A Semismooth Newton-Type Method for the Nearest Doubly Stochastic Matrix Problem

Hao Hu * Haesol Im † Xinxin Li‡ Henry Wolkowicz§

12:12 Noon, July 19, 2021

Department of Combinatorics and Optimization
Faculty of Mathematics, University of Waterloo, Canada.
Research supported by NSERC.

Contents

*Department of Combinatorics and Optimization Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; Research supported by The Natural Sciences and Engineering Research Council of Canada; www.huhao.org.

†Department of Combinatorics and Optimization Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; Research supported by The Natural Sciences and Engineering Research Council of Canada.

‡School of Mathematics, Jilin University, Changchun, China. E-mail: xinxinli@jlu.edu.cn. This work was supported by the National Natural Science Foundation of China (No.11601183) and Natural Science Foundation for Young Scientist of Jilin Province (No. 20180520212JH).

§Department of Combinatorics and Optimization Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; Research supported by The Natural Sciences and Engineering Research Council of Canada; www.math.uwaterloo.ca/~hwolkowi.

List of Figures

List of Tables

List of Algorithms

Abstract

We study a semismooth Newton-type method for the nearest doubly stochastic matrix problem where both differentiability and nonsingularity of the Jacobian can fail. The optimality conditions for this problem are formulated as a system of strongly semismooth functions. We show that the so-called local error bound condition does not hold for this system. Thus the guaranteed convergence rate of Newton-type methods is at most superlinear. By exploiting the problem structure, we construct a modified two step semismooth Newton method that guarantees a nonsingular Jacobian matrix at each iteration, and that converges to the nearest doubly stochastic matrix quadratically. To the best of our knowledge, this is the first Newton-type method which converges Q-quadratically in the absence of the local error bound condition.

Key Words: nearest doubly stochastic matrix, semismooth newton method, strongly semismooth, quadratic convergence, equivalence class.

1 Introduction

Newton's method is a powerful, popular iterative technique for solving systems of nonlinear equations. The popularity arises from its fast asymptotic convergence rate. But this fast convergence requires assumptions such as: nonsingularity of the Jacobian matrix at the solution, or the so-called *local error bound condition*, see Definition 4.2 below, and e.g., [14,18,22,42,45]. These assumptions unfortunately can fail for many interesting applications. Recent extensions when nonsingularity fails in the differentiable case appears in e.g., [32, 33] and the references therein. In this paper, we present a two-step semismooth Newton-type algorithm for the nearest doubly stochastic matrix problem. We illustrate that it is still possible to achieve a Q-quadratic convergence rate even if the above assumptions fail. To our knowledge this is the first Newton-type method to have a provable Q-quadratic convergence rate without the local error bound condition. We include empirical evidence that illustrates the improved speed and accuracy of our algorithm compared to several other methods in the literature.

The proposed algorithm is also closely related to the recent developments for solving semidefinite programming relaxations using alternating direction method of multipliers (ADMM), see [8, 10]. The ADMM is recently proven to be a powerful method for solving facially reduced semidefinite programs. It is currently the most efficient technique to approximately solve the semidefinite relaxations of various hard combinatorial problems, see [25, 30, 31, 43]. For example, the nearest doubly stochastic matrix problem can serve as a subproblem in solving certain relaxations for the quadratic assignment problem, e.g., [25, 36]. An efficient algorithm for solving this subproblem is the key to push the computational limit further. The algorithm presented in this paper is efficient and robust, which serves this purpose.

1.1 Preliminaries

A doubly stochastic matrix is a nonnegative square matrix $X \in \mathbb{R}^{n\times n}$ whose rows and columns sum to one. Doubly stochastic matrices have many applications for example in economics, probability and statistics, quantum mechanics, communication theory and operation research, e.g., [11, 37, 39]. The nearest doubly stochastic matrix, but with a prescribed entry, has been studied in [3]; it is related to the numerical simulation of large circuit networks.

Throughout this paper we assume we are given a matrix $\hat{X} \in \mathbb{R}^{n\times n}$. The problem of computing its nearest doubly stochastic matrix is formally given by

$$\begin{array}{ll} \min & \|X - \hat{X}\|^2 \\ \text{s.t.} & Xe = e, \\ & X^T e = e, \\ & X \geq 0, \end{array} \tag{1.1}$$

dsm:main

where $\|\cdot\|$ is the Frobenius norm, and $e \in \mathbb{R}^n$ is the all-ones vector. Here, the column sum constraints appear first. Moreover, the constraints can be viewed within the family of *network flow problems* as they define the *assignment problem* constraints, e.g., [4, Chap. 7].

1.1.1 A Vectorized Formulation and Optimality Conditions

The nearest doubly stochastic matrix problem (1.1) is defined using the matrix variable $X \in \mathbb{R}^{n\times n}$. It is often more convenient to work with vectors, and therefore we shall derive an equivalent formulation using a vector of variables $x \in \mathbb{R}^{n^2}$.

Let $x = \text{vec}(X) \in \mathbb{R}^{n^2}$ denote the vector obtained by stacking the columns of $X \in \mathbb{R}^{n\times n}$. Conversely, $X = \text{Mat}(x) \in \mathbb{R}^{n\times n}$ is the unique matrix such that $x = \text{vec}(X)$. Recall that the matrix equation $AXB = C$ can be written as $(B^T \otimes A)\,\text{vec}(X) = \text{vec}(C)$, where $\otimes$ denotes the Kronecker product. Therefore, the equality constraints in (1.1) are equivalent to

$$(I \otimes e^T)x = e \text{ and } (e^T \otimes I)x = e.$$

Thus we can express the feasible region of (1.1) in vector form as the set $\{x \in \mathbb{R}^{n^2} : \bar{A}x = \bar{e}, x \geq 0\}$, where $\bar{e} \in \mathbb{R}^{2n}$ is the all-ones vector and the matrix $\bar{A}$ is

$$\bar{A} = \begin{bmatrix} I \otimes e^T \\ e^T \otimes I \end{bmatrix} = \begin{bmatrix} e^T & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & e^T \\ I & \cdots & I \end{bmatrix} \in \mathbb{R}^{2n \times n^2}. \tag{1.2}$$

?fullA?

It is easy to see that one of the equality constraints is redundant. Therefore, we discard the last row in $\bar{A}$, i.e., the $2n$-th constraint that the last row of X sums to one. We denote this by A. We observe that the matrix with all elements $1/n$ is strictly feasible. Therefore, we now have that the Mangasarian-Fromovitz constraint qualification, MFCQ, holds. This further means that the set of optimal dual variables is compact, [23, 38].

Let $\hat{x} = \text{vec}(\hat{X})$. The doubly stochastic matrix problem (1.1) in the vector form is given by the unique minimum of the strictly convex minimization problem

$$x^* = \text{argmin} \left\{ \frac{1}{2} \|x - \hat{x}\|^2 \, : \, Ax = b, \, x \geq 0 \right\}, \tag{1.3}$$

?eq:mainprob?

where $A \in \mathbb{R}^{(2n-1)\times n^2}$ is the first $2n-1$ rows of $\bar{A}$, and $b \in \mathbb{R}^{2n-1}$ is the all-ones vector. The optimal doubly stochastic matrix is denoted by $X^* = \text{Mat}(x^*)$. By abuse of notation, where needed we often use double indices for the vectors $x = (x_{ij}) \in \mathbb{R}^{n^2}$.[1]

The standard Karush-Kuhn-Tucker, KKT, optimality conditions for the primal-dual variables (x, y, z) for (1.3) are:

$$\begin{bmatrix} x - \hat{x} - A^T y - z \\ Ax - b \\ z^T x \end{bmatrix} = 0, \quad x, z \in \mathbb{R}^{n^2}_+, y \in \mathbb{R}^{2n-1}. \tag{1.4}$$

eq:optcondxyz_c

The system (1.4) is a bilinear system of order n^2. Theorem 1.1 below shows that we can simplify the KKT conditions and obtain an elegant characterization of optimality. This new optimality condition is a smaller system of size $2n-1$. However, the new system involves a nonsmooth (metric) projection of a given v onto the nonnegative orthant, denoted $v_+ = \text{argmin}_x\{\|x - v\| : x \geq 0\}$. (The absolute value of the projection onto the nonpositive orthant is denoted v_-.) Therefore we get $v = v_+ - v_-$, $v_+^T v_- = 0$.

⟨thm:optcondG⟩? **Theorem 1.1.** *Let $\hat{x} \in \mathbb{R}^{n^2}$ be given. The optimal solution $x^* \in \mathbb{R}^{n^2}$ for the nearest doubly stochastic problem (1.3) exists and is unique. Moreover, $x^* \in \mathbb{R}^{n^2}$ solves (1.3) if, and only if,*

$$x^* = (\hat{x} + A^T y^*)_+, \quad F(y^*) := A(\hat{x} + A^T y^*)_+ - b = 0, \text{ for some } y^* \in \mathbb{R}^{2n-1}. \tag{1.5}$$

?eq:optcondG?

Proof. The existence and uniqueness of the optimum x follows since (1.3) is a projection onto a closed convex set. The Lagrangian dual of (1.3) is

$$\max_{z \geq 0, y} \min_x L(x, y, z) = \frac{1}{2}\|x - \hat{x}\|^2 - y^T(Ax - b) - z^T x.$$

[1] The perturbation function (optimal value function) is $p^*(\epsilon) = \min\left\{\frac{1}{2}\|x - \hat{x}\|^2 : Ax = b + \epsilon, \, x \geq 0\right\}$. Then $\partial p^*(0) = \{y\}$ is the set of optimal dual multipliers, which is always a compact, convex set since MFCQ holds. So differentiability holds if, and only if, it is a singleton.

For nonnegative vectors $z, x \geq 0$, the optimality is characterized by the KKT conditions (1.4), i.e., from dual feasibility ($\nabla_x L(x,y,z) = (x-\hat{x}) - A^T y - z = 0$), primal feasibility, complementary slackness, respectively. We get

$$0 \leq x = (\hat{x} + A^T y)_+ - (\hat{x} + A^T y)_- + z,\ z \geq 0,\ Ax = b,\ z_i x_i = 0, \forall i. \tag{1.6}$$?eq:optcondplus

This implies

$$x = (\hat{x} + A^T y)_+,\ z = (\hat{x} + A^T y)_-.$$

□

It follows from Theorem 1.1 that if $F(y) = 0$, then $x = (\hat{x} + A^T y)_+$ is the optimal primal point and $z = (\hat{x} + A^T y)_-$ is an optimal dual vector for (1.3). We note that this characterization is well-known, and it can also be derived for more general results for finite dimensional problems e.g., [1, 47], and for infinite dimensional problems, see e.g. [7, 9, 20, 41]. In [44], this reformulation strategy is used for the nearest correlation matrix problem. They also prove that the obtained semismooth system has a nonsingular Jacobian at the optimum and leads to a very competitive algorithm. This is in contrast to our problem, where the generalized Jacobian at the optimum can contain many highly singular matrices.

Remark 1.2. *Our problem is a special case of the linearly constrained linear least squares problem, e.g. [34], that is itself a special case of quadratic programming, e.g. [19]. These problems lie within the class of linear complementarity problems, e.g., [16].*

In contrast to our dual type algorithm that applies a Newton-type method to the optimality conditions, the approaches in the literature include:

1. *active set methods, e.g., [5];*
2. *quadratic cost network flow problems, e.g., [6, 24];*
3. *path following, interior point methods, e.g. [19], that also use a Newton method applied to perturbed optimality conditions;*
4. *classical Lemke and Wolfe type methods, e.g., [16];*
5. *splitting methods such as ADMM, e.g. [25].*

1.1.2 Semi-smooth Newton Methods

In this paper we solve the nearest matrix problem by applying a semismooth Newton method to the nonsmooth optimality conditions of (1.3) in the form $F(y) = 0$. We now present the preliminaries for semismooth Newton methods.

Suppose $F : \mathbb{R}^s \to \mathbb{R}^t$ is locally Lipschitzian. According to Rademacher's Theorem [46], F is Frechét differentiable almost everywhere. Denote by D_F the set of points at which F is differentiable. Let $F'(y)$ be the usual Jacobian matrix at $y \in D_F$. The generalized Jacobian $\partial F(y)$ of F at y in the sense of Clarke [14] is

$$\partial F(y) := \operatorname{conv} \left\{ \lim_{\substack{y_i \to y \\ y_i \in D_F}} F'(y_i) \right\}. \tag{1.7}$$ {?}

The generalized Jacobian $\partial F(y)$ is said to be *nonsingular*, if every $V \in \partial F(y)$ is nonsingular. The Lipschitz continuous function F is *semismooth* at y, if F is directionally differentiable at y and

$$\|F(y+d) - F(y) - Vd\| = \mathcal{O}(\|d\|), \ \forall\, V \in \partial F(y+d) \text{ and } d \to 0.$$

Moreover, F is *strongly semismooth* at y, if F is semismooth at y and

$$\|F(y+d) - F(y) - Vd\| = \mathcal{O}(\|d\|^2) \quad \forall\, V \in \partial F(y+d) \text{ and } d \to 0.$$

We note that the projection operator v_+ in our optimality conditions (1.5) is a special case of a metric projection operator and is strongly semismooth e.g., [13, 48].

Now let y^0 be a given initial point. If $\partial F(y)$ is nonsingular, the semismooth Newton method for solving equation $F(y) = 0$ is defined by the iterations

$$\boxed{y^{k+1} = y^k - V_k^{-1} F(y^k), \text{ with } V_k \in \partial F(y^k).} \tag{1.8}$$?semiNM?

A sequence $\{y^k\}$, is said to *converge Q-quadratically* to y^*, if $y^k \to y^*$ and

$$\limsup_{k\to\infty} \frac{\|y^{k+1} - y^*\|}{\|y^k - y^*\|^2} < M, \text{ for some positive constant } M > 0.$$

The following local convergence result for the semismooth Newton method as applied to a semismooth function F is due to [45].

?⟨semiNMconv⟩? **Theorem 1.3.** *[45] Let $F(y^*) = 0$ and let $\partial F(y^*)$ be nonsingular. If F is (strongly) semismooth at y^*, then the semismooth Newton method (1.8) is (Q-quadratically) convergent in a neighborhood of y^*.*

The nonsingularity assumption for $\partial F(y^*)$ can be a restrictive assumption for the convergence of semismooth (and smooth) Newton methods. This condition is not satisfied by many applications, including our nearest doubly stochastic matrix problem. In these cases, regularization such as the Levenberg-Marquardt method (LMM) could be used to obtain the nonsingularity. If F is differentiable and the local error bound condition is satisfied, see Definition 4.2 below, then the LMM approach achieves quadratic convergence, see [21, 35, 40, 50]. Note that the local error bound condition does not hold for the nearest doubly stochastic matrix problem, see Theorem 4.3.

1.2 Contributions

1. We present a modified two-step semismooth Newton method that exploits the special network structure of our nearest matrix problem.
2. At each iterate y, the first step finds a point (vertex) y' in the same equivalence class so that we can guarantee that the matrix chosen from the generalized Jacobian is nonsingular. Thus a regular Newton step can be taken.
3. This two-step method converges Q-quadratically for the nearest doubly stochastic matrix problem. This is done in the absence of the local error bound condition. The problem structure allows for Q-quadratic convergence to the solution. The main idea of our algorithm is to partition the search space into equivalence classes so that the difficulty of singular generalized Jacobians can be avoided.
4. The numerical tests show that our algorithm outperforms existing methods both in speed and accuracy. The algorithm is also very robust for difficult instances.

2 Semi-smooth Newton Method for Connected X^*

?⟨sec:semi⟩? In this section, we show that the semismooth Newton method (1.8) can be used to find a solution to the optimality conditions (2.5) when the bipartite graph for the optimal primal solution $X^* = \text{Mat}(x^*)$ of (1.3) is *connected*.

2.1 Bipartite Graphs and Connectedness

For every matrix $X \in \mathbb{R}^{m\times n}$ we associate a bipartite graph $G = (V, E)$ with node set divided into two V_1, V_2 corresponding to the rows and columns of X, respectively. The edges $(i, j) \in E$ correspond to *nonzero* entries of X, i.e., $ij \in E \iff i \in V_1, j \in V_2, X_{ij} \neq 0$. The adjacency matrix of G can be written as

$$\begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}. \tag{2.1}$$

?eq:Badj?

We call the zero-one matrix B the *reduced adjacency matrix* of the bipartite graph G. We call B and X *connected* matrices if the graph G is connected. We call them *disconnected* otherwise.

?⟨def:cm⟩? **Lemma 2.1** (*connected matrix*, [12, page 109])**.** *A matrix $X \in \mathbb{R}^{m\times n}$ is connected if there do not exist permutation matrices P and Q such that*

$$PXQ = \begin{bmatrix} X^1 & 0 \\ 0 & X^2 \end{bmatrix},$$

where X^1 is $p \times q$ satisfying $1 \leq p + q \leq m + n - 1$.[2]

Let $N = \{1, \ldots, n\}$. In this paper we consider square matrices $X \in \mathbb{R}^{n\times n}$ so that the associated bipartite graph has edges $ij \in N \times N$. Let $R, C \subseteq N$,be two subsets, with $\bar{R}$ and $\bar{C}$ the respective complements in N. Then $X \in \mathbb{R}^{n\times n}$ can be partitioned and permuted using the two subsets as

$$\begin{bmatrix} X_{\bar{R},\bar{C}} & X_{\bar{R},C} \\ X_{R,\bar{C}} & X_{R,C} \end{bmatrix}. \tag{2.2}$$

?blk_X?

We note that X is connected if both $X_{R,\bar{C}}, X_{\bar{R},C}$ are zero or empty blocks, for some pair of subsets. We emphasize that the diagonal blocks are *not* necessarily square; and if X is disconnected, then one of them can be empty and thus there are zero rows or columns.

We partition the dual variables y corresponding to the column and row sum constraints as

$$y = \begin{pmatrix} c \\ r \end{pmatrix} \in \mathbb{R}^{2n-1}, \quad c \in \mathbb{R}^n, r \in \mathbb{R}^{n-1}. \tag{2.3}$$

?eq:ycr?

The structure of A enables us to write the equation $X = \text{Mat}\left(\hat{x} + A^T y\right)_+$ as

$$X_{ij} = \begin{cases} (\hat{X}_{ij} + r_i + c_j)_+ & \text{if } i \neq n, \forall j, \\ (\hat{X}_{ij} + c_j)_+ & \text{if } i = n, \forall j. \end{cases} \tag{2.4}$$

?Xrc?

[2] A connected matrix is often called indecomposable in the literature.

2.2 The Algorithm for Connected X^*

Recall that the optimality conditions function

$$F(y) = A(\hat{x} + A^T y)_+ - b = 0, \tag{2.5}$$?nm_eq?

is *strongly semismooth*, see [13, 48]. Our algorithm is based on applying a Newton-type method to solve this equation. More precisely, at each iterate y, we have $x = (\hat{x} + A^T y)_+$ and $z = (\hat{x} + A^T y)_-$, and so we guarantee dual feasibility and complementarity:

$$x - (\hat{x} + A^T y) - z = 0,\ x^T z = 0, \quad x, z \geq 0.$$

The Newton algorithm solves $F(y) = 0$ in order to obtain the missing linear primal feasibility $Ax = b$.

Below we provide a sufficient condition for the nonsingularity of the generalized Jacobian $\partial F(y)$ at a $y \in \mathbb{R}^{2n-1}$. From Theorem 1.3, we see that this sufficient condition then guarantees that the semismooth Newton method converges locally to an optimum with a Q-quadratic convergence rate

In order to obtain the generalized Jacobian at $y \in \mathbb{R}^{2n-1}$, we need the following set. Recall that we use double indices for vectors $x = (x_{ij}) = ((\hat{x} + A^T y)_{ij}) \in \mathbb{R}^{n^2}$.

$$\mathcal{M}(y) := \left\{ M \in \mathbb{R}^{n \times n} \mid M_{ij} = \left\{ \begin{array}{cl} 1 & \text{if } (\hat{x} + A^T y)_{ij} > 0 \\ {[0,1]} & \text{if } (\hat{x} + A^T y)_{ij} = 0 \\ 0 & \text{if } (\hat{x} + A^T y)_{ij} < 0 \end{array} \right. \right\}. \tag{2.6}$$?My?

Note that the *minimal* M, elementwise, is the adjacency matrix for $\operatorname{Mat}(\hat{x} + A^T y)_+$.

The generalized Jacobian of the non-linear system (2.5) at y is given by the set

$$\partial F(y) = \{A \operatorname{Diag}(\operatorname{vec}(M)) A^T \mid M \in \mathcal{M}(y)\}. \tag{2.7}$$?eq:jac?

For example, for the case where $\hat{x} + A^T y > 0$ and Δy small, we get

$$F(y + \Delta y) = A(\hat{x} + A^T(y + \Delta y)) - b = F(y) + A \operatorname{Diag}(\operatorname{vec}(M)) A^T \Delta y,\ \operatorname{Diag}(\operatorname{vec}(M)) = I.$$

In the general case, we replace the elements of M with appropriate $M_{ij} \in [0,1]$. In our applications, we choose $M_{ij} \in \{0,1\}$, and in fact, we choose the maximal M as defined below in (3.9). Therefore, in our applications, every $V \in \partial F(y)$ is a sum of rank one zero-one matrices.

The next result shows that the matrices in the generalized Jacobian have a special structure in terms of the matrices M in (2.6).

?⟨Vstructure⟩? **Proposition 2.2.** *Let $y \in \mathbb{R}^{2n-1}$ be given and let $M \in \mathcal{M}(y)$. Then the linear transformation*

$$\mathcal{V}(M) := A \operatorname{Diag}(\operatorname{vec}(M)) A^T \in \partial F(y) \subset \mathbb{S}^{2n-1}_+. \tag{2.8}$$?eq:jac2?

Moreover, $\partial F(y)$ is a nonempty, convex compact set. And $\partial F(y)$ is a singleton if, and only if F is differentiable if, and only if, $\mathcal{M}(y)$ is a singleton.

Now let

$$M \in \mathcal{M}(y) \in \mathbb{R}^{n \times n}, \quad \hat{M} \in \mathbb{R}^{(n-1) \times n},$$

where the latter is formed from the first $n-1$ rows of M. Then the matrix $\mathcal{V}(M)$ has the following structure

$$\mathcal{V}(M) = \begin{bmatrix} \operatorname{Diag}(M^T e) & \hat{M}^T \\ \hat{M} & \operatorname{Diag}(\hat{M} e) \end{bmatrix}.^{3} \tag{2.9}$$?eq:MMhatstruc?

Proof. The convexity and compactness of the generalized Jacobian are well known properties. The singleton property is clear from the definitions. Note that $A \geq 0$ with no zero columns.

The expression for $\mathcal{V}(M)$ follows from the structure of A. □

For $y \in \mathbb{R}^{2n-1}$, the fact that the matrix $\mathcal{V}(M)$ in (2.8) is nonsingular for a given $M \in \mathcal{M}(y)$, is equivalent to linear independence of $2n-1$ columns in A associated to a subset of the positive entries in M. Therefore, it is clear that one should choose as many elements $M_{i,j} > 0$ as possible to obtain a nonsingular element in the generalized Jacobian. In what follows, we derive an alternative characterization that connects the nonsingularity of $\mathcal{V}(M)$ and the connectedness of M, see Lemma 2.3 below.

?⟨Vsing⟩? **Lemma 2.3.** *Let $M \geq 0$. The matrix $\mathcal{V}(M)$ is nonsingular if, and only if, M is connected.*

Proof. This result can be derived easily using [15, Prop. 2.15]. Translated to our framework, The result states that a set of linearly independent columns in our matrix A forms a basis of $\mathbb{R}^{2n-1}$ if, and only if, the associated set of arcs forms a spanning tree. These arcs correspond to the graph of our matrices M. The result follows by noting that $\mathcal{V}(M)$ is positive definite if and only if the bipartite graph associated with M is connected and thus it contains a spanning tree. (M is obtained removing the last row and column of the signless Laplacian of the adjacency matrix of the graph G in (2.1). Results on singularity for signless Laplacians appear in e.g., [17, Prop. 2.1], and [49] for the reduced signless Laplacian.)

We now include an alternative proof for the sake of self-containment. We denote $V = \mathcal{V}(M)$.

First suppose that M is disconnected. We now proof M is singular. We distinguish the following two cases.

1. Now consider the special case that M contains a zero column or a zero row among the first $n-1$ rows. Then there is a zero diagonal entry, see Proposition 2.2. Since V is positive semidefinite, we conclude that V is singular.

   For the case where the last row of M is zero, we have that $M^T e - \hat{M}^T e = 0$ and thus there is an eigenvector that has a 0 eigenvalue, i.e., by abuse of notation and using e of different dimensions, we see that

$$V \begin{bmatrix} e \\ -e \end{bmatrix} = \begin{bmatrix} \operatorname{Diag}(M^T e)e - \hat{M}^T e \\ \hat{M} e - \operatorname{Diag}(\hat{M} e)e \end{bmatrix} = \begin{bmatrix} M^T e - \hat{M}^T e \\ \hat{M} e - \hat{M} e \end{bmatrix} = 0.$$

2. We now assume that M and $\hat{M}$ can be permuted so that they have the form

$$M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix} \text{ and } \hat{M} = \begin{bmatrix} M_1 & 0 \\ 0 & \hat{M}_2 \end{bmatrix}.$$

[3] $\mathcal{V}(M)$ can be obtained by removing the last row and column of the signless Laplacian of the adjacency matrix of the graph G in (2.1), [17, 28].

Using Proposition 2.2, we obtain

$$V = \begin{bmatrix} \mathrm{Diag}(M_1^T e) & 0 & M_1^T & 0 \\ 0 & \mathrm{Diag}(M_2^T e) & 0 & \hat{M}_2^T \\ M_1 & 0 & \mathrm{Diag}(M_1 e) & 0 \\ 0 & \hat{M}_2 & 0 & \mathrm{Diag}(\hat{M}_2 e) \end{bmatrix}.$$

But then the first block column and the third block column of V are linearly dependent. Thus V is singular.

Thus we have shown that V is singular in both cases. (Note that we do not need the nonnegativity condition $s \geq 0$ in this direction.)

Conversely, assume that V is singular. Then there exists a non-zero vector $w = \begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{R}^{2n-1}$, for some $u \in \mathbb{R}^n$, $v \in \mathbb{R}^{n-1}$ such that $w^T V w = 0$. We can rewrite $w^T V w$ as follows.

$$\begin{aligned} w^T V w &= u^T \mathrm{Diag}(M^T e) u + 2u^T \hat{M}^T v + v^T \mathrm{Diag}(\hat{M} e) v \\ &= \textstyle\sum_{j=1}^n u_j^2 M_{n,j} + \sum_{i=1}^{n-1} \sum_{j=1}^n (v_i + u_j)^2 M_{i,j} \\ &= \langle W, M \rangle, \end{aligned} \tag{2.10}$$

?eq:Vsing?

where

$$W := \begin{bmatrix} (v_1 + u_1)^2 & \cdots & (v_1 + u_n)^2 \\ \vdots & \ddots & \vdots \\ (v_{n-1} + u_1)^2 & \cdots & (v_{n-1} + u_n)^2 \\ u_1^2 & \cdots & u_n^2 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Up to permutation, we can assume $v_1, \ldots, v_{k_1}$ and $u_1, \ldots, u_{k_2}$ are the only non-zero entries in w, where k_1, k_2 are nonnegative integers. Note that $k_1 + k_2 > 0$ as $w \neq 0$. We distinguish the following cases based on k_1 and k_2.

1. Suppose that $0 < k_1$ and $0 < k_2 < n$. The matrix W can be partitioned correspondingly as

$$W = \begin{bmatrix} W_1 & W_{12} \\ W_{21} & W_2 \end{bmatrix} \in \mathbb{R}^{n \times n},$$

   with non-trivial off-diagonal blocks $W_{12} \in \mathbb{R}^{k_1 \times (n-k_2)}$ and $W_{21} \in \mathbb{R}^{(n-k_1) \times k_2}$. By assumption, $W_{12} > 0$ and $W_{21} > 0$ are element-wise positive. Partitioning M in the same way yields

$$M = \begin{bmatrix} M_1 & M_{12} \\ M_{21} & M_2 \end{bmatrix}.$$

   Note that $W \geq 0$ and $M \geq 0$. As $\langle W, M \rangle = w^T V w = 0$, this implies that the off-diagonal blocks $M_{12} \in \mathbb{R}^{k_1 \times (n-k_2)}$ and $M_{21} \in \mathbb{R}^{(n-k_1) \times k_2}$ must be zero. Therefore, M is a block-diagonal matrix.

2. Using the same argument as above, the remaining three possibilities lead to a zero row or column in M. They are listed below.

$$\begin{aligned} &k_1 = 0, k_2 = n && \Longrightarrow && M = 0; \\ &k_1 = 0, 0 < k_2 < n && \Longrightarrow && \text{the first } k_2 \text{ columns of } M \text{ are zeros;} \\ &k_1 > 0, k_2 = 0 && \Longrightarrow && \text{the first } k_1 \text{ rows of } M \text{ are zeros.} \end{aligned}$$

This shows that M is disconnected. ∎

We now provide two properties of the generalized Jacobian that show the relationships between nonsingularity, connectedness and also differentiability and strict complementarity.

?⟨sys_sing⟩? **Theorem 2.4.** *Let $y \in \mathbb{R}^{2n-1}$, and set*

$$X := \operatorname{Mat}(\hat{x} + A^T y)_+, \quad Z := \operatorname{Mat}(\hat{x} + A^T y)_-.$$

Then the following holds:

1. *The generalized Jacobian $\partial F(y)$ is nonsingular if, and only if, the matrix X is connected.*
2. *The generalized Jacobian $\partial F(y)$ is a singleton (F is differentiable at y) if, and only if, strict complementarity, $X + Z > 0$ holds.*

Proof. 1. Let $M' \in \mathcal{M}(y)$ be such that $M'_{ij} = 0$ if $X_{ij} = 0$, i.e., M' is the smallest elementwise. Note that $\mathcal{V}(\cdot)$ is a monotonic mapping, i.e., for any $M \in \mathcal{M}(y)$, we have $M' \leq M$ and thus $\mathcal{V}(M') \preceq \mathcal{V}(M)$. Hence we have

$$\begin{aligned}
\partial F(y) \text{ is nonsingular} \quad &\iff \quad \mathcal{V}(M) \text{ is nonsingular } \forall M \in \mathcal{M}(y) && \text{(by definition)} \\
&\iff \quad \mathcal{V}(M') \text{ is nonsingular for smallest } M' && \\
&\iff \quad M' \text{ is connected} && \text{(by Lemma 2.3)} \\
&\iff \quad X \text{ is connected,} &&
\end{aligned}$$

where the last equivalence follows since $X_{ij} > 0 \iff M'_{ij} > 0, \forall ij$ for the smallest M'

2. From the definitions of $\mathcal{M}(y)$ (2.6) and the Jacobian in (2.7), and the fact that $A \geq 0$ with no zero columns, we conclude that $\partial F(y)$ is a singleton (differentiability) holds if, and only if, $\mathcal{M}(y)$ is a singleton. By definition, this is equivalent to strict complementarity. Note that if $M \in \mathcal{M}(y)$, if strict complementarity holds, we have

$$X_{ij} = 0 \implies Z_{ij} > 0 \implies \operatorname{Mat}(\hat{x} + A^T y)_{ij} < 0 \implies M_{ij} = 0.$$

(See also Proposition 2.2.)

∎

?⟨main_conv⟩? **Corollary 2.5.** *Suppose $F(y^*) = 0$ and $X^* = \operatorname{Mat}(\hat{x} + A^T y^*)_+$ is connected. Then the semismooth Newton method (1.8) has local quadratic convergence to y^*.*

Theorem 2.4 and Corollary 2.5 show that if differentiability fails at the optimum, then strict complementarity fails. This type of degeneracy is typically tied to ill-conditioning and slow convergence. Similarly, if the optimum is disconnected, we get problems with singular generalized Jacobians. This motivates the next section that deals with finding nonsingular matrices in the generalized Jacobian.

3 An All-inclusive Semi-smooth Newton Method

In this section we develop an algorithm that allows for the cases where the optimal solution X^* is *disconnected*. In this case the generalized Jacobian $\partial F(y)$ of the non-linear system (2.5) is singular. Hence the iterates of the semismooth Newton method (1.8) are not well-defined, and the convergence result in Corollary 2.5 is not applicable, see e.g., [45]. In fact, the iterate in (1.8) may not even be defined at all, since every matrix $V \in \partial F(y)$ is singular; see Theorem 2.4 below. Note that this now includes the important cases where the optimal solution is a permutation matrix, a matrix that *highly* disconnected. We show that we can move from each iterate y to a point y' in the same equivalence class, see Definition 3.1 below, so that we can find a matrix that is *non*singular in the generalized Jacobian at y'.

We now modify the semismooth Newton method (1.8) so that the iterates in the modified algorithm are well-defined, and the convergence rate is quadratic even if X^* is *disconnected*. The main idea for constructing well-defined iterates is outlined as follows:

> for any vector y, construct an *equivalent* vector y' so that there exists at least one nonsingular matrix in $\partial F(y')$ to obtain a well-defined next iterate.

3.1 Equivalence Classes

This section introduces the notion of *equivalence classes* of y corresponding to a given dual feasible X. This is related to the *normal cone* at X. Our Newton method finds iterates y, but we see below that we are in particular interested in moving between equivalence classes of y. And in particular, we are interested in a special point y in each equivalence class.

3.1.1 Preliminaries

We first define an equivalence relation for a partition of the underlying space $\mathbb{R}^{2n-1}$ to use for our modified Newton method.

Definition 3.1 (*equivalence class,* $[y]$)**.** *Two vectors y and y' in $\mathbb{R}^{2n-1}$ are equivalent, denoted by $y \sim y'$, if*

$$(\hat{x} + A^T y)_+ = (\hat{x} + A^T y')_+.$$

The set of equivalent vectors in $\mathbb{R}^{2n-1}$ is called the equivalence class. We denote the equivalence class to which y belongs to by

$$[y] := \{y' \in \mathbb{R}^{2n-1} \mid y \sim y'\}.$$

Recall that the nonnegative polar cone of a closed convex set C at $w \in C$ is given by $(C-w)^+ = \{v : (c-w)^T v \geq 0, \ \forall c \in C\}$. We can show that each equivalence class is actually a polyhedron that can be viewed in the $y \in \mathbb{R}^{2n-1}$ space, or equivalently in the $x \in \mathbb{R}^{n^2}$ space. The associated linear equations and inequalities are given explicitly in the next result.

Lemma 3.2. *Let $\tilde{y} \in \mathbb{R}^{2n-1}$ and $\tilde{x} = (\hat{x} + A^T \tilde{y})_+$. Then the following are equivalent:*

1. $y \in [\tilde{y}]$
2. $$\begin{array}{rcll} (A^T y)_i &=& (\hat{x} - \tilde{x})_i & \text{if} \quad \tilde{x}_i > 0, \\ (A^T y)_j &\leq& (\hat{x} - \tilde{x})_j & \text{if} \quad \tilde{x}_j = 0. \end{array}$$

3.

$$\tilde{x} - \hat{x} - A^T y \in (\mathbb{R}^{n^2}_+ - \tilde{x})^+.$$

Proof. A vector y is contained in $[\tilde{y}]$ if, and only if, $\tilde{x}$ is the optimal solution of the following optimization problem

$$\tilde{x} = \operatorname{argmin}_x \left\{ \frac{1}{2} \|x - \hat{x} - A^T y\|^2 : x \in \mathbb{R}^{n^2}_+ \right\}. \tag{3.1}$$

?unique_opt2?

It follows from the classical Rockafellar-Pshenichnyi optimality condition for (3.1), that $\tilde{x}$ is an optimal solution if, and only if, the gradient of the objective function at $\tilde{x}$ satisfies

$$\tilde{x} - \hat{x} - A^T y \in (\mathbb{R}^{n^2}_+ - \tilde{x})^+.$$

This yields the third item. The second item follows from the fact that a vector $v \in (\mathbb{R}^{n^2}_+ - \tilde{x})^+$ is equivalent to $v_i = 0$ for $\tilde{x}_i > 0$ and $v_i \geq 0$ for $\tilde{x}_i = 0$. □

We now introduce some notation in order to facilitate the discussions about the disconnected case. Let $y \in \mathbb{R}^{2n-1}$ and $X = \operatorname{Mat}\left(\hat{x} + A^T y\right)_+ \in \mathbb{R}^{n \times n}$. Suppose that X is disconnected with the following block diagonal structure:

$$X = \operatorname{Blkdiag}(X^1, \ldots, X^K), \tag{3.2}$$

?Xblk?

where $X^i \in \mathbb{R}^{m_i \times n_i}$ is connected for all $i = 1, \ldots, K$. We write $y = \begin{pmatrix} c \\ r \end{pmatrix} \in \mathbb{R}^{2n-1}$ correspondingly with the labels

$$c = \begin{pmatrix} c^1 \\ \vdots \\ c^K \end{pmatrix} \in \mathbb{R}^n, \qquad \text{with } c^i \in \mathbb{R}^{n_i}, \quad \text{for } i = 1, \ldots, K,$$
$$r = \begin{pmatrix} r^1 \\ \vdots \\ r^K \end{pmatrix} \in \mathbb{R}^{n-1}, \quad \text{with } r^i \in \mathbb{R}^{m_i}, \quad \text{for } i = 1, \ldots, K-1, \text{ and } r^K \in \mathbb{R}^{m_K - 1}.$$

The partition and its relation with c^i and r^i can be visualized as

$$X = \begin{matrix} \\ r^1 \\ \vdots \\ r^K \end{matrix} \begin{matrix} \begin{matrix} (c^1)^T & \cdots & (c^K)^T \end{matrix} \\ \begin{pmatrix} X^1 & \cdots & X^{1,K} = 0 \\ \vdots & \ddots & \vdots \\ X^{K,1} = 0 & \cdots & X^K \end{pmatrix} \end{matrix}, \tag{3.3}$$

?yrc?

where the off-diagonal blocks X^{ij} $(i \neq j)$ are zero due to the disconnectedness assumption. Each diagonal block X^i may be viewed as a smaller doubly stochastic matrix, if it is a square matrix. This motivates us to define the vectors by pairing c^i and r^i:

$$\begin{aligned} \mathscr{Y}^i &= \begin{pmatrix} c^i \\ r^i \end{pmatrix} \in \mathbb{R}^{m_i + n_i} \qquad \text{for } i = 1, \ldots, K-1, \\ \mathscr{Y}^K &= \begin{pmatrix} c^K \\ r^K \end{pmatrix} \in \mathbb{R}^{m_K + n_K - 1}. \end{aligned} \tag{3.4}$$

?ycr2?

We use calligraphic letter $\mathscr{Y}^i$ to distinguish it from the i-th iterate y^i in the Newton method (1.8) and Algorithm 3.1.

We note that each diagonal block X^k is completely determined by the vector $\mathscr{Y}^k$, i.e.,

$$X^k_{ij} = \left(\hat{X}^k_{ij} + c^k_j + r^k_i\right)_+ = \left(\hat{X}^k_{ij} + \mathscr{Y}^k_j + \mathscr{Y}^k_{n_k+i}\right)_+ .$$

We also note that if two vectors y and $\tilde{y}$ are equivalent, then the corresponding matrices $X = \mathrm{Mat}\left(\hat{x} + A^T y\right)_+$ and $\tilde{X} = \mathrm{Mat}\left(\hat{x} + A^T \tilde{y}\right)_+$ admit the same partition (3.3). Therefore, using the equivalence relation defined in Definition 3.1, it is unambiguous to speak of the (i,j)-th off-diagonal block X^{ij} or (i,i)-th diagonal block X^i when it comes to the same equivalence class.

Given $y \in \mathbb{R}^{2n-1}$, we list the notations to remind readers;

$$\text{Each } y \text{ gives rise to } \begin{cases} Y = Y_y = \mathrm{Mat}\left(\hat{x} + A^T y\right), \\ X = X_y = \mathrm{Mat}\left(\hat{x} + A^T y\right)_+, \\ M \in \mathcal{M}(y), \\ \mathcal{V}(M) = A\,\mathrm{Diag}(\mathrm{vec}(M))A^T \in \partial F(y), \end{cases} \tag{3.5}$$?notations?

where we ignore the subscripts when the meaning is clear. We partition the matrices Y and M in the same way as X in (3.3), respectively. Denote by Y^{ij} and M^{ij} the (i,j)-th block of Y and M, respectively. It is worthwhile to note that the off-diagonal blocks Y^{ij} $(i \neq j)$ are always *non-positive* due to the block-diagonal structure of X.

This notation is extended verbatim to any other vectors in $\mathbb{R}^{2n-1}$. For example, if $\tilde{y} \in \mathbb{R}^{2n-1}$, then the symbols $\tilde{Y}$ and $\tilde{Y}^{ij}$ are unambiguously defined just as for y above. In what follows, we will use these notations directly without defining them again.

3.1.2 Uniqueness

In this section we present sufficient conditions for the equivalence class to be a singleton. We first note that uniqueness of the optimum X^* means that the solution set of the system (2.5) is an equivalence class.

Lemma 3.3. *The solution set $\{y \mid F(y) = 0\}$ of the system (2.5) is an equivalence class.*

Proof. The proof follows by definition, from the fact that the optimum X^* is unique. □

Although the optimal solution of the primal problem (1.3) is unique, the solution of the optimality conditions in (1.5) for the dual variable y is a compact, convex, nonempty set, but is *not* necessarily a singleton set in general. The next result implies that we obtain a unique solution to (1.5) when the unique primal optimal solution to (1.1) is *connected*.

?⟨unique_alg⟩? **Theorem 3.4.** *Let $\tilde{y} \in \mathbb{R}^{2n-1}$ be given. If $\mathrm{Mat}\left(\hat{x} + A^T \tilde{y}\right)_+$ is connected, then the equivalence class $[\tilde{y}]$ is a singleton.*

Proof. Recall that the equivalence class can be defined by the linear equations and inequalities in Lemma 3.2. Applying the first proof in Lemma 2.3, if X is connected, then the columns of A associated with $x_i > 0$ form a basis of $\mathbb{R}^{2n-1}$. Therefore, the equations in Lemma 3.2 determine a unique solution. This implies that $[\tilde{y}]$ is a singleton.

We also provide an elementary proof below for the sake of self-containment. Assume X is connected, and let $y = \begin{pmatrix} c \\ r \end{pmatrix} \in \mathbb{R}^{2n-1}$ be an element in $[\tilde{y}]$. The entry y_i is said to be unique, if $\{y_i \mid y \in [\tilde{y}]\}$ has exactly one element. The subsets $R, C \subseteq \{1, \ldots, n\}$ are called unique, if the entries r_i for $i \in R \backslash \{n\}$ and the entries c_j for $j \in C$ are unique. We show that there exist unique subsets $R, C \subset \{1, \ldots, n\}$ and they can be extended so that $R = C = \{1, \ldots, n\}$.

1. The existence: Let $R = \{n\}$ and $C \subseteq \{1, \ldots, n\}$ be such that $j \in C$ if, and only if, $X_{n,j} > 0$. Since X is connected, C cannot be an empty set. As $X_{n,j} > 0$ for every $j \in C$, we have $X_{n,j} = (\hat{X}_{n,j} + c_j)_+ = \hat{X}_{n,j} + c_j$, see (2.4). Thus the entries c_j for $j \in C$ are uniquely determined. This shows that the subsets R and C are unique.

2. The extension: Let the subsets $R, C \subseteq \{1, \ldots, n\}$ be unique. Since X is connected, there exists at least one non-zero entry in $X_{\bar{R},C}$ or $X_{R,\bar{C}}$, see the paragraph after (2.2). Assume that $X_{\bar{R},C}$ contains a non-zero entry. Let $i \in \bar{R}$ be the row index associated with this non-zero entry. Then $X_{i,j} > 0$ for some $j \in C$, and this yields

$$X_{ij} = (\hat{X}_{ij} + r_i + c_j)_+ = \hat{X}_{ij} + r_i + c_j.$$

As C is unique, c_j is unique as $j \in C$, and thus r_i is also unique. It follows that the subsets $R_+ = R \cup \{i\}$ and $C_+ = C$ are unique. The case when $X_{R,\bar{C}}$ contains some non-zero entries is similar.

Therefore, $R = C = \{1, \ldots, n\}$ are unique, and this shows that $[\tilde{y}]$ has a unique solution. □

Note that Theorem 3.4 does not assume that X is a doubly stochastic matrix. The uniqueness of the solution to the system (1.5) follows directly from Theorem 3.4 as a special case when X is a doubly stochastic matrix.

?⟨unique⟩? **Corollary 3.5.** *If the optimal solution X^* of (1.3) is connected, then the solution y^* to the system (1.5) is unique.* □

Remark 3.6. *The converse direction in Theorem 3.4 doesn't hold. Assume that X and $\hat{X}$ are both 2 by 2 identity matrices. Then $[y]$ contains vectors satisfying the system*

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} y \right)_+ \quad \text{with variable } y \in \mathbb{R}^3.$$

This system can be written equivalently as

$$\begin{aligned} y_1 + y_3 &= 0, \\ y_1 &\leq 0, \\ y_2 + y_3 &\leq 0, \\ y_2 &= 0. \end{aligned}$$

We can easily derive that $y_1 = y_2 = y_3 = 0$. Thus, there is a unique solution y to the system even X is disconnected.

3.1.3 Polyhedron Description

The polyhedron characterization in Lemma 3.2 does not exploit the structures in A. In this section, we provide a different characterization using the blocks in a disconnected X. This alternative characterization enables us to find a vertex of $[y]$ efficiently and prove the convergence of our algorithm.

Consider the vectors c^k and r^k associated with the k-th diagonal block $X^k, k = 1, \ldots, K$. If we add a constant to c^k and subtract the same constant from r^k, then the diagonal block X^k remains the same. We define a matrix U associated with this operation as follows. Let R_k and C_k be the row and column indices corresponding to the k-th diagonal block of X, respectively. Define the matrix

$$U = \begin{bmatrix} u^1 & \cdots & u^K \end{bmatrix} \in \mathbb{R}^{2n-1 \times K}, \tag{3.6}$$?shiftV?

where the non-zero elements in each column $u^k \in \mathbb{R}^{2n-1}$ is given by

$$\begin{aligned} u_i^k &= -1 \quad \text{for } i \in C_k, \\ u_{n+i}^k &= 1 \quad \text{for } i \in R_k. \end{aligned} \tag{3.7}$$?shiftv?

It is clear that the aforementioned operation can be described by $y + \lambda_k u^k$, for some $\lambda_k \in \mathbb{R}$.

In Lemma 3.7 below, we show that equivalent vectors in $[y]$ are in the span of the first $K-1$ columns of the matrix U.

?⟨yV⟩? **Lemma 3.7.** *If $y \sim \tilde{y}$, then*

$$\tilde{y} - y \in \operatorname{range}(\begin{bmatrix} u^1 & \cdots & u^{K-1} \end{bmatrix}),$$

the span of the first $K-1$ columns of U.

Proof. The statement in Theorem 3.4 can be extended trivially to non-square matrices. Note that each diagonal block X^k is connected. If we fix an element in $r^k \in \mathbb{R}^{m_k}$, say the last entry $r^k_{m_k}$, then we can use the same argument in Theorem 3.4 to see that all other entries in $\mathscr{Y}^k = \begin{pmatrix} c^k \\ r^k \end{pmatrix}$ are uniquely determined. From this, we can deduce that $c^k = \tilde{c}^k + \lambda_k e$ and $r^k = \tilde{r}^k - \lambda_k e$ for some constant λ_k for every $k = 1, \ldots, K-1$.

Since the last block X^K is connected, we can apply Theorem 3.4 to X^K. This shows that $\tilde{\mathscr{Y}}^K = \mathscr{Y}^K$ and thus $\lambda_K = 0$. Putting together, this implies that $\tilde{y} = y + U\lambda$ for some $\lambda \in \mathbb{R}^K$ with $\lambda_K = 0$ using the definition of U. □

The following result is a direct consequence of Lemma 3.7. It states that the associated diagonal blocks of Y and $\tilde{Y}$ remain the same for the equivalent vectors $y, \tilde{y}$.

?⟨Yblk⟩? **Corollary 3.8.** *If $y \sim \tilde{y}$, then $\mathscr{Y}^K = \tilde{\mathscr{Y}}^K$ and $Y^k = \tilde{Y}^k$ for $k = 1, \ldots, K$.*

We now show that every equivalence class has a polyhedral representation via U.

?⟨equi⟩? **Theorem 3.9.** *Let $\tilde{y} \in \mathbb{R}^{2n-1}$. The equivalence class $[\tilde{y}]$ is a polyhedron given by*

$$[\tilde{y}] = \{y \in \mathbb{R}^{2n-1} \mid y = \tilde{y} + U\lambda \text{ for some } \lambda \in \mathbb{R}^K, \lambda_K = 0 \text{ and } Y^{ij} \le 0 \text{ for } i \neq j\},[4] \tag{3.8}$$ eq:equi_poly

where Y^{ij} is the (i,j)-the block of $Y = \operatorname{Mat}(\hat{x} + A^T y)$ with respect to the partition associated with $\tilde{y}$ as defined in (3.2).

[4] The redundant variable λ_K in (3.8) is included to simplify the proof in Lemma 3.16.

Proof. Let y be a vector on the right hand side set from (3.8). By the definition of U, the matrices X and $\tilde{X}$ have the same diagonal blocks. Since $X = (Y)_+$ and the off-diagonal blocks $Y^{ij} \leq 0$ are non-positive, the off-diagonal blocks $X^{ij} = (Y^{ij})_+ = 0$. This shows that $X = \tilde{X}$ and thus $y \in [\tilde{y}]$.

Conversely, for any vector $y \in [\tilde{y}]$, we have $y = \tilde{y} + U\lambda$ and $\lambda_K = 0$ by Lemma 3.7. Since $y \in [\tilde{y}]$, we must have $Y^{ij} \leq 0$ for $i \neq j$. Therefore, y is contained in the set on the right-hand-side. □

3.1.4 Vertices

For any equivalence class $[\tilde{y}]$, we aim to find a vector $y \in [\tilde{y}]$ so that $\partial F(y)$ contains at least one nonsingular matrix. For $y \in \mathbb{R}^{2n-1}$, the matrix $M \in \mathcal{M}(y)$ is said to be *maximal*, if

$$(\hat{x} + A^T y)_{ij} \geq 0 \Longrightarrow M_{ij} = 1. \tag{3.9}$$?maxs?

For a maximal M', it is easy to see that $M' \geq M$ and thus $\mathcal{V}(M') \succeq \mathcal{V}(M)$ for every $M \in \mathcal{M}(y)$. Therefore, if $\partial F(y)$ contains a nonsingular matrix, then $\mathcal{V}(M')$ must be nonsingular. In this case, we also call the matrices $\mathcal{V}(M') \in \partial F(y)$ *maximal*.

It turns out that the generalized Jacobian at any vertex of the polyhedron $[\tilde{y}]$ contains at least one nonsingular matrix.

?⟨vertices⟩? **Theorem 3.10.** *Let $\tilde{y} \in \mathbb{R}^{2n-1}$ be given, $y \in [\tilde{y}]$, and let M be maximal for y. The vector y is a vertex of the polyhedron $[\tilde{y}]$ if, and only if, M is connected.*

Proof. Assume M is disconnected. Without loss of generality, we can write

$$M = \begin{bmatrix} M^1 & 0 \\ 0 & M^2 \end{bmatrix} \text{ and } Y = \begin{bmatrix} Y^1 & Y^{12} \\ Y^{21} & Y^2 \end{bmatrix}.$$

It holds that $Y^{12} < 0$ and $Y^{21} < 0$ by the maximality of M. Thus, there exists an $\epsilon > 0$ such that the vectors $y' = y + \epsilon u^1$ and $y'' = y - \epsilon u^1$ are in $[\tilde{y}]$, where the vector u^1 is defined as in (3.7). But then $y = \frac{1}{2}y' + \frac{1}{2}y''$ and thus y is not a vertex.

Conversely, assume that M is connected. If X is connected, then Corollary 3.5 implies that the polyhedron $[\tilde{y}] = y$ and thus y is an extreme point. Therefore, we assume that X is disconnected and consider its partition as given in (3.3). Suppose for the sake of contradiction that $y \in [\tilde{y}]$ is not an extreme point. Then there exist a scalar $\alpha \in (0, 1)$ and vectors $y', y'' \in [\tilde{y}]$ both different from y such that $y = \alpha y' + (1 - \alpha)y''$. Then, $Y = \alpha Y' + (1 - \alpha)Y''$. Note that the diagonal blocks of Y, Y' and Y'' are the same, see Corollary 3.8.

Since M is connected, one of the off-diagonal blocks $M^{i,K}, M^{K,i}$ for $i = 1, \ldots, K-1$ must contain a positive entry. By the maximality of M, we have that $M^{ij} > 0$ if, and only if, $Y^{ij} \geq 0$. This implies that one of the off-diagonal blocks $Y^{i,K}, Y^{K,i}$ for $i = 1, \ldots, K-1$ must contain a nonnegative entry, say the (i,j)-th entry $Y^{K-1,K}_{ij}$ of the $(K-1, K)$-th block. In addition, as X is disconnected and $X = (Y)_+$, the off-diagonal blocks $Y^{i,K}, Y^{K,i}$ for $i = 1, \ldots, K-1$ must be non-positive. Putting together, the entry $Y^{K-1,K}_{ij}$ must be zero.

Similarly, we have that $(Y')^{K-1,K} \leq 0$ and $(Y'')^{K-1,K} \leq 0$, and therefore, the equation

$$0 = Y^{K-1,K}_{i,j} = \alpha (Y')^{K-1,K}_{i,j} + (1-\alpha)(Y'')^{K-1,K}_{i,j}$$

implies that $(Y')^{K-1,K}_{ij} = (Y'')^{K-1,K}_{ij} = 0$. Therefore, it holds that (see (2.4) and (3.5))

$$\begin{aligned} 0 &= (Y')^{K-1,K}_{ij} = \hat{X}^{K-1,K}_{ij} + (r')^{K-1}_i + (c')^K_j, \\ 0 &= (Y'')^{K-1,K}_{ij} = \hat{X}^{K-1,K}_{ij} + (r'')^{K-1}_i + (c'')^K_j. \end{aligned}$$

From Corollary 3.8, we know that $(\mathscr{Y}')^K = (\mathscr{Y}'')^K$ and thus $(c')_j^K = (c'')_j^K$. This implies that $(r')_i^{K-1} = (r'')_i^{K-1}$. It then follows from Theorem 3.9 that $(\mathscr{Y}')^{K-1} = (\mathscr{Y}'')^{K-1}$. This argument can be repeated for all the remaining diagonal blocks until we get $y = y' = y''$. This yields contradiction, and thus y is an extreme point. ☐

It follows from Lemma 2.3 and Theorem 3.10 that $\partial F(y)$ contains a nonsingular matrix whenever y is a vertex of the polyhedron $[\tilde{y}]$. More precisely, the maximal $\mathcal{V}(M)$ is nonsingular when y is a vertex. This result is stated in the next corollary.

?⟨nonsingular⟩? **Corollary 3.11.** *If y is a vertex of the polyhedron $[\tilde{y}]$, then $\partial F(y)$ contains at least one nonsingular matrix. In particular, the maximal matrix $V \in \partial F(y)$ is nonsingular.*

The rest of this section provides a method for finding a vertex efficiently. We start with the existence of a vertex for any polyhedron $[y]$.

?⟨existy⟩? **Lemma 3.12.** *Let $y \in \mathbb{R}^{2n-1}$. The polyhedron $[y] \subset \mathbb{R}^{2n-1}$ contains at least one vertex.*

Proof. A polyhedron contains a line if there exists a vector $y \in \mathbb{R}^{2n-1}$ and a non-zero direction $d \in \mathbb{R}^{2n-1}$ such that $y + \alpha d$ is contained in the polyhedron for all scalars α. It is well known that a polyhedron has at least one vertex if, and only if, it does not contain a line.

If X is connected, then the polyhedron $[y]$ contains only one vector y which is a vertex by Corollary 3.5. Suppose that X is disconnected. Assume, without loss of generality, that we can write X and Y as

$$X = \begin{bmatrix} X^1 & 0 \\ 0 & X^2 \end{bmatrix} \text{ and } Y = \begin{bmatrix} Y^1 & Y^{12} \\ Y^{21} & Y^2 \end{bmatrix},$$

where the diagonal block X^2 are connected. Here, the block X^1 does not have to be connected.

Let $\tilde{y} := y + \alpha d$ for some d and define

$$D := \operatorname{Mat}\left(A^T d\right) = \begin{bmatrix} D^1 & D^{12} \\ D^{21} & D^2 \end{bmatrix}.$$

It follows from Corollary 3.8 that the entries in y and $\tilde{y}$ associated with the last connected block are the same, i.e., $\mathscr{Y}^2 = \tilde{\mathscr{Y}}^2$. Thus, the entries in d associated with D^2 must be zero. From this, we can see that if the direction d is non-zero, then there exists at least one non-zero element in D^{12} or D^{21}.

If $y \sim \tilde{y}$, then applying Corollary 3.8 again yields $Y_2 = \tilde{Y}_2$ and this implies that

$$\tilde{Y} = \begin{bmatrix} \tilde{Y}^1 & \tilde{Y}^{12} \\ \tilde{Y}^{21} & \tilde{Y}^2 \end{bmatrix} = \begin{bmatrix} Y^1 + \alpha D^1 & Y^{12} + \alpha D^{12} \\ Y^{21} + \alpha D^{21} & Y^2 \end{bmatrix}.$$

In both cases, at least one of the entries in the off-diagonal blocks $\tilde{Y}^{12}$ or $\tilde{Y}^{21}$ becomes positive for sufficiently large or small α. This shows that $\tilde{y}$ is not equivalent to y for all α. Thus $[y]$ doesn't contain a line, and it has at least one vertex. ☐

?⟨polytope⟩? **Remark 3.13.** *In Lemma 3.12, if we assume additionally that $X = \operatorname{Mat}\left(\hat{x} + A^T y\right)_+$ does not contain any zero rows or columns, then $[y]$ is even bounded and thus a polytope. We prove this by contradiction. Assume that $[y]$ is not bounded. Then there exists a non-zero direction $d \in \mathbb{R}^{2n-1}$ such that $y + \alpha d \in [y]$ for all $\alpha \geq 0$. By Theorem 3.9, we have that $d = U\lambda$ for some non-zero*

$\lambda \in \mathbb{R}^K$ with $\lambda_K = 0$. Thus, $\tilde{y} := y + \alpha U\lambda \in [y]$ for all $\alpha \geq 0$. The (i,j)-th off-diagonal blocks of Y and $\tilde{Y}$ satisfy

$$\tilde{Y}^{i,j} = Y^{i,j} + (\lambda_i - \lambda_j)J,$$

where J is all-ones matrix of appropriate size. As $\lambda_K = 0$ and $\lambda \neq 0$, there exists an index $i \in \{1, \dots, n-1\}$ such that $\lambda_i - \lambda_K > 0$ or $\lambda_K - \lambda_i > 0$. This implies that the blocks $\tilde{Y}^{i,K}$ or $\tilde{Y}^{K,i}$ contain a positive entry for sufficiently large α. But then y and $\tilde{y}$ are equivalent. This is a contradiction. Therefore, $[y]$ is always bounded.

Finally, the problem (1.3) satisfies Mangasarian-Fromovitz constraint qualification and this implies the set of dual optimal solutions is bounded. This yields an alternative derivation that the optimal set $[y^*]$ is bounded.

We can find a vertex of the polyhedron $[\tilde{y}]$ as follows. In Theorem 3.9, $[\tilde{y}]$ is expressed as the projection of a higher dimensional polyhedron in variables $y \in \mathbb{R}^{2n-1}$ and $\lambda \in \mathbb{R}^{K-1}$. Through the Fourier–Motzkin elimination, we can describe the polyhedron $[\tilde{y}]$ solely using variables $y \in \mathbb{R}^{2n-1}$. Then a vertex of $[\tilde{y}]$ can be obtained via solving a particular linear program. This procedure, however, is very expensive. In what follows, we provide an efficient combinatorial method for finding a vertex of $[\tilde{y}]$.

?⟨thm:shift⟩? **Lemma 3.14.** *Let $y \in \mathbb{R}^{2n-1}$. Let X be disconnected and*

$$X = \begin{bmatrix} X^{R,C} & 0 \\ 0 & X^{\bar{R},\bar{C}} \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Y^{R,C} & Y^{R,\bar{C}} \\ Y^{\bar{R},C} & Y^{\bar{R},\bar{C}} \end{bmatrix},$$

for some subsets R, C as in (2.2). Let $\tilde{y} := y + tu^1$ for some $t \in \mathbb{R}$, where u^1 is defined as in (3.7) for the partition above. Then

$$y \sim \tilde{y} \quad \Longleftrightarrow \quad \max_{i,j} Y^{\bar{R},C}_{i,j} \leq t \leq -\max_{i,j} Y^{R,\bar{C}}_{i,j}. \tag{3.10}$$?thm:shifteq?

Proof. The scalar t is well-defined, as $Y^{\bar{R},C}$ and $Y^{R,\bar{C}}$ are non-positive. Then the matrix $\tilde{Y}$ can be written as

$$\tilde{Y} = \begin{bmatrix} Y^{R,C} & Y^{R,\bar{C}} + tJ \\ Y^{\bar{R},C} - tJ & Y^{\bar{R},\bar{C}} \end{bmatrix}, \tag{3.11}$$?Yform?

where J is the all-ones matrix of appropriate sizes. We see that $y \sim \tilde{y}$ if, and only if, $Y^{R,\bar{C}} + tJ \leq 0$ and $Y^{\bar{R},C} - tJ \leq 0$. The latter is equivalent to the inequalities in (3.10). □

For any vector y, we can find a vertex of $[y]$ efficiently.

?⟨ysing⟩? **Theorem 3.15.** *For any vector $y \in \mathbb{R}^{2n-1}$, there is a polynomial-time algorithm for finding a vertex of the polyhedron $[y]$.*

Proof. Let X, Y and the maximal matrix M defined as in (3.5) and (3.9) associated with y. Denote by $M^{i,j}$ the (i,j)-th block of M $(i, j = 1, \dots, K)$ corresponding to the partition of X in (3.2). If y is not a vertex, then M is disconnected. Thus, there exists a subset $\mathcal{B} \subseteq \{1, \dots, K\}$ such that $K \in \mathcal{B}$,

$$Y^{R,\bar{C}} < 0 \text{ and } Y^{\bar{R},C} < 0,$$

where R and C be the collection of row and column indices of Y associated with the blocks in $\mathcal{B}$. For example, if $\mathcal{B} = \{K\}$, then

$$Y^{R,\bar{C}} = \begin{bmatrix} Y^{1,K} \\ \vdots \\ Y^{K-1,K} \end{bmatrix} < 0 \text{ and } Y^{\bar{R},C} = \begin{bmatrix} Y^{K,1} & \cdots & Y^{K,K-1} \end{bmatrix} < 0.$$

This means we can find a constant $t \neq 0$ such that $\max Y^{\bar{R},C} \leq t \leq -\max Y^{R,\bar{C}}$ as in (3.10). Let $\tilde{y} = y + tw$, where $w \in \mathbb{R}^{2n-1}$ is defined as

$$\begin{aligned} w_i &= -1 \quad \text{for } i \in C, \\ w_{n+i} &= \;\;\, 1 \quad \text{for } i \in R. \end{aligned} \tag{3.12}$$

?shiftw?

By Lemma 3.14, we have $\tilde{y} \sim y$. Recall that $\tilde{Y}$ has the form (3.11). In particular, we distinguish the following two cases depending on t:

1. If we take $t = -\max Y^{R,\bar{C}} > 0$, then $Y^{R,\bar{C}} + tJ$ contains at least one zero entry.
2. Similarly, if $t = \max Y^{\bar{R},C} < 0$, then $Y^{\bar{R},C} - tJ$ contains at least one zero entry.

(3.13) ?rule?

Let $\tilde{M}$ be the maximal matrix defined similarly for $\tilde{y}$. In either case, the number of non-zero elements in $\tilde{M}$ is strictly less than these in M. As $y \sim \tilde{y}$, we can repeat this procedure until M is connected. □

3.2 The Algorithm and its Local Convergence

For any vector $y \in \mathbb{R}^{2n-1}$, we can find a vertex $\tilde{y}$ of the polyhedron $[y]$ using Theorem 3.15. It follows from Corollary 3.11 that the maximal matrix $\tilde{V} \in \partial F(\tilde{y})$ is nonsingular. Thus we can generate well-defined iterates when maximal $\tilde{V} \in \partial F(\tilde{y})$ is used at each iteration. We achieve this by developing a variant of the Semi-smooth Newton method.

Algorithm 3.1 A Modified Semi-smooth Newton Method

Require: y^0 initial point, *tol* tolerance
while $\|F(y^k)\| > tol$ **do**
  Find a vertex $\tilde{y}^k$ of $[y^k]$ using Theorem 3.15
  Compute the maximal $\tilde{V}_k \in \partial F(\tilde{y}^k)$
  Update $y^{k+1} = \tilde{y}^k - \tilde{V}_k^{-1} F(\tilde{y}^k)$
end while

?⟨mo⟩?

We now prove Q-quadratic local convergence of the modified Newton method. Recall that the distance between a vector $y \in \mathbb{R}^{2n-1}$ to a subset $S \subseteq \mathbb{R}^{2n-1}$ is

$$\operatorname{dist}(y, S) := \inf_{s \in S} \|y - s\|. \tag{3.14}$$

eq:ptsetdist

Similarly, we denote the nearest point distance between two subsets $S, T \subset \mathbb{R}^{2n-1}$ as

$$\operatorname{dist}(S, T) := \inf_{s \in S, t \in T} \|s - t\|. \tag{3.15}$$

eq:setsetdist

The main idea behind the proof is that if an equivalence class $[y]$ is sufficiently close to the optimal set $[y^*]$ in the sense of (3.15), then every element in $[y]$ is also close to $[y^*]$ in the sense of (3.14); and this further implies that each vertex in $[y]$ is also close to one of the vertices in $[y^*]$. For any polyhedron $[y]$, we denote by $\mathrm{ext}[y]$ the set of vertices of $[y]$.

?⟨thm_close0⟩? **Lemma 3.16.** *Suppose that $F(y^*) = 0$. Then there exist $\epsilon > 0$ and $\kappa > 0$ such that for any $y \in \mathbb{R}^{2n-1}$ with $\mathrm{dist}(y, [y^*]) < \epsilon$ we have:*

1. *$\mathrm{dist}(\tilde{y}, [y^*]) < \kappa \cdot \epsilon$ for every $\tilde{y} \in [y]$.*
2. *$\mathrm{dist}(\tilde{y}, \mathrm{ext}[y^*]) < \kappa \cdot \epsilon$ for every $\tilde{y} \in \mathrm{ext}[y]$.*

Proof. 1. Let $y \in \mathbb{R}^{2n-1}$. Without loss of generality, we assume that y^* satisfies $\|y - y^*\| = \mathrm{dist}(y, [y^*]) < \epsilon$. It follows from Lemma 4.1 that if $\epsilon > 0$ is sufficiently small, then

$$X^*_{ij} > 0 \quad \Longrightarrow \quad X_{ij} > 0. \tag{3.16}$$?XXstar?

Let X and X^* be partitioned as in (3.2),

$$\begin{array}{rcl} X & = & \mathrm{Blkdiag}\left(X^1, \ldots, X^K\right), \\ X^* & = & \mathrm{Blkdiag}\left((X^*)^1, \ldots, (X^*)^{K^*}\right), \end{array} \tag{3.17}$$?eq_par?

where K and K^* are the number of blocks in X and X^*, respectively. It follows from (3.16) that $K \le K^*$, and moreover, we can view each block $(X^*)^{i,j}$ as a unique sub-block of $X^{k,l}$ for some $k, l = 1, \ldots, K$. As an example, assume we have the following partition for X and X^* into $K = 2$ and $K^* = 3$ blocks, respectively,

$$X = \left[\begin{array}{ccc|cc} X_{1,1} & X_{1,2} & X_{1,3} & 0 & 0 \\ X_{2,1} & X_{2,2} & X_{2,3} & 0 & 0 \\ X_{3,1} & X_{3,2} & X_{3,3} & 0 & 0 \\ \hline 0 & 0 & 0 & X_{4,4} & X_{4,5} \\ 0 & 0 & 0 & X_{5,4} & X_{5,5} \end{array}\right], X^* = \left[\begin{array}{cc|c|cc} X^*_{1,1} & X^*_{1,2} & 0 & 0 & 0 \\ X^*_{2,1} & X^*_{2,2} & 0 & 0 & 0 \\ \hline 0 & 0 & X^*_{3,3} & 0 & 0 \\ \hline 0 & 0 & 0 & X^*_{4,4} & X^*_{4,5} \\ 0 & 0 & 0 & X^*_{5,4} & X^*_{5,5} \end{array}\right].$$

Then the top-left block $(X^*)^1 \in \mathbb{R}^{2\times 2}$ and the mid block $(X^*)^2 \in \mathbb{R}^1$ of X^* on the right hand side are both sub-blocks of $X^1 \in \mathbb{R}^{3\times 3}$ of X on the left hand side.

For convenience, we define a zero-one matrix $P \in \{0,1\}^{K^*\times K}$ such that $P_{ij} = 1$ if, and only if, $(X^*)^i$ is a sub-block of X^j. For instance, the matrix P in the previous example is given by

$$P = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 2}.$$

Denote by $p_i \in \mathbb{R}^K$ the i-th row of P. Note that $p_i = p_j$ if, and only if, the diagonal blocks $(X^*)^i$ and $(X^*)^j$ are the sub-blocks of the same diagonal block in X. We also define the matrices U and U^* as in (3.6) for X and X^*, respectively.

From Theorem 3.9, we know that if $\tilde{y} \in [y]$, then $\tilde{y} = y + U\tilde{\lambda}$ for some $\tilde{\lambda} \in \mathbb{R}^K$ with $\tilde{\lambda}_K = 0$. Define $y' := y^* + U\tilde{\lambda}$. By construction, the distance between $\tilde{y}$ and y' is small as

$$\|\tilde{y} - y'\| = \|y + U\tilde{\lambda} - y^* - U\tilde{\lambda}\| = \|y - y^*\| < \epsilon. \tag{3.18}$$?eq_close1?

We will show that $\operatorname{dist}(y', [y^*])$ is also sufficiently small. The key idea is that y' only slightly violates the set of constraints defining the polyhedron $[y^*]$ in Theorem 3.9. From this, we can establish an upper bound for $\operatorname{dist}(y', [y^*])$ using Hoffman's error bound [29]. Together with (3.18), the first inequality in the statement follows immediately.

Let $\lambda^* = P\tilde{\lambda} \in \mathbb{R}^{K^*}$. Since the last diagonal block $(X^*)^{K^*}$ of X^* must be a sub-block of X^K and $\tilde{\lambda}_K = 0$, we have the equality

$$\lambda^*_{K^*} = 0. \tag{3.19}$$

?eq_close2?

One can easily verify that $U = U^*P$ and thus $U\tilde{\lambda} = U^*\lambda^*$. This means we can write

$$y' := y^* + U\tilde{\lambda} = y^* + U^*\lambda^*. \tag{3.20}$$

?eq_close3?

From Theorem 3.9, the equivalence class $[y^*]$ can be defined as

$$[y^*] = \left\{ y^* + U^*\lambda \in \mathbb{R}^{2n-1} \mid \begin{array}{c} \lambda \in \mathbb{R}^{K^*}, \lambda_{K^*} = 0 \\ (Y^*)^{ij} \leq 0 \text{ for } i \neq j \end{array} \right\}, \tag{3.21}$$

?ystar2?

where $(Y^*)^{ij}$ is the (i,j)-th block of $Y^* = \operatorname{Mat}\left(\hat{x} + A^T y^*\right)$ with respect to the partition of X^* in (3.17). From (3.19) and (3.20), it is clear that y' satisfies the equality constraints in (3.21).

As to the inequality constraints, we partition Y' in the same way as X^*. Each block $(Y')^{ij}$ can also be viewed as a sub-block of $X^{k,l}$ for some $k, l = 1, \ldots, K$. We distinguish the following two cases based on the off-diagonal blocks in Y'.

(a) If $(Y')^{ij}$ $(i \neq j)$ is a sub-block of a diagonal block X^k, then the diagonal blocks $(Y')^i$ and $(Y')^j$ are also sub-blocks of X^k. As $\lambda^* = P\tilde{\lambda}$, this means $\lambda^*_i = \lambda^*_j = \tilde{\lambda}_k$. Since $y' = y^* + U\tilde{\lambda} = U^*\lambda^*$, we obtain that $(Y')^{ij} = (Y^*)^{ij} \leq 0$.

(b) If $(Y')^{ij}$ is a sub-block of an off-diagonal block $Y^{k,l}$, then the constraint $(Y')^{ij} \leq 0$ may not be satisfied. As $\|\tilde{y} - y'\| < \epsilon$ from (3.18), it holds that

$$\|\tilde{Y} - Y'\| = \|\hat{x} + A^T\tilde{y} - \hat{x} - A^T y'\| = \|A^T(\tilde{y} - y')\| < \|A\|\epsilon.$$

In addition, $\tilde{y} \in [y]$ implies that $\tilde{Y}^{k,l} \leq 0$, see Theorem 3.9. This means the largest nonnegative entry in $(Y')^{i,j}$ is at most $\|A\|\epsilon$.

This shows that y' violates the constraints in $[y^*]$ only up to a scalar multiplication of ϵ. The Hoffman's error bound implies that $\operatorname{dist}(y', [y^*]) < c \cdot \epsilon$ for some universal constant c which depends only on the matrix A and the polyhedron $[y^*]$.

We can establish the first inequality now. Let $v \in [y^*]$ such that $\|y' - v\| = \operatorname{dist}(y', [y^*]) < \epsilon$. For any $\tilde{y} \in [y]$, we have that

$$\begin{array}{rcl} \operatorname{dist}(\tilde{y}, [y^*]) & \leq & \|\tilde{y} - v\| \\ & \leq & \|\tilde{y} - y'\| + \|v - y'\| \\ & = & \|\tilde{y} - y'\| + \operatorname{dist}(y', [y^*]) \\ & < & c_1\epsilon, \end{array}$$

where $c_1 = c + 1$.

2. For any $\tilde{y} \in \operatorname{ext}[y]$, we know from the first part that $\operatorname{dist}(\tilde{y}, [y^*]) < c_1 \cdot \epsilon$ for some universal constant c_1. Without loss of generality, we assume that y^* satisfies $\|\tilde{y} - y^*\| = \operatorname{dist}(\tilde{y}, [y^*])$. If $y^* \in \operatorname{ext}[y^*]$, then $\operatorname{dist}(\tilde{y}, \operatorname{ext}[y^*]) < c_1 \cdot \epsilon$. Thus, we assume that $y^* \notin \operatorname{ext}[y^*]$.

   We transform y^* into a vertex $\tilde{y}^* \in \operatorname{ext}[y^*]$ using the procedure in the proof of Theorem 3.15. The obtained vertex $\tilde{y}^*$ depends on the choice in (3.13). This yields a sequence λ_i^* such that $\tilde{y}^* = y^* + \sum_{i=1}^m \lambda_i^* w_i$, where w_i is defined as in (3.13) and m is the number of iterations. Note that $m \leq K^* - 1$. In what follows, we show that it is possible to pick a sufficiently small λ_i^* at each iteration.

   In the first iteration, we identify subsets R and C such that $(Y^*)^{\bar{R},C} < 0$ and $(Y^*)^{R,\bar{C}} < 0$. Then we choose either $\lambda_1^* = \max(Y^*)^{\bar{R},C}$ or $\lambda_1^* = -\max(Y^*)^{R,\bar{C}}$. As $\|\tilde{y} - y^*\| = \operatorname{dist}(\tilde{y}, [y^*]) < c_1 \cdot \epsilon$, it holds that

$$\|\tilde{Y} - Y^*\| = \|A^T(\tilde{y} - y^*)\| < c_1\|A\|\epsilon = \epsilon_1, \tag{3.22}$$?eq_close4?

   where we set $\epsilon_1 := c_1\|A\|\epsilon$. Therefore, we obtain that

$$\|\tilde{Y}^{\bar{R},C} - (Y^*)^{\bar{R},C}\| < \epsilon_1 \text{ and } \|\tilde{Y}^{R,\bar{C}} - (Y^*)^{R,\bar{C}}\| < \epsilon_1. \tag{3.23}$$?eq_close5?

   Since $\tilde{y} \in \operatorname{ext}[y]$ is a vertex, the associated maximal matrix $\tilde{M}$ is connected by Theorem 3.10, see the definition of maximality in (3.9). This implies that

$$\max \tilde{Y}^{\bar{R},C} \geq 0 \text{ or } \max \tilde{Y}^{R,\bar{C}} \geq 0, \tag{3.24}$$?eq_close6?

   as otherwise $\tilde{M}$ is disconnected. Using (3.23) and (3.24), we conclude that

$$\max(Y^*)^{\bar{R},C} > -\epsilon_1 \text{ or } \max(Y^*)^{R,\bar{C}} > -\epsilon_1.$$

   We choose λ_1^* as follows.

   (a) If $\max(Y^*)^{\bar{R},C} > -\epsilon_1$, then $\lambda_1^* = \max(Y^*)^{\bar{R},C}$.

   (b) If $\max(Y^*)^{R,\bar{C}} > -\epsilon_1$, then $\lambda_1^* = -\max(Y^*)^{R,\bar{C}}$.

   As $\lambda_1^* < 0$, we have that $|\lambda_1^*| < \epsilon_1$ in both cases. In the second iteration, we apply the same procedure to $y^* + \lambda_1^* w_1$. The same argument above can be used, except that ϵ_1 is replaced by $2\epsilon_1$, to show that that $|\lambda_2^*| < 2\epsilon_1$. Proceeding in this way, we conclude that $|\lambda_k^*| < 2^k\epsilon$ for $k = 1, \ldots, m$.

   These upper bounds for $\lambda_1^*, \ldots, \lambda_m^*$ imply that

$$\begin{aligned} \|\tilde{y} - \tilde{y}^*\| &= \|(\tilde{y} - y^*) - \textstyle\sum_{k=1}^m \lambda_k^* w^k\| \\ &\leq \|\tilde{y} - y^*\| + \|\textstyle\sum_{k=1}^m \lambda_k^* w^k\| \\ &\leq c_1 \cdot \epsilon + \textstyle\sum_{k=1}^m |\lambda_k^*| \cdot \|w^k\| \\ &\leq c_1 \cdot \epsilon + (\max_k \|w^k\|)(\textstyle\sum_{k=1}^m 2^k)\epsilon_1 \\ &< c_2 \cdot \epsilon, \end{aligned}$$

   for some constant c_2 depending on m. As $\tilde{y}^* \in \operatorname{ext}[y^*]$, we have that $\operatorname{dist}(\tilde{y}, \operatorname{ext}[y^*]) \leq \|\tilde{y} - \tilde{y}^*\| < c_2 \cdot \epsilon$.

Finally, we take $\kappa = \max\{c_1, c_2\}$ and this finishes the proof. □

We provide the convergence of the modified Newton method.

?⟨thmmain⟩? **Theorem 3.17.** *Let the current iterate y^k be sufficiently close to the (compact, convex) solution set $[y^*]$. Then the modified Newton method converges, and at a Q-quadratic rate, to a point in $[y^*]$.*

Proof. For any $y \in \mathbb{R}^{2n-1}$, if $M \in \mathcal{M}(y)$ is maximal in (2.6), then M is an n by n zero-one matrix, see also (3.9). This means that there are at most 2^{n^2} different maximal matrix in $\mathcal{M}(y)$. If $\bar{\mathcal{M}}$ is the collection of different maximal matrices in $\partial F(y)$, i.e.,

$$\bar{\mathcal{M}} := \{M \mid M \in \mathcal{M}(y) \text{ is maximal}\},$$

then $|\mathcal{V}|$ is finite. Therefore, there exists a constant β such that $\|V^{-1}\| \leq \beta$ for every $V \in \bar{\mathcal{M}}$.

Let K^* be the number of blocks in the unique optimal solution X^*. Let $0 < \eta < \min\{1, \frac{1}{\beta\kappa^2}\}$ be a fixed constant, where κ is the constant in Lemma 3.16. Since F is semismooth at any optimal solution y^*, there exists $\epsilon > 0$ such that

$$\|F(y^*) - F(y) - V\,(y^* - y)\,\| \leq \eta\|y^* - y\|^2, \quad \forall y \in B(y^*, \epsilon) \text{ and } V \in \partial F(y), \tag{3.25}$$?local?

where $B(y^*, \epsilon)$ is the ϵ ball around y^*. Recall that the number of vertices of any polytope is finite, and $\bar{\mathcal{M}}$ is a finite set. Thus, for any fixed η, we can assume that the above inequality (3.25) holds for every vertex $\tilde{y}^*$ of $[y^*]$.

Let $\tilde{y}^k$ be any vertex of $[y^k]$ obtained from Theorem 3.15 in the algorithm. Let $\tilde{y}^* \in \text{ext}[y^*]$ be such that $\|\tilde{y}^k - \tilde{y}^*\| = \text{dist}(\tilde{y}^k, \text{ext}[y^*])$. It holds that

$$\begin{array}{rcl} \text{dist}(y^{k+1}, [y^*]) & \leq & \|y^{k+1} - \tilde{y}^*\| \\ & = & \|\tilde{y}^k - \tilde{y}^* - \tilde{V}_k^{-1}F(\tilde{y}^k)\| \\ & = & \|\tilde{V}_k^{-1}(F(\tilde{y}^*) - F(\tilde{y}^k) - \tilde{V}_k(\tilde{y}^* - \tilde{y}^k))\| \\ & \leq & \|\tilde{V}_k^{-1}\| \cdot \|F(\tilde{y}^*) - F(\tilde{y}^k) - \tilde{V}_k(\tilde{y}^* - \tilde{y}^k)\| \\ & \leq & \beta\eta\|\tilde{y}^k - \tilde{y}^*\|^2, \end{array}$$

where the last inequality follows from (3.25). If $\text{dist}(y^k, [y^*]) = \epsilon > 0$ is sufficiently small, then Lemma 3.16 shows that that

$$\|\tilde{y}^k - \tilde{y}^*\|^2 = \text{dist}(\tilde{y}^k, \text{ext}[y^*])^2 < \kappa^2 \cdot \epsilon^2 = \kappa^2 \cdot \text{dist}(y^k, [y^*])^2.$$

Thus, this yields

$$\begin{array}{rcl} \text{dist}(y^{k+1}, [y^*]) & \leq & \beta\eta\|\tilde{y}^k - \tilde{y}^*\|^2 \\ & \leq & \beta\eta\kappa^2\,\text{dist}(y^k, [y^*])^2 \\ & < & \text{dist}(y^k, [y^*])^2. \end{array}$$

This shows that $\text{dist}(y^{k+1}, [y^*]) < \text{dist}(y^k, [y^*])^2$, and thus the modified Newton method converges quadratically to the optimal set $[y^*]$. □

We observe that the performance of Algorithm 3.1 depends on the number of blocks K^* in the optimal solution X^*. In Lemma 3.16, the constant κ depends on K^*. If K^* is large, then the condition for the quadratic convergence in Theorem 3.17 is stricter. This suggests that an instance can be more difficult to solve if the optimal solution X^* contains many blocks. Our numerical experiment verifies this observation, see Figure 1.

Finally we discuss about an undesirable phenomenon called *cycling*. If $y^k = y^{k'}$ for some $k < k'$, then we say the algorithm is cycling. Thus, the algorithm may loop indefinitely. Fortunately, if y^k is sufficiently close to $[y^*]$ as required in Theorem 3.17, then cycling cannot happen as

$\operatorname{dist}([y^{k+1}],[y^*]) < \operatorname{dist}([y^k],[y^*])^2$. In the general case, we can avoid cycling empirically by taking a random choice in the step (3.13) in Theorem 3.15. This generates a random vertex each time. With this simple trick, we never end up in a cycle in our numerical experiments. Therefore we focus on the case when cycling does not occur. (It is worth mentioning that this cycling is similar to the simplex method cycling for degenerate problems, i.e., when the simplex algorithm remains stuck at the same feasible vertex. However, unlike the simplex method, the total number of vertices in our problem is not finite.)

4 Refinement and the Local Error Bound Condition

In this section we show that we can split the problem into smaller problems when the iterate y in the (modified) Newton method is sufficiently close to the solution y^* of (1.3). Under the strict complementarity assumption, we can split the problem recursively until the assumption in Corollary 2.5 holds; we obtain the solutions for each subproblem by the semismooth Newton method (1.8).

Recall that if y^* is a solution to the system (2.5), then $x^* = (\hat{x} + A^T y^*)_+$ is an optimal solution to (1.3) and $z^* = (\hat{x} + A^T y^*)_-$ is an optimal dual variable for (1.3), see Theorem 1.1. We say that *strict complementarity* holds at (x^*, z^*), if $x^* + z^* > 0$.

?⟨pattern⟩? **Lemma 4.1.** *Suppose $F(y^*) = 0$. There exists an $\epsilon > 0$ such that for every y satisfying $||y - y^*|| < \epsilon$, it holds that*

$$x_i^* > 0 \Longrightarrow x_i > 0, \tag{4.1}$$

?pos1?

where $x = (\hat{x} + A^T y)_+$. Moreover, if (x^*, z^*) satisfies strict complementarity, then we can also take ϵ such that

$$x_i^* > 0 \Longleftarrow x_i > 0.$$

Proof. Let A_i denote the i-th column of A. If $x_i^* > 0$, then $x_i^* = \hat{x}_i + A_i^T y^* > 0$ and thus $x_i = (\hat{x}_i + A_i^T y)_+ = \hat{x}_i + A_i^T y > 0$ for small $\epsilon > 0$. Now suppose that the pair (x^*, z^*) satisfies strict complementarity. Assume to the contrary that $x_i^* = 0$. Then we have $z_i^* > 0$, i.e., $\hat{x}_i + A_i^T y^* < 0$, and thus $\hat{x}_i + A_i^T y < 0$ for sufficiently small ϵ. It follows that $x_i = (\hat{x}_i + A_i^T y)_+ = 0$. □

Suppose y is close to y^*. Lemma 4.1 suggests that the $X = \operatorname{Mat}(\hat{x} + A^T y)_+$ and the optimal solution $X^* = \operatorname{Mat}(\hat{x} + A^T y^*)_+$ share the same block-diagonal structure. As a heuristic, we can split the problem into smaller subproblems if the residual is sufficiently small. If strict complementarity holds, then the smaller subproblems will not be disconnected eventually and thus the semismooth Newton method (1.8) can be applied.

The local error bound condition is a sufficient condition for the convergence of Newton-type methods. It is a weaker requirement than the nonsingularity (i.e., connectedness) condition used in Section 2. In this section, we show that the system (2.5) for the nearest doubly stochastic matrix problem does not satisfy the local error bound condition.

localerrorbnd⟩? **Definition 4.2** (*local error bound*)**.** *Let $[y^*]$ be the solution set of (2.5) and let N be a neighbourhood such that $[y^*] \cap N \neq \emptyset$. If there exists a positive constant c such that*

$$c \cdot \operatorname{dist}(y, [y^*]) \leq ||F(y)||, \quad \forall y \in N, \tag{4.2}$$

?localerror?

then we say that F satisfies the local error bound condition *on N for the system (2.5).*

We show that the local error condition does not hold for (2.5), and this implies that $\partial F(y)$ is singular in general. Recall that strict complementarity holds for (2.5) if $x + z > 0$ for optimal primal and dual variables x and z.

?⟨leb⟩? **Theorem 4.3.** *Consider the system* (2.5)*. Assume that strict complementarity holds. Then* $F(y)$ *in* (2.5) *does not satisfy the local error bound condition.*

Proof. Let $y \in \mathbb{R}^{2n-1}$. Define the projection

$$P_C(y) := \operatorname{argmin}_{u \in C} \|u - y\|, \text{ where } C \text{ is a polyhedron.} \tag{4.3}$$?leb_P?

Let $y^* = P$ and $d = y - y^*$. Note that $\operatorname{dist}(y, [y^*]) = ||y - y^*|| = ||d||$.

Let $x = (\hat{x} + A^T y)_+$ and $x^* = (\hat{x} + A^T y^*)_+$. Define $s^* \in \{0,1\}^{n^2}$ such that $s_i^* = 1$ if, and only if, $x_i^* > 0$. Applying (4.1) in Lemma 4.1, we can assume $\|d\|$ is sufficiently small so that $x_i > 0$ if, and only if, $x_i^* > 0$. Therefore, it holds that

$$\begin{array}{rcl}(\hat{x} + A^T y)_+ &=& (\hat{x} + A^T y^* + A^T d)_+ \\ &=& x^* + \operatorname{Diag}(s^*) A^T d\end{array}$$

Thus we have

$$\begin{array}{rcl}||F(y)|| &=& ||A(\hat{x} + A^T y)_+ - b|| \\ &=& ||Ax^* + A\operatorname{Diag}(s) A^T d - b|| \\ &=& ||A\operatorname{Diag}(s) A^T d||.\end{array}$$

If X^* is disconnected, then $A\operatorname{Diag}(s)A^T \in \partial F(y^*)$ is singular. Let $\epsilon > 0$. Let $\{y^i\}$ be a sequence in $\mathbb{R}^{2n-1}$ such that $d^i = y^i - y^*$ with , and

Let $\{d^i\}$ be a sequence in $\mathbb{R}^{2n-1}$ such that $\|d^i\| = \|d\|$. Assume that the sequence $\{d^i\}$ converges to a vector in the null space of $A\operatorname{Diag}(s)A^T$. The normal fan of $[y^*]$ is complete, see Definition 7.1 and Example 7.3 in [51]. It follows from the classical Rockafellar-Pshenichnyi optimality condition for the minimization problem (4.3) that there always exists a vector $y^i \in \mathbb{R}^{2n-1}$ such that $d^i = y^i - P_{[y^*]}(y^i)$ for each d^i. This yields

$$\frac{||F(y^i)||}{\operatorname{dist}(y^i, [y^*])} = \frac{||A\operatorname{Diag}(s)A^T d^i||}{||d||} \to 0.$$

This shows that there exists no positive constant c such that (4.2) holds, and thus the local error bound condition fails. □

5 Numerical Experiments

In this section, we present numerical tests for the modified Newton algorithm. The main purpose is to illustrate empirically the correctness of our proposed algorithm. (Further extensive testing of these semismooth methods are given in [3, 44].)

First, we compare Algorithm 3.1 with the standard interior point method (IPM), and the alternating direction method of multipliers (ADMM). For the interior point method, the problem is modelled in CVX [26] and then solved using MOSEK solver [2]. For ADMM, we transform (1.3) to the equivalent problem $\min\left\{\frac{1}{2}\|x - \hat{x}\|^2 \,:\, x = y, Ay = b, \, x \geq 0\right\}$. The coupling constraint $x = y$

induces a standard splitting in the polyhedral cone variable x and the linear equality variable y. For more details about the implementation of ADMM applied to the least square problems, we refer to [27].

For the numerical experiments, we generate the data $\hat{X}$ from the standard normal distribution. Throughout Tables 5.1 and 5.2, n refers to the size of $\hat{X} \in \mathbb{R}^{n\times n}$; *iteration* refers to the number of iterations; *opt.cond.* refers to the sum of the norms of the optimality conditions in (1.4) at termination, i.e., primal and dual feasibility and complementary slackness; and *time* refers to the total running time in seconds.

Table 5.1 displays the numerical results for one instance of sizes $n = 100, \ldots, 500$, respectively. We compare the three methods. It is clear that the modified semismooth Newton method has a superior running time to ADMM and IPM. It also does better with respect to the optimality conditions. The tolerance for the optimality conditions for IPM is from the default obtained from MOSEK. As expected, interior point methods have difficulty obtaining more than square root of machine epsilon accuracy. The accuracy for ADMM methods take significantly longer if more accuracy is requested. In addition from (1.6), we see that both dual feasibility and complementary slackness hold exactly for the NM algorithm, and the optimality conditions error is totally from the primal feasibility residual $\|Ax - b\|$.

	The modified NM Algorithm 3.1			IPM			ADMM		
n	iteration	opt. cond.	time	iteration	opt. cond.	time	iteration	opt. cond.	time
100	9	1.2e-14	0.1	25	4.0e-10	0.51	941	9.9e-13	0.21
200	13	1.8e-14	0.1	26	1.4e-06	1.5	1735	9.9e-13	1.3
300	12	7.5e-15	0.18	22	6.8e-07	2.2	2746	1.0e-12	4.3
400	12	7.8e-15	0.33	22	1.3e-05	4.3	3834	1.0e-12	17
500	13	5.3e-15	0.55	25	4.9e-07	8.1	4634	1.0e-12	30

Table 5.1: Small instances

?⟨table1⟩?

In Table 5.2 below we present the numerical results of larger instances, $n = 1000, 1500, 2000$. We did not include results for IPM or ADMM as they took significantly longer.

	The modified NM Algorithm 3.1		
n	iteration	opt. cond.	time
1000	11	1.4e-15	0.47
2000	11	1.1e-15	1.6
3000	12	6.8e-16	3.9
4000	13	3.6e-16	7.6
5000	13	4.3e-16	12

Table 5.2: Medium and large instances

?⟨table2⟩?

Next, we compare Algorithm 3.1 with the semismooth Newton-CG algorithm (SSNCG1) presented in [36]. Roughly speaking, SSNCG1 avoids the singularity of the Jacobian matrix by adding a scaled identity matrix ϵI at each iteration for some $\epsilon > 0$. The scalar ϵ is determined by the residual at the current iteration and a number of parameters. The SSNCG1 also involves a line-search to determine its step length. We also note that the problem formulation in [36] does not remove the redundant constraint as (1.3).

We generate test instances whose optimal solution X^* has many blocks. As discussed in the paragraph after Theorem 3.17, we expect that an instance is difficult to solve if there are many blocks in X^*. This is substantiated in the numerical results in Figure 1.

In addition, we observe that Algorithm 3.1 consistently takes less iterations than SSNCG1. This may be explained by our fast quadratic convergence. However, the running time of Algorithm 3.1 is longer than SSNCG1 due to the costly vertex finding step in Algorithm 3.1. More specifically, the update in (3.11) takes a significant amount of the running time. The update (3.11) can be computed much more efficiently if the algorithm is implemented in C.

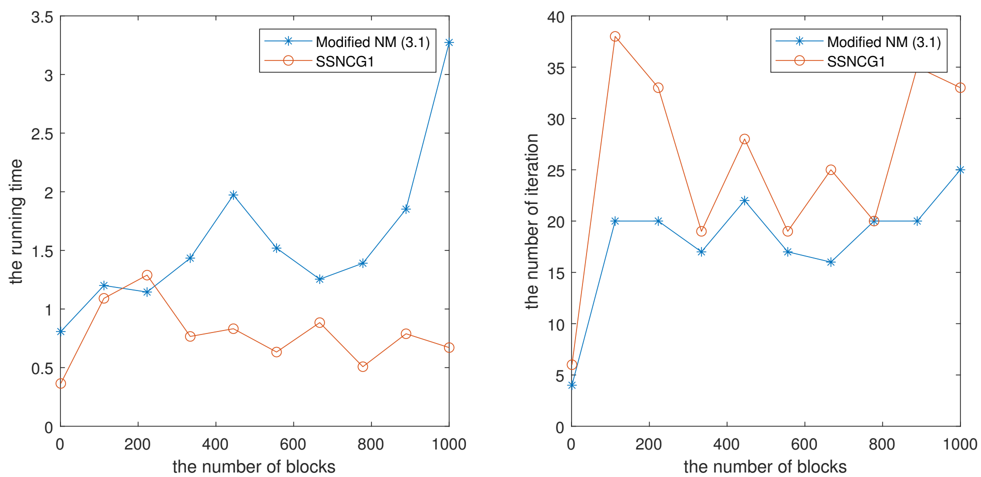

Figure 1: Problem instances of size $n = 1000$, but with different number of blocks in the optimal solution X^*

?⟨fig:blocks⟩?

6 Conclusion

The nearest doubly stochastic matrix problem is formulated as a system of strongly semismooth equations. We show that this system does not satisfy the so-called local error bound condition, and therefore, the quadratic convergence of a Newton-type method may not be guaranteed. We exploit the problem structure to construct a modified Newton method that converges to the solution at a quadratic rate. The novelty of the proposed algorithm is that the search space is partitioned into equivalence classes to overcome degeneracy. This partitioning strategy can be extended to more general problems. This is also the first known Newton-type method which enjoys quadratic convergence in the absence of the local error bound condition.

7 Acknowledgement

We would like to thank Xudong Li, Defeng Sun and Kim-Chuan Toh for sharing their codes in [36] with us.

?⟨ind:index⟩?

References

[1] S. Al-Homidan and H. Wolkowicz, *Approximate and exact completion problems for Euclidean distance matrices using semidefinite programming*, Linear Algebra Appl., 406 (2005), pp. 109–141. 5

[2] M. ApS, *Mosek optimization toolbox for MATLAB*, User's Guide and Reference Manual, version, 4 (2019). 26

[3] Z. Bai, D. Chu, and R. Tan, *Computing the nearest doubly stochastic matrix with a prescribed entry*, SIAM Journal on Scientific Computing, 29 (2007), pp. 635–655. 3, 26

[4] D. Bertsimas and J. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Belmont, MA, 1997. 3

[5] M. Best, *Quadratic programming with computer programs*, Advances in Applied Mathematics, CRC Press, Boca Raton, FL, 2017. 5

[6] N. Boland, C. Goh, and A. Mees, *An algorithm for solving quadratic network flow problems*, Appl. Math. Lett., 4 (1991), pp. 61–64. 5

[7] J. Borwein and A. Lewis, *Partially finite convex programming, part I, duality theory*, Math. Program., 57 (1992), pp. 15–48. 5

[8] J. Borwein and H. Wolkowicz, *Facial reduction for a cone-convex programming problem*, J. Austral. Math. Soc. Ser. A, 30 (1980/81), pp. 369–380. 3

[9] ——, *A simple constraint qualification in infinite-dimensional programming*, Math. Programming, 35 (1986), pp. 83–96. 5

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Machine Learning, 3 (2011), pp. 1–122. 3

[11] R. Brualdi, *Some applications of doubly stochastic matrices*, Linear Algebra and its Applications, 107 (1988), pp. 77–100. 3

[12] R. Brualdi and H. Ryser, *Combinatorial Matrix Theory*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1991. 7

[13] X. Chen, H. Qi, and P. Tseng, *Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems*, SIAM Journal on Optimization, 13 (2003), pp. 960–985. 6, 8

[14] F. Clarke, *Optimization and nonsmooth analysis*, vol. 5 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 1990. 2, 5

[15] W. Cook, W. Cunningham, W. Pulleyblank, and A. Schrijver, *Combinatorial optimization*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication. 9

[16] R. Cottle, J.-S. Pang, and R. Stone, *The linear complementarity problem*, vol. 60 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009. Corrected reprint of the 1992 original [MR1150683]. 5

[17] D. Cvetković, P. Rowlinson, and S. K. Simić, *Signless Laplacians of finite graphs*, Linear Algebra Appl., 423 (2007), pp. 155–171. 9

[18] H. Dan, N. Yamashita, and M. Fukushima, *Convergence properties of the inexact levenberg-marquardt method under local error bound conditions*, Optimization methods and software, 17 (2002), pp. 605–626. 2

[19] D. den Hertog, *Interior point approach to linear, quadratic and convex programming*, vol. 277 of Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1994. Algorithms and complexity. 5

[20] F. Deutsch, W. Li, and J. D. Ward, *A dual approach to constrained interpolation from a convex subset of hilbert space*, Journal of Approximation Theory, 90 (1997). 5

[21] J. Fan and Y. Yuan, *On the quadratic convergence of the Levenberg-Marquardt method without nonsingularity assumption*, Computing, 74 (2005), pp. 23–39. 6

[22] J. Fernández and M. Verón, *Newton's method: an updated approach of Kantorovich's theory*, Birkhäuser, 2017. 2

[23] J. Gauvin, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138. 4

[24] C. Goh, N. Boland, and A. Mees, *An algorithm for solving quadratic cost network flow optimization problems*, in Optimization, Vol. 1, 2 (Singapore, 1992), World Sci. Publ., River Edge, NJ, 1992, pp. 284–293. 5

[25] N. Graham, H. Hu, H. Im, X. Li, and H. Wolkowicz, *A restricted dual Peaceman-Rachford splitting method for QAP*, tech. rep., University of Waterloo, Waterloo, Ontario, 2020. 29 pages, submitted, research report. 3, 5

[26] M. Grant, S. Boyd, and Y. Ye, *Disciplined convex programming*, in Global optimization, vol. 84 of Nonconvex Optim. Appl., Springer, New York, 2006, pp. 155–210. 26

[27] B. He, M. Xu, and X. Yuan, *Solving large-scale least squares semidefinite programming by alternating direction methods*, SIAM Journal on Matrix Analysis and Applications, 32 (2011), pp. 136–152. 27

[28] R. Hessert and S. Mallik, *Moore-Penrose inverses of the signless Laplacian and edge-Laplacian of graphs*, Discrete Mathematics, 344 (2021), p. 112451. 9

[29] A. Hoffman, *On approximate solutions of systems of linear inequalities*, J. of Research of the National Bureau of Standards, 49 (1952), pp. 263–265. 22

[30] H. Hu and R. Sotirov, *On solving the quadratic shortest path problem*, INFORMS Journal on Computing, 32 (2020), pp. 219–233. 3

[31] H. Hu, R. Sotirov, and H. Wolkowicz, *Facial reduction for symmetry reduced semidefinite programs*, 2019. last revision Oct. 2020; under review for publication. 3

[32] J. Hueso, E. Martinez, and J. Torregrosa, *Modified Newton's method for systems of nonlinear equations with singular jacobian*, Journal of Computational and Applied Mathematics, 224 (2009), pp. 77–83. 3

[33] C. Kelley and Z. Xue, *Inexact newton methods for singular problems*, Optimization Methods and Software, 2 (1993), pp. 249–267. 3

[34] C. Lawson and R. Hanson, *Solving least squares problems*, vol. 15 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. Revised reprint of the 1974 original. 5

[35] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quarterly of applied mathematics, 2 (1944), pp. 164–168. 6

[36] X. Li, D. Sun, and K. Toh, *On the efficient computation of a generalized jacobian of the projector over the birkhoff polytope*, Mathematical Programming, 179 (2020), pp. 419–446. 3, 27, 28

[37] J. D. Louck, *Doubly stochastic matrices in quantum mechanics*, Foundations of Physics, 27 (1997), pp. 1085–1104. 3

[38] O. L. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47. 4

[39] M. Marcus, *Some properties and applications of doubly stochastic matrices*, The American Mathematical Monthly, 67 (1960), pp. 215–221. 3

[40] D. Marquardt, *An algorithm for least-squares estimation of nonlinear inequalities*, SIAM J. Appl. Math. v11, (1963), pp. 431–441. 6

[41] C. Micchelli, P. Smith, J. Swetits, and J. Ward, *Constrained l_p approximation*, Journal of Constructive Approximation, 1 (1985), pp. 93–102. 5

[42] B. S. Mordukhovich, *Variational analysis and generalized differentiation I: Basic theory*, vol. 330, Springer Science & Business Media, 2006. 2

[43] D. Oliveira, H. Wolkowicz, and Y. Xu, *ADMM for the SDP relaxation of the QAP*, Math. Program. Comput., 10 (2018), pp. 631–658. 3

[44] H. Qi and D. Sun, *A quadratically convergent Newton method for computing the nearest correlation matrix*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 360–385. 5, 26

[45] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Mathematical programming, 58 (1993), pp. 353–367. 2, 6, 12

[46] H. Rademacher, *Uber partielle und totale differenzierbarkeit i.*, Math. Ann., 89 (1919), pp. 340–359. 5

[47] P. Smith and H. Wolkowicz, *A nonlinear equation for linear programming*, Math. Programming, 34 (1986), pp. 235–238. 5

[48] D. Sun and J. Sun, *Semismooth matrix-valued functions*, Mathematics of Operations Research, 27 (2002), pp. 150–169. 6, 8

[49] B. Tam and S. Wu, *On the reduced signless Laplacian spectrum of a degree maximal graph*, Linear Algebra and its Applications, 432 (2010), pp. 1734–1756. 9

[50] N. Yamashita and M. Fukushima, *On the rate of convergence of the Levenberg-Marquardt method*, in Topics in numerical analysis, Springer, 2001, pp. 239–249. 6

[51] G. Ziegler, *Lectures on polytopes*, Springer-Verlag, New York, 1995. 26