Computing the Nearest Doubly Stochastic Matrix with A Prescribed Entry *

Zheng-Jian Bai † Delin Chu ‡ Roger C. E. Tan §

June 2006

Abstract

In this paper a nearest doubly stochastic matrix problem is studied. This problem is to find the closest doubly stochastic matrix with the prescribed $(1, 1)$ entry to a given matrix. According to the well-established dual theory in optimization, the dual of the underlying problem is an unconstrained differentiable but not twice differentiable convex optimization problem. A Newton-type method is used for solving the associated dual problem and then the desired nearest doubly stochastic matrix is obtained. Under some mild assumptions, the quadratic convergence of the proposed Newton's method is proved. The numerical performance of the method is also demonstrated by numerical examples.

Keywords. Doubly stochastic matrix, generalized Jacobian, Newton's method, quadratic convergence.

1 Introduction

A matrix $A \in \mathbb{R}^{n\times n}$ is called *doubly stochastic* if it is non-negative and all its row and column sums equal to one. Doubly stochastic matrices have found many important applications in probability and statistics, quantum mechanics, the study of hypergroups, economics and operation research, physical chemistry, communication theory and graph theory, etc., see [3, 5, 14, 15, 22] and the references therein.

In this paper, we are interested in the best approximation problem related to doubly stochastic matrices: Given a matrix $T \in \mathbb{R}^{n\times n}$, find its nearest doubly stochastic matrix with the same $(1, 1)$ entry as the given matrix T. This problem can be mathematically stated as follows:

$$\left\{\begin{array}{ll} \min & \frac{1}{2}\|M - T\|_F^2 \\ \text{s.t.} & M\mathbf{e} = \mathbf{e}, \quad \mathbf{e}^T M = \mathbf{e}^T, \\ & \mathbf{e}_1^T M\,\mathbf{e}_1 = \mathbf{e}_1^T T\mathbf{e}_1, \\ & M \geq 0, \end{array}\right. \tag{1}$$

where

$$\mathbf{e} = (1, \dots, 1)^T \in \mathbb{R}^n, \quad \mathbf{e}_1 = (1, 0, \dots, 0)^T \in \mathbb{R}^n,$$

*This Work was partially supported by Research Grant R-146-000-047-112 of National University of Singapore and Research Grant 0000-X07152 of Xiamen University

†Department of Information and Computational Mathematics, Xiamen University, Xiamen 361005, P.R. China. Email: zjbai@xmu.edu.cn

‡Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543. Email: matchudl@nus.edu.sg

§Department of Mathematics, National University of Singapore, 2 Science Drive 2, ingapore 117543. Email: scitance@nus.edu.sg

and $M \geq 0$ means that M is non-negative. Problem (1) was originally suggested by Professor Zhaojun Bai (Department of Computer Science, UC Davis). It arose from numerical simulation of large (semi-conductor, electronic) circuit networks. Padé approximation technique using the Lanczos process is very powerful for computing a lower order approximation to the linear system matrix describing the large linear network [1, 2]. The matrix T produced by the Lanczos process is in general not a doubly stochastic matrix. Suppose the original system matrix is doubly stochastic, then we need to find the nearest doubly stochastic matrix M to T and at the same time match the moments.

Problem (1) has been studied in [8] based on alterating projection method [4]. In [8, 11], Problem (1) is simplified by removing the requirements on the $(1,1)$ entry and the non-negativity of the matrix M. In this case, the solution can be obtained explicitly. We will revisit Problem (1). Based on the dual approach in optimization [13], we will first reformulate (1) as an unconstrained differentiable but not twice differentiable convex optimization problem, next apply Newton's method to solve this convex problem, and then obtain the desired nearest doubly stochastic matrix. Under some mild assumptions, we will show that the proposed Newton's method is quadratically convergent. We will also demonstrate the numerical performance of the method by numerical examples.

Throughout this paper, the following notation will be used:

- $$T = \begin{bmatrix} t_{1,1} & \cdots & t_{1,n} \\ \vdots & \ldots & \vdots \\ t_{n,1} & \cdots & t_{n,n} \end{bmatrix}.$$
- $A \geq 0$ $(A > 0)$ means that A is non-negative (positive).
- $$\mathcal{K} = \{A :\ A \in \mathbb{R}^{n\times n},\ A \geq 0\}, \quad (z)_+ = \max\{0, z\}.$$
- $\Pi_{\mathcal{K}}(X)$ denotes the metric projection of X onto $\mathcal{K}$, i.e.,

$$\Pi_{\mathcal{K}}(X) = \begin{bmatrix} (x_{1,1})_+ & \cdots & (x_{1,n})_+ \\ \vdots & \ldots & \vdots \\ (x_{n,1})_+ & \cdots & (x_{n,n})_+ \end{bmatrix}, \quad \forall\ X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ldots & \vdots \\ x_{n,1} & \cdots & x_{n,n} \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

2 Newton's Method

In this section we consider a Newton-type method for computing the solution of Problem (1).

Let

$$f(M) := \frac{1}{2}\|M - T\|_F^2, \quad \mathcal{A}(M) := \begin{bmatrix} M\mathbf{e} \\ \left[\ I_{n-1} \quad 0\ \right] M^T\mathbf{e} \\ \mathbf{e}_1^T M \mathbf{e}_1 \end{bmatrix}, \quad b := \begin{bmatrix} \mathbf{e} \\ \left[\ I_{n-1} \quad 0\ \right] \mathbf{e} \\ \mathbf{e}_1^T T \mathbf{e}_1 \end{bmatrix},$$

then Problem (1) is equivalent to

$$\begin{cases} \min & f(M) \\ \text{s.t.} & \mathcal{A}(M) = b \\ & M \in \mathcal{K} \end{cases} \tag{2}$$

The dual problem [13] of (2) is

$$\begin{cases} \sup & -\theta(x) \\ \text{s.t.} & x \in \mathbb{R}^{2n}, \end{cases} \tag{3}$$

where

$$\theta(x) = \frac{1}{2}\|\Pi_{\mathcal{K}}(T + \mathcal{A}^\star(x))\|_F^2 - x^T b - \frac{1}{2}\|T\|_F^2,$$

and $\mathcal{A}^\star$ is the adjoint of A and is defined by

$$\begin{aligned}\mathcal{A}^\star(x) &= \begin{bmatrix} I_n & 0 \end{bmatrix} x\mathbf{e}^T + \mathbf{e}x^T \begin{bmatrix} 0_{n\times(n-1)} & 0 \\ I_{n-1} & 0 \\ 0 & 0 \end{bmatrix} + \mathbf{e}_1 \begin{bmatrix} 0 & 1 \end{bmatrix} x\mathbf{e}_1^T \\ &= \begin{bmatrix} x_1 + x_{n+1} + x_{2n} & x_1 + x_{n+2} & \cdots & x_1 + x_{2n-1} & x_1 \\ x_2 + x_{n+1} & x_2 + x_{n+2} & \cdots & x_2 + x_{2n-1} & x_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_n + x_{n+1} & x_n + x_{n+2} & \cdots & x_n + x_{2n-1} & x_n \end{bmatrix}, \quad \forall x = \begin{bmatrix} x_1 \\ \vdots \\ x_{2n} \end{bmatrix} \in \mathbb{R}^{2n}.\end{aligned}$$

The relation between the values of (2) at its minimum and of the dual (3) at its maximum is stated in the following theorem.

Theorem 1 *There exists a matrix $M \in \mathbb{R}^{n\times n}$ in the topological interior of $\mathcal{K}$ such that $\mathcal{A}(M) = b$, if and only if*

$$0 < \mathbf{e}_1^T T \mathbf{e}_1 < 1. \tag{4}$$

Under the condition (4),

(i) Problem (2) has a unique solution, denoted by $M^\star$;

(ii) The supremum of dual problem (3) is actually a maximum. Let this maximum be achieved at $x^\star$. Then

$$M^\star = \Pi_{\mathcal{K}}(T + \mathcal{A}^\star(x^\star)) \tag{5}$$

Proof. If M is in the topological interior of $\mathcal{K}$ and $\mathcal{A}(M) = b$, then (4) follows directly from the properties that

$$\mathbf{e}_1^T T \mathbf{e}_1 = \mathbf{e}_1^T M \mathbf{e}_1 > 0, \quad \mathbf{e}_1^T M \mathbf{e} = 1,$$

and all entires of $\mathbf{e}_1^T M$ are positive. Conversely, if (4) holds, then it is clear that the matrix M defined by

$$M := \frac{1}{n-1} \begin{bmatrix} r_0 & r & \cdots & r & r \\ r & r_0 & \ddots & & r \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ r & & \ddots & \ddots & r \\ r & r & \cdots & r & r_0 \end{bmatrix} \text{ with } r_0 = (n-1)\,\mathbf{e}_1^T T \mathbf{e}_1 > 0,\ r = 1 - \mathbf{e}_1^T T \mathbf{e}_1 > 0 \tag{6}$$

satisfies that M is in the topological interior of $\mathcal{K}$ and $\mathcal{A}(M) = b$. Hence Theorem 1 follows.

Under the condition (4), Parts (i) and (ii) are now well-known, see [10, 13]. □

Remark 1 *In [10], the condition ensuring that there exists a matrix $M \in \mathbb{R}^{n\times n}$ in the topological interior of $\mathcal{K}$ such that $\mathcal{A}(M) = b$ is called the the Slater condition for 2. Hence, we can regard (4) as the Slater condition for (2).*

According to Theorem 1, once we can compute an optimal solution $x^\star$ of the dual problem (3), then we can obtain the optimal solution $M^\star$ of Problem (2) by using (5).

Define

$$
\begin{aligned}
F(x): &= \mathcal{A}(\Pi_{\mathcal{K}}(T+\mathcal{A}^\star(x))) - b \\
&= \begin{bmatrix}
(t_{1,1}+x_1+x_{n+1}+x_{2n})_+ + \sum_{i=2}^{n-1}(t_{1,i}+x_1+x_{n+i})_+ + (t_{1,n}+x_1)_+ \\
\sum_{i=1}^{n-1}(t_{2,i}+x_2+x_{n+i})_+ + (t_{2,n}+x_2)_+ \\
\vdots \\
\sum_{i=1}^{n-1}(t_{n,i}+x_n+x_{n+i})_+ + (t_{n,n}+x_n)_+ \\
(t_{1,1}+x_1+x_{n+1}+x_{2n})_+ + \sum_{j=2}^{n}(t_{j,1}+x_j+x_{n+1})_+ \\
\sum_{j=1}^{n}(t_{j,2}+x_j+x_{n+2})_+ \\
\vdots \\
\sum_{j=1}^{n}(t_{j,n-1}+x_j+x_{2n-1})_+ \\
(t_{1,1}+x_1+x_{n+1}+x_{2n})_+
\end{bmatrix}
- \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ t_{11} \end{bmatrix}
\end{aligned}
\tag{7}
$$

for any $x = \begin{bmatrix} x_1 \\ \vdots \\ x_{2n} \end{bmatrix} \in \mathbb{R}^{2n}$. It is easy to know that the function $\theta(x)$ is continuously differentiable and its gradient $\nabla\theta(x) = F(x)$ is globally Lipschitz continuous. So, both gradient-type methods and quasi-Newton methods can be directly employed to solve (3). However, since, $\theta(x)$ is not twice continuously differentiable, the convergence rates of these methods are at most linear.

Since $\theta(x)$ is convex and differentiable, so, at solution $x^\star$ of (3),

$$
\nabla\theta(x^\star) = 0, \qquad \text{i.e.,}\ \ F(x^\star) = 0.
$$

This indicates that we can obtain a solution of (3) by solving the equation $F(x) = 0$. $F(x)$ is globally Lipschitz continuous. According to Rademacher's theorem [20, Chapter 9.J], $F(x)$ is Fréchet differentiable almost everywhere. Let Ω_{F} be the set of points at which F is Fréchet differentiable. Denote the Jacobian of $F(x)$ at $x \in \Omega_{\mathrm{F}}$ by $F'(x)$. The generalized Jacobian $\partial F(x)$ of F at $x \in \mathbb{R}^{2n}$ in the sense of Clarke is defined by

$$
\partial F(x) := \mathrm{conv}\{\partial_B F(x)\},
$$

where "conv" denotes the convex hull and

$$
\partial_B F(x) := \left\{ V \in \mathbb{R}^{2n\times 2n} :\ V \text{ is an accumulation point of } F'(x^{(k)}),\ x^{(k)} \to x,\ x^{(k)} \in \Omega_{\mathrm{F}} \right\}.
$$

The nonsmooth Newton's method for solving equation

$$
F(x) = 0 \tag{8}
$$

is given by

$$
x^{(k+1)} = x^{(k)} - V_k^{-1} F(x^{(k)}), \quad V_k \in \partial F(x^{(k)}). \tag{9}
$$

The following result has been established in [17].

Theorem 2 *[17] Let $x^\star$ be a solution of the equation $F(x) = 0$. If all $V \in \partial F(x^\star)$ are nonsingular and F is semismooth at $x^\star$, i.e., F is directionally differentiable at $x^\star$ and for any $V \in \partial F(x^\star + \delta x)$ and $\delta x \to 0$,*

$$
F(x^\star + \delta x) - F(x^\star) - V(\delta x) = o(\|\delta x\|_F),
$$

then every sequence generalized by (9) is superlinearly convergent to $x^\star$ provided that the starting point $x^{(0)}$ is sufficiently close to $x^\star$. Moreover, if F is strongly semismooth at $x^\star$, i.e., F is semismooth at $x^\star$ and

$$F(x^\star+\delta x)-F(x^\star)-V(\delta x)=o(\|\delta x\|_F^2),\quad \forall V\in\partial F(x^\star+\delta x),\ \delta x\to 0,$$

then the convergence rate is quadratic.

Motivated by Theorem 2, in the following we discuss the strong semismoothness of F and the nonsingularity of all $V\in\partial F(x^\star)$ at a solution $x^\star$ of $F(x)=0$.

Since $x=\begin{bmatrix} x_1\\ \vdots\\ x_{2n}\end{bmatrix}\in\Omega_F$, i.e., $F'(x)$ exists, if and only if

$$\begin{cases} t_{11}+x_1+x_{n+1}+x_{2n}\neq 0,\\ t_{1,j}+x_1+x_{n+j}\neq 0,\quad j=2,\cdots,n-1,\\ t_{i,j}+x_i+x_{n+j}\neq 0,\quad i=2,\cdots,n,\ j=1,\cdots,n-1,\\ t_{i,n}+x_i\neq 0,\quad i=1,\cdots,n,\end{cases}$$

in the case that the inequalities above hold,

$$\begin{aligned}
a_{1,1}: &= \frac{\partial(t_{11}+x_1+x_{n+1}+x_{2n})_+}{\partial x_1}=\frac{\partial(t_{11}+x_1+x_{n+1}+x_{2n})_+}{\partial x_{n+1}}\\
&= \frac{\partial(t_{11}+x_1+x_{n+1}+x_{2n})_+}{\partial x_{2n}}=\begin{cases}1 & \text{if } t_{11}+x_1+x_{n+1}+x_{2n}>0\\ 0 & \text{if } t_{11}+x_1+x_{n+1}+x_{2n}<0\end{cases}\\
a_{1,j}: &= \frac{\partial(t_{1,j}+x_1+x_{n+j})_+}{\partial x_1}=\frac{\partial(t_{1,j}+x_1+x_{n+j})_+}{\partial x_{n+j}}=\begin{cases}1 & \text{if } t_{1,j}+x_1+x_{n+j}>0\\ 0 & \text{if } t_{1,j}+x_1+x_{n+j}<0\end{cases}\quad j=2,\cdots,n-1,\\
a_{i,j}: &= \frac{\partial(t_{i,j}+x_i+x_{n+j})_+}{\partial x_i}=\frac{\partial(t_{i,j}+x_i+x_{n+j})_+}{\partial x_{n+j}}=\begin{cases}1 & \text{if } t_{i,j}+x_i+x_{n+j}>0\\ 0 & \text{if } t_{i,j}+x_i+x_{n+j}<0\end{cases}\quad \begin{matrix} i=2,\cdots,n\\ j=1,\cdots,n-1\end{matrix}\\
a_{i,n}: &= \frac{\partial(t_{i,n}+x_i)_+}{x_i}=\begin{cases}1 & \text{if } t_{i,n}+x_i>0\\ 0 & \text{if } t_{i,n}+x_i<0\end{cases}\quad i=1,\cdots,n,
\end{aligned}$$

and

$$F'(x)=\left[\begin{array}{cccc|cccc|c}
\sum_{i=1}^n a_{1,i} & & & & a_{1,1} & a_{1,2} & \cdots & a_{1,n-1} & a_{1,1}\\
 & \sum_{i=1}^n a_{2,i} & & & a_{2,1} & a_{2,2} & \cdots & a_{2,n-1} & 0\\
 & & \ddots & & \vdots & \vdots & \cdots & \vdots & \vdots\\
 & & & \sum_{i=1}^n a_{n,i} & a_{n,1} & a_{n,2} & \cdots & a_{n,n-1} & 0\\ \hline
a_{1,1} & a_{2,1} & \cdots & a_{n,1} & \sum_{i=1}^n a_{i,1} & & & & a_{1,1}\\
a_{1,2} & a_{2,2} & \cdots & a_{n,2} & & \sum_{i=1}^n a_{i,2} & & & 0\\
\vdots & \vdots & \cdots & \vdots & & & \ddots & & \vdots\\
a_{1,n-1} & a_{2,n-1} & \cdots & a_{n,n-1} & & & & \sum_{i=1}^n a_{i,n-1} & 0\\ \hline
a_{1,1} & 0 & \cdots & 0 & a_{1,1} & 0 & \cdots & 0 & a_{1,1}
\end{array}\right],$$

thus, for any $x = \begin{bmatrix} x_1 \\ \vdots \\ x_{2n} \end{bmatrix} \in \mathbb{R}^{2n}$

$$
\begin{aligned}
& V \in \partial_B F(x) \\
\Leftrightarrow \quad & V = \left[\begin{array}{cccc|cccc|c}
\sum_{i=1}^n b_{1,i} & & & & b_{1,1} & b_{1,2} & \cdots & b_{1,n-1} & b_{1,1} \\
& \sum_{i=1}^n b_{2,i} & & & b_{2,1} & b_{2,2} & \cdots & b_{2,n-1} & 0 \\
& & \ddots & & \vdots & \vdots & \cdots & \vdots & \vdots \\
& & & \sum_{i=1}^n b_{n,i} & b_{n,1} & b_{n,2} & \cdots & b_{n,n-1} & 0 \\
\hline
b_{1,1} & b_{2,1} & \cdots & b_{n,1} & \sum_{i=1}^n b_{i,1} & & & & b_{1,1} \\
b_{1,2} & b_{2,2} & \cdots & b_{n,2} & & \sum_{i=1}^n b_{i,2} & & & 0 \\
\vdots & \vdots & \cdots & \vdots & & & \ddots & & \vdots \\
b_{1,n-1} & b_{2,n-1} & \cdots & b_{n,n-1} & & & & \sum_{i=1}^n b_{i,n-1} & 0 \\
\hline
b_{1,1} & 0 & \cdots & 0 & b_{1,1} & 0 & \cdots & 0 & b_{1,1}
\end{array}\right], \quad (10)
\end{aligned}
$$

where

$$
\begin{cases}
b_{1,1} = 1 & \text{if } t_{11} + x_1 + x_{n+1} + x_{2n} > 0 \\
b_{1,1} \in \{0,1\} & \text{if } t_{11} + x_1 + x_{n+1} + x_{2n} = 0, \\
b_{1,1} = 0 & \text{if } t_{11} + x_1 + x_{n+1} + x_{2n} < 0
\end{cases}
$$
$$
\begin{cases}
b_{1,j} = 1 & \text{if } t_{1,j} + x_1 + x_{n+j} > 0 \\
b_{1,j} \in \{0,1\} & \text{if } t_{1,j} + x_1 + x_{n+j} = 0 \quad j = 2, \cdots, n-1, \\
b_{1,j} = 0 & \text{if } t_{1,j} + x_1 + x_{n+j} < 0
\end{cases}
$$
$$
\begin{cases}
b_{i,j} = 1 & \text{if } t_{i,j} + x_i + x_{n+j} > 0 \\
b_{i,j} \in \{0,1\} & \text{if } t_{i,j} + x_i + x_{n+j} = 0 \\
b_{i,j} = 0 & \text{if } t_{i,j} + x_i + x_{n+j} < 0
\end{cases}
\quad \begin{array}{l} i = 2, \cdots, n \\ j = 1, \cdots, n-1 \end{array} \qquad (11)
$$
$$
\begin{cases}
b_{i,n} = 1 & \text{if } t_{i,n} + x_i > 0 \\
b_{i,n} \in \{0,1\} & \text{if } t_{i,n} + x_i = 0 \quad i = 1, \cdots, n, \\
b_{i,n} = 0 & \text{if } t_{i,n} + x_i < 0
\end{cases}
$$

As a result, we obtain

Theorem 3 *$V \in \partial F(x)$ if and only if V is of the form*

$$
V = \left[\begin{array}{cccc|cccc|c}
\sum_{i=1}^n v_{1,i} & & & & v_{1,1} & v_{1,2} & \cdots & v_{1,n-1} & v_{1,1} \\
& \sum_{i=1}^n v_{2,i} & & & v_{2,1} & v_{2,2} & \cdots & v_{2,n-1} & 0 \\
& & \ddots & & \vdots & \vdots & \cdots & \vdots & \vdots \\
& & & \sum_{i=1}^n v_{n,i} & v_{n,1} & v_{n,2} & \cdots & v_{n,n-1} & 0 \\
\hline
v_{1,1} & v_{2,1} & \cdots & v_{n,1} & \sum_{i=1}^n v_{i,1} & & & & v_{1,1} \\
v_{1,2} & v_{2,2} & \cdots & v_{n,2} & & \sum_{i=1}^n v_{i,2} & & & 0 \\
\vdots & \vdots & \cdots & \vdots & & & \ddots & & \vdots \\
v_{1,n-1} & v_{2,n-1} & \cdots & v_{n,n-1} & & & & \sum_{i=1}^n v_{i,n-1} & 0 \\
\hline
v_{1,1} & 0 & \cdots & 0 & v_{1,1} & 0 & \cdots & 0 & v_{1,1}
\end{array}\right], \qquad (12)
$$

where

$$\begin{cases} v_{1,1} = 1 & \text{if } t_{11} + x_1 + x_{n+1} + x_{2n} > 0 \\ v_{1,1} \in [0,\ 1] & \text{if } t_{11} + x_1 + x_{n+1} + x_{2n} = 0 \\ v_{1,1} = 0 & \text{if } t_{11} + x_1 + x_{n+1} + x_{2n} < 0 \end{cases}$$
$$\begin{cases} v_{1,j} = 1 & \text{if } t_{1,j} + x_1 + x_{n+j} > 0 \\ v_{1,j} \in [0,\ 1] & \text{if } t_{1,j} + x_1 + x_{n+j} = 0 \quad j = 2, \cdots, n-1, \\ v_{1,j} = 0 & \text{if } t_{1,j} + x_1 + x_{n+j} < 0 \end{cases}$$
$$\begin{cases} v_{i,j} = 1 & \text{if } t_{i,j} + x_i + x_{n+j} > 0 \\ v_{i,j} \in [0,\ 1] & \text{if } t_{i,j} + x_i + x_{n+j} = 0 \\ v_{i,j} = 0 & \text{if } t_{i,j} + x_i + x_{n+j} < 0 \end{cases} \quad \begin{array}{l} i = 2, \cdots, n \\ j = 1, \cdots, n-1 \end{array} \tag{13}$$
$$\begin{cases} v_{i,n} = 1 & \text{if } t_{i,n} + x_i > 0 \\ v_{i,n} \in [0,\ 1] & \text{if } t_{i,n} + x_i = 0 \quad i = 1, \cdots, n. \\ v_{i,n} = 0 & \text{if } t_{i,n} + x_i < 0 \end{cases}$$

We are now ready to present our results on the strong semismoothness of F and the nonsingularity of all $V \in \partial F(x)$.

Theorem 4 *At any point $x \in \mathbb{R}^{2n}$, $F(x)$ is directionally differentiable and*

$$F(x + \delta x) - F(x) - V\delta x = 0, \quad \forall\ V \in \partial F(x + \delta x),\ \delta x \to 0. \tag{14}$$

Hence, F is strongly semismooth at any $x \in \mathbb{R}^{2n}$.

Proof. A simple calculation yields that

$$\lim_{t \to 0^+} \frac{F(x + th) - F(x)}{t}$$

exists for any $x, h \in \mathbb{R}^{2n}$, and so $F(x)$ is directionally differentiable at any point $x \in \mathbb{R}^{2n}$. In addition, it can be verified using (10) and (11) that

$$F(x + \delta x) - F(x) - V\delta x = 0, \quad \forall\ V \in \partial_B F(x + \delta x),\ \delta x \to 0.$$

Since any $V \in \partial F(x + \delta x)$ is just a convex combination of elements in $\partial_B F(x + \delta x)$, so, (14) holds. □

Theorem 5 *For any $x = \begin{bmatrix} x_1 \\ \vdots \\ x_{2n} \end{bmatrix} \in \mathbb{R}^{2n}$, let*

$$\begin{aligned} M: &= \begin{bmatrix} m_{1,1} & \cdots & m_{1,n} \\ \vdots & \cdots & \vdots \\ m_{n,1} & \cdots & m_{n,n} \end{bmatrix} = \Pi_{\mathcal{K}}(T + \mathcal{A}^\star(x)) \\ &= \Pi_{\mathcal{K}}\left(\begin{bmatrix} t_{1,1} + x_1 + x_{n+1} + x_{2n} & t_{1,2} + x_1 + x_{n+2} & \cdots & t_{1,n-1} + x_1 + x_{2n-1} & t_{1,n} + x_1 \\ t_{2,1} + x_2 + x_{n+1} & t_{2,2} + x_2 + x_{n+2} & \cdots & t_{2,n-1} + x_2 + x_{2n-1} & t_{2,n} + x_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ t_{n,1} + x_n + x_{n+1} & t_{n,2} + x_n + x_{n+2} & \cdots & t_{n,n-1} + x_n + x_{2n-1} & t_{n,n} + x_n \end{bmatrix} \right) \end{aligned}$$

and

$$N_M = \left[\begin{array}{cccc|cccc|c}
\sum_{i=1}^n m_{1,i} & & & & m_{1,1} & m_{1,2} & \cdots & m_{1,n-1} & m_{1,1} \\
 & \sum_{i=1}^n m_{2,i} & & & m_{2,1} & m_{2,2} & \cdots & m_{2,n-1} & 0 \\
 & & \ddots & & \vdots & \vdots & \cdots & \vdots & \vdots \\
 & & & \sum_{i=1}^n m_{n,i} & m_{n,1} & m_{n,2} & \cdots & m_{n,n-1} & 0 \\
\hline
m_{1,1} & m_{2,1} & \cdots & m_{n,1} & \sum_{i=1}^n m_{i,1} & & & & m_{1,1} \\
m_{1,2} & m_{2,2} & \cdots & m_{n,2} & & \sum_{i=1}^n m_{i,2} & & & 0 \\
\vdots & \vdots & \cdots & \vdots & & & \ddots & & \vdots \\
m_{1,n-1} & m_{2,n-1} & \cdots & m_{n,n-1} & & & & \sum_{i=1}^n m_{i,n-1} & 0 \\
\hline
m_{1,1} & 0 & \cdots & 0 & m_{1,1} & 0 & \cdots & 0 & m_{1,1}
\end{array}\right],$$

Then

(i) N_M is symmetric and positive semi-definite.

(ii) All $V \in \partial F(x)$ are nonsingular if and only if N_M is positive definite.

Proof. (i) Since

$$m_{i,j} \geq 0, \ i,j = 1, \cdots, n,$$

for any $h = \begin{bmatrix} h_1 \\ \vdots \\ h_{2n} \end{bmatrix} \in \mathbb{R}^{2n}$

$$h^T N_M h = m_{1,1}(h_1 + h_{n+1} + h_{2n})^2 + \sum_{i=2}^{n} m_{i,1}(h_i + h_{n+1})^2 + \sum_{j=2}^{n-1}\sum_{i=1}^{n} m_{i,j}(h_i + h_{n+j})^2 + \sum_{i=1}^{n} m_{i,n} h_i^2 \geq 0$$

and N_M is symmetric, so N_M is symmetric and positive semi-definite.

(ii) Among all $V \in \partial F(x)$, we consider the following one:

$$V_{min} =$$

$$\left[\begin{array}{cccc|cccc|c}
\sum_{i=1}^n v_{1,i}^{(min)} & & & & v_{1,1}^{(min)} & v_{1,2}^{(min)} & \cdots & v_{1,n-1}^{(min)} & v_{1,1}^{(min)} \\
 & \sum_{i=1}^n v_{2,i}^{(min)} & & & v_{2,1}^{(min)} & v_{2,2}^{(min)} & \cdots & v_{2,n-1}^{(min)} & 0 \\
 & & \ddots & & \vdots & \vdots & \cdots & \vdots & \vdots \\
 & & & \sum_{i=1}^n v_{n,i}^{(min)} & v_{n,1}^{(min)} & v_{n,2}^{(min)} & \cdots & v_{n,n-1}^{(min)} & 0 \\
\hline
v_{1,1}^{(min)} & v_{2,1}^{(min)} & \cdots & v_{n,1}^{(min)} & \sum_{i=1}^n v_{i,1}^{(min)} & & & & v_{1,1}^{(min)} \\
v_{1,2}^{(min)} & v_{2,2}^{(min)} & \cdots & v_{n,2}^{(min)} & & \sum_{i=1}^n v_{i,2}^{(min)} & & & 0 \\
\vdots & \vdots & \cdots & \vdots & & & \ddots & & \vdots \\
v_{1,n-1}^{(min)} & v_{2,n-1}^{(min)} & \cdots & v_{n,n-1}^{(min)} & & & & \sum_{i=1}^n v_{i,n-1}^{(min)} & 0 \\
\hline
v_{1,1}^{(min)} & 0 & \cdots & 0 & v_{1,1}^{(min)} & 0 & \cdots & 0 & v_{1,1}^{(min)}
\end{array}\right], \tag{15}$$

where

$$\begin{cases} v_{1,1}^{(min)} = 1 & \text{if } m_{1,1} = (t_{11} + x_1 + x_{n+1} + x_{2n})_+ > 0 \\ v_{1,1}^{(min)} = 0 & \text{if } m_{1,1} = (t_{11} + x_1 + x_{n+1} + x_{2n})_+ = 0 \end{cases}$$
$$\begin{cases} v_{1,j}^{(min)} = 1 & \text{if } m_{1,j} = (t_{1,j} + x_1 + x_{n+j})_+ > 0 \\ v_{1,j}^{(min)} = 0 & \text{if } m_{1,j} = (t_{1,j} + x_1 + x_{n+j})_+ = 0 \end{cases} \quad j = 2, \cdots, n-1,$$
$$\begin{cases} v_{i,j}^{(min)} = 1 & \text{if } m_{i,j} = (t_{i,j} + x_i + x_{n+j})_+ > 0 \\ v_{i,j}^{(min)} = 0 & \text{if } m_{i,j} = (t_{i,j} + x_i + x_{n+j})_+ = 0 \end{cases} \quad \begin{matrix} i = 2, \cdots, n \\ j = 1, \cdots, n-1 \end{matrix} \tag{16}$$
$$\begin{cases} v_{i,n}^{(min)} = 1 & \text{if } m_{i,n} = (t_{i,n} + x_i)_+ > 0 \\ v_{i,n}^{(min)} = 0 & \text{if } m_{i,n} = (t_{i,n} + x_i)_+ = 0 \end{cases} \quad i = 1, \cdots, n,$$

V_{min} and $V - V_{min}$ are symmetric and positive semi-definite since all $V \in \partial F(x)$ are given by (12) and (13),

$$v_{i,j}^{(min)} \geq 0, \quad v_{i,j} - v_{i,j}^{(min)} \geq 0,$$

and for any $h = \begin{bmatrix} h_1 \\ \vdots \\ h_{2n} \end{bmatrix} \in \mathbb{R}^{2n}$

$$h^T V_{min} h = v_{1,1}^{(min)}(h_1 + h_{n+1} + h_{2n})^2 + \sum_{i=2}^{n} v_{i,1}^{(min)}(h_i + h_{n+1})^2 + \sum_{j=2}^{n-1}\sum_{i=1}^{n} v_{i,j}^{(min)}(h_i + h_{n+j})^2 + \sum_{i=1}^{n} v_{i,n}^{(min)} h_i^2 \geq 0,$$

$$\begin{aligned} h^T(V - V_{min})h &= (v_{1,1} - v_{1,1}^{(min)})(h_1 + h_{n+1} + h_{2n})^2 + \sum_{i=2}^{n}(v_{i,1} - v_{i,1}^{(min)})(h_i + h_{n+1})^2 \\ &\quad + \sum_{j=2}^{n-1}\sum_{i=1}^{n}(v_{i,j} - v_{i,j}^{(min)})(h_i + h_{n+j})^2 + \sum_{i=1}^{n}(v_{i,n} - v_{i,n}^{(min)})h_i^2 \geq 0. \end{aligned}$$

Thus, all $V \in \partial F(x)$ are nonsingular if and only if V_{min} is positive definite.

Recall that $v_{i,j}^{(min)} > 0$ if and only if $m_{i,j} > 0$, $i, j = 1, \cdots, n$ and for any $h = \begin{bmatrix} h_1 \\ \vdots \\ h_{2n} \end{bmatrix} \in \mathbb{R}^{2n}$

$$h^T V_{min} h = v_{1,1}^{(min)}(h_1 + h_{n+1} + h_{2n})^2 + \sum_{i=2}^{n} v_{i,1}^{(min)}(h_i + h_{n+1})^2 + \sum_{j=2}^{n-1}\sum_{i=1}^{n} v_{i,j}^{(min)}(h_i + h_{n+j})^2 + \sum_{i=1}^{n} v_{i,n}^{(min)} h_i^2$$

$$h^T N_M h = m_{1,1}(h_1 + h_{n+1} + h_{2n})^2 + \sum_{i=2}^{n} m_{i,1}(h_i + h_{n+1})^2 + \sum_{j=2}^{n-1}\sum_{i=1}^{n} m_{i,j}(h_i + h_{n+j})^2 + \sum_{i=1}^{n} m_{i,n} h_i^2,$$

so, we get that $h^T V_{min} h > 0$ if and only if $h^T N_M h > 0$. This implies that V_{min} is positive definite if and only if N_M is positive definite. Hence, Part (ii) is proved. □

Remark 2 *The positive semi-definiteness of N_M can be proved*[1] *alternatively as follows: Let*

$$\mathcal{N}_1 = \left[\begin{array}{cccc|cccc|c}
\sum_{i=2}^n m_{1,i} & & & & 0 & m_{1,2} & \cdots & m_{1,n-1} & 0 \\
 & \sum_{i=1}^n m_{2,i} & & & m_{2,1} & m_{2,2} & \cdots & m_{2,n-1} & 0 \\
 & & \ddots & & \vdots & \vdots & \cdots & \vdots & \vdots \\
 & & & \sum_{i=1}^n m_{n,i} & m_{n,1} & m_{n,2} & \cdots & m_{n,n-1} & 0 \\
\hline
0 & m_{2,1} & \cdots & m_{n,1} & \sum_{i=2}^n m_{i,1} & & & & 0 \\
m_{1,2} & m_{2,2} & \cdots & m_{n,2} & & \sum_{i=1}^n m_{i,2} & & & 0 \\
\vdots & \vdots & \cdots & \vdots & & & \ddots & & \vdots \\
m_{1,n-1} & m_{2,n-1} & \cdots & m_{n,n-1} & & & & \sum_{i=1}^n m_{i,n-1} & 0 \\
\hline
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0
\end{array}\right],$$

and

$$\mathcal{N}_2 = \left[\begin{array}{cccc|cccc|c}
m_{1,1} & & & & m_{1,1} & 0 & \cdots & 0 & m_{1,1} \\
 & 0 & & & 0 & 0 & \cdots & 0 & 0 \\
 & & \ddots & & \vdots & \vdots & \cdots & \vdots & \vdots \\
 & & & 0 & 0 & 0 & \cdots & 0 & 0 \\
\hline
m_{1,1} & 0 & \cdots & 0 & m_{1,1} & & & & m_{1,1} \\
0 & 0 & \cdots & 0 & & 0 & & & 0 \\
\vdots & \vdots & \cdots & \vdots & & & \ddots & & \vdots \\
0 & 0 & \cdots & 0 & & & & 0 & 0 \\
\hline
m_{1,1} & 0 & \cdots & 0 & m_{1,1} & 0 & \cdots & 0 & m_{1,1}
\end{array}\right].$$

Then

$$N_M = \mathcal{N}_1 + \mathcal{N}_2.$$

Obviously, $\mathcal{N}_2$ is positive semi-definite. Furthermore, $\mathcal{N}_1$ is non-negative, symmetric and weakly diagonally dominant, so the well-known Gershgorin's theorem [24] gives that all eigenvalues of $\mathcal{N}_1$ are non-negative. Thus, $\mathcal{N}_1$ is positive semi-definite. Hence, $N_M = \mathcal{N}_1 + \mathcal{N}_2$ is also positive semi-definite.

If $x = x^\star$ with $F(x^\star) = 0$, then Theorem 5 (ii) can be simplified significantly, as shown in the next result.

Theorem 6 *Let $M^\star$ be the (unique) solution of Problem (2) and $x^\star \in \mathbb{R}^{2n}$ satisfy $F(x^\star) = 0$. Denote*

$$M^\star =: \begin{bmatrix} t_{1,1} & m^\star_{1,2} & \cdots & m^\star_{1,n-1} & m^\star_{1,n} \\ m^\star_{2,1} & m^\star_{2,2} & \cdots & m^\star_{2,n-1} & m^\star_{2,n} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ m^\star_{n,1} & m^\star_{n,2} & \cdots & m^\star_{n,n-1} & m^\star_{n,n} \end{bmatrix}, \tag{17}$$

$$L^\star := \begin{bmatrix} \frac{1}{\sqrt{1-t_{1,1}}} & \\ & I \end{bmatrix} \begin{bmatrix} 0 & m^\star_{1,2} & \cdots & m^\star_{1,n-1} \\ m^\star_{2,1} & m^\star_{2,2} & \cdots & m^\star_{2,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ m^\star_{n,1} & m^\star_{n,2} & \cdots & m^\star_{n,n-1} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1-t_{1,1}}} & \\ & I \end{bmatrix}. \tag{18}$$

Then

(i) It is true that

$$\|L^\star\|_2 \le 1. \tag{19}$$

[1]This alternative proof is given by an anonmous referee.

(ii) All $V \in \partial F(x^\star)$ are nonsingular if and only if

$$\|L^\star\|_2 < 1. \tag{20}$$

Proof. We have from Theorem 5 (i) that $N_{M^\star}$ is symmetric and positive semi-definite. Now, $0 < t_{1,1} < 1$, and $M^\star$ satisfies that

$$\begin{cases} 0 < m^\star_{1,1} = t_{1,1} < 1, \ \sum_{i=2}^n m^\star_{1,i} = 1 - t_{1,1}, \ \sum_{i=2}^n m^\star_{i,1} = 1 - t_{1,1}, \\ \sum_{i=1}^n m^\star_{j,i} = 1, \ j = 2, \cdots, n, \\ \sum_{i=1}^n m^\star_{i,j} = 1, \ j = 2, \cdots, n-1, \end{cases} \tag{21}$$

we obtain by using the positive semi-definiteness of $N_{M^\star}$ that the matrix

$$\mathcal{N}_{M^\star} := \left[\begin{array}{cccc|cccc} 1 - t_{1,1} & & & & 0 & m^\star_{1,2} & \cdots & m^\star_{1,n-1} \\ & 1 & & & m^\star_{2,1} & m^\star_{2,2} & \cdots & m^\star_{2,n-1} \\ & & \ddots & & \vdots & \vdots & \cdots & \vdots \\ & & & 1 & m^\star_{n,1} & m^\star_{n,2} & \cdots & m^\star_{n,n-1} \\ \hline 0 & m^\star_{2,1} & \cdots & m^\star_{n,1} & 1 - t_{1,1} & & & \\ m^\star_{1,2} & m^\star_{2,2} & \cdots & m^\star_{n,2} & & 1 & & \\ \vdots & \vdots & \cdots & \vdots & & & \ddots & \\ m^\star_{1,n-1} & m^\star_{2,n-1} & \cdots & m^\star_{n,n-1} & & & & 1 \end{array}\right] \tag{22}$$

is positive semi-definite. Equivalently, (19) holds.

(ii) By Theorem 5 (ii) we know that all $V \in \partial F(x^\star)$ are nonsingular if and only if $N_{M^\star}$ is positive definite, which is equivalent to that the matrix $\mathcal{N}_{M^\star}$ defined by (22) is positive definite. Therefore, Part (ii) follows directly from the property that $\mathcal{N}_{M^\star}$ is positive definite if and only if (20) holds. □

Theorem 6 is very pleasant because it indicates that for almost all $T \in \mathbb{R}^{n\times n}$, all $V \in \partial F(x^\star)$ are nonsingular for the solution $x^\star$ of the equation $F(x) = 0$.

The following corollary contains two important sufficient conditions ensuring that all $V \in \partial F(x^\star)$ are nonsingular for the solution $x^\star$ of the equation $F(x) = 0$.

Corollary 7 *With the notation in Theorem 6, if $M^\star e_i > 0$ for some $1 \le i \le n$, or $e_j^T M^\star > 0$ for some $1 \le j \le n$, then all $V \in \partial F(x^\star)$ are nonsingular. Here e_i and e_j are the ith and jth columns of I_n, respectively.*

Proof. By Theorem 6 and its proof we only need to show that $\mathcal{N}_{M^\star}$ defined by (22) is positive definite provided $M^\star e_i > 0$ for some $1 \le i \le n$ (or $e_j^T M^\star > 0$ for some $1 \le j \le n$).

In the following we only assume that $M^\star e_i > 0$ for some $1 \le i \le n$ because the case that $e_j^T M^\star > 0$ for some $1 \le j \le n$) can be discussed similarly.

First, we have

$$0 < t_{1,1} < 1, \ 0 \le m^\star_{1,j} < 1, \ 0 \le m^\star_{j,1} < 1, \ j = 2, \cdots, n,$$

and for any $h = \begin{bmatrix} h_1 \\ \vdots \\ h_{2n-1} \end{bmatrix} \in \mathbb{R}^{2n-1}$,

$$h^T \mathcal{N}_{M^\star} h = \sum_{i=2}^n m^\star_{i,1}(h_i + h_{n+1})^2 + \sum_{j=2}^{n-1}\sum_{i=1}^n m^\star_{i,j}(h_i + h_{n+j})^2 + \sum_{i=1}^n m^\star_{i,n} h_i^2. \tag{23}$$

Next, we show by considering three different cases that $h^T \mathcal{N}_{M^\star} h = 0$ only if $h = 0$, as follows.

Case 1: $m^\star_{2,1} \neq 0, \cdots, m^\star_{n,1} \neq 0$. In this case

$$
\begin{aligned}
& h^T \mathcal{N}_{M^\star} h = 0 \\
\Longrightarrow\ & \left\{ \begin{array}{l} h_2 = \cdots = h_n = -h_{n+1}, \\ h^T \mathcal{N}_{M^\star} h = \sum_{j=2}^{n-1} (h_{n+1} - h_{n+j})^2 \sum_{i=2}^{n} m^\star_{i,j} + h_{n+1}^2 \sum_{i=2}^{n} m^\star_{i,n} \\ \qquad\qquad + \sum_{j=2}^{n-1} m^\star_{1,j} (h_1 + h_{n+j})^2 + m^\star_{1,n} h_1^2 = 0 \end{array} \right. \\
\Longrightarrow\ & \left\{ \begin{array}{l} h_2 = \cdots = h_n = -h_{n+1} = \cdots = -h_{2n-1} = 0, \\ h^T \mathcal{N}_{M^\star} h = h_1^2 \sum_{j=2}^{n} m^\star_{1,j} = 0 \end{array} \right. \\
& \qquad (\text{since } 0 < \sum_{i=2}^{n} m^\star_{i,j} = 1 - m^\star_{1,j} \leq 1,\ j = 2, \cdots, n) \\
\Longrightarrow\ & h_1 = \cdots = h_{2n-1} = 0, \quad (\text{since } 0 < \sum_{j=2}^{n} m^\star_{1,j} = 1 - t_{1,1} < 1), \\
\Longrightarrow\ & h = 0.
\end{aligned}
$$

Case 2: $m^\star_{1,k} \neq 0, \cdots, m^\star_{n,k} \neq 0$ for some k with $2 \leq k \leq n-1$. In this case, we have

$$
\begin{aligned}
& h^T \mathcal{N}_{M^\star} h = 0 \\
\Longrightarrow\ & \left\{ \begin{array}{l} h_1 = \cdots = h_n = -h_{n+k} \\ h^T \mathcal{N}_{M^\star} h = (h_{n+k} - h_{n+1})^2 \sum_{i=2}^{n} m^\star_{i,1} + \sum_{j=2}^{k-1} (h_{n+k} - h_{n+j})^2 \sum_{i=1}^{n} m^\star_{i,j} \\ \qquad\qquad + \sum_{j=k+1}^{n-1} (h_{n+k} - h_{n+j})^2 \sum_{i=1}^{n} m^\star_{i,j} + h_{n+k}^2 \sum_{i=1}^{n} m^\star_{i,n} \end{array} \right. \\
\Longrightarrow\ & h_1 = \cdots = h_{2n-1} = 0 \\
& (\text{since } \quad 0 < \sum_{i=2}^{n} m^\star_{i,1} = 1 - t_{1,1} \neq 0,\ \sum_{i=1}^{n} m^\star_{i,n} = 1,\ \sum_{i=1}^{n} m^\star_{i,j} = 1,\ j = 2, \cdots, k-1, k+1, \cdots, n-1) \\
\Longrightarrow\ & h = 0.
\end{aligned}
$$

Case 3: $m^\star_{1,n} \neq 0, \cdots, m^\star_{n,n} \neq 0$. In this case, we have

$$
\begin{aligned}
& h^T \mathcal{N}_{M^\star} h = 0 \\
\Longrightarrow\ & \left\{ \begin{array}{l} h_1 = \cdots = h_n = 0 \\ h^T \mathcal{N}_{M^\star} h = h_{n+1}^2 \sum_{i=2}^{n} m^\star_{i,1} + \sum_{j=2}^{n-1} h_{n+j}^2 \sum_{i=1}^{n} m^\star_{i,j} \end{array} \right. \\
\Longrightarrow\ & h_1 = \cdots = h_n = 0,\ h_{n+1} = \cdots = h_{2n-1} = 0 \\
& (\text{since } 0 < \sum_{i=2}^{n} m^\star_{i,1} = 1 - t_{1,1} \neq 0,\ \sum_{i=1}^{n} m^\star_{i,j} = 1,\ j = 2, \cdots, n-1) \\
\Longrightarrow\ & h = 0.
\end{aligned}
$$

Now we have shown that $h^T \mathcal{N}_{M^\star} h = 0$ only if $h = 0$. This means that $\mathcal{N}_{M^\star}$ is positive definite. □

Remark 3 *Corollary 7 can be proved*[2] *alternatively as follows: Consider non-negative matrix*

$$
Z = \begin{bmatrix} D & R \\ R^T & D \end{bmatrix},
$$

[2]This alternative proof is given by an anonymous referee.

where $D = \begin{bmatrix} t_{1,1} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}$ *and* R *is obtained from the matrix* $M^\star$ *by replacing its* $(1,1)$ *entry with* 0*. Then* Z *is symmetrix doubly stochastic with the largest eigenvalue equal to* 1*. If* $M^\star$ *has a nonzero row or a nonzero column, then* Z *is irreducible. Thus, up to a multiple,* $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^{2n}$ *is its only eigenvector corresponding to the largest eigenvalue* 1*. Now, let us remove the last row and the last column of* Z *to get the matrix* $\tilde{Z} \in \mathbb{R}^{(2n-1)\times(2n-1)}$*, then there is no positive vector* $v \in \mathbb{R}^{2n-1}$ *such that* $\tilde{v} = v$*. In other words, the largest eigenvalue of* $\tilde{Z}$ *is less than* 1*. Hence,* $I_{2n-1} - \tilde{Z}$ *is positive definite, and so is*

$$\mathcal{N}_{M^\star} = (-I_n \oplus I_{n-1})(I_{2n-1} - \tilde{Z})(-I_n \oplus I_{n+1}).$$

An important consequence of Theorems 4 and 6 is on the convergence of Newton's method (9).

Theorem 8 *Let* $x^\star \in \mathbb{R}^{2n}$ *be solution of the equation* $F(x) = 0$*. If (20) holds, then Newton's method (9) is quadratically convergent provided that* $x^{(0)}$ *is sufficiently close to* $x^\star$*.*

3 Numerical Algorithm

In our numerical implementation,, we use the following globalized version of Newton's method for solving the dual problem (3). Recall that $\nabla\theta(x) = F(x)$ for any $x \in \mathbb{R}^{2n}$.

Algorithm 1 (Nonsmooth Newton's Method)

Step 0. *Given* $x^{(0)} \in \mathbb{R}^{2n}$*,* $\eta \in (0,1)$*,* $\rho, \delta \in (0, 1/2)$*.* $k := 0$*.*

Step 1 (Newton's Iteration). *Let* $V^{(k)}_{min}$ *be defined by (15) and (16) with* x *being replaced by* $x^{(k)}$*, and compute an approximate solution* $\Delta x^{(k)} \in \mathbb{R}^{2n}$ *using the conjugate gradient (CG) method [9, Algorithm 10.2.1] to*

$$F(x^{(k)}) + V^{(k)}_{min}\Delta x = 0 \tag{24}$$

such that

$$\|F(x^{(k)}) + V^{(k)}_{min}\Delta x^{(k)}\|_F \le \min\{\eta, \|F(x^{(k)}\|_F\}\|F(x^{(k)})\|_F \tag{25}$$

if $V^{(k)}_{min}$ *is nonsingular. If (25) is not achieved, or if the condition*

$$(\Delta x^{(k)})^T F(x^{(k)}) \le -\min\{\eta, \|F(x^{(k)}\|_F\}(\Delta x^{(k)})^T \Delta x^{(k)}, \tag{26}$$

is not satisfied, or $V^{(k)}_{min}$ *is singular, let*

$$\Delta x^{(k)} = -F(x^{(k)}).$$

Step 2 (Line Search in the Descent Direction $\Delta x^{(k)}$ of $\Theta(x)$ at $x^{(k)}$). *Let* s_k *be the smallest nonnegative integer* s *such that*

$$\theta(x^{(k)} + \rho^s \Delta x^{(k)}) - \theta(x^{(k)}) \le \delta\rho^s(\Delta x^{(k)})^T F(x^{(k)}).$$

Set

$$x^{(k+1)} := x^{(k)} + \rho^{s_k}\Delta x^{(k)}.$$

Step 3. *Replace k by $k+1$ and go to Step 1.*

In Algorithm 1, we choose the starting point $x^{(0)}$ as the solution of the following simplified version of (2)

$$\begin{array}{ll} \min & \frac{1}{2}\|M-T\|_F^2 \\ \text{s.t.} & \mathcal{A}(M)=b. \end{array} \tag{27}$$

This simplified problem has been studied in [8]. As in Section 2, by the dual approach, we know that the unique solution M^0 to problem (27) is given by

$$M^0 = T + \mathcal{A}^*(x^{(0)}), \tag{28}$$

where $x^{(0)} \in R^{2n}$ is a solution of

$$\mathcal{A}\left(T + \mathcal{A}^*(x)\right) = b. \tag{29}$$

$x^{(0)}$ can be obtained by applying the CG method to

$$\begin{bmatrix} n & & & & 1 & \cdots & 1 & 1 \\ & n & & & 1 & \cdots & 1 & 0 \\ & & \ddots & & \vdots & \cdots & \vdots & \vdots \\ & & & n & 1 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & n & & & 1 \\ \vdots & \vdots & \cdots & \vdots & & \ddots & & \vdots \\ 1 & 1 & \cdots & 1 & & & n & \\ 1 & 1 & \cdots & 1 & 1 & & & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ x_{n+1} \\ \vdots \\ x_{2n-1} \\ x_{2n} \end{bmatrix} = \begin{bmatrix} 1-\sum_{i=1}^n t_{1,i} \\ 1-\sum_{i=1}^n t_{2,i} \\ \vdots \\ 1-\sum_{i=1}^n t_{n,i} \\ 1-\sum_{i=1}^n t_{i,1} \\ \vdots \\ 1-\sum_{i=1}^n t_{i,n-1} \\ 0 \end{bmatrix}. \tag{30}$$

Theorem 9 *Assume that the inequality (20) holds. Then the sequence $\{x^{(k)}\}$ generated by Algorithm 1 converges to the solution $x^\star$ of $F(x)=0$ quadratically.*

Proof. Since for any $k \geq 0$, $\Delta x^{(k)}$ is always a descent direction of $\theta(x)$ at $x = x^{(k)}$, and $\theta(x)$ is convex, we know that $\{x^{(k)}\}$ is bounded. So, we obtain by using Theorem 6.3.3 in [18] that

$$\lim_{k\to\infty} \nabla\theta(x^{(k)}) = 0,$$

which, in return, together with the convexity of $\theta(x)$ and the boundedness of $\{x^{(k)}\}$, yields that $x^{(k)} \to x^\star$ for some $x^\star$ satisfying $F(x^\star)=0$.

Note that (20) holds, by Theorem 6 (ii), all $V \in \partial F(x^\star)$ are nonsingular. Since $x^{(k)} \to x^\star$, by Proposition 3.1 in [17], for all k sufficiently large, $V_{min}^{(k)}$ is positive definite, $\{\|V_{min}^{(k)}\|_F\}$ and $\{\|(V_{min}^{(k)})^{-1}\|_F\}$ are uniformly bounded. Thus, for all k sufficiently large, $\Delta x^{(k)}$ can satisfy (25) and (26), and moreover, (14) and that $F(x^\star)=0$ yield

$$F(x^{(k)}) - F(x^\star) - V_{min}^{(k)}(x^{(k)} - x^\star) = 0, \text{ i.e., } F(x^{(k)}) = V_{min}^{(k)}(x^{(k)} - x^\star).$$

Hence, for all k sufficiently large,

$$\begin{aligned} \|x^{(k)} + \Delta x^{(k)} - x^\star\|_F &= \|(V_{min}^{(k)})^{-1}(F(x^{(k)}) + V_{min}^{(k)}\Delta x^{(k)})\|_F \\ &\leq \|(V_{min}^{(k)})^{-1}\|_F \|F(x^{(k)}) + V_{min}^{(k)}\Delta x^{(k)}\|_F \\ &\leq \|(V_{min}^{(k)})^{-1}\|_F \min\{\eta, \|F(x^{(k)})\|_F\} \|F(x^{(k)})\|_F \\ &\leq \|(V_{min}^{(k)})^{-1}\|_F \|F(x^{(k)}\|_F^2 \\ &\leq \|(V_{min}^{(k)})^{-1}\|_F \|V_{min}^{(k)}\|_F^2 \|x^{(k)} - x^\star\|_F^2 \\ &= O(\|x^{(k)} - x^\star\|_F^2). \end{aligned}$$

Then, for all k sufficiently large, $s_k = 0$, $\rho^{s_k} = 1$ and

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)}.$$

Therefore, the quadratic convergence follows. □

In the rest of this section, we report our numerical results for solving (1) by Algorithm 1. All the tests are implemented in `MATLAB 7.0.1` running on a P4 PC of 2.40 GHz CPU. We also compare the performance of our method with that of the alternating projection method proposed in [8].

In our experiments, we tested the following two classes of problems.

Example 1 *Let $\widetilde{M}$ be given by (6). Set*

$$T := \widetilde{M} + \tau R$$

where R is a random $n \times n$ real matrix with entries in $[-1.0, 1.0]$ and $\tau \in \mathbb{R}$ is a perturbed parameter. Here, we set $t_{1,1} = 0.5 < 1$ to ensure that (4) holds. We report our numerical results for $n = 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000$, and $\tau = 0.1, 1.0, 10.0$.

Example 2 *The matrix T is generated randomly with entries uniformly distributed between -10.0 and 10.0, but we set $t_{1,1} = 0.5 < 1$. We give our numerical results for $n = 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000$.*

To demonstrate the performance of Algorithm 1, the linear systems (24) and (30) are solved with provision for lower (inexact) and higher (approx. exactly) accuracy requirements[3]. To do so, in our numerical experiments, we set the parameters used in our algorithm as either

(a). `Tol` $= 10^{-6}$, $\eta = 10^{-6}$, $\rho = 0.5$, and $\delta = 10^{-4}$, or

(b). `Tol` $= 10^{-10}$, $\eta = 10^{-15}$, $\rho = 0.5$, and $\delta = 10^{-4}$.

Here, `Tol` is the required tolerance used in the stopping criterion defined by

$$\|\nabla\theta(x^{(k)})\|_F = \|F(x^{(k)}\|_F \leq \texttt{Tol}.$$

Our numerical results are given in Tables 1–4, where `Time`, `Iter.`, `Res0.`, `Res*.` and `Err*.` stand for the CPU times required for convergence, the number of iterations, the residuals $\|\nabla\theta(\cdot)\|_F$ at the starting point $x^{(0)}$ and the final iterate of Algorithm 1, and the error $\| \begin{bmatrix} M\mathbf{e} \\ M^T\mathbf{e} \\ \mathbf{e}_1^T M \mathbf{e}_1 \end{bmatrix} - \begin{bmatrix} \mathbf{e} \\ \mathbf{e} \\ \mathbf{e}_1^T T \mathbf{e}_1 \end{bmatrix} \|_F$ at the computed solution M^*, respectively.

In our experiments, the quadratic convergence of Algorithm 1 has been observed. From Tables 1–4, we note that if we solve the linear system (24) with a lower accuracy, it needs less CPU time while we can obtain a coarser solution. Conversely, if we solve the linear system (24) with a higher accuracy, we can obtain a relatively more precise solution while it needs relatively more CPU time. Finally, in our experiments, the largest numerical examples contain $25,000,000$ unknowns in the primal problem (1) and $10,000$ unknowns in the dual problem (3). This shows that Algorithm 1 is very efficient for large scale problems.

Next, we compare the performance of our Algorithm 1 with that of the alternating projection method in [8]. For the purpose of comparison, we set the stopping tolerance for both algorithms as

$$\| \begin{bmatrix} M\mathbf{e} \\ M^T\mathbf{e} \\ \mathbf{e}_1^T M \mathbf{e}_1 \end{bmatrix} - \begin{bmatrix} \mathbf{e} \\ \mathbf{e} \\ \mathbf{e}_1^T T \mathbf{e}_1 \end{bmatrix} \|_F \leq \texttt{Tol}.$$

[3] As an anonymous referee pointed, the linear system (30) can also be solved in linear time by using the direct methods exploiting the structure of the coefficient matrix of (30) [25, 26, 27, 28]

Tol $= 10^{-6}$, $\eta = 10^{-6}$, $\rho = 0.5$, and $\delta = 10^{-4}$						
τ	n	Time	Iter.	Res0.	Res*.	Err*.
0.1	500	5.6 s	6	3.9×10^2	7.8×10^{-8}	7.8×10^{-8}
	1,000	26.2 s	7	1.1×10^3	7.2×10^{-13}	7.2×10^{-13}
	1,500	1 m 01 s	7	2.0×10^3	4.6×10^{-12}	4.6×10^{-12}
	2,000	1 m 55 s	7	3.1×10^3	1.1×10^{-7}	1.1×10^{-7}
	2,500	3 m 34 s	8	4.4×10^3	1.3×10^{-13}	1.3×10^{-13}
	3,000	5 m 15 s	8	5.8×10^3	2.8×10^{-13}	2.8×10^{-13}
	3,500	7 m 20 s	8	7.3×10^3	6.4×10^{-13}	6.4×10^{-13}
	4,000	10 m 07 s	8	8.9×10^3	2.0×10^{-12}	2.0×10^{-12}
	4,500	13 m 18 s	8	1.1×10^4	3.8×10^{-12}	3.8×10^{-12}
	5,000	16 m 52 s	9	1.2×10^4	1.2×10^{-8}	1.2×10^{-8}
1.0	500	8.5 s	8	3.9×10^3	1.3×10^{-11}	1.3×10^{-11}
	1,000	37.3 s	9	1.1×10^4	8.0×10^{-12}	8.0×10^{-12}
	1,500	1 m 27 s	9	2.1×10^4	8.3×10^{-13}	8.3×10^{-13}
	2,000	2 m 39 s	9	3.2×10^4	2.7×10^{-11}	2.7×10^{-11}
	2,500	4 m 15 s	9	4.4×10^4	2.0×10^{-11}	2.0×10^{-11}
	3,000	6 m 16 s	9	5.8×10^4	9.3×10^{-11}	9.3×10^{-11}
	3,500	8 m 50 s	9	7.3×10^4	8.5×10^{-11}	8.5×10^{-11}
	4,000	11 m 52 s	9	8.9×10^4	4.4×10^{-7}	4.4×10^{-7}
	4,500	17 m 16 s	10	1.1×10^5	1.5×10^{-12}	1.5×10^{-12}
	5,000	23 m 34 s	10	1.2×10^5	6.2×10^{-12}	6.2×10^{-12}
10.0	500	12.5 s	10	3.9×10^4	1.6×10^{-11}	1.6×10^{-11}
	1,000	46.4 s	10	1.1×10^5	1.2×10^{-10}	1.2×10^{-10}
	1,500	1 m 44 s	10	2.1×10^5	5.0×10^{-10}	5.0×10^{-10}
	2,000	3 m 11 s	10	3.2×10^5	2.1×10^{-9}	2.3×10^{-9}
	2,500	5 m 38 s	11	4.4×10^5	2.4×10^{-12}	2.4×10^{-12}
	3,000	8 m 11 s	11	5.8×10^5	1.7×10^{-12}	1.7×10^{-12}
	3,500	11 m 23 s	11	7.3×10^5	3.7×10^{-11}	3.7×10^{-11}
	4,000	15 m 24 s	11	8.9×10^5	2.6×10^{-11}	2.6×10^{-11}
	4,500	19 m 46 s	11	1.1×10^6	1.3×10^{-9}	1.3×10^{-9}
	5,000	27 m 14 s	11	1.2×10^6	2.5×10^{-10}	2.5×10^{-10}

Table 1: Numerical results of Example 1 (a)

$\texttt{Tol} = 10^{-10}$, $\eta = 10^{-15}$, $\rho = 0.5$, and $\delta = 10^{-4}$						
τ	n	Time	Iter.	Res0.	Res*.	Err*.
0.1	500	13.5 s	6	3.9×10^2	1.3×10^{-14}	1.3×10^{-14}
	1,000	1 m 05 s	7	1.1×10^3	2.0×10^{-14}	2.0×10^{-14}
	1,500	2 m 30 s	7	2.0×10^3	3.0×10^{-14}	3.0×10^{-14}
	2,000	4 m 18 s	7	3.1×10^3	3.9×10^{-14}	3.9×10^{-14}
	2,500	5 m 03 s	8	4.4×10^3	4.9×10^{-14}	4.9×10^{-14}
	3,000	11 m 53 s	8	5.8×10^3	5.8×10^{-14}	5.8×10^{-14}
	3,500	16 m 33 s	8	7.3×10^3	6.7×10^{-14}	6.7×10^{-14}
	4,000	24 m 52 s	9	8.9×10^3	7.7×10^{-14}	7.7×10^{-14}
	4,500	28 m 07 s	8	1.1×10^4	8.4×10^{-14}	8.4×10^{-14}
	5,000	41 m 23 s	9	1.2×10^4	9.3×10^{-14}	9.3×10^{-14}
1.0	500	20.6 s	8	3.9×10^3	2.8×10^{-14}	2.8×10^{-14}
	1,000	1 m 19 s	8	1.1×10^4	8.0×10^{-14}	8.0×10^{-14}
	1,500	3 m 30 s	9	2.0×10^4	8.3×10^{-14}	8.3×10^{-14}
	2,000	6 m 28 s	9	3.2×10^4	1.1×10^{-13}	1.1×10^{-13}
	2,500	7 m 08 s	10	4.4×10^4	1.4×10^{-13}	1.4×10^{-13}
	3,000	15 m 16 s	10	5.8×10^4	1.6×10^{-13}	1.6×10^{-13}
	3,500	23 m 24 s	10	7.3×10^4	1.9×10^{-13}	1.9×10^{-13}
	4,000	30 m 02 s	10	8.9×10^4	2.2×10^{-13}	2.2×10^{-13}
	4,500	38 m 21 s	10	1.1×10^5	2.5×10^{-13}	2.5×10^{-13}
	5,000	48 m 54 s	10	1.2×10^5	2.7×10^{-13}	2.7×10^{-13}
10.0	500	29.1 s	9	3.9×10^4	3.3×10^{-11}	3.3×10^{-11}
	1,000	2 m 04 s	10	1.1×10^5	3.2×10^{-13}	3.2×10^{-13}
	1,500	4 m 29 s	10	2.1×10^5	6.5×10^{-13}	6.5×10^{-13}
	2,000	8 m 00 s	10	3.2×10^5	2.4×10^{-12}	2.4×10^{-12}
	2,500	8 m 35 s	11	4.4×10^5	7.7×10^{-13}	7.7×10^{-13}
	3,000	12 m 42 s	11	5.8×10^5	9.1×10^{-13}	9.1×10^{-13}
	3,500	17 m 02 s	11	7.3×10^5	1.1×10^{-12}	1.1×10^{-12}
	4,000	22 m 22 s	11	8.9×10^5	1.2×10^{-12}	1.2×10^{-12}
	4,500	29 m 25 s	11	1.1×10^6	1.3×10^{-12}	1.3×10^{-12}
	5,000	51 m 57 s	11	1.2×10^6	1.5×10^{-12}	1.5×10^{-12}

Table 2: Numerical results of Example 1 (b)

`Tol` $= 10^{-6}$, $\eta = 10^{-6}$, $\rho = 0.5$, and $\delta = 10^{-4}$					
n	Time	Iter.	Res0.	Res*.	Err*.
500	13.0 s	10	3.9×10^4	4.4×10^{-11}	4.4×10^{-11}
1,000	46.7 s	10	1.1×10^5	7.4×10^{-11}	7.4×10^{-11}
1,500	1 m 48 s	10	2.1×10^5	4.5×10^{-10}	4.5×10^{-10}
2,000	3 m 19 s	10	3.2×10^5	7.1×10^{-10}	7.1×10^{-10}
2,500	5 m 20 s	10	4.4×10^5	2.2×10^{-9}	2.2×10^{-9}
3,000	8 m 39 s	10	5.8×10^5	2.1×10^{-11}	2.1×10^{-11}
3,500	12 m 06 s	11	7.3×10^5	3.9×10^{-11}	3.9×10^{-11}
4,000	15 m 54 s	11	8.9×10^5	8.0×10^{-11}	8.0×10^{-11}
4,500	21 m 15 s	11	1.1×10^6	5.0×10^{-11}	5.0×10^{-11}
5,000	27 m 12 s	11	1.2×10^6	1.1×10^{-10}	1.1×10^{-10}

Table 3: Numerical results of Example 2 (a)

`Tol` $= 10^{-10}$, $\eta = 10^{-15}$, $\rho = 0.5$, and $\delta = 10^{-4}$					
n	Time	Iter.	Res0.	Res*.	Err*.
500	29.7 s	9	3.9×10^4	3.9×10^{-12}	3.9×10^{-12}
1,000	2 m 03 s	10	1.1×10^5	3.2×10^{-13}	3.2×10^{-13}
1,500	4 m 25 s	10	2.1×10^5	7.8×10^{-13}	7.8×10^{-13}
2,000	8 m 22 s	11	3.2×10^5	6.0×10^{-13}	6.0×10^{-13}
2,500	14 m 26 s	11	4.4×10^5	7.5×10^{-13}	7.5×10^{-13}
3,000	19 m 57 s	11	5.8×10^5	9.2×10^{-13}	9.2×10^{-13}
3,500	26 m 45 s	11	7.3×10^5	1.0×10^{-12}	1.0×10^{-12}
4,000	33 m 12 s	11	8.9×10^5	1.2×10^{-12}	1.2×10^{-12}
4,500	49 m 47 s	12	1.1×10^6	1.4×10^{-12}	1.4×10^{-12}
5,000	56 m 41 s	11	1.2×10^6	1.5×10^{-12}	1.5×10^{-12}

Table 4: Numerical results of Example 2 (b)

Tol $= 10^{-6}$				
	n	Time	Iter.	Dist
Algorithm 1	500	6.7 s	6	28.028746
	1,000	26.2 s	7	56.855379
	1,500	1 m 05 s	7	85.690588
	2,000	2 m 01 s	7	114.556635
$\eta = 10^{-6}$	2,500	3 m 08 s	7	143.421146
	3,000	5 m 16 s	8	172.273970
	3,500	7 m 24 s	8	201.138311
Algorithm 1	500	7.0 s	6	28.028746
	1,000	31.7 s	7	56.855379
	1,500	1 m 17 s	7	85.690588
	2,000	2 m 25 s	7	114.556635
$\eta = 10^{-10}$	2,500	3 m 42 s	7	143.421146
	3,000	6 m 21 s	8	172.273970
	3,500	8 m 40 s	8	201.138311
Algorithm 1	500	8.6 s	6	28.028746
	1,000	39.0 s	7	56.855379
	1,500	1 m 32 s	7	85.690588
	2,000	2 m 51 s	7	114.556635
$\eta = 10^{-15}$	2,500	4 m 16 s	7	143.421146
	3,000	7 m 22 s	8	172.273970
	3,500	10 m 10 s	8	201.138311
Method in [8]	500	21.1 s	181	28.028746
	1,000	2 m 53 s	263	56.855379
	1,500	10 m 51 s	321	85.690588
	2,000	25 m 55 s	371	114.556635
	2,500	53 m 19 s	417	143.421146
	3,000	1 h 59 m 01 s	457	172.273970
	3,500	2 h 58 m 54 s	495	201.138311

Table 5: Numerical results of Example 1 (a)

Here, we choose Tol to be different values, e.g. Tol $= 10^{-6}, 10^{-10}$, etc. Consequently, in Algorithm 1 the equation (24) is solved with varying accuracies, see $\eta = 10^{-6}, 10^{-15}$, etc. The values of remaining parameters used in Algorithm 1 are set as above. Tables 5–6 list the numerical results for Example 1 with varying n, Tol, and η, where Dist is the distance between T and the computed closest matrix M in the Frobenius norm. Here, we only report the numerical results for $\tau = 0.1$ and $n = 500, 1000, 1500, 2000$, 2500, 3000 and 3500 as the other cases behavior similarly.

From Tables 5–6 we observe that Algorithm 1 is much more efficient than the alternating projection method in [8].

$\texttt{Tol} = 10^{-10}$				
	n	Time	Iter.	Dist
Algorithm 1	500	7.3 s	6	28.0045624140
	1,000	33.6 s	7	56.8449830344
	1,500	1 m 18 s	7	85.6644919002
	2,000	2 m 47 s	8	114.5446535050
$\eta = 10^{-10}$	2,500	4 m 27 s	8	143.4251826139
	3,000	6 m 31 s	8	172.3135425545
	3,500	9 m 08 s	8	201.0862903056
Algorithm 1	500	9.0 s	6	28.0045624140
	1,000	40.9 s	7	56.8449830344
	1,500	1 m 35 s	7	85.6644919002
	2,000	3 m 17 s	8	114.5446535050
$\eta = 10^{-15}$	2,500	5 m 13 s	8	143.4251826139
	3,000	7 m 32 s	8	172.3135425545
	3,500	10 m 16 s	8	201.0862903056
Method in [8]	500	35.1 s	294	28.0045624140
	1,000	4 m 36 s	413	56.8449830344
	1,500	16 m 10 s	503	85.6644919002
	2,000	40 m 01 s	581	114.5446535050
	2,500	1 h 30 m 24 s	643	143.4251826139
	3,000	2 h 47 m 51 s	709	172.3135425545
	3,500	4 h 20 m 10 s	770	201.0862903056

Table 6: Numerical results of Example 1 (b)

4 Conclusions

In this paper we proposed to solve the dual problem (3) by Algorithm 1 in order to obtain the solution of the nearest doubly stochastic matrix problem (1). Under the mild assumptions (4) and (20), we have shown that Algorithm 1 is quadratically convergent. We have also demonstrated its numerical performance by some examples.

In problem (1), only the $(1,1)$ entry of the matrix M is fixed to be identical to the given matrix T. This is an assumption without loss of generality, since the framework we establish in this paper can be easily applied to the nearest doubly stochastic matrix problem with k prescribed entries. In fact, consider the following problem

$$\begin{cases} \min & \frac{1}{2}\|M-T\|_F^2 \\ \text{s.t.} & M\mathbf{e}=\mathbf{e}, \quad \mathbf{e}^T M=\mathbf{e}^T, \\ & \mathbf{e}_{i_1}^T M\,\mathbf{e}_{j_1} = \mathbf{e}_{i_1}^T T\mathbf{e}_{j_1}, \\ & \qquad \vdots \\ & \mathbf{e}_{i_k}^T M\,\mathbf{e}_{j_k} = \mathbf{e}_{i_k}^T T\mathbf{e}_{j_k}, \\ & M \geq 0, \end{cases} \tag{31}$$

where k, i_1, $\cdots$, i_k, j_1, $\cdots$, j_k are integers,

$$1 \leq k \leq n, \quad 1 \leq i_1 \leq \cdots \leq i_k \leq n, \quad 1 \leq j_1 \leq \cdots \leq j_k \leq n,$$

(i_1, j_1), $\cdots$, (i_k, j_k) are distinct, e_j denotes the jth column of the n-by-n identity matrix. Let

$$f(M) := \frac{1}{2}\|M-T\|_F^2, \quad \mathcal{B}(M) := \begin{bmatrix} M\mathbf{e} \\ \begin{bmatrix} I_{n-1} & 0 \end{bmatrix} M^T\mathbf{e} \\ \mathbf{e}_{i_1}^T M\mathbf{e}_{j_1} \\ \vdots \\ \mathbf{e}_{i_k}^T M\mathbf{e}_{j_k} \end{bmatrix}, \quad c := \begin{bmatrix} \mathbf{e} \\ \begin{bmatrix} I_{n-1} & 0 \end{bmatrix} \mathbf{e} \\ \mathbf{e}_{i_1}^T T\mathbf{e}_{j_1} \\ \vdots \\ \mathbf{e}_{i_k}^T T\mathbf{e}_{j_k} \end{bmatrix},$$

then Problem (31) is equivalent to

$$\begin{cases} \min & f(M) \\ \text{s.t.} & \mathcal{B}(M) = c \\ & M \in \mathcal{K} \end{cases} \tag{32}$$

The dual problem of (32) is

$$\begin{cases} \sup & -\theta(x) \\ \text{s.t.} & x \in \mathbb{R}^{2n}, \end{cases} \tag{33}$$

where

$$\theta(x) = \frac{1}{2}\|\Pi_{\mathcal{K}}(T+\mathcal{B}^\star(x))\|_F^2 - x^T c - \frac{1}{2}\|T\|_F^2,$$

and $\mathcal{B}^\star$ is the adjoint of $\mathcal{B}$. Hence, we can extend all results that we derived for the problem (1) to the problem (31), and apply Newton's method to solve the dual problem (33) and then obtain the desired solution of the problem (31).

Acknowledgments

We are very grateful to Dr. Defeng Sun for his valuable suggestions and comments on the reported work.

References

[1] Zhaojun Bai and R.W. Freund, *A partial Padé-via-Lanczos method for reduced-order modeling*, Linear Algebra Appl., 332-334 (2001), pp. 139–164.

[2] R. Feldmann and R.W. Freund, *Efficient linear circuit analysis by Padé approximation via the Lanczos process*, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, 14 (1995), pp. 639–649.

[3] J. D. Louck, *Doubly stochastic matrices in quantum mechanics*, Found. Plys., 27 (1997), pp. 1085–1104.

[4] J.P. Boyle and R.L. Dykstra, *A method for finding projections onto the intersections of convex sets in Hilbert spaces*, In Advances in Order Restricted Statistical Inference, Eds: R. Dykstra, T. Robertson and F. T. Wright, Lecture Notes in Statistics, Vol. 37, pp. 28–47, Springer- Verlag, Berlin, New York, 1986.

[5] R.A. Brualdi, *Some applications of doubly stochastic matrices*, Linear Algebra Appl., 107 (1988), pp. 77–100.

[6] H. Chiang and C.K. Li, *Generalized doubly stochastic matrices and linear preservers*, Linear Multilinear Algebra, 53 (2005), pp. 1–11.

[7] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[8] W. Glunt, T. L. Hayden, and R. Reams, *The nearest doubly stochastic matrix to a real matrix with the same first moment*, Numer. Linear Algebra Appl., 5 (1998), pp. 475–482.

[9] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd edition, Johns Hopkins University Press, Baltimore and London, 1996.

[10] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[11] R.N. Khoury, *Closest matrices in the space of generalized doubly stochastic matrices*, J. Math. Anal. Appl., 222 (1998), pp, 562–568.

[12] B. Kummer, *Newton's method for nondifferentiable functions*, Advances in Mathematical Optimization, 114–125, Math. Res., 45, Akademie-Verlag, Berlin, 1988.

[13] J. Malick, *A dual approach to semidefinite least-squares problems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 272–284.

[14] M. Marcus, *Some properties and applications of doubly stochastic matrices*, Amer. Math. Monthly, 67 (1960), pp. 215–221.

[15] H. Minc, *Nonnegative Matrices*, Wiley, New York, 1988.

[16] H.-D. Qi and D. Sun, *Quadratic convergence and numerical experiments of Newton's method for computing the nearest correlation matrix*, Preprint, Department of Mathematics, National University of Singapore, Singapore, 2005.

[17] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[18] J. E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, New Jersey, 1983.

[19] R.T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.

[20] R.T. Rockafellar and R.J.B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[21] A. Shapiro, *Sensitivity analysis of generalized equations*, J. Math. Sci., 115 (2003), pp. 2554–2565.

[22] V.S. Sunder and N.J. Wildberger, *Action of finite hypergroups and examples harmonic analysis and hypergroups*, (Boston, MA: Birkhauser), 1998.

[23] E.H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory I and II.* In E.H. Zarantonello, editor, *Contributions to Nonlinear Functional Analysis*, Academic Press (New York, 1971) pp. 237–424.

[24] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[25] S. Chandrasekaran, P. Dewilde, M. Gu, T. Pals, X. Sun, A.J. van der Veen, and D. White, *Some fast algorithms for sequentially semiseparable representations*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 341–364.

[26] S. Chandrasekaran and M. Gu, *Fast and stable algorithms for banded plus semiseparable systems of linear systems*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 373–384.

[27] Y. Eidelman and I. Gohberg, *A modification of the Dewilde-van der Veen method for inversion of finite structured matrices*, Linear Algebra Appl, 343/344 (2002), pp. 419–450.

[28] N. Mastronardi, S. Chandrasekaran and S. van Huffel, *Fast and stable two-way chasing algorithms for diagonal plus semi-separable systems of linear equations*, Numer. Linear Algebra Appl., 38 (2000), pp. 7–12.