Matrix Scaling, Entropy Minimization, and Conjugate Duality. I. Existence Conditions

Michael H. Schneider*
Department of Mathematical Sciences
Johns Hopkins University
Baltimore, Maryland 21218

Dedicated to Alan J. Hoffman on the occasion of his 65th birthday.

Submitted by Richard A. Brualdi

ABSTRACT

We have four primary objectives in this paper. First, we introduce a problem called *truncated matrix scaling* that generalizes two well-studied matrix scaling problems—diagonal similarity scaling and fixed row-column equivalence scaling. Second, we derive necessary and sufficient conditions for the attainment of the infimum in the Fenchel dual of a class of convex optimization problems. Third, we show that existence of a solution for truncated matrix scaling is equivalent to the attainment of the infimum in a corresponding dual optimization problem. We thereby derive necessary and sufficient conditions for the existence of a solution for truncated matrix scaling. Fourth, we derive known existence conditions for similarity and equivalence scaling from the conditions for truncated matrix scaling.

1. INTRODUCTION

Given a nonnegative matrix A, a *scaling* of A is a matrix of the form D_1AD_2 where D_1 and D_2 are diagonal matrices with positive diagonal entries whose dimensions are compatible with A's. Matrix scaling problems have been studied extensively in linear algebra, economics, statistics, urban planning, and demography. In this paper, we introduce a common generalization of two classical scaling problems—diagonal similarity scaling and diagonal equivalence scaling. We describe each of these problems.

*Research supported in part by NSF grant ECS 87-18971.

655 Avenue of the Americas, New York, NY 10010 0024-3795/89/$3.50

Let A be an $n \times n$ nonnegative matrix; then we call A *balanced* if

$$\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji} \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

The diagonal similarity scaling problem is:

Find a diagonal matrix D with positive diagonal entries such that DAD^{-1} is balanced.

This problem has been studied by Osborne [23], Grad [15], Eaves et al. [9], and Schneider and Zenios [29]. It occurs, for example, in developmental economics when a social accounting matrix is estimated and used for economic equilibrium modeling.

Now let A be an $m \times n$ nonnegative matrix, and let α_1 and α_2 be given positive vectors of length m and n, respectively. The diagonal equivalence scaling problem is:

Find diagonal matrices D_1 and D_2 with positive diagonal entries such that the matrix D_1AD_2 has row sums equal to α_1 and column sums equal to α_2.

This problem has been studied by numerous researchers in many diverse fields. The paper [29] contains a survey of algorithms and applications of matrix balancing.

We call our common generalization of similarity and equivalence scaling the *truncated matrix scaling problem* (or *truncated scaling*). Let (A, L, U) be a triple of $n \times n$ nonnegative matrices. Then, the truncated scaling problem is:

Find a diagonal matrix D with positive diagonal entries and a nonnegative matrix Λ such that X, the Hadamard product of Λ and DAD^{-1}, is balanced and lies between L and U, elementwise. In addition, each pair (i, j) must satisfy

(a) $x_{ij} = l_{ij}$ whenever $\lambda_{ij} > 1$, and
(b) $x_{ij} = u_{ij}$ whenever $\lambda_{ij} < 1$.

The Hadamard product of two matrices A and B (of the same size) is the matrix C such that $c_{ij} = a_{ij}b_{ij}$.

Intuitively, the matrix X is formed by first performing the similarity scaling DAD^{-1} and then truncating any element lying outside of its bound. The resulting truncated matrix is required to be balanced. Balinksi and

Demange [3, 4] have considered a related generalization of equivalence scaling in which logical conditions such as (a) and (b) are incorporated to take account of lower and upper bounds on the row and column sums.

In this paper we show that necessary and sufficient conditions for the existence of a solution to the truncated scaling problem can be derived from a duality theorem giving necessary and sufficient conditions for the attainment of the infimum in the dual of a convex optimization problem. We then apply the duality theorem to the truncated scaling problem, thereby deriving conditions for the existence of a solution for truncated scaling. In the special cases of similarity scaling and equivalence scaling, we derive results of Brualdi [7], Eaves et al. [9], Menon [21], Menon and Schneider [22], Rothblum and Schneider [27], and Sinkhorn [30]. The duality theorem and the truncated scaling problem are, we believe, new.

The truncated scaling problem is an important modeling extension beyond equivalence and diagonal scaling. For example, in applications of diagonal scaling in development economics, the A matrix is a raw estimate of a *flow-of-funds matrix* between sectors of an economy. The balance conditions are the *a priori* accounting identities that each sector's receipts and expenditures must be equal. Since the data are imperfect, the raw estimate never satisfies the balance conditions, and therefore some numerical procedure must be used to generate a balanced estimate. Typically, some of the data are known to be quite accurate (such as government statistics or data from industries consisting of a few large firms), whereas other data is extremely unreliable (consumer or agricultural data). Thus a better estimate is obtained if lower and upper bounds are introduced so that reliable data are constrained to lie near their initial values, whereas unreliable data are allowed to vary substantially. See [29] for a discussion of other models which incorporate measures of data reliability.

The technique of analyzing matrix scaling problems by studying equivalent optimization problems has produced many important results. See, for example, [6, 9, 11, 12, 18, 20]. Other researchers have studied the relationship between entropy minimization and matrix scaling [8, 10, 11, 13, 17]. Seven interrelated papers have appeared recently dealing with generalizations of scaling problem. These are the papers of Balinski and Demange [3, 4], Bapat and Raghavan [5], Franklin and Lorenz [14], Rothblum [26], and Rothblum and Schneider [27], as well as this paper. Each paper contains necessary and sufficient condition for the existence of a solution to a scaling problem (along with other interesting material related to scaling problems). These papers differ with respect to the particular scaling problems considered and with respect to the techniques used to prove the results. We will give a brief summary of the differences between these papers within the optimization framework described in this paper.

TABLE 1

PRIMAL PROBLEMS CORRESPONDING TO SCALING PROBLEMS

Problem	Constraint set	Comments
Similarity scaling	$\{x \geqslant 0 \mid Mx = 0\}$	M is the incidence matrix of a directed graph
Equivalence scaling	$\{x \geqslant 0 \mid Mx = b\}$	M is the incidence matrix of an undirected bipartite graph
Balinski-Demange [3, 4]	$\{x \geqslant 0 \mid b^- \leqslant Mx \leqslant b^+\}$	M is the incidence matrix of an undirected bipartite graph
Bapat-Raghavan [5]	$\{x \geqslant 0 \mid Mx = b\}$	M is an arbitrary matrix
Rothblum [26]	$\{x \geqslant 0 \mid Mx = b\}$	M is an arbitrary matrix
Franklin-Lorenz [14]	$\{x \geqslant 0 \mid Mx = b\}$	M is an arbitrary matrix
Rothblum-Schneider [27]	$\{x \geqslant 0 \mid Mx = b\}$	M is the incidence matrix of an undirected bipartite graph
This paper	$\{x \mid Mx = 0,\ l \leqslant x \leqslant u\}$	M is the incidence matrix of a directed graph
This paper	$x \in S \cap C$, S a subspace, C a polyhedron	General convex functions considered

The existence of a solution to a scaling problem is equivalent to the attainment of the infimum in the dual of a linearly constrained entropy minimization problem (see, for example, Section 5). We can associate to each scaling problem considered in these seven papers an entropy minimization problem over a convex polyhedral set. We call this the *primal problem* corresponding to the scaling problem. The scaling problem can now be analyzed using the primal or the dual optimization problem, or by using the optimality conditions. For the scaling problems considered in these papers, it suffices to consider an entropy function of the form $\Sigma_j x_j[\ln(x_j/c_j) - 1]$ for given positive constants c_j. The differences between the scaling problems considered in each paper can be summarized by describing the differences between the constraint sets over which the entropy function is minimized. We have summarized these differences in Table 1.

A variety of proof techniques are used in these papers. Basically, the existence results show that a positivity condition related to a Slater-type constraint qualification condition is necessary and sufficient for the existence of a solution to the scaling problem. Among the papers using optimization techniques (all except Bapat and Raghavan), the general approach is to use sufficient conditions from convex analysis for the existence of a solution to the dual problem or for the existence of a multiplier vector for the primal problem. Then the sufficient condition is shown to be necessary, using the property of the entropy function that it is differentiable at precisely those points where every coordinate is strictly positive.

Balinski and Demange [3] in Theorem 1 use a sufficient condition for the existence of a Kuhn-tucker vector for a convex program. Bapat and Raghavan [5] in Lemma 1 and Theorem 1 use degree theory for continuous mappings to prove a preliminary existence result, which together with the duality theory for linear programming is used to prove their main existence result. Franklin and Lorenz [14] show directly that the solution of their primal problem must be strictly positive, thereby implying the existence of a multiplier vector. Rothblum and Schneider [27] in Theorem 2 use a sufficient condition based on directions of recession for the attainment of the infimum of the dual problem to derive existence results for the equivalence scaling problem. This technique is used by Rothblum [26] for the generalized scaling problem considered by Bapat and Raghavan and by Franklin and Lorenz.

In this paper, we are interested in a generalization of matrix scaling incorporating lower and upper bounds. The existence result, however, is proved in a more general setting of nonseparable convex optimization over polyhedral convex sets. Our Theorem 7 is not restricted to entropy functions; rather we prove the result for arbitrary convex functions with the property that the domain of the subdifferential and the relative interior of the domain of the function coincide. We use a sufficient condition for the existence of a multiplier vector and show that for functions with this property the sufficient condition is also necessary. Thus, we have derived a somewhat more general existence result which contains all the existence results described here as special cases.

In Section 2 we give a precise statement of truncated scaling and show that it generalizes both similarity and equivalence scaling. We analyze the truncated scaling problem using the theory of conjugate duality as described in Rockafellar [25]. Alternatively, some of our results could be derived from the theory of monotropic programming as described in [24]. In Section 3, we summarize the duality results we need and develop necessary and sufficient conditions for the attainment of the infimum in the dual of a convex optimization problem (Theorem 7). In Section 4 we summarize some elementary properties of the entropy function.

In Section 5 we define the primal and dual optimization problems corresponding to the truncated scaling problem and show that the existence of a solution for truncated scaling is equivalent to the attainment of the infimum in the dual optimization problem. We then apply Theorem 7 and derive necessary and sufficient conditions for the existence of a solution to truncated scaling (Theorem 14). Finally, in Section 6 we show that previous existence results for similarity and equivalence scaling can be derived from Theorem 14.

2. TRUNCATED SCALING

2.1. Problem Statement

Let A be an $n \times n$ nonnegative matrix; then we call A *balanced* if

$$\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji} \qquad \text{for} \quad i = 1,2,\dots,n.$$

We will frequently want to refer to the set of matrices X with the property that $x_{ij} = 0$ whenever $a_{ij} = 0$. We will call such an X a matrix *whose pattern is contained in A's.*

Let L be an $n \times n$ nonnegative matrix, and let U be an $n \times n$ matrix whose entries—which will always be upper bounds—are either nonnegative or the symbol $+\infty$. Whenever the symbol $+\infty$ occurs on the right-hand side of an inequality (e.g., $x_{ij} \leqslant u_{ij} = +\infty$), it means there is no bound on that constraint. The ordered triple (A, L, U) is called *consistent* if

(i) $0 \leqslant L \leqslant U \leqslant \infty$,

(ii) there is an $n \times n$ matrix X whose pattern is contained in A's such that X is balanced and $L \leqslant X \leqslant U$, and

(iii) $u_{ij} > 0$ whenever $a_{ij} > 0$.

The notation $D = \operatorname{diag}(d_1, d_2, \dots, d_n)$ will denote the $n \times n$ diagonal matrix whose diagonal entries are $d_1, d_2, \dots, d_n$.

We develop necessary and sufficient conditions for the existence of a solution to the following problem, which we call *truncated matrix scaling.*

PROBLEM 1 (Truncated matrix scaling). Given a consistent triple of $n \times n$ matrices (A, L, U), find a matrix $D = \operatorname{diag}(d_1, d_2, \dots, d_n)$ with positive diagonal entries and an $n \times n$ nonnegative matrix Λ such that for X defined by $x_{ij} = \lambda_{ij} d_i a_{ij} d_j^{-1}$ for $i = 1,2,\dots,n$ and $j = 1,2,\dots,n$ the following conditions are satisfied:

(i) the matrix X is balanced and satisfies $L \leqslant X \leqslant U$;

(ii) the matrices X and Λ satisfy:

(a) if $\lambda_{ij} > 1$, then $x_{ij} = l_{ij}$, and

(b) if $\lambda_{ij} < 1$, then $x_{ij} = u_{ij}$.

The requirement that (A, L, U) be consistent eliminates infeasible and redundant instances of truncated scaling and therefore is introduced without loss of generality. Clearly, if X is a solution to the truncated-scaling problem, then X must be a matrix whose pattern is contained in A's. Thus, consistency conditions (i) and (ii) are obviously necessary for the existence of a solution. If consistency condition (iii) were violated, then any such elements a_{ij} could be replaced by 0, thereby producing an equivalent problem.

In the truncated scaling problem, X is derived from A, D, and Λ by forming the product DAD^{-1} and then multiplying elementwise the entries of Λ and DAD^{-1}. Loosely speaking, condition (ii) states that the multiplier λ_{ij} is used (i.e., not equal to 1) only when it is necessary to force the element x_{ij} to lie within its bounds. Equivalently, X is derived from DAD^{-1} by truncating each entry above its upper bound or below its lower bound to the respective bound. The entries that lie within the bounds are left intact. The problem requires that after truncation the matrix must be balanced.

2.2. *The Correspondence between Matrices and Graphs*

Condition (i) of Problem 1 can be represented as a feasible circulation condition for a sparse network associated with the matrix A. We describe this equivalent formulation, since we will use it to formulate a truncated scaling problem as an optimization problem.

Let A be an $n \times n$ nonnegative matrix; we define the *graph of* A, written $G(A)$, to be the weighted directed graph (V, E, a) where

$$V = \{1,2,\dots,n\},$$

$$E = \{e = (i,j) | a_{ij} > 0\},$$

$$a_e = a_{ij} \qquad \text{for} \quad e = (i,j) \in E.$$

We will use a_{ij} for the value of the ijth element of the matrix A, and a_e for the value of the weight function a at edge e. It is easy to see that the correspondence between nonnegative matrices A and weighted directed graph with nonnegative weight functions and no parallel edges is a bijection.

There is an obvious correspondence between $n \times n$ matrices X and vectors x defined on the edges of $G(A)$; namely,

$$x_e = x_{ij} \qquad \text{for} \quad e = (i,j) \in E.$$

This correspondence is, of course, a bijection when restricted to matrices whose pattern is contained in A's. We will use this correspondence by referring, for example, to the *vector x corresponding to the matrix X*.

Let (A, L, U) be a consistent triple of $n \times n$ matrices. We define the *incidence matrix for* $G(A)$ to be the $|V| \times |E|$ matrix $[\chi_e]_{e \in E}$, where χ_e is the incidence (column) vector for e. That is, if $e = (i, j)$, then $\chi_e \in \Re^V$ has a -1 and a $+1$ in rows i and j, respectively, and 0's in every other row. Let B be the incidence matrix for $G(A)$, and let l and u be the vectors in $\Re^E$ corresponding to L and U, respectively. Consider the linear constraints:

$$\begin{aligned} Bx &= 0, \\ l \leqslant x &\leqslant u. \end{aligned} \tag{1}$$

A vector x satisfying $Bx = 0$ is called a *circulation* in $G(A)$; if x satisfies $l \leqslant x \leqslant u$ as well, then x is called a *feasible circulation* in $G(A)$ (with respect to l and u).

Under the correspondence between vectors and matrices, a matrix X whose pattern is contained in A's is balanced if and only if the corresponding vector x is a circulation in $G(A)$. For such X it is easy to see that condition (i) of the truncated scaling problem is equivalent to the constraints (1). We will use this equivalence when formulating the truncated scaling problem as an optimization problem.

2.3. *Examples*

We mention two special classes of truncated scaling problems, as they have been extensively investigated in the literature.

2.3.1. Diagonal Similarity Scaling. Consider the special case of truncated scaling in which $L = 0$ and $U = +\infty$. If the required D and Λ exist, it follows from consistency condition (iii) that $\lambda_{ij} = 1$ for all i and j for which $a_{ij} > 0$. The truncated scaling problem then reduces to the following scaling problem:

PROBLEM 2 (Similarity scaling). Given an $n \times n$ nonnegative matrix A, find a matrix $D = \operatorname{diag}(d_1, d_2, \dots, d_n)$ with positive diagonal entries such that $X = DAD^{-1}$ is balanced.

Existence results for similarity scaling directly related to this paper have been developed by Osborne [23] and Eaves et al. [9]. In Section 6 we show that some of these results can be derived from the existence theory for truncated scaling.

2.3.2. *Fixed Row-Column Equivalence Scaling.* Let A' be an $m \times n$ nonnegative matrix, and let α be a strictly positive vector in $\Re^{m+n}$. We call the pair (A', α) *consistent* if

(i) $\sum_{i=1}^{m} \alpha_i = \sum_{i=m+1}^{m+n} \alpha_i$, and

(ii) there is some $m \times n$ matrix X whose pattern is contained in A' such that

$$\begin{aligned} \sum_{j=1}^{m} x_{ij} &= \alpha_i \qquad \text{for} \quad i = 1,2,\ldots,m, \\ \sum_{i=1}^{n} x_{ij} &= \alpha_{m+j} \qquad \text{for} \quad j = 1,2,\ldots,n. \end{aligned} \tag{2}$$

We will say that a matrix X satisfying the equations (2) has *row and column sums given by* α.

Consider the following matrix scaling problem:

PROBLEM 3 (Equivalence scaling). Given a consistent pair (A', α), find matrices

$$D_1 = \operatorname{diag}(d_1, d_2, \ldots, d_m),$$

$$D_2 = \operatorname{diag}(d_{m+1}, d_{m+2}, \ldots, d_{m+n})$$

with positive diagonal entries such that the matrix $X = D_1 A' D_2$ has row and column sums given by α.

Equivalence scaling has been studied extensively in many different fields. Some algorithmic and applied aspects of the problem are described in [29]. Some existence results directly related to this paper have been developed by Menon [21], Brualdi [7], Menon and Schneider [22], and Rothblum and Schneider [27]. In Section 6 we will derive some of these results as corollaries of the existence theory for truncated scaling.

We need to transform the equivalence scaling problem into a truncated scaling problem. This is a standard transformation which is used in network flow theory to convert a transshipment problem to a circulation problem (see, for example [19, p. 113]). First, we define a directed bipartite graph from A', (V', E'), by

$$V' = \{1,2,\ldots,m+n\},$$

$$E' = \{(i,j) \mid 1 \leqslant i \leqslant m, m+1 \leqslant j \leqslant m+n, \text{and } a_{i(j-m)} > 0\}.$$

Then we define (V, E) by

$$V = \{0, 1, 2, \ldots, m+n\},$$

$$E = E' \cup \{(0, i) \mid i = 1, 2, \ldots, m\} \cup \{(i, 0) \mid i = m+1, m+2, \ldots, m+n\},$$

and a weight function a on the edges $e = (i, j)$ of E by

$$a_e = \begin{cases} a_{i(j-m)} & \text{if } e \in E', \\ \alpha_j & \text{if } i = 0, \\ \alpha_i & \text{if } j = 0. \end{cases}$$

We call the resulting weighted directed graph (V, E, a) the *graph corresponding to* (A', α).

We define lower- and upper-bound vectors $l = (l_e; e \in E)$ and $u = (u_e; e \in E)$ as follows. Let $e = (i, j) \in E$; then

$$l_e = \begin{cases} 0 & \text{if } e \in E', \\ \alpha_j & \text{if } i = 0, \\ \alpha_i & \text{if } j = 0, \end{cases}$$

and

$$u_e = \begin{cases} +\infty & \text{if } e \in E', \\ \alpha_j & \text{if } i = 0, \\ \alpha_i & \text{if } j = 0. \end{cases}$$

See Figure 1.

We now have a weighted directed graph (V, E, a) together with lower and upper bounds l and u. Let A be the $(m+n+1) \times (m+n+1)$ matrix whose graph is (V, E, a), and let L and U be the matrices corresponding to l and u, respectively. We call (A, L, U) the *ordered triple corresponding to* (A', α). It is not hard to show that (A', α) is consistent if and only if (A, L, U) is consistent.

It is not hard to show that the equivalence scaling problem for (A', α) is equivalent to the truncated scaling problem for the corresponding triple (A, L, U) in the sense that a solution for one can be transformed immediately into a solution for the other.

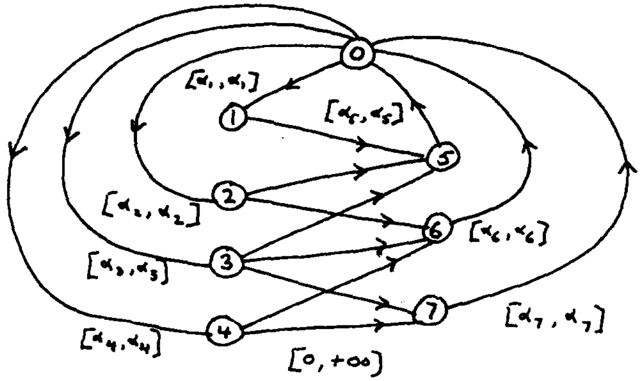

FIG. 1. The graph for equivalence scaling.

LEMMA 1. *Let A' be an $m \times n$ nonnegative matrix, and let α be a strictly positive vector in $\Re^{m+n}$. Let (A', α) be consistent, and let (A, L, U) be the corresponding ordered triple. Then*

(i) *if $D = \operatorname{diag}(d_1, d_2, \ldots, d_{m+n})$ and Λ solve the truncated scaling problem for (A, L, U), then D_1 and D_2 solve the equivalence scaling problem for (A', α), where*

$$D_1 = \operatorname{diag}(d_1, d_2, \ldots, d_m),$$
$$D_2 = \operatorname{diag}\left(d_{m+1}^{-1}, d_{m+2}^{-1}, \ldots, d_{m+n}^{-1}\right); \tag{3}$$

(ii) *conversely, if D_1 and D_2 in* (3) *solve the equivalence scaling problem for (A', α), then D and Λ solve the truncated scaling problem for (A, L, U), where*

$$D = \operatorname{diag}(1, d_1, d_2, \ldots, d_{m+n}), \tag{4}$$

and Λ is the matrix of all 1's.

Proof. The result follows from a straightforward calculation. ∎

3. DUALITY FOR CONVEX OPTIMIZATION

In this section we derive necessary and sufficient conditions for the attainment of the infimum in the dual problem for a class of convex optimization problems. The main result of this section is Theorem 7. We use the notation and definitions in Rockafellar [25].

Let f be a closed proper convex (extended) real-valued function defined on $\mathfrak{R}^n$. The function f is *closed* if its *epigraph* defined by

$$\operatorname{epi} f = \{(x, \mu) | x \in \mathfrak{R}^n, \mu \in \mathfrak{R}, \mu \geqslant f(x)\}$$

is a closed subset of $\mathfrak{R}^{n+1}$. The (*effective domain* of f), written dom f, is defined by

$$\operatorname{dom} f = \{x \in \mathfrak{R}^n | f(x) < +\infty\}.$$

A function f is *proper* if it never assumes the value $-\infty$ and is not identically equal to $+\infty$. Also, f is *convex* if epi f is a convex subset of $\mathfrak{R}^{n+1}$.

For convex f the (*subdifferential* of f at x) written $\partial f(x)$, is the subset of $\mathfrak{R}^n$ defined by $x^* \in \partial f(x)$ if

$$f(y) \geqslant f(x) + \langle x^*, y - x\rangle \qquad \text{for all} \quad y \in \mathfrak{R}^n,$$

where $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner product defined on $\mathfrak{R}^n \times \mathfrak{R}^n$. If $\partial f(x)$ is not empty, then f is (*subdifferentiable* at x). The (*domain* of ∂f), written dom ∂f, is the set of $x \in \mathfrak{R}^n$ where $\partial f(x) \neq \varnothing$.

For a subset $K \subseteq \mathfrak{R}^n$ the (*indicator function of* K), written $\delta(x|K)$, is defined by

$$\delta(x|K) = \begin{cases} 0 & \text{if} \quad x \in K, \\ +\infty & \text{otherwise}. \end{cases}$$

The function f is *polyhedral* if epi f is a polyhedral convex subset of $\mathfrak{R}^{n+1}$. The *relative interior* of a convex set C, written ri C, is the interior of C relative to the *affine hull* of C (see [25] for the exact definition).

We will use the following standard results concerning subdifferentials, normal cones, and indicator functions (see [25, pp. 215, 223]).

LEMMA 2. *Let f and g be proper convex functions on $\Re^n$. If* ri(dom f) $\cap$ ri(dom g) $\neq \varnothing$, *then*

$$\partial(f+g)(x) = \partial f(x) + \partial g(x) \qquad \textit{for all} \quad x \in \Re^n.$$

Further, whenever f (and/or g) is polyhedral, ri(dom f) *can be replaced by* dom f (*similarly for* g).

LEMMA 3. *Let C_1 and C_2 be polyedral convex subsets of $\Re^n$. Then*

(i) $\partial\delta(x|C_i)$ *is the normal cone to C_i at x, $i=1,2$;*

(ii) *if $C = C_1 \cap C_2 \neq \varnothing$, then the normal cone to C at x is the sum of the normal cones to C_1 and C_2 at x. That is,*

$$\partial\delta(x|C) = \partial\delta(x|C_1) + \partial\delta(x|C_2).$$

We are interested in the following optimization problem:

$$\inf_x f(x)$$
$$\text{subject to} \quad x \in K = S \cap C, \tag{5}$$

where f is a closed proper convex function on $\Re^n$, S is a subspace of $\Re^n$, and C is a polyhedral convex subset of $\Re^n$. The problem (5) is equivalent to

$$\inf_{x \in \Re^n} \{ f(x) + \delta(x|K) \}.$$

It is easy to see that a necessary and sufficient condition for the infimum of a convex function g of $\Re^n$ to be attained at a point x is that $0 \in \partial g(x)$. Since $\partial\delta(x|K)$ is the normal cone to K at x, the next theorem follows directly from Lemma 2. (See [25, Theorem 27.4].)

THEOREM 4. *Let f be a closed proper convex function on $\Re^n$, and let K be a polyhedral convex subset of $\Re^n$ such that $K \cap$* ri(dom f) $\neq \varnothing$. *Then for a vector $x \in K$ the following are equivalent:*

(i) *the infimum of f over K is attained at the point x;*

(ii) *there exists $x^* \in \partial f(x)$ such that $-x^*$ is in the normal cone to K at x.*

We now wish to define a dualization of (5). For a convex function f on $\Re^n$, the *conjugate of f*, written f^*, is the (closed convex) function on $\Re^n$

defined by

$$f^*(x^*) = \sup_{x \in \Re^n} \{\langle x, x^* \rangle - f(x)\}.$$

Clearly, (5) is equivalent to

$$\inf_x g(x)$$
$$\text{subject to} \quad x \in S, \tag{6}$$

where

$$g(x) = f(x) + \delta(x|C). \tag{7}$$

Then the specialization of the Fenchel dual of (6) is

$$\inf_x g^*(x^*)$$
$$\text{subject to} \quad x^* \in S^{\perp}, \tag{8}$$

where $S^{\perp}$ is the orthogonal complement of the subspace S.

We quote a standard duality result for convex optimization (see [25, Corollary 31.4.2]).

THEOREM 5. *Let g be a closed proper convex function on $\Re^n$, and let S be a subspace of $\Re^n$. Then the following are equivalent*:

(i) *the vectors x and x^* satisfy*

$$g(x) = \inf_S g = -\inf_{S^{\perp}} g^* = -g^*(x^*);$$

(ii) *the vectors x and x^* satisfy $x \in S$, $x^* \in S^{\perp}$, and $x^* \in \partial g(x)$.*

A sufficient condition for the infimum in the dual problem (8) to be attained is that $S \cap \mathrm{ri}(\mathrm{dom}\, g) \neq \varnothing$. We will show that for a restricted class of convex functions and g defined by (7), the weaker condition $S \cap C \cap \mathrm{ri}(\mathrm{dom}\, f) \neq \varnothing$ is both necessary and sufficient.

LEMMA 6. *Let f be a closed proper convex function on $\Re^n$, and let S and C be, respectively, a subspace and a polyhedral convex subset of $\Re^n$*

whose intersection is nonempty. Let $K = S \cap C$, *and assume that* $C \cap \operatorname{ri}(\operatorname{dom} f) \neq \varnothing$. *Then for a given vector* x *the following are equivalent*:

(i) $x \in K$ *and there exists* $x^* \in \partial f(x)$ *such that* $-x^*$ *is in the normal cone to* K *at* x;

(ii) $x \in S$ *and there exists* $y^* \in S^{\perp}$ *such that* $y^* \in \partial[f + \delta(\cdot|C)](x)$.

Moreover, when these equivalent conditions are satisfied, the vectors x, x^*, *and* y^* *are related by* $x^* = y^* - z^*$ *for some* $z^* \in \partial\delta(x|C)$.

Proof. (i) ⇒ (ii): Suppose $x \in K$, $x^* \in \partial f(x)$, and $-x^* \in \partial\delta(x|K)$. Then it follows from Lemma 3 that

$$-x^* = z^* - y^*$$

for some $z^* \in \partial\delta(x|C)$ and $y^* \in \partial\delta(x|S) = S^{\perp}$. Thus

$$\begin{aligned} y^* &= x^* + z^* \\ &\in \partial f(x) + \partial\delta(x|C) \\ &= \partial[f + \delta(\cdot|C)](x). \end{aligned}$$

(ii) ⇒ (i): Similarly, for $x \in S$, $y^* \in S^{\perp}$, and $y^* \in \partial[f + \delta(\cdot|C)](x)$, we have

$$y^* = x^* + z^*$$

for some $x^* \in \partial f(x)$ and $z^* \in \partial\delta(x|C)$. Note that $x \in C$, since $\partial\delta(x|C) = \varnothing$ for $x \notin C$. Therefore $x \in K$, $x^* \in \partial f(x)$, and $-x^* = z^* - y^*$ is in the normal cone to $K = S \cap C$ at x. ■

It is well known that for a closed proper convex function f,

$$\operatorname{ri}(\operatorname{dom} f) \subseteq \operatorname{dom} \partial f \subseteq \operatorname{dom} f.$$

That is, f is always subdifferentiable on the relative interior of its domain, but may not be subdifferentiable on its relative boundary. We want to consider the special case in which the $\operatorname{dom} \partial f$ and $\operatorname{ri}(\operatorname{dom} f)$ coincide—namely, those f for which the subdifferential is empty at every point of the relative boundary.

We can now state the main theorem of this section.

THEOREM 7. *Let* f *be a closed proper convex function on* $\Re^n$, *and let* S *and* C *be, respectively, a subspace and a polyhedral convex subset of* $\Re^n$

whose intersection is nonempty. Let $K = S \cap C$, and let $g(x) = f(x) + \delta(x|C)$. Suppose that $\operatorname{dom} \partial f = \operatorname{ri}(\operatorname{dom} f)$, *that* $C \cap \operatorname{ri}(\operatorname{dom} f) \neq \varnothing$, *and that the infimum of f over K is attained. Then the following are equivalent*:

(i) *the set* $K \cap \operatorname{ri}(\operatorname{dom} f) \neq \varnothing$:

(ii) *there exist* $x \in K$, $x^* \in \partial f(x)$ *such that* $-x^*$ *is in the normal cone to K at x*;

(iii) *there exist* $x \in S$ *and* $y^* \in S^\perp$ *such that* $y^* \in \partial g(x)$;

(iv) *there exist* $x \in S$ *and* $y^* \in S^\perp$ *such that*

$$g(x) = \inf_S g = -\inf_{S^\perp} g^* = -g^*(y^*).$$

Moreover, whenever the equivalent conditions are satisfied, the vectors x in (ii), (iii), *and* (iv) *and* y^* *in* (iii) *and* (iv) *can be chosen to coincide. The vectors* x, x^*, *and* y^* *are related by* $x^* = y^* - z^*$ *for some* $z^* \in \partial\delta(x|C)$.

Proof. (i) ⇒ (ii): This follows from Theorem 4.

(ii) ⇒ (iii): This follows from Lemma 6.

(iii) ⇒ (i): It follows from Lemma 2 that

$$\partial g(x) = \partial f(x) + \partial\delta(x|C).$$

Thus, $y^* \in \partial g(x)$ implies that both $\partial f(x)$ and $\partial\delta(x|C)$ are nonempty. Therefore, $x \in C \cap \operatorname{ri}(\operatorname{dom} f)$, and (since $x \in S$) the implication follows.

(iii) ⇔ (iv): This follows from Theorem 5. ■

It is useful (and sufficient for our application) to state a sufficient condition for the infimum of f over K in Theorem 7 to be attained. Thus, a closed proper convex function is called *cofinite* if $\operatorname{dom} f^* = \Re^n$, that is, if the conjugate of f is finite everywhere. It follows directly from (7) and the definition of the conjugate function that g is cofinite whenever f is cofinite. It follows from [25, Corollary 31.4.2] that the infimum in (6) is attained whenever g is cofinite. Since (5) and (6) are equivalent, the infimum of f over K must also be attained whenever f is cofinite and K is nonempty.

4. THE ENTROPY FUNCTION

For fixed constant $c > 0$, the *entropy function with parameter c*, written $\operatorname{Ent}(x; c)$, is the convex function on $\Re$ defined by

$$\operatorname{Ent}(x;c) = \begin{cases} x[\ln(x/c) - 1] & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ +\infty & \text{if } x < 0. \end{cases}$$

A simple calculation shows that for $x > 0$

$$\mathrm{Ent}'(x; c) = \ln(x/c).$$

Therefore, $\mathrm{Ent}^*(t; c)$, the conjugate of $\mathrm{Ent}(x; c)$, is

$$\mathrm{Ent}^*(t; c) = \sup_{x \geqslant 0} \{tx - \mathrm{Ent}(x; c)\}, \tag{9}$$

and the supremum is attained at x satisfying

$$t - \ln(x/c) = 0,$$

or, equivalently, at $x = ce^t$. Substituting $x = ce^t$ into (9) produces

$$\mathrm{Ent}^*(t; c) = cte^t - ce^t(t-1) = ce^t. \tag{10}$$

For $0 \leqslant l \leqslant u$, define the function $\phi(x)$ on $\Re$ by

$$\phi(x) = \begin{cases} x[\ln(x/c) - 1] & \text{if} \quad x \in [l, u], \\ +\infty & \text{otherwise}. \end{cases}$$

It follows from (10) by a direct calculation that

$$\phi^*(t) = \begin{cases} tl - l[\ln(l/c) - 1] & \text{if} \quad ce^t \leqslant l, \\ ce^t & \text{if} \quad l \leqslant ce^t \leqslant u, \\ tu - u[\ln(u/c) - 1] & \text{if} \quad ce^t \geqslant u. \end{cases} \tag{11}$$

We will use this to derive a dual optimization problem for which the attainment of the infimum is equivalent to the existence of a solution to the truncated scaling problem.

5. SOLUTIONS TO TRUNCATED SCALING

In this section we describe the relationship between truncated scaling and the duality theory for convex optimization. We associate primal and dual optimization problems to be truncated scaling problem and show that attainment of the infimum in the dual problem is equivalent to the existence of a solution for truncated scaling. We specialize condition (i) of Theorem 7 to the truncated scaling problem and derive existence conditions by applying Alan J. Hoffman's circulation conditions to the graph corresponding to the truncated scaling problem. The application of Hoffman's theorem was also used by Brualdi [7] to derive existence conditions for equivalence scaling. The primal-dual pair of optimization problems is an example of the duality theory for monotropic programming [24].

Let (A, L, U) be a consistent triple of $n \times n$ matrices. Let $G(A) = (V, E, a)$ be the graph corresponding to A, and let $Bx = 0$, $l \leqslant x \leqslant u$, be the equivalent network formulation of truncated scaling condition (i) described in Section (2.2). We define the *primal optimization problem corresponding to* (A, L, U) to be the convex optimization problem

$$\begin{aligned} &\inf_x f(x) \\ \text{subject to}\quad &Bx = 0, \\ &l \leqslant x \leqslant u, \end{aligned} \tag{12}$$

where $f(x)$ is the convex function on $\Re^E$ defined by

$$f(x) = \sum_{e \in E} \operatorname{Ent}(x_e; a_e). \tag{13}$$

Since f is separable, it follows from (10) and a direct calculation that

$$f^*(x^*) = \sum_{e \in E} a_e e^{x_e^*}. \tag{14}$$

Define

$$\begin{aligned} S &= \{x \in \Re^E | Bx = 0\} \\ C &= \{x \in \Re^E | l \leqslant x \leqslant u\}, \\ K &= S \cap C. \end{aligned} \tag{15}$$

Note that $S^\perp = \{B^T p \,|\, p \in \Re^V\}$.

We want to define a dualization of the problem (12). Clearly, (12) is equivalent to

$$\begin{aligned} &\inf_x g(x) \\ \text{subject to}\quad &x \in S, \end{aligned} \tag{16}$$

where $g(x) = f(x) + \delta(x|C)$.

Then we define the *dual optimization problem corresponding to* (A, L, U) to be the convex optimization problem

$$\inf_{x^*} g^*(x^*)$$

$$\text{subject to} \quad x^* \in S^{\perp}. \tag{17}$$

In [28] we derive the dual function g^* explicitly from (11).

The next three lemmas summarize elementary results concerning the correspondence between optimization and scaling.

LEMMA 8. *Let (A, L, U) be a consistent triple of $n \times n$ matrices, and suppose that K defined by* (15) *is nonempty. Then x^* is in the normal cone to K at x if and only if $x^* = B^T p + y$ for some $p \in \Re^V$ and $y \in \Re^E$ such that*

$$y_e < 0 \;\Rightarrow\; x_e = l_e,$$

$$y_e > 0 \;\Rightarrow\; x_e = u_e. \tag{18}$$

Proof. Since (A, L, U) is consistent, the lemma follows directly from Lemma 3 and a direct calculation showing that y is in the normal cone to C at x if and only if y satisfies (18). ∎

LEMMA 9. *Let (A, L, U) be a consistent triple of $n \times n$ matrices, and let f and K be defined by* (13) *and* (15), *respectively. Let x, x^*, $y \in \Re^E$ and $p \in \Re^V$ be given vectors. Then the following are equivalent:*

(i) *$x \in K$, $x^* \in \partial f(x)$, and $-x^* = B^T p + y$ is in the normal cone to K at x;*

(ii) *the matrices $D = \operatorname{diag}(d_1, d_2, \dots, d_n)$ and Λ solve the truncated scaling problem for (A, L, U) where $d_i = e^{p_i}$ and*

$$\lambda_{ij} = \begin{cases} e^{-y_e} & \textit{if } \; e = (i, j) \in E, \\ 0 & \textit{otherwise}. \end{cases}$$

Proof. By Lemma 8 and properties of the entropy function, for $x \in K$ condition (i) is equivalent to

$$\ln\left(\frac{x_e}{a_e}\right) = p_i - y_e - p_j \qquad \text{for} \quad e = (i, j) \in E \tag{19}$$

and some $y \in \Re^E$ satisfying (18). It follows by exponentiating (19) that Equations (19) and (18) are equivalent to

$$x_e = \lambda_e d_i a_e d_j^{-1} \qquad \text{for} \quad e = (i, j) \in E,$$

where

$$\lambda_e > 1 \;\Rightarrow\; a_e = l_e,$$

$$\lambda_e < 1 \;\Rightarrow\; a_e = u_e.$$

The result now follows from the correspondence between graphs and matrices described in Section 2.2. ■

Lemma 10. *Let* (A, L, U) *be a consistent triple of* $n \times n$ *matrices, and let* f *and* K *be defined by* (13) *and* (15), *respectively. Then the following are true*:

(i) $\operatorname{dom} \partial f = \operatorname{ri}(\operatorname{dom} f) = \{x \in \Re^E \mid x > 0\}$,
(ii) $C \cap \operatorname{ri}(\operatorname{dom} f) \neq \varnothing$,
(iii) f *is cofinite, and*
(iv) *the infimum of* f *over* K *is attained.*

Proof. Part (i) follows directly from (13). Part (ii) follows directly from consistency condition (iii). Parts (iii) and (iv) follow directly from (14). ■

Necessary and sufficient conditions for the existence of a solution for the truncated scaling problem now follow directly from Lemmas 8, 9, and 10 and Theorem 7.

Theorem 11. *Let* (A, L, U) *be a consistent triple of* $n \times n$ *matrices. Then the following are equivalent*:

(i) *the truncated scaling problem for* (A, L, U) *has a solution*;

(ii) *there exists a balanced matrix X such that*

(a) $L \leqslant X \leqslant U$ *and*

(b) $x_{ij} > 0$ *if and only if* $a_{ij} > 0$ *for* $i, j = 1,2,\ldots,n$.

Proof. (i) ⇒ (ii): Let (D, Λ) solve the truncated scaling problem for (A, L, U), and define X by $x_{ij} = \lambda_{ij} d_i a_{ij} d_j^{-1}$ for $i, j = 1,2,\ldots,n$. Clearly, $L \leqslant X \leqslant U$, and $x_{ij} = 0$ whenever $a_{ij} = 0$. Thus, it suffices to show that $x_{ij} > 0$ whenever $a_{ij} > 0$.

Suppose that $a_{ij} > 0$. For a given pair (i, j), if $l_{ij} > 0$, then (b) follows from (a). Thus, suppose $x_{ij} = l_{ij} = 0$. Then $\lambda_{ij} = 0 < 1$, and we must have $0 = x_{ij} = u_{ij}$, which violates consistency condition (iii).

(ii) ⇒ (i): Consider the primal optimization problem corresponding to (A, L, U) defined in (12). It follows from Lemma 10 and the definition of consistency that the assumptions of Theorem 7 are satisfied. Therefore parts (i) and (ii) of Theorem 7 are equivalent. The result now follows from Lemma 9 together with the observation that the condition $K \cap \mathrm{ri}(\mathrm{dom}\, f) \neq \varnothing$ in Theorem 7 is equivalent to conditions (a) and (b) of the present theorem. ■

We next consider when a matrix X satisfying part (ii) of Theorem 11 exists. We use an adaptation of Brualdi's technique [7] for equivalence scaling to derive conditions for truncated matrix scaling based on the structure of the underlying graph $G(A)$. First, we need to introduce some notation to describe the edges directed into and out of a subset of the vertices of a directed graph.

Given a directed graph (V, E) and a subset W of V, define the subsets of the edges $\delta^-(W)$ and $\delta^+(W)$ by

$$\delta^-(W) = \{e = (i, j) \in E \mid i \in W \text{ and } j \in V \setminus W\}$$

$$\delta^+(W) = \{e = (i, j) \in E \mid i \in V \setminus W \text{ and } j \in V\}.$$

(See Figure 2.)

We state without proof a well-known theorem of Alan J. Hoffman giving necessary and sufficient conditions for the existence of a feasible circulation in a directed graph. See [16] or [19, Theorem 3.1].

THEOREM 12. *Let (V, E) be a directed graph, and let l and u be vectors in $\Re^E$ such that $0 \leqslant l \leqslant u$. (As before, the entries of u can be $+\infty$.) Then the following are equivalent:*

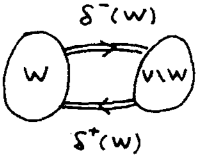

FIG. 2. The sets $\delta^-(W)$ and $\delta^+(W)$.

(i) *there exists a circulation* x *for* (V, E) *satisfying* $l \leqslant x \leqslant u$;
(ii) *for every* $W \subset V$, $\varnothing \neq W \neq V$,

$$\sum_{e \in \delta^-(W)} l_e \leqslant \sum_{e \in \delta^+(W)} u_e.$$

We now use Theorem 12 to derive necessary and sufficient conditions under which there exists a matrix satisfying part (ii) of Theorem 11.

LEMMA 13. *Let* (V, E) *be a directed graph, and let* l *and* u *be vectors in* $\Re^E$ *such that* $0 \leqslant l \leqslant u$. *Suppose there exists a feasible circulation for* (V, E). *Then the following are equivalent*:

(i) *there exists a strictly positive feasible circulation* x *for* (V, E) (*i.e.*, $x_{ij} > 0$ *for all* $e \in E$);

(ii) *for every* $W \subset V$ *for which there exists an edge* $e' \in \delta^-(W)$ *with* $l_{e'} = 0$, *we have*

$$\sum_{e \in \delta^-(W)} l_e < \sum_{e \in \delta^+(W)} u_e. \tag{20}$$

Proof. (i) ⇒ (ii): First, we note that since there exists a feasible circulation for (V, E), it follows from Theorem 12 that (20) is satisfied weakly. Suppose $W \subset V$ and $e' \in \delta^-(W)$ satisfy

$$l_{e'} = 0,$$

$$\sum_{e \in \delta^-(W)} l_e = \sum_{e \in \delta^+(W)} u_e.$$

It follows that any feasible circulation x for (V, E) must satisfy $x_e = l_e$ $e \in \delta^-(W)$ and $x_e = u_e$ for $e \in \delta^+(W)$. In particular, $x_{e'} = 0$, and therefore there cannot exist a strictly positive feasible circulation for (V, E).

(ii) ⇒ (i): Of course, the result is vacuous whenever $l > 0$. We must show that (ii) is a sufficient condition for the existence of a strictly positive feasible circulation. For an arbitrary $\epsilon > 0$ we define the vector $l' \in \Re^E$ by

$$l'_e = \begin{cases} l_e & \text{if} \quad l_e > 0, \\ \epsilon & \text{if} \quad l_e = 0. \end{cases}$$

Then

$$\sum_{e \in \delta^-(W)} l'_e = \sum_{e \in \delta^-(W)} l_e + n(W)\epsilon, \tag{21}$$

where $n(W)$ is the number of edges $e \in \delta^-(W)$ with $l_e = 0$.

If (20) holds, then it follows from (21) (and the existence of a feasible circulation with respect to l and u) that for some $\epsilon > 0$ we have

$$\sum_{e \in \delta^-(W)} l'_e \leqslant \sum_{e \in \delta^+(W)} u_e \tag{22}$$

for every $W \subset V$, $\varnothing \neq W \neq V$. It follows from Theorem 12 that there exists a circulation x for (V, E) satisfying $l' \leqslant x \leqslant u$. This proves the theorem. ■

In summary, we can combine Theorems 7 and 11 and Lemma 13 to derive the following result.

THEOREM 14. *Let (A, L, U) be a consistent triple of $n \times n$ matrices. Let (V, E, a) be the graph corresponding to A, and let l and u be the vectors corresponding to L and U, respectively. Then the following are equivalent:*

(i) *the truncated scaling problem for (A, L, U) has a solution;*

(ii) *there exists a balanced matrix X such that*

(a) *$L \leqslant X \leqslant U$ and*

(b) *$x_{ij} > 0$ if and only if $a_{ij} > 0$ for $i, j = 1, 2, \dots, n$;*

(iii) *for every subset $W \subset V$ for which $\delta^-(W)$ contains an edge e' with $l_{e'} = 0$, we have*

$$\sum_{e \in \delta^-(W)} l_e < \sum_{e \in \delta^+(W)} u_e; \tag{23}$$

(iv) *the infimum in (17), the dual optimization problem corresponding to (A, L, U), is attained.*

When a solution to the truncated scaling problem for (A, L, U) exists, then the uniqueness of the matrix X derived by truncating DAD^{-1} follows directly from the strict convexity of the entropy function.

6. RELATIONSHIP TO PREVIOUS RESEARCH

We first consider results for the diagonal similarity scaling problem described in Section 2.3.1. We derive the following result as an direct consequence of Theorem 14 for the case of $L = 0$ and $U = +\infty$.

COROLLARY 15 (Eaves et al. [9, Theorem 7]). *Let A be an $n \times n$ nonnegative matrix. Then the following are equivalent:*

(i) *there exists a matrix $D = \operatorname{diag}(d_1, d_2, \dots, d_n)$ with positive diagonal entries such that DAD^{-1} is balanced,*

(ii) *there is some nonnegative balanced matrix X satisfying $x_{ij} > 0$ if and only if $a_{ij} > 0$;*

(iii) *the infimum in the dual optimization problem corresponding to the triple $(A, 0, +\infty)$ is attained.*

We now show that a characterization of *completely reducible matrices* of Eaves et al. [9] also follows from Theorem 5. Let (V, E) be a directed graph, and let i and j be any two vertices of V. Then a *directed path from i to j* is a sequence of edges $\{e_1, e_2, \dots, e_p\}$ of E such that

(i) $e_k = (i_k, j_k)$ for $k = 1, 2, \dots, p$,
(ii) $i_1 = i$ and $j_p = j$, and
(iii) $j_k = i_{k+1}$ for $k = 1, 2, \dots, p-1$.

A graph (V, E) is called *completely reducible* if for every pair of vertices i and j, there exists a directed path from i to j if and only if there exists a directed path from j to i. Note, a graph is completely reducible if and only if it is the disjoint union of its strongly connected components. A matrix A is called *completely reducible* if its graph $G(A)$ is completely reducible.

It is easy to see that condition (ii) of Corollary 15 is satisfied if and only if A is completely reducible. Thus we derive the following result as a corollary of Theorem 14:

COROLLARY 16 (Eaves et al. [9, Theorem 2]). *Let A be an $n \times n$ nonnegative matrix. Then the following are equivalent:*

(i) *there exists a matrix* $D = \operatorname{diag}(d_1, d_2, \dots, d_n)$ *with positive diagonal entries such that* DAD^{-1} *is balanced;*

(ii) A *is completely reducible.*

Next, we show that results of Brualdi [7], Menon [21], Menon and Schneider [22], and Rothblum and Schneider [27] for fixed row-column equivalence scaling can be derived from Theorem 14.

First, we need some notation. Given positive integers m and n, we define the sets M and N by

$$M = \{1, 2, \dots, m\}$$

$$N = \{m+1, m+2, \dots, m+n\}.$$

We will use I and J for nonempty subsets of M and N, respectively, and I' and J' for the complements of I and J in M and N, respectively (i.e., $I' = M \setminus I$, and $J' = N \setminus J$).

Let A' be a $m \times n$ nonnegative matrix, and let I and J be nonempty subsets of M and N, respectively. We define the set $A'[I \mid J]$ by

$$A'[I \mid J] = \{(i, j) \mid i \in I, m + j \in J, \text{and } a'_{ij} > 0\}.$$

We use the symbol $\subset$ to mean strict containment.

Corollary 17. *Let* A' *be a* $m \times n$ *nonnegative matrix, and let* α *be a strictly positive vector in* $\Re^{m+n}$. *Let* (A', α) *be consistent. Then the following are equivalent:*

(a) *the equivalence scaling problem for* (A', α) *has a solution;*

(b) *there is some* $m \times n$ *nonnegative matrix* X' *with row and column sums given by* α *such that*

$$x'_{ij} > 0 \quad \Leftrightarrow \quad a'_{ij} > 0;$$

(c) *for every pair of nonempty subsets* $I \subset M$ *and* $J \subset N$ *for which* $A'[I' \mid J] = \varnothing$ *and* $A'[I \mid J'] \neq \varnothing$,

$$\sum_{i \in J} \alpha_i < \sum_{i \in I} \alpha_i;$$

(d) *the infimum is attained in the dual optimization problem for the truncated scaling problem corresponding to* (A', α).

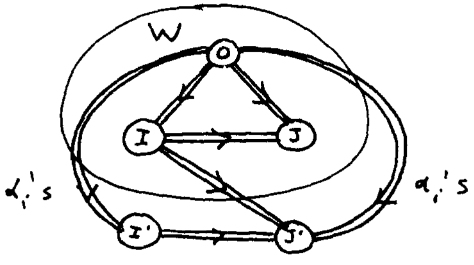

FIG. 3. Conditions for equivalence scaling.

Proof. Let (A, L, U) be the ordered triple corresponding to (A', α) as described in Section 2.3.2. The method of proof we use is to show that each part of the theorem is equivalent to the corresponding part of Theorem 14 when applied to (A, L, U).

The equivalence between (a) and (i) follows from Lemma 1. Let (V, E, a) be the graph corresponding to (A, L, U). The equivalence between (b) and (ii) follows directly from the construction of (V, E, a) described in Section 2.3.2 and the correspondence between graphs and matrices in Section 2.2.

To show the equivalence between (c) and (iii), let W be a subset of V satisfying the conditions of (iii). There are two cases depending on whether or not W contains the vertex 0. Suppose that W contains the vertex 0, and let I and J be the vertices of M and N, respectively, contained in W. (See Figure 3.) First, we observe that for any W for which $A'[I' \mid J] \neq \varnothing$ the right-hand side of (23) is $+\infty$, and it follows that we will always have a strict inequality. Thus, it suffices to consider those W for which $A'[I' \mid J] = \varnothing$.

Since α is strictly positive, the requirement that $\delta^-(W)$ contain an edge whose lower bound is 0 is equivalent to the requirement that $A'[I \mid J'] \neq \varnothing$. Thus, the condition

$$\sum_{e \in \delta^-(W)} l_e < \sum_{e \in \delta^-(W)} u_e$$

reduces to

$$\sum_{i \in I'} \alpha_i < \sum_{i \in J'} \alpha_i,$$

which is equivalent to

$$\sum_{i \in J} \alpha_i < \sum_{i \in I} \alpha_i.$$

(Recall, $\sum_{i \in M} \alpha_i = \sum_{i \in N} \alpha_i$.)

The case of W not containing vertex 0 is handled similarly. The equivalence between (c) and (iii) now follows.

The equivalence between (d) and (iv) is immediate, since they are both specializations of Theorem 7, part (iv). ■

The implication (a) ⇔ (b) of Corollary 17 is equivalent to Theorem 2 of Menon [21]. The implication (b) ⇔ (c) is essentially equivalent to theorem (2.1) of Brualdi [7]. The implication (a) ⇔ (c) is equivalent to Theorem 4.1 of Menon and Schneider [22]. The implications (a) ⇒ (b) and (b) ⇒ (c) are elementary. Therefore, the results of Menon and Brualdi follows directly from the result of Menon and Schneider. See [22, p. 333] for the details. Theorem 17 is also contained in Rothblum and Schneider [27], where the result is derived using optimization techniques and the theory of linear inequalities.

The following result of Sinkhorn is a special case of Corollary 17 and therefore is also a consequence of Theorem 14.

In a companion paper [28] we derive explicitly the dual optimization problem for truncated scaling. Further, we show that in the special cases of similarity and equivalence scaling, the dual problem either reduces to or is equivalent to optimization problems used by Bacharach [1], Bachem and Korte [2], Eaves et al. [9], Marshall and Olkin [20], and Rothblum and Schneider [27] to study matrix scaling.

COROLLARY 18 [30, Theorem 1]. *Let A be an $n \times n$ (strictly) positive matrix; then there exists a unique doubly stochastic X which can be expressed in the form $X = D_1 A D_2$ where D_1 and D_2 are positive definite diagonal matrices.*

I want to thank Hans Schneider for stimulating my interest in matrix scaling problems and encouraging the work contained in this paper. I also want to thank Mark Hartmann and Jong-Shi Pang for reading a previous version of this paper.

REFERENCES

1 M. Bacharach, *Biproportional Matrices and Input-Output Change*, Cambridge U.P., London, 1970.

2 Achim Bachem and Bernhard Korte, On the *RAS* algorithm, *Computing* 23:189–198 (1979).

3 M. L. Balinski and G. Demange, Algorithms for Proportional Matrices in Reals and Integers, Technical report, École Polytechnique, Lab. D'Econometrie, Apr. 1987; *Math. Programming*, to appear.

4 M. L. Balinski and G. Demange, An Axiomatic Approach to Proportionality between Matrices, Technical Report, S.U.N.Y., Inst. for Decision Sciences, Stony Brook, N.Y., Aug. 1987.

5 R. Bapat and T. E. S. Raghavan, An extension of a Theorem of Darroch and Ratcliff in Loglinear Models and Its Application to Scaling Multidimensional Matrices, Statistical Lab. Technical Report 87-02, Indian Statistical Inst., New Delhi, India, Mar. 1987; *Linear Algebra Appl.*, to appear.

6 L. M. Bregman, Proof of the convergence of Sheleikhovskii's method for a problem with transportion constraints, *USSR Comput. Math. and Math. Phys.* 1(1):191–204 (1967).

7 Richard A. Brualdi, Convex sets of non-negative matrices, *Canad. J. Math.* 20:144–157 (1968).

8 Yair Censor, On linearly constrained entropy maximization, *Linear Algebra Appl.* 80:191–195 (1986).

9 B. Curtis Eaves, Alan J. Hoffman, Uri G. Rothblum, and Hans Schneider, Line-sum-symmetric scalings of square nonnegative matrices, *Math. Programming Stud.* 25:124–141 (1985).

10 Tommy Elfving, On some methods for entropy maximization and matrix scaling, *Linear Algebra Appl.* 34:321–339 (1980).

11 Jan Eriksson, A note on solution of large sparse maximum entropy problems with linear equality constraints, *Math. Programming* 18:146–154 (1980).

12 Sven Erlander, Entropy in linear programs, *Math. Programming* 21:137–151 (1981).

13 Sven Erlander, *Optimal Spatial Interaction and the Gravity Model*, Lecture Notes in Econom. and Math. Systems, Vol. 173, Springer-Verlag, 1980.

14 Joel Franklin and Jens Lorenz, On the Scaling of Multidimensional Matrices, Technical Report, Applied Mathematics Dept., California Inst. of Technology, undated; *Linear Algebra Appl.*, to appear.

15 J. Grad, Matrix balancing, *Comput. J.* 14(3):280–284 (Aug. 1971).

16 Alan J. Hoffman, Some recent applications of the theory of linear inequalities to extremal combinatorial analysis, in *Proc. Sympos. Appl. Math.* Vol. 10, Amer. Math. Soc., 1960.

17 R. S. Krupp, Properties of Kruithof's projection method, *Bell System Tech. J.* 58(2):517–538 (Feb. 1979).

18 B. Lamond and N. F. Stewart, Bregman's balancing method, *Transportation Res. Part B* 15B(4):239–248 (Aug. 1981).

19 Lester R. Ford, Jr., and D. R. Fulkerson, *Flows in Networks*, Princeton U.P., Princeton, N.J., 1962.

20 A. W. Marshall and I. Olkin, Scaling of matrices to achieve specified row and column sums, *Numer. Math.* 12:83–90 (1968).

21 M. V. Menon, Matrix links, an extremisation problem and the reduction of a non-negative matrix to one with prescribed row and column sums, *Canad. J. Math.* 20:225–232 (1968).

22 M. V. Menon and Hans Schneider, The spectrum of a nonlinear operator associated with a matrix, *Linear Algebra Appl.* 2:321–334 (1969).

23 E. E. Osborne, On pre-conditioning of matrices, *J. Assoc. Comput. Mach.* 7:338–345 (1960).

24 R. Tyrrell Rockafellar, *Network Flows and Monotropic Optimization*, Wiley, 1984.

25 R. Tyrrell Rockafellar, *Convex Analysis*, Princeton U.P., Princeton, N.J., 1970.

26 Uriel G. Rothblum, Generalized Scalings Satisfying Linear Equations, Technical Report, Technion—Israel Inst. of Technology, Industrial Engineering and Management, 1988; *Linear Algebra Appl.*, to appear.

27 Uriel G. Rothblum and Hans Schneider, Scaling of Matrices Which Have Prespecified Row Sums and Column Sums via Optimization, CMS Technical Suppary Report 88-28, Center for the Mathematical Sciences, Univ. of Wisconsin —Madison, 1987; *Linear Algebra Appl.*, to appear.

28 Michael H. Schneider, Matrix Scaling, Entropy Minimization, and Conjugate Duality (II): The Dual Problem, Technical Report, Dept. of Mathematical Sciences, Johns Hopkins Univ., Aug. 1988.

29 Michael H. Schneider and Stavros Zenios, A Comparative Study of Algorithms for Matrix Balancing, OR Group Report Series 88-02, Dept. of Mathematical Sciences, Johns Hopkins Univ., 1987.

30 Richard Sinkhorn, A relationship between arbitrary positive matrices and doubly stochastic matrices, *Ann. Math. Statist.* 35:876–879 (1964).

Received 25 August 1988; final manuscript accepted 6 October 1988