Discrepancy of Matrices of Zeros and Ones

Richard A. Brualdi* and Jian Shen†
Department of Mathematics
University of Wisconsin
Madison, Wisconsin 53706

brualdi@math.wisc.edu jshen@math.wisc.edu

AMS Subject Classification: 05B20

Submitted: January 18, 1999; Accepted: February 10, 1999

Abstract

Let m and n be positive integers, and let $R = (r_1, \ldots, r_m)$ and $S = (s_1, \ldots, s_n)$ be non-negative integral vectors. Let $\mathcal{A}(R, S)$ be the set of all $m \times n$ $(0,1)$-matrices with row sum vector R and column vector S, and let $\bar{A}$ be the $m \times n$ $(0,1)$-matrix where for each i, $1 \leq i \leq m$, row i consists of r_i 1's followed by $n - r_i$ 0's. If S is monotone, the *discrepancy* $d(A)$ of A is the number of positions in which $\bar{A}$ has a 1 and A has a 0. It equals the number of 1's in $\bar{A}$ which have to be shifted in rows to obtain A. In this paper, we study the minimum and maximum $d(A)$ among all matrices $A \in \mathcal{A}(R, S)$. We completely solve the minimum discrepancy problem by giving an explicit formula in terms of R and S for it. On the other hand, the problem of finding an explicit formula for the maximum discrepancy turns out to be very difficult. Instead, we find an algorithm to compute the maximum discrepancy.

*Partially supported by NSF Grant DMS-9424346.
†Supported by an NSERC Postdoctoral Fellowship.

1 Introduction

Let m and n be positive integers, and let $R = (r_1, \ldots, r_m)$ and $S = (s_1, \ldots, s_n)$ be non-negative integral vectors. The vector R is called *monotone* if $r_1 \geq \cdots \geq r_m$. Let $\mathcal{A}(R,S)$ be the set of all $m \times n$ $(0,1)$-matrices with row sum vector R and column vector S, and let $\bar{A}$ be the $m \times n$ $(0,1)$-matrix where for each i, $1 \leq i \leq m$, row i consists of r_i 1's followed by $n - r_i$ 0's. Let the column sum vector of $\bar{A}$ be $R^* = (r_1^*, \ldots, r_n^*)$. It follows that R^* is monotone and

$$r_j^* = |\{i : r_i \geq j, i = 1, \ldots, m\}| \text{ for } j = 1, \ldots, n.$$

R and R^* are called *conjugate partitions* of $\tau = r_1 + \cdots + r_m = r_1^* + \cdots + r_n^*$.

Let $S = (s_1, \ldots, s_n)$ and $T = (t_1, \ldots, t_n)$ be two non-negative integral vectors. For convenience, we write

$$|T - S| := \sum_{i=1}^{n} \max\{0, t_i - s_i\}.$$

(Notice that $|T - S|$ is, in general, not equal to $|S - T|$.) In particular,

$$|T| := \sum_{i=1}^{n} t_i.$$

The vector S is said to be *majorized* by T, written $S \prec T$, if

$$\sum_{i=1}^{j} s_i \leq \sum_{i=1}^{j} t_i \text{ for all } j = 1, 2, \ldots, n$$

with equality when $j = n$. We emphasis here that we do not assume the monotone properties of S and T in our definition of majorization throughout the paper. This generalizes the traditional definition of majorization in the literature. To avoid any ambiguity, we will specify in each of the lemmas and theorems which vectors are assumed to be monotone.

The set $\mathcal{A}(R,S)$ was the subject of intensive study during the late 1950s and early 1960s by many researchers. (See [1] for a survey paper.) For example, the following lemma of Gale-Ryser stated the conditions for the existence of a matrix in $\mathcal{A}(R,S)$. It was originally stated under the condition that both R and S were monotone. It is clear that the monotone property of R can be dropped from the lemma since any reordering of rows in a matrix in $\mathcal{A}(R,S)$ does not affect the vectors R^* and S.

Lemma 1 (Gale [3], Ryser [4]) *Suppose S is monotone. Then $\mathcal{A}(R,S) \neq \emptyset$ if and only if $S \prec R^*$ and $r_i \leq n$ for all $i = 1, \ldots, m$.*

If $\mathcal{A}(R,S) \neq \emptyset$, then each $A \in \mathcal{A}(R,S)$ can be obtained from $\bar{A}$ by shifting 1's in each row. Throughout the paper, by shifting 1's we always means shifting 1's to the right. If S is monotone, Brualdi and Sanderson [2] defined the *discrepancy* $d(A)$ of A to be the number of positions in which $\bar{A}$ has a 1 and A has a 0. It equals the number of 1's in $\bar{A}$ which have to be shifted to obtain A. We are interested in the discrepancy set $\{d(A) : A \in \mathcal{A}(R,S)\}$. Let

$$\tilde{d} = \tilde{d}(R,S) = \min\{d(A) : A \in \mathcal{A}(R,S)\}$$

and

$$\bar{d} = \bar{d}(R,S) = \max\{d(A) : A \in \mathcal{A}(R,S)\}.$$

In 1957, Ryser [4] defined an *interchange* to be a transformation which replaces the 2×2 submatrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

of a matrix A of 0's and 1's with the 2×2 submatrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

or vice versa. Clearly an interchange (and hence any sequence of interchanges) does not alter the row and column sum vectors of a matrix, and therefore transforms a matrix in $\mathcal{A}(R,S)$ into another matrix in $\mathcal{A}(R,S)$. Ryser [4] proved the converse of the result by inductively showing that given $A, B \in \mathcal{A}(R,S)$ there is a sequence of interchanges which transforms A into B. In particular, if $d(A) = \tilde{d}$ and $d(B) = \bar{d}$, then there is a sequence of interchanges which transforms A into B. Thus for each integer d with $\tilde{d} \leq d \leq \bar{d}$, there is a matrix in $\mathcal{A}(R,S)$ having discrepancy d, since an interchange can only change the discrepancy of a matrix by at most 1. Therefore

$$\{d(A) : A \in \mathcal{A}(R,S)\} = \{d : \tilde{d} \leq d \leq \bar{d}\};$$

in other words, to determine the discrepancy set $\{d(A) : A \in \mathcal{A}(R,S)\}$, it suffices to determine the minimum and maximum discrepancies among all matrices in $\mathcal{A}(R,S)$. Since $d(A)$ is defined under the assumption that S is monotone, we assume that S is monotone throughout the rest of the paper.

In Section 2, we show that the minimum discrepancy of all matrices in $\mathcal{A}(R,S)$ is $|R^* - S|$. On the other hand, the problem of finding an explicit formula for the maximum discrepancy turns out to be very difficult. We find an algorithm to compute the maximum discrepancy in Section 3.

2 Minimum Discrepancy

We prove in this section an explicit formula in terms of R and S for the minimum discrepancy of all matrices in $\mathcal{A}(R,S)$. We begin with the following lemma.

Lemma 2 *Suppose $S=(s_1,\ldots,s_n)$ and $T=(t_1,\ldots,t_n)$ are monotone vectors such that $S\prec T$. Then there exist $k=|T-S|+1$ monotone vectors $S_i=(s_1^{(i)},\ldots,s_n^{(i)})$, $1\le i\le k$, such that*

1. *$S=S_1\prec S_2\prec\cdots\prec S_k=T$, and*
2. *$|S_{i+1}-S_i|=1$ for all $1\le i\le k-1$.*

Proof. Set $S_1=S$ and $S_k=T$. Lemma 2 is trivial if $k\le 2$. Now suppose $k\ge 3$. Since $S\ne T$, there exists a smallest index l_0 satisfying $s_{l_0}>t_{l_0}$. If $l_0\le n-1$, then either $s_{l_0}>s_{l_0+1}$ or $s_{l_0+1}=s_{l_0}>t_{l_0}\ge t_{l_0+1}$. Thus there exists a smallest index l_1 satisfying $s_{l_1}>t_{l_1}$, and satisfying $s_{l_1}>s_{l_1+1}$ if $l_1\le n-1$. Thus $l_0\le l_1$ and

$$s_i\begin{cases} >t_i & \text{if } l_0\le i\le l_1,\\ \le t_i & \text{if } i\le l_0-1.\end{cases}$$

Since $S\prec T$, we have $s_1\le t_1$ and $l_0>1$. Let l_2 be the smallest index i satisfying $1\le i<l_0$ and $s_i<t_i$. (Such an i exists since $S\prec T$ and $S\ne T$.) Since $S\prec T$,

$$\sum_{i=1}^{l_2}t_i=\sum_{i=1}^{l_2-1}t_i+t_{l_2}>\sum_{i=1}^{l_2-1}s_i+s_{l_2}=\sum_{i=1}^{l_2}s_i.$$

Let S_2 be defined by

$$s_j^{(2)}=\begin{cases} s_j-1 & \text{if } j=l_1,\\ s_j+1 & \text{if } j=l_2,\\ s_j & \text{otherwise.}\end{cases}$$

Thus, for all l such that $l_2\le l\le l_0-1$,

$$\sum_{i=1}^{l}t_i=\sum_{i=1}^{l_2}t_i+\sum_{i=l_2+1}^{l}t_i>\sum_{i=1}^{l_2}s_i+\sum_{i=l_2+1}^{l}s_i=\sum_{i=1}^{l}s_i \tag{1}$$

and, for all l such that $l_0\le l\le l_1-1$,

$$\sum_{i=1}^{l}t_i=\sum_{i=1}^{l_1}t_i-\sum_{i=l+1}^{l_1}t_i>\sum_{i=1}^{l_1}s_i-\sum_{i=l+1}^{l_1}s_i=\sum_{i=1}^{l}s_i. \tag{2}$$

Since $S \prec T$, it follows from (1) and (2) that $S_2 \prec T$. By the choices of l_1 and l_2, we have

$$s^{(2)}_{l_1} = s_{l_1} - 1 \geq s_{l_1+1} = s^{(2)}_{l_1+1} \text{ if } l_1 + 1 \leq n,$$

and

$$s^{(2)}_{l_2-1} = s_{l_2-1} = t_{l_2-1} \geq t_{l_2} \geq s_{l_2} + 1 = s^{(2)}_{l_2} \text{ if } l_2 - 1 \geq 1.$$

Thus S_2 is monotone. Also it can be checked that $S_1 \prec S_2$, $|S_2 - S_1| = 1$ and $|T - S_2| = k - 2$. By replacing S_2 with S, Lemma 2 follows by induction. □

Theorem 1 *Suppose S_1 and S_2 are monotone vectors such that $S_1 \prec S_2$. If $A \in \mathcal{A}(R, S_2)$, then a matrix in $\mathcal{A}(R, S_1)$ can be obtained from A by shifting at most $|S_2 - S_1|$ $1's$ in rows.*

Proof. By Lemma 2, it may be supposed that $|S_2 - S_1| = 1$. Since $S_1 \prec S_2$, there are l_1, l_2 such that $l_2 < l_1$ and

$$s^{(2)}_j = \begin{cases} s^{(1)}_j - 1 & \text{if } j = l_1, \\ s^{(1)}_j + 1 & \text{if } j = l_2, \\ s^{(1)}_j & \text{otherwise.} \end{cases}$$

Thus $s^{(2)}_{l_2} = s^{(1)}_{l_2} + 1 \geq s^{(1)}_{l_1} + 1 = s^{(2)}_{l_1} + 2$; in other words, column l_2 of A contains at least 2 more 1's than column l_1 of A. Thus a 1 can be shifted in a row from column l_2 to column l_1, and so a matrix in $\mathcal{A}(R, S_1)$ is obtained. □

Corollary 1 *Suppose S is monotone. If $\mathcal{A}(R, S) \neq \emptyset$, then*

$$\min_{A \in \mathcal{A}(R,S)} d(A) = |R^* - S|.$$

Proof. Since $\mathcal{A}(R, S) \neq \emptyset$, we have $S \prec R^*$ by Lemma 1. Suppose that $A \in \mathcal{A}(R, S)$. Since columns i of $\bar{A}$ and A have r^*_i and s_i 1's, respectively, at least $\max\{0, r^*_i - s_i\}$ 1's in column i must be shifted in rows in order to obtain A from $\bar{A}$. This implies $d(A) \geq \sum_i \max\{0, r^*_i - s_i\} = |R^* - S|$. On the other hand, by applying Theorem 1 in the case $S_1 = S$ and $S_2 = R^*$, a matrix in $\mathcal{A}(R, S)$ can be obtained from $\bar{A} \in \mathcal{A}(R, R^*)$ by shifting at most $|R^* - S|$ $1's$ in rows; that is, $\min_{A \in \mathcal{A}(R,S)} d(A) \leq |R^* - S|$, from which Corollary 1 follows. □

3 Maximum Discrepancy

In this section, we find an algorithm to compute the maximum discrepancy of all matrices in $\mathcal{A}(R,S)$. We begin with the following lemma which will be used in Lemma 4. We comment here that Lemma 3, under the weaker condition that only S is monotone, is weaker than Theorem 1.

Lemma 3 *Suppose S is monotone and $A \in \mathcal{A}(R,T)$. If $S \prec T$, then $\mathcal{A}(R,S) \neq \emptyset$ and some matrix in $\mathcal{A}(R,S)$ can be obtained from A by shifting 1's in rows.*

Proof. We use induction on n, the number of components of S. If T is monotone, then the lemma follows from Theorem 1; in particular, the lemma holds for $n = 2$, since $n = 2$, S monotone and $S \prec T$ imply that T is monotone. We now assume that $n \geq 3$ and T is not monotone, and proceed by induction on n. We define $S' = (s'_1, \ldots, s'_n)$ to be a maximal monotone vector in the sense of majorization satisfying

$$S \prec S' \prec T$$

By the choice of S', there is an l, $1 \leq l < n$, such that $\sum_{i=1}^{l} s'_i = \sum_{i=1}^{l} t_i$. We can partition S' and T such that $S' = (S'_1, S'_2)$ and $T = (T_1, T_2)$, where S'_1, T_1 are vectors with l components and S'_2, T_2 are vectors with $n-l$ components. It can be seen that $S'_1 \prec T_1$ and $S'_2 \prec T_2$ since $\sum_{i=1}^{l} s'_i = \sum_{i=1}^{l} t_i$. Also S'_1 and S'_2 are monotone since $S' = (S'_1, S'_2)$ is.

Now we consider the partition of $A = [A_1\ A_2]$, where A_1 and A_2 are $m \times l$ and $m \times (n-l)$ matrices, respectively. Then $A_1 \in \mathcal{A}(R_1, T_1)$ and $A_2 \in \mathcal{A}(R_2, T_2)$ for some R_1 and R_2 satisfying $R_1 + R_2 = R$. By the induction hypothesis, some matrices $B_1 \in \mathcal{A}(R_1, S'_1)$ and $B_2 \in \mathcal{A}(R_2, S'_2)$ can be obtained by shifting 1's in rows from A_1 and A_2, respectively. Then the matrix $[B_1\ B_2] \in \mathcal{A}(R, S')$ can be obtained from $[A_1\ A_2] = A$ by shifting 1's in rows. Since $S \prec S'$ and S' is monotone, by Theorem 1, some matrix in $\mathcal{A}(R,S)$ can be obtained by shifting 1's in rows from $[B_1\ B_2]$ and so from A. This completes the proof of the lemma. $\square$

Suppose S is monotone. For each $A \in \mathcal{A}(R,S)$, we can partition A into two regions according to the shape of $\bar{A}$; that is, region 1 consists of positions in $\{(i,j) : 1 \leq i \leq m, 1 \leq j \leq r_i\}$, while region 2 consists of positions in $\{(i,j) : 1 \leq i \leq m, r_i < j \leq n\}$.

Suppose $R^{(1)} = (r_1^{(1)}, \ldots, r_m^{(1)})$, $R^{(2)} = (r_1^{(2)}, \ldots, r_m^{(2)})$ are two non-negative integral vectors such that $r_i^{(1)} \leq r_i$ and $r_i^{(2)} \leq n - r_i$ for all i, $1 \leq i \leq m$. Define $\bar{A}(R^{(1)}, R^{(2)}) =$

(a_{ij}) to be the $m \times n$ matrix defined, for each i, by

$$a_{ij} = \begin{cases} 1 & \text{if } 1 \le j \le r_i^{(1)} \text{ or } r_i + 1 \le j \le r_i + r_i^{(2)}, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $\bar{A}(R^{(1)}, R^{(2)})$ is the matrix with row sum vectors $R^{(i)}$ in region i, $i = 1, 2$, and with all 1's in the leftmost possible positions. Let $(R^{(1)}, R^{(2)})^*$ denote the column sum vector of $\bar{A}(R^{(1)}, R^{(2)})$. If $R^{(1)} = O$, a zero vector, and $R^{(2)} = R$, then $\bar{A}(R^{(1)}, R^{(2)})$ is the matrix $\bar{A}(O, R)$, and $(O, R)^*$ is the column sum vector of $\bar{A}(O, R)$. Let

$$\mathcal{J} = \mathcal{J}(R, S) := \{\bar{A}(R^{(1)}, R^{(2)}) : R^{(1)} + R^{(2)} = R \text{ and } S \prec (R^{(1)}, R^{(2)})^*\}.$$

Lemma 4 *Suppose S is monotone. Then*

$$\max_{A \in \mathcal{A}(R,S)} d(A) = \max_{\bar{A}(R-T,T) \in \mathcal{J}} |T|.$$

Proof. Let $A \in \mathcal{A}(R, S)$ with maximum $d(A)$. Let B be the matrix obtained from A by moving all 1's in rows to the leftmost possible positions within each of the two regions. Then the column sum vector of B majorizes S and so $B \in \mathcal{J}$. Let $B = \bar{A}(R - T_A, T_A)$. Then $d(A) = |T_A|$. This implies that

$$\max_{A \in \mathcal{A}(R,S)} d(A) \le \max_{\bar{A}(R-T,T) \in \mathcal{J}} |T|.$$

Now suppose that $B = \bar{A}(R - T, T) \in \mathcal{J}$ has maximum $|T|$ among all matrices in $\mathcal{J}$. Since $S \prec (R - T, T)^*$, by Lemma 3, some matrix $A \in \mathcal{A}(R, S)$ can be obtained from B by shifting 1's in rows. Since shifting 1's in rows does not decrease the number of 1's in region 2 (recall that shifting 1's means shifting 1's to the right), we have $|T| \le d(A)$. Thus

$$\max_{\bar{A}(R-T,T) \in \mathcal{J}} |T| \le \max_{A \in \mathcal{A}(R,S)} d(A),$$

from which Lemma 4 follows. □

For two vectors $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_n)$, we define $U < V$ in the sense of lexicography; that is, there is some j such that $u_j < v_j$ and $u_i = v_i$ for all $i < j$. Similarly, we can define $U \le V$ in the sense of lexicography; that is, either $U = V$ or $U < V$ holds.

Throughout the rest of the section, we select $C := \bar{A}(R - U, U) \in \mathcal{J}$ with priority in the order: (1.) $(O, U)^*$ is lexically maximum, (2.) maximal $(R - U, U)^*$ in the sense of majorization. In other words, among all candidates $\bar{A}(R-U,U)$ with the property that $(O,U)^*$ is lexically maximum, we select C with maximal $(R-U,U)^*$ in the sense of majorization. We also select $D := \bar{A}(R-V,V) \in \mathcal{J}$ with priority in the order: (1.) maximum $|V|$, (2.) $(O,V)^*$ is lexically maximum, (3.) maximal $(R-V,V)^*$ in the sense of majorization.

Now we focus on the structure of C and D. It is known that C, D can be obtained from $\bar{A}$ by shifting 1's in rows. We may assume the following rule when shifting 1's in rows to obtain C, D from $\bar{A}$:

Shifting Rule: For each i, let (i,j_i) be the rightmost position having a 1 in row i in region 1, and let (i,k_i) be the leftmost position having a 0 in row i in region 2. If a shift takes place in row i, then the 1 at the (i,j_i) position is moved to the (i,k_i) position.

It is trivial that every matrix in $\mathcal{J}$ can be obtained from $\bar{A}$ by a sequence of 0-1 shifts satisfying the above Shifting Rule. For each position (i,j) in region 2 (thus $j \geq r_i+1$), we assign to it a weight $w(i,j)$ as follow:

$$w(i,j) = \begin{cases} 2j-2r_i-1 & \text{if } r_i+1 \leq j \leq 2r_i, \\ \infty & \text{if } 2r_i+1 \leq j \leq n, \end{cases}$$

Indeed, it can be checked that $w(i,j)$ is the distance that a 1 has to be moved from region 1 to the position (i,j) in region 2 by the Shifting Rule. (In the case that $2r_i+1 \leq j \leq n$, the (i,j) position must have a 0 for any matrix in $\mathcal{J}$. Thus it is natural to define the distance that a 1 has to be moved from region 1 to the position (i,j) as infinity.)

Lemma 5 *Both matrices C and D satisfy the following: For each fixed j, the 1's in column j that lie in region 2 appear in the positions (i,j) with $w(i,j)$ as small as possible.*

Proof. We only prove the lemma for $C=(c_{ij})$. A similar proof works for D. Suppose the lemma fails for C. Then there are i,j,k such that (i,j), (k,j) are in region 2, and $c_{ij}=1$, $c_{kj}=0$ and $w(i,j) > w(k,j)$. By the Shifting Rule, the positions $(i,j-w(i,j))$ and $(k,j-w(k,j))$ have a 0 and a 1, respectively. Let C_1 be obtained from C by making 0-1 switches at the four positions $(i,j),(i,j-w(i,j)),(k,j),(k,j-w(k,j))$. Then the column sum vector of C_1 majorizes $(R-U,U)^*$ since $j-w(i,j) < j-w(k,j)$. Let $C_2=\bar{A}(R-U_1,U_1)$ be obtained from C_1 by moving all 1's in rows within each of the two regions to the leftmost possible positions. Then $|U|=|U_1|$ and $(O,U)^* \leq (O,U_1)^*$. Also $(R-U,U)^* \prec (R-U_1,U_1)^*$ since $(R-U_1,U_1)^*$ majorizes the column sum vector of C_1. Thus $C \neq C_2 \in \mathcal{J}$. This contradicts the choice of C. $\square$

Theorem 2

$$|U| = |V|.$$

Proof. Let $D = (d_{ij})$. By the choice of D, we have $|U| \leq |V|$. Now suppose $|U| < |V|$. Let $(O, U)^* = (u_1, \ldots, u_n)$ and $(O, V)^* = (v_1, \ldots, v_n)$. Since $(O, U)^*$ is lexically maximum, there is a j such that $u_j > v_j$ and $u_i = v_i$ for all $i \leq j - 1$. Let

$$\mathcal{P} := \{\text{positions } (i, k) \text{ in region } 2 : d_{ik} = 1 \text{ and } k \leq j\}.$$

By Lemma 5, we may properly choose the matrix C such that $c_{ik} = 1$ whenever $(i, k) \in \mathcal{P}$. Since $u_j > v_j$, there is a position (i, j) in region 2 such that $c_{ij} = 1$ and $d_{ij} = 0$. Let $k = j - w(i, j)$. Then $c_{ik} = 0$ and $d_{ik} = 1$ by the Shifting Rule. Let $(R - U, U)^* = (c_1^*, \ldots, c_n^*)$, $(R - V, V)^* = (d_1^*, \ldots, d_n^*)$.

Claim 1: There is some l, $k \leq l < j$, such that

$$\sum_{t=1}^{l} s_t = \sum_{t=1}^{l} d_t^*.$$

Proof of Claim 1: Otherwise $\sum_{t=1}^{l} s_t < \sum_{t=1}^{l} d_t^*$ for all l, $k \leq l < j$, since $S \prec (R - V, V)^*$. Let D_1 be obtained from D by making a 0-1 switch at positions (i, j) and (i, k). Then the number of 1's that lie in region 2 in D_1 is $|V| + 1$. Since $S \prec (R - V, V)^*$, it can be checked that S is majorized by the column sum vector of D_1. By moving all 1's in rows within each of the two regions to the leftmost possible positions in D_1, we can obtain a matrix in $\mathcal{J}$ contradicting the choice of D with maximum $|V|$. Thus Claim 1 holds.

Now we may choose l to be the smallest index satisfying Claim 1.

Claim 2: There exists in region 2 a position $(i', j') \notin \mathcal{P}$ such that $d_{i'j'} = 1$ and $d_{i'k'} = 0$ with $k' = j' - w(i', j') \leq l$.

Proof of Claim 2: Otherwise no 1 with column index less than or equal to l is shifted in a row to a position outside of $\mathcal{P}$ in D. But in C, the 1 in the (i, k) position is shifted in row i to the (i, j) position which is outside of $\mathcal{P}$. Thus $\sum_{t=1}^{l} s_t = \sum_{t=1}^{l} d_t^* > \sum_{t=1}^{l} c_t^*$, which contradicts $S \prec (R - U, U)^*$. Thus Claim 2 holds.

Since $d_{i'j'} = 1$, by the definition of $\mathcal{P}$, we have $j' > j$. Let D_2 be obtained from D by making 0-1 switches at positions $(i, j), (i, k), (i', j')$ and (i', k'). Let $D_3 =$

$\bar{A}(R - V_3, V_3)$ be obtained from D_2 by moving all 1's in rows within each of the two regions to the leftmost possible positions. Then $|V| = |V_3|$ and $(O, V)^* < (O, V_3)^*$.

Case 1: $k' \le k$. Then it is easy to see that $(R - V, V)^* \prec (R - V_3, V_3)^*$ since $j' > j$. Thus $S \prec (R - V_3, V_3)^*$ since $S \prec (R - V, V)^*$.

Case 2: $k < k' \le l$. Let $(R - V_3, V_3)^* = (e_1^*, \ldots, e_n^*)$. Since l is the smallest index satisfying Claim 1,

$$\sum_{t=1}^{l'} s_t \le \sum_{t=1}^{l'} d_t^* - 1 \le \sum_{t=1}^{l'} e_t^*$$

for all $l', k \le l' < l$. Then it can be verified that $S \prec (R - V_3, V_3)^*$ since $j' > j$.

Since $S \prec (R - V_3, V_3)^*$ is always true in both cases above, we have $D_3 \in \mathcal{J}$. This contradicts the choice of D since $|V| = |V_3|$ and $(O, V)^* \prec (O, V_3)^*$. This completes the proof of $|U| \ge |V|$. Therefore $|U| = |V|$. □

By Lemma 4 and Theorem 2, we have the following Corollary.

Corollary 2 *Suppose S is monotone. Then*

$$\max_{A \in \mathcal{A}(R,S)} d(A) = |U|.$$

Since $(O, U)^*$ is lexically maximum, we can use the following greedy algorithm to construct a $C = \bar{A}(R - U, U)$. By Corollary 2, this yields an algorithm to compute $\max_{A \in \mathcal{A}(R,S)} d(A)$.

Algorithm to construct a matrix $C = \bar{A}(R - U, U) \in \mathcal{A}(R, S)$ with $d(C) = \bar{d}(R, S)$:

Begin with the matrix $\bar{A}$ with row sum vector R.

1. Let j be the smallest index i such that column i has a non-empty intersection with region 2.

2. Apply the Shifting Rule to shift a 1 to the position (i, j) in region 2 with the smallest weight $w(i, j)$ among all positions in column j that lie in region 2 and contain a 0, under the condition that the column sum vector of the ending matrix majorizes S. If more than one shift is possible, arbitrarily choose one.

3. Repeat Step 2, shifting to the positions in column j in region 2 as many 1's as possible. If no more shifts are possible, then go to Step 4.

4. $j := j + 1$.

5. If $j \leq n$, then go back to Step 2; otherwise, output the current matrix.

4 Concluding Discussion

We may generalize the minimum and maximum discrepancy problems by allowing regions 1 and 2 to have a general shape not necessarily determined by the shape of $\bar{A}$. For example, if we only assume that regions 1 and 2 satisfy the following:

1. Region i is connected for each $i = 1, 2$, and

2. The intersection of each row of A with region i is connected for each $i = 1, 2$,

and define, for each $A \in \mathcal{A}(R, S)$, the discrepancy $d(A)$ of A to be the number of 1's of A in region 2, then we have the following

Generalized Problems: Suppose S is monotone. For any two regions satisfying the above conditions, find

$$\min_{A \in \mathcal{A}(R,S)} d(A) \quad \text{and} \quad \max_{A \in \mathcal{A}(R,S)} d(A).$$

The above two generalized problems are equally difficult since a matrix $A \in \mathcal{A}(R, S)$ having the maximum number of 1's in region 2 clearly has the minimum number of 1's in region 1. By slightly modifying our techniques in Section 3, we can give similar algorithms to compute the minimum and maximum discrepancies. However, we believe that to give explicit formulas for the minimum and maximum discrepancies is almost hopeless for the general case.

Acknowledgment. We are grateful to a referee for pointing out some small mistakes in our original manuscript and for providing us with a number of helpful suggestions leading to a clearer presentation of the paper.

References

[1] R. A. Brualdi, Matrices of zeros and ones with fixed row and column sum vectors, *Linear Algebra Appli.* **33**:159-231 (1980).

[2] R. A. Brualdi and J. G. Sanderson, Nested species subsets, gaps, and discrepancy, *Oecologia*, to appear.

[3] D. Gale, A theorem on flows in networks, *Pacific J. Math.* **7**:1073-1082 (1957).

[4] H. J. Ryser, Combinatorial properties of matrices of zeros and ones, *Canada. J. Math.* **9**:371-377 (1957).