This fact suggests that taking the barrier constant μ too small would be a mistake for nonconvex $f(x)$.

Much of the convergence analysis for the exact version of the method carries over without change. The local convergence analysis presented in Theorem 4 only requires the existence of the differential $dM(x)$ of the algorithm map $M(x)$ in a neighborhood of the minimum point x^∞. Existence of $dM(x)$ locally is assured by the implicit function theorem under the hypotheses of the theorem. Thus, Theorem 4 extends to nonconvex $f(x)$ subject to the interpretation that x^{n+1} may only supply a local minimum of (2). Many of the features of global convergence also extend naturally. If the key assumption that $f(x)$ has at most a finite number of stationary points on each nonempty manifold M_S is retained, then the proof of Theorem 3 shows that the iterates x^n of the algorithm converge to a point x^∞ satisfying the first-order Karush-Kuhn-Tucker conditions. Of course, in the absence of convexity, we now have no guarantee that x^∞ furnishes either the global or even a local minimum of $f(x)$. This theoretical handicap should not deter practical application of the method.

Finally, taking a cue from the work of Censor and Zenios (1992), one can generalize the convex programming algorithm to problems with concave inequality constraints $g_i(x) \geq 0$ instead of the nonnegativity constraints $x_i \geq 0$. A surrogate function for $f(x)$ in this more general setting is

$$R(x \mid x^n) = f(x) - \mu \sum_i \big[g_i(x^n)\ln g_i(x) - dg_i(x^n)x\big]. \qquad (19)$$

The point $x = x^n$ is the maximum point of

$$\mu \sum_i \big[g_i(x^n)\ln g_i(x) - dg_i(x^n)x\big]$$

by virtue of the concavity of $\ln g_i(x)$. If an initial interior point x^1 can be found in the sense that all $g_i(x^1) > 0$, then every subsequent point x^{n+1} generated by minimizing (19) is an interior point that decreases $f(x)$. This algorithm is potentially useful for the broad class of geometric programming problems [17].

References

1. Y. Censor and S. A. Zenios, *Proximal minimization with D-functions*, J. Optimization Theory Appl. **73** (1992), 451–464.
2. A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm* (with discussion), J. R. Stat. Soc. B. **39** (1977), 1–38.
3. J. Ding and T. Y. Li, *An algorithm based on weighted logarithmic barrier functions for linear complementarity problems*, Arabian J. Sci. Engin. **15** (1990), 679–685.
4. S.-C. Fang and S. Puthenpura, *Linear Optimization and Extensions: Theory and Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1993.
5. A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.
6. D. Goldfarb and S. Liu, *An $O(n^3L)$ primal interior point algorithm for convex quadratic programming*, Math Programming **49** (1991), 325–340.
7. C. C. Gonzaga, *An algorithm for solving linear programming problems in $O(n^3L)$ operations*, In: Progress in Mathematical Programming (N. Megiddo, ed.), Springer, New York, (1988), 1–28.
8. M. R. Hestenes, *Optimization Theory: The Finite Dimensional Case*, Kreiger, Huntington, NY, 1981.
9. N. Karmarkar, *A new polynomial time algorithm for linear programming*, Combinatorica **4** (1984), 373–395.
10. V. Klee and G. L. Minty, *How good is the simplex algorithm?*, In: Inequalities III (O. Shisha, ed.), Academic Press, New York, (1972), 159–179.
11. R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
12. D. G. Luenberger, *Linear and Nonlinear Programming, 2nd Ed.*, Addison-Wesley, Reading, MA, 1984.
13. N. Megiddo and M. Shub, *Boundary behavior of interior point algorithms in linear programming*, Math. Operations Res. **14** (1989), 97–146.
14. Y. E. Nesterov and A. S. Nemirovsky, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*, SIAM, Philadelphia, 1994.
15. J. M. Ortega, *Numerical Analysis: A Second Course*, SIAM, Philadelphia, 1990.
16. A. M. Ostrowski, *Solutions of Equations in Euclidean and Banach Spaces*, Academic, New York, 1973.
17. A. L. Peressini, F. E. Sullivan, and J. J. Uhl Jr., *The Mathematics of Nonlinear Programming*, Springer, New York, 1988.
18. R. Polyak, *Modified barrier functions (theory and methods)*, Math Programming, **54** (1992), 177–222.
19. C. R. Rao, *Linear Statistical Inference and its Applications, 2nd Ed.*, Wiley, New York, 1973.
20. C. Roos and J.-P. Vial, *Long steps with the logarithmic penalty barrier function in linear programming*, In: Economic Decision Making: Games, Economics, and Optimization (J. Gabszevwicz , J.-F. Richard, and L. Wolsey, eds.), Elsevier, Amsterdam, (1990), 433–441.

DEPARTMENT OF BIOSTATISTICS, SCHOOL OF PUBLIC HEALTH, UNIVERSITY OF MICHIGAN, ANN ARBOR, MI 48109-2029, U.S.A.