A partial proximal point algorithm for nuclear norm regularized matrix least squares problems

Kaifeng Jiang*, Defeng Sun†, and Kim-Chuan Toh‡

January 10, 2012

Abstract

We introduce a partial proximal point algorithm for solving nuclear norm regularized matrix least squares problems with equality and inequality constraints. The inner subproblems, reformulated as a system of semismooth equations, are solved by an inexact smoothing Newton method, which is proved to be quadratically convergent under a constraint non-degeneracy condition, together with the strong semi-smoothness property of the singular value thresholding operator. Numerical experiments on a variety of problems including those arising from low-rank approximations of transition matrices show that our algorithm is efficient and robust.

1 Introduction

Let $\Re^{p\times q}$ be the space of all $p \times q$ matrices equipped with the standard trace inner product $\langle X, Y\rangle = \mathrm{Tr}(X^T Y)$ and its induced Frobenius norm $\|\cdot\|$. Without loss of generality, we assume $p \leq q$ throughout this paper. For a given $X \in \Re^{p\times q}$, its nuclear norm $\|X\|_*$ is defined as the sum of all its singular values and its operator norm $\|X\|_2$ is the largest singular value. Let $\mathcal{S}^n$ be the space of all $n \times n$ symmetric matrices and $\mathcal{S}^n_+$ be the cone of symmetric positive semidefinite matrices. We use the notation $X \succeq 0$ to denote that X is a symmetric positive semidefinite matrix.

In this paper, we are interested in solving the following nuclear norm regularized matrix least squares problem with linear equality and inequality constraints:

$$\min_{X\in\Re^{p\times q}} \Big\{ \frac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho\|X\|_* + \langle C, X\rangle \ : \ \mathcal{B}(X) \in d + \mathcal{Q} \Big\}, \tag{1}$$

*Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076 (kaifengjiang@nus.edu.sg).

†Department of Mathematics and Risk Management Institute, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076 (matsundf@nus.edu.sg).

‡Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076 (mattohkc@nus.edu.sg); and Singapore-MIT Alliance, 4 Engineering Drive 3, Singapore 117576.

where $\mathcal{A} : \Re^{p\times q} \to \Re^m$ and $\mathcal{B} : \Re^{p\times q} \to \Re^s$ are linear maps, $C \in \Re^{p\times q}, b \in \Re^m, d \in \Re^s$, ρ is a given positive parameter, and $\mathcal{Q} = \{0\}^{s_1} \times \Re^{s_2}_+$ is a polyhedral convex cone with $s = s_1 + s_2$. In many applications, we need to find a low rank approximation of a given target matrix while preserving certain structures. The nuclear norm function has been widely used as a regularizer which favors a low rank solution of (1). In [12], Chu et al. addressed some theoretical and numerical issues concerning structured low rank approximation problems. In many data analysis problems, the collected empirical data, possibly contaminated by noise, usually do not have the specified structure or the desired low rank. So it is important to find the nearest low rank approximation of the given matrix while maintaining the underlying structure of the original system. In practice, the data to be analyzed is very often nonnegative such as those corresponding to concentrations or intensity values, and it would be preferable to take into account such structural constraints.

Our nuclear norm regularized matrix least squares problem (1) arises from the recent intensive studies of the following affine rank minimization problem:

$$\min\Big\{\mathrm{rank}(X) \;:\; \mathcal{A}(X) \;=\; b,\; X \in \Re^{p\times q}\Big\}. \tag{2}$$

The problem (2) has many applications in diverse fields, see, e.g., [1, 2, 10, 44]. However, this affine rank minimization problem is generally an NP-hard nonconvex optimization problem. A tractable heuristic introduced in [19, 20] is to minimize the nuclear norm over the same constraints as in (2):

$$\min\Big\{\|X\|_* \;:\; \mathcal{A}(X) \;=\; b,\; X \in \Re^{p\times q}\Big\}. \tag{3}$$

The nuclear norm function is the greatest convex function majorized by the rank function over the unit ball of matrices with operator norm at most one, and it has been widely used as a surrogate for promoting the low rank structure. A frequently used alternative to (3) for accommodating problems with noisy data is the following matrix least squares problem with nuclear norm regularization (see [36, 52]):

$$\min\Big\{\frac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho\|X\|_* \;:\; X \in \Re^{p\times q}\Big\}. \tag{4}$$

It is known that (3) can be equivalently reformulated as a semidefinite programming (SDP) problem (see [44]), which has one $(p+q)\times(p+q)$ semidefinite constraint and m linear equality constraints. One can use standard interior-point method based semidefinite programming solvers such as SeDuMi [49] and SDPT3 [51] to solve this SDP problem. However, these solvers are not suitable for problems with large $p+q$ or m since in each iteration of these solvers, a large and dense Schur complement equation must be solved even when the data is sparse.

To overcome the difficulties faced by interior-point methods, several algorithms have been proposed to solve (3) or (4) directly. In [9], a singular value thresholding (SVT) algorithm is proposed for solving the following Tikhonov regularized version of (3):

$$\min\Big\{\tau\|X\|_* + \frac{1}{2}\|X\|^2 \;:\; \mathcal{A}(X) \;=\; b,\; X \in \Re^{p\times q}\Big\}, \tag{5}$$

where τ is a given positive regularization parameter. The SVT algorithm is a gradient method applied to the dual problem of (5). In [36], a fixed point continuation (FPC) algorithm is proposed for solving (4), together with a Bregman iterative algorithm for solving (3). Their numerical results on randomly generated matrix completion problems demonstrated that the FPC algorithm is much more efficient than the semidefinite programming solver SDPT3 when low accuracy solutions are sought. In [52], Toh and Yun proposed an accelerated proximal gradient algorithm (APG), which terminates in $O(1/\sqrt{\varepsilon})$ iterations for achieving ε-optimality, to solve the unconstrained matrix least squares problem (4). Their numerical results show that the APG algorithm is highly efficient and robust in solving large-scale random matrix completion problems. In [33], Liu et al. considered the following nuclear norm minimization problem with linear and second order cone constraints:

$$\min \Big\{ \|X\|_* \; : \; \mathcal{A}(X) \in b + \mathcal{K}, \; X \in \Re^{p \times q} \Big\}, \tag{6}$$

where $\mathcal{K} = \{0\}^{m_1} \times \mathcal{K}^{m_2}$, and $\mathcal{K}^{m_2}$ stands for the m_2-dimensional second order cone. They developed three inexact proximal point algorithms (PPA) in primal, dual and primal-dual forms with comprehensive convergence analysis, built upon the classic results of the general PPA established by Rockafellar [46]. Their numerical results demonstrated the efficiency and robustness of these PPAs in solving randomly generated and real matrix completion problems. Moreover, they showed that the SVT algorithm [9] is just one outer iteration of the exact primal PPA, and the Bregman iterative method [36] is a special case of the exact dual PPA. However, all the above mentioned models and related algorithms cannot address the following goal: given an observed data matrix (possibly contaminated by noise), find the nearest low rank approximation of the target matrix while maintaining certain prescribed structures. In particular, the APG method considered in [52] cannot be applied directly to solve the problem (1).

A strong motivation for proposing the model (1) arises from finding the nearest low rank approximation of transition matrices. For a given data matrix P which describes the full distribution of a random walk through the entire data set, the problem of finding the low rank approximation of P can be stated as follows:

$$\min \Big\{ \frac{1}{2}\|X - P\|^2 + \rho\|X\|_* \; : \; Xe = e, \; X \geq 0, \;\; X \in \Re^{n \times n} \Big\}, \tag{7}$$

where $e \in \Re^n$ is the vector of all ones and $X \geq 0$ denotes the condition that all entries of X are nonnegative. In [32], Lin proposed the Latent Markov Analysis (LMA) approach for finding the reduced rank approximations of transition matrices. The LMA is applied to clustering such that the inferred cluster relationships can be described probabilistically by the reduced-rank transition matrix. In [11], Chennubhotla exploited the spectral properties of the Markov transition matrix to obtain low rank approximation of the original transition matrix in order to develop a fast eigen-solver for spectral clustering. Furthermore, in many applications, since only partial information of the original transition matrix is available, it is also important to estimate the missing entries of P. For example, transition probabilities between different credit ratings play a crucial role in credit portfolio management. If our primary interest is in a specific group, then the number of observed rating transitions might be very small. Due to the lack of rating data, it is important to estimate the rating transition matrix in the presence of missing data [4].

In this paper, we study a partial proximal point algorithm (PPA) proposed by Ha [28] for solving (1), in which only some of the variables appear in the quadratic proximal term. The partial PPA was further analyzed by Bertsekas and Tseng [6], in which the close relation between the partial PPA and some parallel algorithms in convex programming was revealed. Given a sequence of parameters σ_k such that

$$0 < \sigma_k \uparrow \sigma_\infty \leq +\infty, \tag{8}$$

and an initial point $X^0 \in \Re^{p\times q}$, a sequence $\{(u^k, X^k)\} \subseteq \Re^m \times \Re^{p\times q}$ is generated by the partial PPA for solving (1) via the following scheme:

$$(u^{k+1}, X^{k+1}) \approx \arg\min \Big\{ f_\rho(u, X) + \frac{1}{2\sigma_k}\|X - X^k\|^2 \; : \; \mathcal{A}(X) + u = b, \, \mathcal{B}(X) \in d + \mathcal{Q} \Big\}, \tag{9}$$

where $f_\rho(u,X) := \dfrac{1}{2}\|u\|^2 + \rho\|X\|_* + \langle C, X\rangle$. Since the nuclear norm $\|\cdot\|_*$ is a nonsmooth function, an important issue we must address is how to solve (9) efficiently. The strong convexity of the objective function suggests to us to apply an indirect method for solving (9) based on the duality theory for convex programming. We note that the proposed partial PPA requires solving an inner subproblem with linear inequality constraints at each iteration. To handle the inequality constraints, Gao and Sun [23] recently designed a quadratically convergent inexact smoothing Newton method to solve semidefinite least squares problems with equality and inequality constraints but with simple quadratic objective functions of the form $\frac{1}{2}\|X - G\|^2$. Their numerical results demonstrated the high efficiency of the inexact smoothing Newton method. This strongly motivated us to use an inexact smoothing Newton method to solve our inner subproblems for achieving fast convergence. For the inner subproblem, due to the presence of inequality constraints, we reformulate the problem as a system of semismooth equations. By defining a smoothing function of the soft thresholding operator, we introduce an inexact smoothing Newton method to solve the semismooth system. The quadratic convergence of the inexact smoothing Newton method is proved under a constraint nondegeneracy condition, together with the strong semismoothness property of the soft thresholding operator.

The remaining parts of this paper are organized as follows. In section 2, we present some preliminaries about semismooth functions. We show that the soft thresholding operator is strongly semismooth everywhere, and define a smoothing function of the soft thresholding operator. In section 3, we design a partial PPA for solving the nuclear norm regularized matrix least squares problem (1), and establish its global and local convergence. Section 4 follows with the design of an inexact smoothing Newton method for solving the inner subproblems. The quadratic convergence of the method is established under a constraint nondegeneracy condition, which we also characterized. We report the numerical implementation and performance of our algorithm in section 5 and give the conclusion in section 6.

2 Preliminaries

In this section, we give a brief introduction on some basic concepts such as semismooth functions, the B-subdifferential and Clarke's generalized Jacobian of Lipschitz functions. We shall show that the soft thresholding operator is strongly semismooth everywhere. These concepts and properties will be critical for us to develop an inexact smoothing Newton method for solving the inner subproblems in our partial PPA.

Let $F : \Re^m \longrightarrow \Re^l$ be a locally Lipschitz function. By Redemacher's theorem, F is Fréchet differentiable almost everywhere. Let D_F denote the set of points in $\Re^m$ where F is differentiable. The B-subdifferential of F at $x \in \Re^m$ is defined by

$$\partial_B F(x) := \{V : \ V = \lim_{k\to\infty} F'(x^k), \ x^k \longrightarrow x, \ x^k \in D_F\},$$

where $F'(x)$ denotes the Jacobian of F at $x \in D_F$. Then Clarke's [13] generalized Jacobian of F at $x \in \Re^m$ is defined as the convex hull of $\partial_B F(x)$, i.e.,

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

From [43, Lemma 2.2], we know that if F is directionally differentiable in a neighborhood of $x \in \Re^m$, then for any $h \in \Re^m$, there exists $\mathcal{V} \in \partial F(x)$ such that $F'(x; h) = \mathcal{V}h$. The following concept of semismoothness was first introduced by Mifflin [38] for functionals and was extended by Qi and Sun [43] to vector-valued functions.

Definition 2.1. *We say that a locally Lipschitz function $F : \Re^m \longrightarrow \Re^l$ is semismooth at x if*
1. F is directionally differentiable at x; and
2. for any $h \in \Re^m$ and $V \in \partial F(x + h)$ with $h \to 0$,

$$F(x + h) - F(x) - Vh = o(\|h\|).$$

Furthermore, F is said to be strongly semismooth at x if F is semismooth at x and for any $h \in \Re^m$ and $V \in \partial F(x + h)$ with $h \to 0$,

$$F(x + h) - F(x) - Vh = O(\|h\|^2).$$

Let K be a closed convex cone in a finite dimensional real Euclidean space $\mathcal{X}$ equipped with a scalar inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|$. Let $\Pi_K : \mathcal{X} \to \mathcal{X}$ denote the metric projector over K, i.e., for any $y \in \mathcal{X}$, $\Pi_K(y)$ is the unique optimal solution to the following convex optimization problem:

$$\min\Big\{\frac{1}{2}\langle x - y, x - y\rangle \ : \ x \in K\Big\}.$$

It is well known [57] that the metric projector $\Pi_K(\cdot)$ is Lipschitz continuous with modulus 1 and $\|\Pi_K(\cdot)\|^2$ is continuously differentiable. Hence, $\Pi_K(\cdot)$ is almost everywhere Fréchet differentiable in $\mathcal{X}$ and for every $y \in \mathcal{X}$, $\partial\Pi_K(y)$ is well defined. For any $X \in \mathcal{S}^n$, let

$X_+ = \Pi_{S^n_+}(X)$ be the metric projection of X onto S^n_+ under the standard trace inner product. Assume that X has the following spectral decomposition

$$X = P\Lambda P^T,$$

where Λ is the diagonal matrix of eigenvalues and P is a corresponding orthogonal matrix of eigenvectors. Then

$$X_+ = P\Lambda_+ P^T,$$

where Λ_+ is a diagonal matrix whose diagonal entries are the nonnegative parts of the respective diagonal entries of Λ. The strong semismoothness of $\Pi_{S^n_+}(\cdot)$ has been proved by Sun and Sun in [50].

Next, we shall show that the soft thresholding operator [9, 33] is strongly semismooth everywhere. Let $Y \in \Re^{p\times q}$ admit the following singular value decomposition (SVD):

$$Y = U[\Sigma\ 0]V^T, \tag{10}$$

where $U \in \Re^{p\times p}$ and $V \in \Re^{q\times q}$ are orthogonal matrices, $\Sigma = \mathrm{Diag}(\sigma_1, \cdots, \sigma_p)$ is the diagonal matrix of singular values of Y, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Define $s : \Re \to \Re$ by

$$s(t) := (t-\rho)_+ - (-t-\rho)_+. \tag{11}$$

For each threshold $\rho > 0$, the soft thresholding operator D_ρ is defined as follows:

$$D_\rho(Y) = U[\Sigma_\rho\ 0]V^T, \tag{12}$$

where $\Sigma_\rho = \mathrm{Diag}(s(\sigma_1), \ldots, s(\sigma_p))$. Decompose $V \in \Re^{q\times q}$ as $V = [V_1\ \ V_2]$, where $V_1 \in R^{q\times p}$ and $V_2 \in \Re^{q\times(q-p)}$. Let the orthogonal matrix $Q \in \Re^{(p+q)\times(p+q)}$ be defined by

$$Q := \frac{1}{\sqrt{2}}\begin{bmatrix} U & U & 0 \\ V_1 & -V_1 & \sqrt{2}V_2 \end{bmatrix}, \tag{13}$$

and $\Xi : \Re^{p\times q} \to \mathcal{S}^{p+q}$ be defined by

$$\Xi(Y) := \begin{bmatrix} 0 & Y \\ Y^T & 0 \end{bmatrix}, \quad Y \in \Re^{p\times q}. \tag{14}$$

By [26, Section 8.6], we know that $\Xi(Y)$ has the following spectral decomposition:

$$\Xi(Y) = Q\begin{bmatrix} \Sigma & 0 & 0 \\ 0 & -\Sigma & 0 \\ 0 & 0 & 0 \end{bmatrix}Q^T, \tag{15}$$

i.e., the eigenvalues of $\Xi(Y)$ are $\pm\sigma_i, i = 1, \ldots, p$, and 0 of multiplicity $q-p$.

For any $W = P\mathrm{Diag}(\lambda_1, \cdots, \lambda_{p+q})P^T \in \mathcal{S}^{p+q}$, define $\widehat{S} : \mathcal{S}^{p+q} \to \mathcal{S}^{p+q}$ by

$$\widehat{S}(W) := P\mathrm{Diag}(s(\lambda_1), \cdots, s(\lambda_{p+q}))P^T = (W-\rho I)_+ - (-W-\rho I)_+.$$

Then, by the strong semismoothness property of $(\cdot)_+ : \mathcal{S}^{p+q} \to \mathcal{S}^{p+q}$, we have that $\widehat{S}(\cdot)$ is strongly semismooth everywhere in $\mathcal{S}^{p+q}$. By direct calculations, we have

$$S(Y) := \widehat{S}(\Xi(Y)) \; = Q \begin{bmatrix} \Sigma_\rho & 0 & 0 \\ 0 & -\Sigma_\rho & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T = \begin{bmatrix} 0 & D_\rho(Y) \\ D_\rho(Y)^T & 0 \end{bmatrix}. \tag{16}$$

Thus we have the following theorem.

Theorem 2.1. *The function $D_\rho(\cdot)$ is strongly semismooth everywhere in $\Re^{p\times q}$.*

Proof. Let $Y \in \Re^{p\times q}$ admit the SVD as in (10). Notice that $\widehat{S}(\cdot)$ is strongly semismooth in $\mathcal{S}^{p+q}$. This, together with (16), proves that $D_\rho(\cdot)$ is strongly semismooth at Y. Since Y is arbitrarily chosen, we have that $D_\rho(\cdot)$ is strongly semismooth everywhere in $\Re^{p\times q}$. $\square$

For the convenience of later discussion, we define the following three index sets:

$$\alpha := \{1, \ldots, p\}, \quad \gamma := \{p+1, \ldots, 2p\}, \quad \beta := \{2p+1, \ldots, p+q\}. \tag{17}$$

For each $\rho > 0$, the index set α is further decomposed into the following subindex sets:

$$\alpha_1 := \{i \,|\sigma_i(Y) > \rho, i \in \alpha\}, \alpha_2 := \{i \,|\sigma_i(Y) = \rho, i \in \alpha\}, \alpha_3 := \{i \,|\sigma_i(Y) < \rho, i \in \alpha\}. \tag{18}$$

In the following, we show that all the elements of the generalized Jacobian $\partial D_\rho(\cdot)$ are self-adjoint and positive semidefinite. First we prove the following lemma.

Lemma 2.1. *Let $Y \in \Re^{p\times q}$ admit the SVD as in (10). Then the unique minimizer of the following problem*

$$\min \Big\{ \|X - Y\|^2 \; : \; X \in B_\rho := \{ Z \in \Re^{p\times q} \; : \; \|Z\|_2 \le \rho \} \Big\} \tag{19}$$

is $X^* = \Pi_{B_\rho}(Y) = U[\min(\Sigma, \rho)\ 0]V^T$, where $\min(\Sigma,\rho) = \mathrm{Diag}(\min(\sigma_1, \rho), \ldots, \min(\sigma_p, \rho))$.

Proof. Obviously problem (19) has an unique optimal solution which is equal to $\Pi_{B_\rho}(Y)$. For any feasible Z with the SVD as in (10), we have $\sigma_i(Z) \le \rho,\ i = 1, \ldots, p$. Since $\|\cdot\|$ is unitarily invariant, by [7, Exercise IV.3.5], we have that

$$\|Y - Z\|^2 \ge \sum_{i\in\alpha_1} (\sigma_i(Y) - \sigma_i(Z))^2 + \sum_{i\in\alpha_2\cup\alpha_3} (\sigma_i(Y) - \sigma_i(Z))^2 \ge \sum_{i\in\alpha_1} (\sigma_i(Y) - \rho)^2.$$

Since $\|Y - X^*\|^2 = \sum_{i\in\alpha_1} (\sigma_i(Y) - \rho)^2$, we have $\|Y - Z\|^2 \ge \|Y - X^*\|^2$ for any $Z \in B_\rho$. Thus $X^* = U[\min(\Sigma, \rho)\ 0]V^T$ is the unique optimal solution. $\square$

Note that the above lemma has also been proved in [42] with a different proof. From the above lemma, we have that $D_\rho(Y) = Y - \Pi_{B_\rho}(Y)$, which implies that $\Pi_{B_\rho}(\cdot)$ is also strongly semismooth everywhere in $\Re^{p\times q}$. Now we have the following proposition.

Proposition 2.1. *For any $Y \in \Re^{p\times q}$ and $\mathcal{V} \in \partial D_\rho(Y)$, it holds that*
(a) $\mathcal{V}$ is self-adjoint.

(b) $\langle H, \mathcal{V}H\rangle \geq 0 \quad \forall H \in \Re^{p\times q}$.

(c) $\langle \mathcal{V}H, H - \mathcal{V}H\rangle \geq 0 \quad \forall H \in \Re^{p\times q}$.

Proof. (a) Since $D_\rho(Y) = Y - \Pi_{B_\rho}(Y)$, for any $\mathcal{V} \in \partial D_\rho(Y)$, there exists $\mathcal{W} \in \partial\Pi_{B_\rho}(Y)$ such that for any $H \in \Re^{p\times q}$, $\mathcal{V}H = H - \mathcal{W}H$. Since $\mathcal{W}$ is self-adjoint [37, Proposition 1], we have that $\mathcal{V}$ is also self-adjoint.

(b) It is a simple conclusion of (c).

(c) Since for any $H \in \Re^{p\times q}$, $\langle \mathcal{V}H, H - \mathcal{V}H\rangle = \langle H - \mathcal{W}H, \mathcal{W}H\rangle \geq 0$, where the last inequality follows from [37, Proposition 1], the third inequality holds. □

Next, we shall show that even though the soft thresholding operator $D_\rho(\cdot)$ is not differentiable everywhere, $\|D_\rho(\cdot)\|^2$ is continuously differentiable. First we summarize some well-known properties of the Moreau-Yosida [40, 56] regularization. Assume that $\mathcal{Y}$ is a finite-dimensional real Euclidean space. Let $f : \mathcal{Y} \to (-\infty, +\infty]$ be a proper lower semi-continuous convex function. For a given $\sigma > 0$, the Moreau-Yosida regularization of f is defined by

$$F_\sigma(y) = \min\left\{ f(x) + \frac{1}{2\sigma}\|x - y\|^2 \ : \ x \ \in \ \mathcal{Y} \right\}. \tag{20}$$

It is well known that F_σ is a continuously differentiable convex function on $\mathcal{Y}$ and

$$\nabla F_\sigma(y) = \frac{1}{\sigma}(y - x(y)),$$

where $x(y)$ denotes the unique optimal solution of (20). It is well known that $x(\cdot)$ is globally Lipschitz continuous with modulus 1 and ∇F_σ is globally Lipschitz continuous with modulus $1/\sigma$.

Proposition 2.2. *Let $\Theta(Y) = \frac{1}{2}\|D_\rho(Y)\|^2$, where $Y \in \Re^{p\times q}$. Then $\Theta(Y)$ is continuously differentiable and*

$$\nabla\Theta(Y) = D_\rho(Y). \tag{21}$$

Proof. It is already known that the following minimization problem

$$F(Y) \ = \ \min\left\{ \rho\|X\|_* + \frac{1}{2}\|X - Y\|^2 \ : \ X \in \Re^{p\times q} \right\},$$

has the unique optimal solution $X = D_\rho(Y)$ (see, [9, 36]). From the properties of the Moreau-Yosida regularization, we know that $D_\rho(\cdot)$ is globally Lipschitz continuous with modulus 1 and $F(Y)$ is continuously differentiable with

$$\nabla F(Y) = Y - D_\rho(Y). \tag{22}$$

Since $D_\rho(Y)$ is the unique optimal solution, we have that

$$F(Y) = \rho\|D_\rho(Y)\|_* + \frac{1}{2}\|D_\rho(Y) - Y\|^2 = \frac{1}{2}\|Y\|^2 - \frac{1}{2}\|D_\rho(Y)\|^2. \tag{23}$$

This, together with (22), implies that $\Theta(\cdot)$ is continuously differentiable with $\nabla\Theta(Y) = D_\rho(Y)$. □

Next, we shall discuss the smoothing counterpart of the soft thresholding operator $D_\rho(\cdot)$. Let $\chi(\varepsilon, t) : \Re \times \Re \to \Re$ be the following Huber smoothing function for $(t)_+$:

$$\chi(\varepsilon, t) = \begin{cases} t & \text{if } t \geq \frac{|\varepsilon|}{2}, \\ \dfrac{1}{2|\varepsilon|}\Big(t + \dfrac{|\varepsilon|}{2}\Big)^2 & \text{if } -\frac{|\varepsilon|}{2} < t < \frac{|\varepsilon|}{2}, \\ 0 & \text{if } t \leq -\frac{|\varepsilon|}{2}, \end{cases} \quad (\varepsilon, t) \in \Re \times \Re. \tag{24}$$

Then the smoothing function for $s(\cdot)$ in (11) is defined as follows:

$$\overline{s}(\varepsilon, t) = \chi(\varepsilon, t - \rho) - \chi(\varepsilon, -t - \rho), \quad (\varepsilon, t) \in \Re \times \Re. \tag{25}$$

Note that $\overline{s}(\varepsilon, \cdot)$ is an odd function with respect to $t \in \Re$. Let $Y \in \Re^{p\times q}$ admit the SVD as in (10). For any $\varepsilon \in \Re$, the smoothing function for $S(Y)$ in (16) is defined as follows:

$$\overline{S}(\varepsilon, Y) := Q \begin{bmatrix} \overline{\Sigma}_\rho & 0 & 0 \\ 0 & -\overline{\Sigma}_\rho & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T, \tag{26}$$

where $\overline{\Sigma}_\rho = \mathrm{Diag}(\overline{s}(\varepsilon, \sigma_1), \ldots, \overline{s}(\varepsilon, \sigma_p))$. By direct calculations, we have

$$\overline{S}(\varepsilon, Y) = \begin{bmatrix} 0 & \overline{D}_\rho(\varepsilon, Y) \\ (\overline{D}_\rho(\varepsilon, Y))^T & 0 \end{bmatrix} = \Xi\Big(\overline{D}_\rho(\varepsilon, Y)\Big),$$

where

$$\overline{D}_\rho(\varepsilon, Y) = U\big[\overline{\Sigma}_\rho \ \ 0\big]V^T \tag{27}$$

is a smoothing function for $D_\rho(Y)$. Note that when $\varepsilon = 0$, $\overline{S}(0, Y) = S(Y)$ and $\overline{D}_\rho(0, Y) = D_\rho(Y)$. For any $\lambda = (\lambda_1, \ldots, \lambda_{p+q})^T \in \Re^{p+q}$, let $\lambda_i = \sigma_i$ for $i \in \alpha$, $\lambda_i = -\sigma_{i-p}$ for $i \in \gamma$, and $\lambda_i = 0$ for $i \in \beta$. When $\varepsilon \neq 0$ or $\sigma_i \neq \rho, i = 1, \ldots, p$, we use $\Lambda(\varepsilon, \lambda) \in \mathcal{S}^{p+q}$ to denote the following first divided difference symmetric matrix for $\overline{s}(\varepsilon, \cdot)$ at λ

$$\Lambda(\varepsilon, \lambda) = \begin{bmatrix} \Lambda_{\alpha\alpha} & \Lambda_{\alpha\gamma} & \Lambda_{\alpha\beta} \\ \Lambda_{\alpha\gamma}^T & \Lambda_{\gamma\gamma} & \Lambda_{\gamma\beta} \\ \Lambda_{\alpha\beta}^T & \Lambda_{\gamma\beta}^T & \Lambda_{\beta\beta} \end{bmatrix}, \tag{28}$$

whose (i, j)-th entry is given by

$$[\Lambda(\varepsilon, \lambda)]_{ij} = \begin{cases} \dfrac{\overline{s}(\varepsilon, \lambda_i) - \overline{s}(\varepsilon, \lambda_j)}{\lambda_i - \lambda_j} \in [0, 1] & \text{if } \lambda_i \neq \lambda_j \\ (\overline{s})'_{\lambda_i}(\varepsilon, \lambda_i) \in [0, 1] & \text{if } \lambda_i = \lambda_j \end{cases}, \ i, j = 1, \ldots p + q.$$

Since $\overline{s}(\varepsilon, \cdot)$ is an odd function, we have the following results:

$$\Lambda_{\alpha\alpha} = \Lambda_{\gamma\gamma}, \quad \Lambda_{\alpha\gamma} = (\Lambda_{\alpha\gamma})^T, \quad \Lambda_{\gamma\beta} = \Lambda_{\alpha\beta},$$

and $(\Lambda(\varepsilon,\lambda))_{ij} \in [0,1]$ for all $i,j = 1,\ldots,p+q$. By the well known result of Löwner [34], we know that for any $H \in \Re^{p\times q}$,

$$(\overline{S})'_Y(\varepsilon, Y)H = Q[\Lambda(\varepsilon,\lambda) \circ (Q^T \Xi(H) Q)]Q^T, \tag{29}$$

where " $\circ$ " denotes the Hadamard product. Since

$$Q^T\Xi(H)Q \quad = \quad \frac{1}{2}\begin{bmatrix} H_1 + H_1^T & H_1^T - H_1 & \sqrt{2}H_2 \\ H_1 - H_1^T & -(H_1 + H_1^T) & \sqrt{2}H_2 \\ \sqrt{2}H_2^T & \sqrt{2}H_2^T & 0 \end{bmatrix}, \tag{30}$$

where $H_1 = U^T H V_1$ and $H_2 = U^T H V_2$, by simple algebraic calculations, we have that

$$(\overline{S})'_Y(\varepsilon, Y)H = Q[\Lambda(\varepsilon,\lambda) \circ (Q^T \Xi(H) Q)]Q^T = \begin{bmatrix} 0 & A_{12} \\ A_{12}^T & 0 \end{bmatrix},$$

where

$$A_{12} = U\Big(\Lambda_{\alpha\alpha} \circ H_1^s + \Lambda_{\alpha\gamma} \circ H_1^a\Big)V_1^T + U(\Lambda_{\alpha\beta} \ \circ H_2)V_2^T,$$

and $H_1^s = (H_1 + H_1^T)/2$, $H_1^a = (H_1 - H_1^T)/2$. When $\varepsilon \neq 0$ or $\sigma_i \neq \rho, i = 1,\ldots,p$, the partial derivative of $\overline{S}(\cdot,\cdot)$ with respect to ε can be computed by

$$(\overline{S})'_\varepsilon(\varepsilon, Y) = Q\left[\begin{array}{ccc} D(\varepsilon,\Sigma) & 0 & 0 \\ 0 & -D(\varepsilon,\Sigma) & 0 \\ 0 & 0 & 0 \end{array}\right]Q^T = \left[\begin{array}{cc} 0 & UD(\varepsilon,\Sigma)V_1^T \\ V_1 D(\varepsilon,\Sigma)U^T & 0 \end{array}\right], \tag{31}$$

where

$$D(\varepsilon,\Sigma) = \mathrm{Diag}((\overline{s})'_\varepsilon(\varepsilon,\sigma_1),\ldots,(\overline{s})'_\varepsilon(\varepsilon,\sigma_p)). \tag{32}$$

Since

$$(\overline{S})'(\varepsilon, Y)(\tau, H) = \left[\begin{array}{cc} 0 & (\overline{D}_\rho)'(\varepsilon, Y)(\tau, H) \\ ((\overline{D}_\rho)'(\varepsilon, Y)(\tau, H))^T & 0 \end{array}\right]$$

for any $(\tau, H) \in \Re \times \Re^{p\times q}$, we have

$$(\overline{D}_\rho)'(\varepsilon, Y)(\tau, H) = U\Big(\Lambda_{\alpha\alpha} \circ H_1^s + \Lambda_{\alpha\gamma} \circ H_1^a + \tau D(\varepsilon,\Sigma)\Big)V_1^T + U(\Lambda_{\alpha\beta} \ \circ H_2)V_2^T. \tag{33}$$

Thus, $\overline{D}_\rho(\cdot,\cdot)$ is continuously differentiable around $(\varepsilon, Y) \in \Re \times \Re^{p\times q}$ if $\varepsilon \neq 0$ or $\sigma_i \neq \rho, i = 1,\ldots,p$. Furthermore, $\overline{D}_\rho(\cdot,\cdot)$ is globally Lipschitz continuous and strongly semismooth at any $(0, Y) \in \Re \times \Re^{p\times q}$ [35].

Next, we will give a characterization of the generalized Jacobian $\partial \overline{D}_\rho(0, Y)$ for $(0, Y) \in \Re \times \Re^{p\times q}$. Let $\mathcal{D}$ be the set of points in $\Re \times \Re^{p\times q}$ at which $\overline{D}_\rho$ is differentiable. Suppose that $\mathcal{N}$ is any set of Lebesgue measure zero in $\Re \times \Re^{p\times q}$. Then

$$\partial \overline{D}_\rho(0, Y) = \mathrm{conv}\Big\{ \lim_{(\varepsilon^k, Y^k)\to(0,Y)} (\overline{D}_\rho)'(\varepsilon^k, Y^k) \ : \ (\varepsilon^k, Y^k) \in \mathcal{D}, \ (\varepsilon^k, Y^k) \notin \mathcal{N} \Big\}. \tag{34}$$

Note that $\partial \overline{D}_\rho(0, Y)$ does not depend on the choice of the null set $\mathcal{N}$ [54, Theorem 4].

Before characterizing $\partial \overline{D}_\rho(0,Y)$ in the next proposition, we need to introduce a "submap" of $\overline{D}_\rho$ as follows. Define $(\overline{D}_\rho)_{|\alpha_2|} : \Re \times \Re^{|\alpha_2|\times|\alpha_2|} \to \Re^{|\alpha_2|\times|\alpha_2|}$ by replacing the dimension p and q in the definition of $\overline{D}_\rho : \Re \times \Re^{p\times q} \to \Re^{p\times q}$ with $|\alpha_2|$, respectively, where the index set α_2 is defined as in (18). As in the case for $\overline{D}_\rho(\cdot,\cdot)$, the mapping $(\overline{D}_\rho)_{|\alpha_2|}(\cdot,\cdot)$ is also Lipschitz continuous. Thus Clarke's generalized Jacobian $\partial(\overline{D}_\rho)_{|\alpha_2|}(0,Z)$ for $(0,Z) \in \Re \times \Re^{|\alpha_2|\times|\alpha_2|}$ is well defined. We let $I_{|\alpha_2|}$ be the identity matrix of size $|\alpha_2|$.

Proposition 2.3. *Let $Y \in \Re^{p\times q}$ admit the SVD as in (10). Then, for any $\mathcal{V} \in \partial \overline{D}_\rho(0,Y)$, there exists $\mathcal{V}_{|\alpha_2|} \in \partial(\overline{D}_\rho)_{|\alpha_2|}(0,\rho I_{|\alpha_2|})$ such that*

$$\begin{aligned}\mathcal{V}(\tau,H) &= U \begin{bmatrix} (H_1^s)_{\alpha_1\alpha_1} & (H_1^s)_{\alpha_1\alpha_2} & \Omega_{\alpha_1\alpha_3} \circ (H_1^s)_{\alpha_1\alpha_3} \\ (H_1^s)^T_{\alpha_1\alpha_2} & \mathcal{V}_{|\alpha_2|}(\tau,(H_1^s)_{\alpha_2\alpha_2}) & 0 \\ \Omega^T_{\alpha_1\alpha_3} \circ (H_1^s)^T_{\alpha_1\alpha_3} & 0 & 0 \end{bmatrix} V_1^T \\ &\quad + U\Big[(\Gamma_{\alpha\gamma} \circ H_1^a)V_1^T + (\Gamma_{\alpha\beta} \circ H_2)V_2^T\Big] \end{aligned} \tag{35}$$

for all $(\tau,H) \in \Re \times \Re^{p\times q}$, where $\Omega_{\alpha_1\alpha_3}, \Gamma_{\alpha\gamma}$ and $\Gamma_{\alpha\beta}$ are defined as follows,

$$(\Omega_{\alpha_1\alpha_3})_{ij} := \frac{\sigma_i - \rho}{\sigma_i - \sigma_j}, \quad \text{for } i \in \alpha_1,\ j \in \alpha_3, \tag{36}$$

$$\Gamma_{\alpha\gamma} := \begin{bmatrix} \omega_{\alpha_1\alpha_1} & \omega_{\alpha_1\alpha_2} & \omega_{\alpha_1\alpha_3} \\ \omega^T_{\alpha_1\alpha_2} & 0 & 0 \\ \omega^T_{\alpha_1\alpha_3} & 0 & 0 \end{bmatrix},\ \omega_{ij} := \frac{(\sigma_i-\rho)_+ + (\sigma_j-\rho)_+}{\sigma_i+\sigma_j},\ \text{for } i \in \alpha_1, j \in \alpha, \tag{37}$$

$$\Gamma_{\alpha\beta} := \begin{bmatrix} \mu_{\alpha_1\bar{\beta}} \\ 0 \end{bmatrix},\ \bar{\beta} := \beta - 2p = \{1,\ldots,q-p\},\ \mu_{ij} := \frac{\sigma_i-\rho}{\sigma_i},\ \text{for } i \in \alpha_1,\ j \in \bar{\beta}, \tag{38}$$

$H_1 = U^T H V_1, H_2 = U^T H V_2$, and $H_1^s = (H_1 + H_1^T)/2,\ H_1^a = (H_1 - H_1^T)/2$.

Proof. Let $\mathcal{N} := \{0\} \times \Re^{p\times q}$ which has Lebesgue measure zero in $\Re \times \Re^{p\times q}$ and

$$\partial_{\mathcal{N}} \overline{D}_\rho(0,Y) := \Big\{ \lim_{k\to\infty} (\overline{D}_\rho)'(\varepsilon^k,Y^k) \ :\ (\varepsilon^k,Y^k) \to (0,Y),\ \varepsilon^k \neq 0 \Big\}. \tag{39}$$

Then, from (34), we have

$$\partial \overline{D}_\rho(0,Y) = \mathrm{conv}(\partial_{\mathcal{N}} \overline{D}_\rho(0,Y)).$$

First, we give a characterization of all elements in the set $\partial_{\mathcal{N}} \overline{D}_\rho(0,Y)$. For any $\mathcal{V} \in \partial_{\mathcal{N}} \overline{D}_\rho(0,Y)$, there exists a sequence $\{(\varepsilon^k,Y^k)\} \to (0,Y)$ with $\varepsilon^k \neq 0$ such that $\overline{D}_\rho$ is differential at (ε^k,Y^k) and for any $(\tau,H) \in \Re \times \Re^{p\times q}$,

$$\mathcal{V}(\tau,H) = \lim_{k\to\infty} (\overline{D}_\rho)'(\varepsilon^k,Y^k)(\tau,H).$$

Since $\varepsilon^k \neq 0$, we have that $\overline{S}$ defined by (26) is differential at (ε^k, Y^k) and

$$\begin{aligned}\lim_{k\to\infty}(\overline{S})'(\varepsilon^k, Y^k)(\tau, H) &= \left[\begin{array}{cc} 0 & \lim\limits_{k\to\infty}(\overline{D}_\rho)'(\varepsilon^k, Y^k)(\tau, H) \\ (\lim\limits_{k\to\infty}(\overline{D}_\rho)'(\varepsilon^k, Y^k)(\tau, H))^T & 0 \end{array}\right] \\ &= \left[\begin{array}{cc} 0 & \mathcal{V}(\tau, H) \\ (\mathcal{V}(\tau, H))^T & 0 \end{array}\right].\end{aligned}$$

Let $Y^k = U^k[\Sigma^k\ 0](V^k)^T$ be the SVD of Y^k, where $U^k \in \Re^{p\times p}$ and $V^k \in \Re^{q\times q}$ are orthogonal matrices, $\Sigma^k = \mathrm{Diag}(\sigma_1^k, \cdots, \sigma_p^k)$, and $\sigma_1^k \geq \sigma_2^k \geq \cdots \geq \sigma_p^k \geq 0$ are singular values of Y^k. Writing each Σ^k in the same format as Σ:

$$\Sigma^k = \left[\begin{array}{ccc} \Sigma_{\alpha_1}^k & 0 & 0 \\ 0 & \Sigma_{\alpha_2}^k & 0 \\ 0 & 0 & \Sigma_{\alpha_3}^k \end{array}\right],$$

we have $\Sigma = \lim_{k\to\infty}\Sigma^k$, which implies that $\Sigma_{\alpha_1}^k - \rho I_{|\alpha_1|}$ and $\Sigma_{\alpha_3}^k - \rho I_{|\alpha_3|}$ are nonsingular matrices for all k sufficiently large and $\lim_{k\to\infty}\Sigma_{\alpha_2}^k = \Sigma_{\alpha_2} = \rho I_{|\alpha_2|}$. Let $\lambda^k = (\lambda_1^k, \ldots, \lambda_{p+q}^k)^T \in \Re^{p+q}$, where $\lambda_i^k = \sigma_i^k$ for $i \in \alpha$, $\lambda_i^k = -\sigma_{i-p}^k$ for $i \in \gamma$, and $\lambda_i^k = 0$ for $i \in \beta$. Let $\Lambda^k \equiv \Lambda^k(\varepsilon^k, \lambda^k)$ be defined by (28) and $D^k \equiv D(\varepsilon^k, \Sigma^k)$ be defined by (32). Then, for any $(\tau, H) \in \Re \times \Re^{p\times q}$, we obtain from (29) and (31) that

$$(\overline{S})'(\varepsilon^k, Y^k)(\tau, H) = Q^k\left[\Lambda^k \circ \left((Q^k)^T\Xi(H)Q^k\right) + \tau\left(\begin{array}{ccc} D^k & 0 & 0 \\ 0 & -D^k & 0 \\ 0 & 0 & 0 \end{array}\right)\right](Q^k)^T, \qquad (40)$$

where Q^k has the form as in (13). By taking subsequences if necessary, we may assume that $\{U^k\}$ and $\{V^k\}$ are both convergent sequences with limits $U = \lim_{k\to\infty}U^k$ and $V = \lim_{k\to\infty}V^k$ (clearly $Y = U[\Sigma\ 0]V^T$). Since both $\{\Lambda^k\}$ and $\{D^k\}$ are uniformly bounded, by taking subsequences further if necessary, we may assume that both $\{\Lambda^k\}$ and $\{D^k\}$ converge. Let $M = \lim_{k\to\infty}\Lambda^k \circ ((Q^k)^T\Xi(H)Q^k) = \lim_{k\to\infty}\Lambda^k \circ (Q^T\Xi(H)Q)$. Taking limits on both sides of (40), we obtain that

$$Q^T\left[\begin{array}{cc} 0 & \mathcal{V}(\tau, H) \\ (\mathcal{V}(\tau, H))^T & 0 \end{array}\right]Q = \left[\begin{array}{ccc} M_{\alpha\alpha} + \tau\lim\limits_{k\to\infty}D^k & M_{\alpha\gamma} & M_{\alpha\beta} \\ M_{\alpha\gamma}^T & M_{\gamma\gamma} - \tau\lim\limits_{k\to\infty}D^k & M_{\gamma\beta} \\ M_{\alpha\beta}^T & M_{\gamma\beta}^T & M_{\beta\beta} \end{array}\right]. \qquad (41)$$

By simple calculations, we have $\lim_{k\to\infty}\Lambda_{\alpha\gamma}^k = \Gamma_{\alpha\gamma}$, $\lim_{k\to\infty}\Lambda_{\alpha\beta}^k = \Gamma_{\alpha\beta}$ and $\lim_{k\to\infty}\Lambda_{\beta\beta}^k = 0$, where $\Gamma_{\alpha\gamma}$ and $\Gamma_{\alpha\beta}$ have the forms as in (37) and (38), respectively. Thus we have

$$M_{\alpha\alpha} = \left[\begin{array}{ccc} (H_1^s)_{\alpha_1\alpha_1} & (H_1^s)_{\alpha_1\alpha_2} & \Omega_{\alpha_1\alpha_3} \circ (H_1^s)_{\alpha_1\alpha_3} \\ (H_1^s)_{\alpha_1\alpha_2}^T & \lim\limits_{k\to\infty}(\Lambda_{\alpha\alpha}^k)_{\alpha_2\alpha_2} \circ (H_1^s)_{\alpha_2\alpha_2} & 0 \\ \Omega_{\alpha_1\alpha_3}^T \circ (H_1^s)_{\alpha_1\alpha_3}^T & 0 & 0 \end{array}\right],$$

$$M_{\alpha\gamma} = \Gamma_{\alpha\gamma} \circ (-H_1^a),\ M_{\alpha\beta} = \Gamma_{\alpha\beta} \circ (\frac{1}{\sqrt{2}}H_2),\ M_{\gamma\gamma} = -M_{\alpha\alpha},\ M_{\gamma\beta} = M_{\alpha\beta},\ M_{\beta\beta} = 0,$$

and

$$\lim_{k\to\infty} D^k = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lim\limits_{k\to\infty} D^k_{\alpha_2\alpha_2} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $\Omega_{\alpha_1\alpha_3}$ has the form as in (36), and

$$D^k_{\alpha_2\alpha_2} = \mathrm{Diag}\big((\overline{s})'_\varepsilon(\varepsilon_k, \sigma^k_{|\alpha_1|+1}), \ldots, (\overline{s})'_\varepsilon(\varepsilon_k, \sigma^k_{|\alpha_1|+|\alpha_2|})\big).$$

By applying (33) to $(\overline{D}_\rho)_{|\alpha_2|}$ at $(\varepsilon^k, \Sigma^k_{\alpha_2})$, for any $(\tau, \Delta) \in \Re \times \Re^{|\alpha_2|\times|\alpha_2|}$, we have

$$(\overline{D}_\rho)'_{|\alpha_2|}(\varepsilon^k, \Sigma^k_{\alpha_2})(\tau, \Delta) = (\Lambda^k_{\alpha\alpha})_{\alpha_2\alpha_2} \circ \frac{\Delta + \Delta^T}{2} + (\Lambda^k_{\alpha\gamma})_{\alpha_2\alpha_2} \circ \frac{\Delta - \Delta^T}{2} + \tau D^k_{\alpha_2\alpha_2}.$$

Since both $\{\Lambda^k\}$ and $\{D^k\}$ converge, we obtain that $\lim_{k\to\infty}(\overline{D}_\rho)'_{|\alpha_2|}(\varepsilon^k, \Sigma^k_{\alpha_2})(\tau, \Delta)$ exists for any $(\tau, \Delta) \in \Re \times \Re^{|\alpha_2|\times|\alpha_2|}$, which implies that $\lim_{k\to\infty}(\overline{D}_\rho)'_{|\alpha_2|}(\varepsilon^k, \Sigma^k_{\alpha_2})$ exists. By the definition of $\partial_{\mathcal{N}}(\overline{D}_\rho)_{|\alpha_2|}(0, \rho I_{|\alpha_2|})$, which is analogous to the one defined in (39), there exists $\mathcal{V}_{|\alpha_2|} \in \partial_{\mathcal{N}}(\overline{D}_\rho)_{|\alpha_2|}(0, \rho I_{|\alpha_2|})$ such that

$$\mathcal{V}_{|\alpha_2|}(\tau, \Delta) = \lim_{k\to\infty}(\overline{D}_\rho)'_{|\alpha_2|}(\varepsilon^k, \Sigma^k_{\alpha_2})(\tau, \Delta) \quad \forall\, (\tau, \Delta) \in \Re \times \Re^{|\alpha_2|\times|\alpha_2|}.$$

In particular, let $\Delta = (H_1^s)_{\alpha_2\alpha_2}$, we have

$$\mathcal{V}_{|\alpha_2|}(\tau, (H_1^s)_{\alpha_2\alpha_2}) = \lim_{k\to\infty}(\overline{D}_\rho)'_{|\alpha_2|}(\varepsilon^k, \Sigma^k_{\alpha_2})(\tau, (H_1^s)_{\alpha_2\alpha_2}) = \lim_{k\to\infty}(\Lambda^k_{\alpha\alpha})_{\alpha_2\alpha_2} \circ (H_1^s)_{\alpha_2\alpha_2} + \tau D^k_{\alpha_2\alpha_2}.$$

Hence

$$M_{\alpha\alpha} + \tau \lim_{k\to\infty} D^k = \begin{bmatrix} (H_1^s)_{\alpha_1\alpha_1} & (H_1^s)_{\alpha_1\alpha_2} & \Omega_{\alpha_1\alpha_3} \circ (H_1^s)_{\alpha_1\alpha_3} \\ (H_1^s)^T_{\alpha_1\alpha_2} & \mathcal{V}_{|\alpha_2|}(\tau, (H_1^s)_{\alpha_2\alpha_2}) & 0 \\ \Omega^T_{\alpha_1\alpha_3} \circ (H_1^s)^T_{\alpha_1\alpha_3} & 0 & 0 \end{bmatrix}. \tag{42}$$

By simple algebraic calculations, we obtain from (41) that

$$\begin{aligned} \mathcal{V}(\tau, H) \;&=\; U \begin{bmatrix} (H_1^s)_{\alpha_1\alpha_1} & (H_1^s)_{\alpha_1\alpha_2} & \Omega_{\alpha_1\alpha_3} \circ (H_1^s)_{\alpha_1\alpha_3} \\ (H_1^s)^T_{\alpha_1\alpha_2} & \mathcal{V}_{|\alpha_2|}(\tau, (H_1^s)_{\alpha_2\alpha_2}) & 0 \\ \Omega^T_{\alpha_1\alpha_3} \circ (H_1^s)^T_{\alpha_1\alpha_3} & 0 & 0 \end{bmatrix} V_1^T \\ &\quad + U\Big[(\Gamma_{\alpha\gamma} \circ H_1^a)V_1^T + (\Gamma_{\alpha\beta} \circ H_2)V_2^T\Big]. \end{aligned}$$

Since $\partial\overline{D}_\rho(0, Y) = \mathrm{conv}(\partial_{\mathcal{N}}\overline{D}_\rho(0, Y))$ and $\partial(\overline{D}_\rho)_{|\alpha_2|}(0, \rho I_{|\alpha_2|}) = \mathrm{conv}(\partial_{\mathcal{N}}(\overline{D}_\rho)_{|\alpha_2|}(0, \rho I_{|\alpha_2|}))$, from the above equality, we conclude that (35) holds. $\square$

Next, we present a useful inequality for elements in $\partial\overline{D}_\rho(0, Y)$, which is analogous to (c) of Proposition 2.1 for the soft thresholding operator $D_\rho(\cdot)$.

Proposition 2.4. *For any $\mathcal{V} \in \partial\overline{D}_\rho(0, Y)$, it holds that*

$$\langle H - \mathcal{V}(0, H), \mathcal{V}(0, H)\rangle \geq 0 \quad \forall H \in \Re^{p\times q}. \tag{43}$$

Proof. First, we show that for any $\mathcal{V} \in \partial_{\mathcal{N}} \overline{D}_\rho(0, Y)$ defined by (39), inequality (43) holds. For any $\mathcal{V} \in \partial_{\mathcal{N}} \overline{D}_\rho(0, Y)$, there exists a sequence $\{(\varepsilon^k, Y^k)\} \to (0, Y), \varepsilon^k \neq 0$ such that $\overline{D}_\rho$ is differentiable at (ε^k, Y^k) and for any $H \in \Re^{p\times q}$, $\mathcal{V}(0, H) = \lim_{k\to\infty}(\overline{D}_\rho)'(\varepsilon^k, Y^k)(0, H)$. Since $\varepsilon^k \neq 0$, we know that $\overline{S}$ defined by (26) is differentiable at (ε^k, Y^k). Now

$$
\begin{aligned}
&2\langle H, (\overline{D}_\rho)'(\varepsilon^k, Y^k)(0, H)\rangle = \Big\langle \Xi(H),\ \Xi\big((\overline{D}_\rho)'(\varepsilon^k, Y^k)(0, H)\big)\Big\rangle \\
&= \langle \Xi(H), (\overline{S})'(\varepsilon^k, Y^k)(0, H)\rangle = \langle \Xi(H),\ Q^k\Big(\Lambda^k \circ ((Q^k)^T \Xi(H) Q^k)\Big)(Q^k)^T\rangle \\
&= \langle \widetilde{H}_k, \Lambda^k \circ \widetilde{H}_k\rangle = \sum_{i=1}^{p+q}\sum_{j=1}^{p+q} \Lambda^k_{ij}\ (\widetilde{H}_k)^2_{ij},
\end{aligned}
$$

where $\widetilde{H}_k = (Q^k)^T \Xi(H) Q^k$. Similarly,

$$
\langle (\overline{D}_\rho)'(\varepsilon^k, Y^k)(0, H), (\overline{D}_\rho)'(\varepsilon^k, Y^k)(0, H)\rangle = \frac{1}{2}\sum_{i=1}^{p+q}\sum_{j=1}^{p+q} (\Lambda^k_{ij})^2\ (\widetilde{H}_k)^2_{ij}.
$$

Since $\Lambda^k_{ij} \in [0, 1]$ for all $i, j = 1, \ldots, p + q$, we have

$$
\langle H - (\overline{D}_\rho)'(\varepsilon^k, Y^k)(0, H), (\overline{D}_\rho)'(\varepsilon^k, Y^k)(0, H)\rangle = \frac{1}{2}\sum_{i=1}^{p+q}\sum_{j=1}^{p+q} (\Lambda^k_{ij} - (\Lambda^k_{ij})^2)(\widetilde{H}_k)^2_{ij} \geq 0.
$$

Hence $\langle H - \mathcal{V}(0, H), \mathcal{V}(0, H)\rangle \geq 0 \quad \forall H \in \Re^{p\times q}$.

Let $\mathcal{V} \in \partial \overline{D}_\rho(0, Y)$. By Carathéodory's theorem, there exists a positive κ and $\mathcal{V}^i \in \partial_{\mathcal{N}} \overline{D}_\rho(0, Y), i = 1, \ldots, \kappa$ such that $\mathcal{V} = \sum_{i=1}^{\kappa} t_i \mathcal{V}^i$, where $t_i \geq 0\ \forall\ i$, and $\sum_{i=1}^{\kappa} t_i = 1$. By the convexity of $\|\cdot\|^2$, we have that for any $H \in \Re^{p\times q}$

$$
\|(\mathcal{V}(0, H))\|^2 = \|\sum_{i=1}^{\kappa} t_i \mathcal{V}^i(0, H)\|^2 \leq \sum_{i=1}^{\kappa} t_i \|\mathcal{V}^i(0, H)\|^2 = \sum_{i=1}^{\kappa} t_i \langle \mathcal{V}^i(0, H), \mathcal{V}^i(0, H)\rangle,
$$

which implies

$$
\langle \mathcal{V}(0, H), \mathcal{V}(0, H)\rangle \leq \sum_{i=1}^{\kappa} t_i \langle H, \mathcal{V}^i(0, H)\rangle = \langle H, \sum_{i=1}^{\kappa} t_i \mathcal{V}^i(0, H)\rangle = \langle H, \mathcal{V}(0, H)\rangle.
$$

Hence, (43) holds. □

3 A partial proximal point algorithm for matrix least squares problems

In this section, we will describe how to design a partial proximal point algorithm (PPA) to solve the problem (1).

Let $\mathcal{Z}$ be a finite dimensional real Euclidean space with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\|\cdot\|$. Let $\mathcal{T} : \mathcal{Z} \to \mathcal{Z}$ be a maximal monotone operator. We define its domain and image, respectively, as follows: $\mathrm{Dom}\,(\mathcal{T}) := \{z \in \mathcal{Z} \mid \mathcal{T}(z) \neq \emptyset\}$ and $\mathrm{Im}\,(\mathcal{T}) := \bigcup_{z \in \mathcal{Z}} \mathcal{T}(z)$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are two finite dimensional real Euclidean spaces each equipped with a scalar product $\langle \cdot, \cdot \rangle$ and its induced norm $\|\cdot\|$. Suppose now that $z \in \mathcal{Z}$ is partitioned into two components $z = (x, y)$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In [28], Ha proposed a partial PPA to solve the inclusion problem in two variables $0 \in \mathcal{T}(x, y)$, in which only one of the variables is involved in the proposed iterative procedure. Below we give a brief review of the partial PPA proposed by Ha [28]. Let $\Pi : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ be the orthogonal projection of $\mathcal{X} \times \mathcal{Y}$ onto $\{0\} \times \mathcal{Y}$, i.e., $\Pi(x, y) = (0, y)$. To solve the inclusion problem $0 \in \mathcal{T}(x, y)$, from a given initial point $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$, the exact partial PPA generates a sequence $\{(x^k, y^k)\}$ by the following scheme:

$$(x^{k+1}, y^{k+1}) \in P_{\sigma_k}(x^k, y^k), \tag{44}$$

where $P_{\sigma_k} := (\Pi + \sigma_k \mathcal{T})^{-1} \Pi$ and the sequence $\{\sigma_k\}$ satisfies (8). In general, the mapping P_{σ_k} is neither single-valued nor nonexpansive. However, by [28, Proposition 2], we know that the second component of $P_{\sigma_k}(x^k, y^k)$ is uniquely determined and nonexpansive. For practical purpose, the following general approximation criteria were introduced in [28]:

$$\|(x^{k+1}, y^{k+1}) - (u^{k+1}, v^{k+1})\| \le \varepsilon_k, \quad \varepsilon_k > 0, \ \sum_{k=0}^{\infty} \varepsilon_k < \infty, \tag{45a}$$

$$\|(x^{k+1}, y^{k+1}) - (u^{k+1}, v^{k+1})\| \le \delta_k \|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|, \tag{45b}$$

$$\|y^{k+1} - v^{k+1}\| \le \delta_k \|y^{k+1} - y^k\|, \quad \delta_k > 0, \ \sum_{k=0}^{\infty} \delta_k < \infty, \tag{45c}$$

where $(u^{k+1}, v^{k+1}) \in P_{\sigma_k}(x^k, y^k)$. In [28], Ha showed that under mild assumptions, any cluster point of the sequence $\{(x^k, y^k)\}$ generated by the partial PPA under criterion (45a) is a solution to $0 \in \mathcal{T}(x, y)$. Moreover, the sequence $\{y^k\}$ converges weakly to $\bar{y}$, which is the second component of a solution to $0 \in \mathcal{T}(x, y)$. If, in addition, (45b) and (45c) are also satisfied and $\mathcal{T}^{-1}$ is Lipschitz continuous at the origin, then the sequence $\{(x^k, y^k)\}$ converges locally at least at a linear rate which tends to zero as $\sigma_k \to +\infty$. For more discussions of the convergence analysis of the partial PPA, see [28, Theorem 1 & 2].

Next, we shall show how to use the partial PPA to solve (1). It is easy to see that (1) can be rewritten as follows:

$$\min_{u \in \Re^m, X \in \Re^{p \times q}} \Big\{ f_\rho(u, X) := \frac{1}{2}\|u\|^2 + \rho \|X\|_* + \langle C, X \rangle \ : \ \mathcal{A}(X) + u = b, \mathcal{B}(X) \in d + \mathcal{Q} \Big\}. \tag{46}$$

Note that the objective function $f_\rho(u, X)$ is strongly convex in u for all $X \in \Re^{p \times q}$. For the convergence analysis, we assume that the following Slater condition holds:

$$\begin{cases} \{\mathcal{B}_i\}_{i=1}^{s_1} \text{ are linearly independent and } \exists\, X^0 \in \Re^{p \times q} \\ \text{such that } \mathcal{B}_i(X^0) = d_i, i = 1, \ldots, s_1 \text{ and } \mathcal{B}_i(X^0) > d_i, i = s_1 + 1, \ldots, s. \end{cases} \tag{47}$$

Let $l(u,X;\zeta,\xi):\Re^m\times\Re^{p\times q}\times\Re^m\times\Re^s\to\Re$ be the ordinary Lagrangian function for (46) in the extended form:

$$l(u,X;\zeta,\xi):=\begin{cases} f_\rho(u,X)+\langle\zeta,b-\mathcal{A}(X)-u\rangle+\langle\xi,d-\mathcal{B}(X)\rangle & \text{if } \xi\in\mathcal{Q}^*,\\ -\infty & \text{if } \xi\notin\mathcal{Q}^*,\end{cases}\tag{48}$$

where $\mathcal{Q}^*=\Re^{s_1}\times\Re^{s_2}_+$ is the dual cone of $\mathcal{Q}$. The essential objective function in (46) is

$$f(u,X)\ :=\ \sup_{\zeta\in\Re^m,\ \xi\in\Re^s} l(u,X;\zeta,\xi)=\begin{cases} f_\rho(u,X) & \text{if } (u,X)\in\mathcal{F}_P,\\ +\infty & \text{if } (u,X)\notin\mathcal{F}_P,\end{cases}\tag{49}$$

where $\mathcal{F}_P=\{(u,X)\in\Re^m\times\Re^{p\times q}\mid \mathcal{A}(X)+u=b,\mathcal{B}(X)\in d+\mathcal{Q}\}$ is the feasible set of (46). The dual problem of (46) is given by:

$$\max\Big\{g_\rho(\zeta,\xi)\ :\ \mathcal{A}^*(\zeta)+\mathcal{B}^*(\xi)+Z=C,\ \|Z\|_2\le\rho,\ \zeta\in\Re^m,\ \xi\in\mathcal{Q}^*,\ Z\in\Re^{p\times q}\Big\},\tag{50}$$

where $g_\rho(\zeta,\xi):=-\frac{1}{2}\|\zeta\|^2+\langle b,\zeta\rangle+\langle d,\xi\rangle$. As in Rockafellar [47], we define the following two maximal monotone operators

$$\begin{cases}\mathcal{T}_f(u,X)=\{(v,Y)\in\Re^m\times\Re^{p\times q}:\ (v,Y)\in\partial f(u,X)\},\\ \mathcal{T}_l(u,X;\zeta,\xi)=\{(v,Y,y,z)\in\Re^m\times\Re^{p\times q}\times\Re^m\times\Re^s:(v,Y,-y,-z)\in\partial l(u,X;\zeta,\xi)\},\end{cases}$$

where $u\in\Re^m, X\in\Re^{p\times q},\zeta\in\Re^m$, and $\xi\in\Re^s$. Note that since $f(u,X)$ is strongly convex in u with modulus 1 for all $X\in\Re^{p\times q}$, $\mathcal{T}_f$ is strongly monotone with modulus 1 with respect to the variable u [46, Proposition 6], i.e.,

$$\langle(u,X)-(u',X'),\,(v,Y)-(v',Y')\rangle\ge\|u-u'\|^2\tag{51}$$

for all $(v,Y)\in\mathcal{T}_f(u,X)$ and $(v',Y')\in\mathcal{T}_f(u',X')$. From the definition of $\mathcal{T}_f$, we know that for any $(v,Y)\in\Re^m\times\Re^{p\times q}$,

$$\mathcal{T}_f^{-1}(v,Y)=\arg\min_{u\in\Re^m,X\in\Re^{p\times q}}\Big\{f(u,X)-\langle v,u\rangle-\langle Y,X\rangle\Big\}.$$

Similarly, for any $(v,Y,y,z)\in\Re^m\times\Re^{p\times q}\times\Re^m\times\Re^s$,

$$\mathcal{T}_l^{-1}(v,Y,y,z)=\arg\min_{\substack{u\in\Re^m\\ X\in\Re^{p\times q}}}\max_{\substack{\zeta\in\Re^m\\ \xi\in\Re^s}}\Big\{l(u,X;\zeta,\xi)-\langle v,u\rangle-\langle Y,X\rangle+\langle y,\zeta\rangle+\langle z,\xi\rangle\Big\}.$$

Since $f(u,X)$ is strongly convex in u for all $X\in\Re^{p\times q}$, we apply the partial PPA proposed by Ha [28] to the maximal monotone operator $\mathcal{T}_f$, in which only the variable X appears in the quadratic proximal term. Given a starting point $(u^0,X^0)\in\Re^m\times\Re^{p\times q}$, the inexact partial PPA generates a sequence $\{(u^k,X^k)\}$ by approximately solving the following problem

$$\min_{u\in\Re^m,X\in\Re^{p\times q}}\Big\{f(u,X)+\frac{1}{2\sigma_k}\|X-X^k\|\Big\}.\tag{52}$$

We can easily see that any minimizer (u, X) of problem (52) satisfies

$$(0, X^k) \in (0, X) + \sigma_k \mathcal{T}_f(u, X). \tag{53}$$

Let $\Pi : \Re^m \times \Re^{p\times q} \to \Re^m \times \Re^{p\times q}$ be the orthogonal projector of $\Re^m \times \Re^{p\times q}$ onto $\{0\} \times \Re^{p\times q}$, i.e., $\Pi(u, X) = (0, X)$. Then (53) can be equivalently written as

$$(u, X) \in (\Pi + \sigma_k \mathcal{T}_f)^{-1}\Pi(u^k, X^k).$$

It follows that the set of minimizers of (52) can be expressed as $(\Pi + \sigma_k \mathcal{T}_f)^{-1}\Pi(u^k, X^k)$.

Next, for any parameter $\sigma > 0$, we show some properties of the mappings $Q_\sigma := (\Pi + \sigma\mathcal{T}_f)^{-1}$ and $P_\sigma := Q_\sigma\Pi = (\Pi + \sigma\mathcal{T}_f)^{-1}\Pi$. The proofs essentially follow the ideas in [30, Proposition 2 & 3] together with [15, Theorem 2.7], and we shall omit them.

Proposition 3.1. *For any given parameter $\sigma > 0$, let $Q_\sigma = (\Pi+\sigma\mathcal{T}_f)^{-1}$, and $P_\sigma = Q_\sigma\Pi = (\Pi + \sigma\mathcal{T}_f)^{-1}\Pi$. Suppose that* $\mathrm{Dom}\,(\mathcal{T}_f) \neq \emptyset$. *Then we have the following properties:*

(i) The mapping Q_σ and P_σ are single-valued in $\Re^m \times \Re^{p\times q}$.

(ii) Let $\beta = \min\{1, \sigma\}$. For any $(u, X), (u', X') \in \Re^m \times \Re^{p\times q}$,

$$\|Q_\sigma(u, X) - Q_\sigma(u', X')\| \leq \tfrac{1}{\beta}\|(u, X) - (u', X')\|.$$
$$\|P_\sigma(u, X) - P_\sigma(u', X')\| \leq \tfrac{1}{\beta}\|X - X'\|.$$

Since the operator P_{σ_k} is single-valued, the approximate rule of the partial PPA for solving problem (46) can be expressed as

$$(u^{k+1}, X^{k+1}) \;\approx\; P_{\sigma_k}(u^k, X^k) \;:=\; (\Pi + \sigma_k\mathcal{T}_f)^{-1}\Pi(u^k, X^k), \tag{54}$$

where $P_{\sigma_k}(u^k, X^k)$ is defined by

$$P_{\sigma_k}(u^k, X^k) \;=\; \mathrm{argmin}_{u\in\Re^m, X\in\Re^{p\times q}} \left\{ f(u, X) + \frac{1}{2\sigma_k}\|X - X^k\|^2 \right\}. \tag{55}$$

Now we calculate the partial quadratic regularization of f in (55), which plays a key role in the study of the partial PPA for solving (46). For a given parameter $\sigma > 0$, the partial quadratic regularization of f in (49) associated with σ is given by

$$F_\sigma(X) \;=\; \min_{u\in\Re^m, Y\in\Re^{p\times q}} \Big\{ f(u, Y) + \frac{1}{2\sigma}\|Y - X\|^2 \Big\}. \tag{56}$$

From (49), we have

$$\begin{aligned} F_\sigma(X) &= \min_{\substack{u\in\Re^m \\ Y\in\Re^{p\times q}}} \sup_{\substack{\zeta\in\Re^m \\ \xi\in\Re^s}} l(u, Y; \zeta, \xi) + \frac{1}{2\sigma}\|Y - X\|^2 = \sup_{\substack{\zeta\in\Re^m \\ \xi\in\Re^s}} \min_{\substack{u\in\Re^m \\ Y\in\Re^{p\times q}}} l(u, Y; \zeta, \xi) + \frac{1}{2\sigma}\|Y - X\|^2 \\ &= \sup_{\substack{\zeta\in\Re^m \\ \xi\in\mathcal{Q}^*}} \min_{\substack{u\in\Re^m \\ Y\in\Re^{p\times q}}} \Big\{ f_\rho(u, Y) + \langle \zeta, b - \mathcal{A}(Y) - u\rangle + \langle \xi, d - \mathcal{B}(Y)\rangle + \frac{1}{2\sigma}\|Y - X\|^2 \Big\}, \end{aligned} \tag{57}$$

where the interchange of $\min_{u,Y}$ and $\sup_{\zeta,\xi}$ follows from the growth properties in (u, Y) [45, Theorem 37.3] and the third equality follows from (48). Thus we have

$$F_\sigma(X) \;=\; \sup_{\zeta\in\Re^m,\ \xi\in\mathcal{Q}^*} \Theta_\sigma(\zeta,\xi;X),$$

where

$$\begin{aligned}
&\Theta_\sigma(\zeta,\xi;X) := \min_{u\in\Re^m, Y\in\Re^{p\times q}} \Big\{ f_\rho(u,Y) + \langle \zeta, b - \mathcal{A}(Y) - u\rangle + \langle \xi, d - \mathcal{B}(Y)\rangle + \frac{1}{2\sigma}\|Y - X\|^2 \Big\} \\
&= \;\; -\frac{1}{2}\|\zeta\|^2 + \langle b,\zeta\rangle + \langle d,\xi\rangle + \min_{Y\in\Re^{p\times q}} \Big\{ \rho\|Y\|_* + \frac{1}{2\sigma}\|Y - W(\zeta,\xi;X)\|^2 \Big\} \\
&\qquad + \frac{1}{2\sigma}\Big(\|X\|^2 - \|W(\zeta,\xi;X)\|^2 \Big) \\
&= \;\; -\frac{1}{2}\|\zeta\|^2 + \langle b,\zeta\rangle + \langle d,\xi\rangle + \frac{1}{2\sigma}\|X\|^2 - \frac{1}{2\sigma}\|D_{\rho\sigma}(W(\zeta,\xi;X))\|^2, \qquad (58)
\end{aligned}$$

where $W(\zeta,\xi;X) = X - \sigma(C - \mathcal{A}^*\zeta - \mathcal{B}^*\xi)$ and the last equality follows from (23). By the saddle point theorem [45, Theorem 28.3] and (57), we have that for any $(\zeta(X), \xi(X))$ such that

$$(\zeta(X), \xi(X)) \in \mathrm{argsup}_{\zeta\in\Re^m,\ \xi\in\mathcal{Q}^*} \Theta_\sigma(\zeta,\xi;X),$$

$(\zeta(X), D_{\rho\sigma}(W(\zeta(X),\xi(X);X)))$ is the unique solution to (56).

Now we formally present the partial PPA for solving (46).

Algorithm 1. Given a tolerance $\varepsilon > 0$, $(u^0, X^0) \in \Re^m \times \Re^{p\times q}$, $\sigma_0 > 0$. Set $k = 0$. Iterate:

Step 1. Compute an approximate maximizer

$$\Re^m \times \mathcal{Q}^* \ni (\zeta^{k+1}, \xi^{k+1}) \;\approx\; \arg\sup_{\zeta\in\Re^m,\, \xi\in\Re^s} \{\theta_{\sigma_k}(\zeta,\xi) := \Theta_{\sigma_k}(\zeta,\xi;X^k) - \delta(\xi \mid \mathcal{Q}^*)\}, \quad (59)$$

where $\Theta_{\sigma_k}(\zeta,\xi;X^k)$ is defined in (58) and $\delta(\cdot|\mathcal{Q}^*)$ is the indicator function over $\mathcal{Q}^*$.

Step 2. Compute $W^{k+1} := W(\zeta^{k+1},\xi^{k+1};X^k)$. Set

$$u^{k+1} = \zeta^{k+1},\ X^{k+1} = D_{\rho\sigma_k}(W^{k+1}), \quad Z^{k+1} = \frac{1}{\sigma_k}(D_{\rho\sigma_k}(W^{k+1}) - W^{k+1}).$$

Step 3. If $\|(X^k - X^{k+1})/\sigma_k\| \le \varepsilon$; stop; else; update σ_k; end.

Suppose that $(\bar\zeta(X^k), \bar\xi(X^k))$ is an optimal solution of the inner subproblem (59) for each X^k. Let P_{σ_k} be defined as in (55). In order to terminate (59) in the above PPA, we introduce the following stopping criteria:

$$\sup \theta_k(\zeta,\xi) - \theta_k(\zeta^{k+1},\xi^{k+1}) \le \frac{\varepsilon_k^2}{4\sigma_k}, \tag{60a}$$

$$\|\zeta^{k+1} - \bar{\zeta}(X^k)\|^2 \le \frac{1}{2}\varepsilon_k^2, \quad \varepsilon_k > 0,\ \sum_{k=0}^{\infty}\varepsilon_k < \infty, \tag{60b}$$

$$\sup \theta_k(\zeta,\xi) - \theta_k(\zeta^{k+1},\xi^{k+1}) \le \frac{\delta_k^2}{2\sigma_k}\|X^{k+1} - X^k\|^2, \tag{60c}$$

$$\|\zeta^{k+1} - \bar{\zeta}(X^k)\|^2 \le \delta_k^2\|\zeta^{k+1} - \zeta^k\|^2, \quad \delta_k > 0,\ \sum_{k=0}^{\infty}\delta_k < \infty, \tag{60d}$$

$$\operatorname{dist}(0, \partial\theta_k(\zeta^{k+1},\xi^{k+1})) \le \frac{\delta_k'}{\sigma_k}\|X^{k+1} - X^k\|, \quad 0 \le \delta_k' \to 0. \tag{60e}$$

Note that $F_{\sigma_k}(X^k) = \sup \theta_k(\zeta,\xi)$ and $\theta_k(\zeta^{k+1},\xi^{k+1}) = \Theta_{\sigma_k}(\zeta^{k+1},\xi^{k+1};X^k)$. The following result reveals the relation between the estimation (60) and (45), which enables us to apply the convergence results of the partial PPA in [28, Theorem 1 & 2] to our partial PPA. The proof essentially follows the idea in [47, Proposition 6].

Proposition 3.2. *Suppose that $(\bar{\zeta}(X^k), \bar{\xi}(X^k))$ is an optimal solution of (59). Let $(\bar{u}^{k+1}, \overline{X}^{k+1}) = (\bar{\zeta}(X^k), D_{\rho\sigma_k}(W(\bar{\zeta}(X^k), \bar{\xi}(X^k); X^k)))$ and $X^{k+1} = D_{\rho\sigma_k}(W(\zeta^{k+1},\xi^{k+1};X^k))$. Then one has*

$$\frac{1}{2\sigma_k}\|X^{k+1} - \overline{X}^{k+1}\|^2 \le \sup \theta_k(\zeta,\xi) - \theta_k(\zeta^{k+1},\xi^{k+1}). \tag{61}$$

Proof. Since $\Theta_\sigma(\zeta,\xi;X)$ is convex in X and $\nabla_X \Theta_{\sigma_k}(\zeta^{k+1},\xi^{k+1};X^k) = (X^k - X^{k+1})/\sigma_k$, the following inequality holds for any $Y \in \Re^{p\times q}$:

$$\begin{aligned}
& \Theta_{\sigma_k}(\zeta^{k+1},\xi^{k+1};X^k) + \langle \sigma_k^{-1}(X^k - X^{k+1}), Y - X^k\rangle \\
\le\ & \Theta_{\sigma_k}(\zeta^{k+1},\xi^{k+1};Y) \ \le\ \sup\left\{\Theta_{\sigma_k}(\zeta,\xi;Y) \mid \zeta \in \Re^m, \xi \in \mathcal{Q}^*\right\} \ =\ F_{\sigma_k}(Y) \\
=\ & \min_{\substack{u\in\Re^m \\ X\in\Re^{p\times q}}} \{f(u,X) + \frac{1}{2\sigma_k}\|X - Y\|^2\} \le f(\bar{u}^{k+1}, \overline{X}^{k+1}) + \frac{1}{2\sigma_k}\|\overline{X}^{k+1} - Y\|^2.
\end{aligned} \tag{62}$$

We also know that

$$\begin{aligned}
\sup \theta_k(\zeta,\xi) &= F_{\sigma_k}(X^k) = \min_{u\in\Re^m, X\in\Re^{p\times q}} \{f(u,X) + \frac{1}{2\sigma_k}\|X - X^k\|^2\} \\
&= f(\bar{u}^{k+1}, \overline{X}^{k+1}) + \frac{1}{2\sigma_k}\|\overline{X}^{k+1} - X^k\|^2,
\end{aligned} \tag{63}$$

which together with (62) and the fact that $\theta_k(\zeta^{k+1},\xi^{k+1}) = \Theta_{\sigma_k}(\zeta^{k+1},\xi^{k+1};X^k)$, implies that

$$\begin{aligned}
& \sup \theta_k(\zeta,\xi) - \theta_k(\zeta^{k+1},\xi^{k+1}) \\
\ge\ & \frac{1}{2\sigma_k}\Big[\|\overline{X}^{k+1} - X^k\|^2 - \|\overline{X}^{k+1} - Y\|^2 - 2\langle X^{k+1} - X^k, Y - X^k\rangle\Big] \\
=\ & \frac{1}{2\sigma_k}\Big[- \|(\overline{X}^{k+1} + X^k - X^{k+1}) - Y\|^2 + \|\overline{X}^{k+1} - X^{k+1}\|^2\Big].
\end{aligned} \tag{64}$$

Since this inequality holds for all $Y \in \Re^{p\times q}$, by taking the maximum of (64) in Y, we have

$$\sup \theta_k(\zeta,\xi) - \theta_k(\zeta^{k+1},\xi^{k+1}) \;\geq\; \frac{1}{2\sigma_k}\|\overline{X}^{k+1} - X^{k+1}\|^2,$$

which proves our assertion. $\square$

3.1 Convergence analysis of the partial PPA

In this subsection, we show the global convergence and local convergence of the partial PPA for solving (46), mainly based upon the convergence results of Ha [28, Theorem 1 & 2], which require the condition (66) in the following proposition. The purpose of this proposition is to give a sufficient condition for (66) to hold.

Proposition 3.3. *Consider the function $f(u,X)$ defined in (49). Suppose that for some $\lambda > 0$, the following parameterized problem perturbed by $(v,Y) \in \Re^m \times \Re^{p\times q}$*

$$\min_{u\in\Re^m, X\in\Re^{p\times q}} \Big\{ f(u,X) - \langle u,\, v\rangle - \langle X,\, Y\rangle \Big\} \tag{65}$$

has an optimal solution whenever $\max\{\|v\|, \|Y\|\} \leq \lambda$. Then we have

$$0 \in \operatorname{int}\operatorname{Im}(\mathcal{T}_f). \tag{66}$$

Proof. Since for each $(v,Y) \in \Re^m \times \Re^{p\times q}$ such that $\max\{\|v\|, \|Y\|\} \leq \lambda$, the parameterized problem (65) has an optimal solution $(\bar{u}, \overline{X})$, we have that

$$0 \in \partial f(\bar{u}, \overline{X}) - (v, Y),$$

which implies that $(v,Y) \in \partial f(\bar{u},\overline{X}) \subseteq \operatorname{Im}(\mathcal{T}_f)$. Therefore, we have $0 \in \operatorname{int}\operatorname{Im}(\mathcal{T}_f)$. $\square$

Remark 3.1. *In many applications, we have $C = 0$ in the objective function $f_\rho(u,X)$ (see the examples in Section 5). In this case, the function $f_\rho(u,X)$ perturbed by $(v,Y) \in \Re^m \times \Re^{p\times q}$ with $\|Y\|_2 < \rho$, is coercive. This is because we have*

$$\begin{aligned} & f_\rho(u,X) - \langle u,\, v\rangle - \langle X,\, Y\rangle = (\frac{1}{2}\|u\|^2 - \langle u,\, v\rangle) + (\rho\|X\|_* - \langle X,\, Y\rangle) \\ \geq\;\; & \frac{1}{2}\|u - v\|^2 - \frac{1}{2}\|v\|^2 + (\rho - \|Y\|_2)\|X\|_*. \end{aligned}$$

Therefore, if $C = 0$, the parameterized problem (65) has an optimal solution for any $(v,Y) \in \Re^m \times \Re^{p\times q}$ such that $\max\{\|v\|, \|Y\|\} \leq \lambda$ for some $0 < \lambda < \rho$.

Theorem 3.1. (**Global convergence**) *Suppose that the condition in Proposition 3.3 is satisfied. Let the partial PPA be executed with the stopping criterion (60a) and (60b). Then the generated sequence $\{(u^k, X^k)\}$ is bounded and converges to an optimal solution $(\bar{u}, \overline{X})$ of (46), and $\{(\zeta^k, \xi^k)\}$ is asymptotically minimizing for the dual problem (50) with*

$$\|C - \mathcal{A}^*(\zeta^{k+1}) - \mathcal{B}^*(\xi^{k+1}) - Z^{k+1}\| = \frac{1}{\sigma_k}\|X^{k+1} - X^k\| \to 0, \tag{67}$$

$$\text{asym } \sup(D) - g_\rho(\zeta^k, \xi^k) \leq \frac{1}{2\sigma_k}\Big[\,\frac{1}{2}\varepsilon_k^2 + \|X^k\|^2 - \|X^{k+1}\|^2\,\Big], \tag{68}$$

where asym sup(D) *is the asymptotic supremum of (50). If the problem* (46) *satisfies the Slater condition (47), then the sequence* $\{(\zeta^k, \xi^k)\}$ *is also bounded, and all of its accumulation points are optimal solutions for the problem (50).*

Proof. Under the given assumption, we have from Proposition 3.3 that $0 \in \operatorname{int} \operatorname{Im}(\mathcal{T}_f)$. Moreover, we know from Proposition 3.2 that (60a) and (60b) implies the general stopping criterion (45a) for $\mathcal{T}_f$. It follows from [28, Theorem 1] that the sequence $\{(u^k, X^k)\}$ is bounded and any of its weak cluster point is an optimal solution to (46) and $X^k \to \overline{X}$. Since $f_\rho(u, X)$ is strongly convex with respect to u, the u-component of the optimal solution is uniquely determined, which implies that $\{u^k\} \to \bar{u}$. Thus the whole sequence $\{(u^k, X^k)\}$ converges to an optimal solution $(\bar{u}, \overline{X})$ of (46). The rest of the proof follows the arguments as in [46, Theorem 4] and we omit it here.

□

Theorem 3.2. (**Local convergence**) *Suppose that the hypotheses in Proposition 3.3 are satisfied. Let the partial PPA be executed with the stopping criterion (60a), (60b), (60c) and (60d). If* $\mathcal{T}_f^{-1}$ *is Lipschitz continuous at the origin with modulus* a_f*, then* $\{(u^k, X^k)\}$ *converges to the unique optimal solution* $(\bar{u}, \overline{X})$ *of* (46)*, and*

$$\|X^{k+1} - \overline{X}\| \leq \eta_k \|X^k - \overline{X}\|, \quad \text{for all } k \text{ sufficiently large}, \tag{69}$$

where

$$\eta_k = [a_f(a_f^2 + \sigma_k^2)^{-1/2} + \delta_k](1 - \delta_k)^{-1} \to \eta_\infty = a_f(a_f^2 + \sigma_\infty^2)^{-1/2} < 1.$$

Moreover, the conclusions of Theorem 3.1 about $\{(\zeta^k, \xi^k)\}$ *are valid.*

If in addition to (60c), (60d) and the condition on $\mathcal{T}_f^{-1}$*, one has (60e) and* $\mathcal{T}_l^{-1}$ *is Lipschitz continuous at the origin with modulus* $a_l\,(\geq a_f)$*, then* $(\zeta^k, \xi^k) \to (\bar{\zeta}, \bar{\xi})$*, where* $(\bar{\zeta}, \bar{\xi})$ *is the unique optimal solution for (50), and one has*

$$\|(\zeta^{k+1}, \xi^{k+1}) - (\bar{\zeta}, \bar{\xi})\| \leq \eta'_k \|X^{k+1} - X^k\|, \quad \text{for all } k \text{ sufficiently large}, \tag{70}$$

where $\eta'_k = a_l(1 + \delta'_k)/\sigma_k \to \eta'_\infty = a_l/\sigma_\infty$.

Proof. Since it follows from Proposition 3.2 that (60c) and (60d) implies the general stopping criterion (45b) and (45c), we can easily obtain the first part of the theorem from Theorem 3.1 and the general results in [28, Theorem 2]. The second part of the theorem can similarly be obtained by following the argument in [47, Theorem 5]. We omit it here. □

4 An inexact smoothing Newton method for solving inner subproblems

In this section, we design an inexact smoothing Newton method for solving the inner subproblem (59), which is the most expensive step in each iteration of the partial PPA. Here we assume that there are inequality constraints in (46). If there are only equality constraints, then (59) is an unconstrained problem, and a semismooth Newton-CG method similar to the one proposed in [58] can be designed to solve (59); we refer the reader to [31] for the details.

For later convenience, we let

$$\widehat{\mathcal{A}} = \begin{pmatrix} \mathcal{A} \\ \mathcal{B} \end{pmatrix},\ \hat{b} = (b; d) \in \Re^{m+s},\ \mathcal{K} = \Re^m \times \mathcal{Q}^* \subseteq \Re^m \times \Re^s, \text{ and } y = (\zeta; \xi) \in \mathcal{K}. \tag{71}$$

In our partial PPA, for some fixed $X \in \Re^{p\times q}$ and $\sigma > 0$, we need to solve the following form of the inner subproblem:

$$\min_{y\in\mathcal{K}} \Big\{ \varphi(y) := \frac{1}{2}\langle y, Ty\rangle + \frac{1}{2\sigma}\|D_{\rho\sigma}(W(y;X))\|^2 - \langle \hat{b}, y\rangle - \frac{1}{2\sigma}\|X\|^2 \Big\}, \tag{72}$$

where $T = [I_m, 0; 0, 0] \in \Re^{(m+s)\times(m+s)}, W(y;X) = X - \sigma(C - \widehat{\mathcal{A}}^* y)$ and $\widehat{\mathcal{A}}^* = (\mathcal{A}^*,\, \mathcal{B}^*)$ is the adjoint of $\widehat{\mathcal{A}}$. Note that $-\varphi(\cdot)$ is the objective function of the inner subproblem (59). The function $\varphi(\cdot)$ in (72) is continuously differentiable with

$$\nabla\varphi(y) = Ty + \widehat{\mathcal{A}} D_{\rho\sigma}(W(y;X)) - \widehat{b}, \quad y \in \Re^{m+s}.$$

Since $\varphi(\cdot)$ is a convex function, $\bar{y} = (\bar{\zeta}; \bar{\xi}) \in \mathcal{K}$ solves problem (72) if and only if it satisfies the following variational inequality

$$\langle y - \bar{y}, \nabla\varphi(\bar{y})\rangle \geq 0 \quad \forall y \in \mathcal{K}. \tag{73}$$

Define $F : \Re^{m+s} \to \Re^{m+s}$ by

$$F(y) \ := \ y - \Pi_{\mathcal{K}}(y - \nabla\varphi(y)), \quad y \in \Re^{m+s}. \tag{74}$$

Then one can easily prove that $\bar{y} \in \mathcal{K}$ solves (73) if and only if $F(\bar{y}) = 0$ [16]. Thus, solving the inner problem (72) is equivalent to solving the following equation

$$F(y) = 0, \quad y \in \Re^{m+s}. \tag{75}$$

Since both $\Pi_{\mathcal{K}}(\cdot)$ and $D_{\rho\sigma}(\cdot)$ are globally Lipschitz continuous, F is globally Lipschitz continuous. For the purpose of introducing an inexact smoothing Newton method, we need to define a smoothing function for $F(\cdot)$.

The smoothing function for the soft thresholding operator $D_{\rho\sigma}(\cdot)$ has been defined by (27) where the threshold value is equal to $\rho\sigma$. Next, we need to define the smoothing function for $\Pi_{\mathcal{K}}(\cdot)$. For simplicity, we shall use the Huber smoothing function χ defined in (24). Let $\pi : \Re \times \Re^{m+s} \to \Re^{m+s}$ be defined by

$$\pi_i(\varepsilon, z) = \left\{ \begin{array}{ll} z_i & \text{if } 1 \leq i \leq m + s_1 \\ \chi(\varepsilon, z_i) & \text{if } m + s_1 + 1 \leq i \leq m + s \end{array} \right. , \quad (\varepsilon, z) \in \Re \times \Re^{m+s}. \tag{76}$$

The function π is obviously continuously differentiable around any $(\varepsilon, z) \in \Re \times \Re^{m+s}$ as long as $\varepsilon \neq 0$ and is strongly semismooth everywhere.

Now, we are ready to define a smoothing function for $F(\cdot)$. Let

$$\Upsilon(\varepsilon, y) := y - \pi(\varepsilon, y - (Ty + \widehat{\mathcal{A}}\overline{D}_{\rho\sigma}(\varepsilon, W(y;X)) - \hat{b})), \quad (\varepsilon, y) \in \Re \times \Re^{m+s}. \tag{77}$$

From the definitions of Υ, π, and $\overline{D}_{\rho\sigma}$, we have that $F(y) = \Upsilon(0, y)$ for any $y \in \Re^{m+s}$.

Proposition 4.1. *The mapping Υ be defined by (77) has the following properties:*

(i) Υ *is globally Lipschitz continuous on* $\Re \times \Re^{m+s}$.

(ii) Υ *is continuously differentiable around* (ε, y) *when* $\varepsilon \neq 0$. *For any fixed* $\varepsilon \in \Re$, $\Upsilon(\varepsilon, \cdot)$ *is a* P_0*-function, i.e., for any* $(y, z) \in \Re^{m+s} \times \Re^{m+s}$ *with* $y \neq z$,

$$\max_{y_i \neq z_i} (y_i - z_i)(\Upsilon_i(\varepsilon, y) - \Upsilon_i(\varepsilon, z)) \geq 0, \tag{78}$$

and thus for any fixed $\varepsilon \neq 0$, $\Upsilon'_y(\varepsilon, y)$ *is a* P_0*-matrix (i.e., all its principal minors are nonnegative).*

(iii) Υ *is strongly semismooth at* $(0, y)$. *In particular, for any* $\varepsilon \downarrow 0$ *and* $\Re^{m+s} \ni h \to 0$ *we have*

$$\Upsilon(\varepsilon, y + h) - \Upsilon(0, y) - \Upsilon'(\varepsilon, y + h)(\varepsilon, h) = O(\|(\varepsilon, h)\|^2).$$

(iv) For any $h \in \Re^{m+s}$,

$$\partial \Upsilon(0, y)(0, h) \subseteq h - \partial \pi(0, y - \nabla \varphi(y))(0, h - (Th + \sigma \widehat{\mathcal{A}} \partial \overline{D}_{\rho\sigma}(0, W(y; X))(0, \widehat{\mathcal{A}}^* h))).$$

Proof. (i) Since both π and $\overline{D}_{\rho\sigma}$ are globally Lipschitz continuous, Υ is also globally Lipschitz continuous.

(ii) By the definition of π and $\overline{D}_{\rho\sigma}$, we know that Υ is continuously differentiable around $(\varepsilon, y) \in \Re \times \Re^{m+s}$ when $\varepsilon \neq 0$. From part (i), we know Υ is continuous on $\Re \times \Re^{m+s}$, it is enough to show that for any $0 \neq \varepsilon \in \Re$, $\Upsilon(\varepsilon, \cdot)$ is a P_0-function. For any fixed $\varepsilon \neq 0$. Define $g_\varepsilon : \Re^{m+s} \to \Re^{m+s}$ by

$$g_\varepsilon(y) = Ty + \widehat{\mathcal{A}} \overline{D}_{\rho\sigma}(\varepsilon, W(y; X)) - \hat{b}, \quad y \in \Re^{m+s}.$$

Then g_ε is continuously differentiable on $\Re^{m+s}$. For any $h \in \Re^{m+s}$, we have

$$\langle h, (g_\varepsilon)'(y) h \rangle = \langle h, Th \rangle + \sigma \langle h, \widehat{\mathcal{A}} (\overline{D}_{\rho\sigma})'_W(\varepsilon, W) \widehat{\mathcal{A}}^* h \rangle = \langle h, Th \rangle + \sigma \langle \widehat{\mathcal{A}}^* h, (\overline{D}_{\rho\sigma})'_W(\varepsilon, W) \widehat{\mathcal{A}}^* h \rangle \geq 0,$$

which implies that g_ε is a P_0-function on $\Re^{m+s}$. Let $(y, z) \in \Re^{m+s} \times \Re^{m+s}$ with $y \neq z$. Then there exists $i \in \{1, \ldots, m+s\}$ with $y_i \neq z_i$ such that

$$(y_i - z_i)((g_\varepsilon)_i(y) - (g_\varepsilon)_i(z)) \geq 0.$$

By noting that for any $h \in \Re^{m+s}$, $(\chi)'_{h_j}(\varepsilon, h_j) \in [0, 1], j = 1, \ldots, m+s$, we have

$$(y_i - z_i)(\Upsilon_i(\varepsilon, y) - \Upsilon_i(\varepsilon, z)) \geq 0.$$

This shows that (78) holds. Hence, $\Upsilon'_y(\varepsilon, y)$ is P_0-matrix for any fixed $\varepsilon \neq 0$.

(iii) Since the composition of strongly semismooth functions is still strongly semismooth [21], Υ is strongly semismooth at $(0, y)$.

(iv) Let $\mathcal{N} = \{0\} \times \Re^{m+s}$ be a set in $\Re \times \Re^{m+s}$ with Lebesgue measure zero, and

$$\partial_{\mathcal{N}} \Upsilon(0, y) := \Big\{ \lim_{k \to \infty} \Upsilon'(\varepsilon^k, y^k) \; : \; (\varepsilon^k, y^k) \to (0, y), \; \varepsilon^k \neq 0 \Big\}.$$

Then we have $\partial\Upsilon(0,y) = \mathrm{conv}(\partial_{\mathcal{N}}\Upsilon(0,y))$. It is enough to show that the inclusion is true where the term on the left-hand side is $\partial_{\mathcal{N}}\Upsilon(0,y)(0,h)$. Since both π and $\overline{D}_{\rho\sigma}$ are directionally differentiable, for any $(\varepsilon \neq 0, y') \in \Re \times \Re^{m+s}$,

$$\Upsilon'(\varepsilon, y')(0,h) = h - \pi'\Big((\varepsilon, z'); (0, h - (Th + \sigma\widehat{\mathcal{A}}\overline{D}'_{\rho\sigma}((\varepsilon, W); (0, \widehat{\mathcal{A}}^* h))))\Big),$$

where $z' = y' - (Ty' + \widehat{\mathcal{A}}\overline{D}_{\rho\sigma}(\varepsilon, W) - \hat{b})$, from which we have

$$\Upsilon'(\varepsilon, y')(0,h) \in h - \partial\pi(\varepsilon, z')(0, h - (Th + \sigma\widehat{\mathcal{A}}\partial\overline{D}_{\rho\sigma}(\varepsilon, W)(0, \widehat{\mathcal{A}}^* h))).$$

By letting $(\varepsilon, y') \to (0, y)$ in the above inclusion, we obtain the required result. $\square$

Now we are ready to introduce the inexact smoothing Newton method, which was developed by Gao and Sun in [23], for solving the nonsmooth equation of the form (75). Let $\kappa \in (0, \infty)$ be a constant. Define $G : \Re \times \Re^{m+s} \to \Re^{m+s}$ by

$$G(\varepsilon, y) := \Upsilon(\varepsilon, y) + \kappa|\varepsilon| y, \quad (\varepsilon, y) \in \Re \times \Re^{m+s}, \tag{79}$$

where $\Upsilon : \Re \times \Re^{m+s} \to \Re^{m+s}$ is defined by (77). Note that $G'_y(\varepsilon, y)$ is a P-matrix (i.e., all its principal minors are positive) for any $(\varepsilon \neq 0, y) \in \Re \times \Re^{m+s}$, and hence it is nonsingular, while by part *(ii)* of Proposition 4.1, $\Upsilon'_y(\varepsilon, y)$ is only a P_0-matrix which may be singular. Define $E : \Re \times \Re^{m+s} \to \Re \times \Re^{m+s}$ by

$$E(\varepsilon, y) := \begin{bmatrix} \varepsilon \\ G(\varepsilon, y) \end{bmatrix} = \begin{bmatrix} \varepsilon \\ \Upsilon(\varepsilon, y) + \kappa|\varepsilon| y \end{bmatrix}, \quad (\varepsilon, y) \in \Re \times \Re^{m+s}.$$

For any $(\varepsilon \neq 0, y) \in \Re^{m+s}$, $E'(\varepsilon, y)$ is a P-matrix, and hence it is nonsingular. Then solving the nonsmooth equation $F(y) = 0$ is equivalent to solving the following smoothed equation

$$E(\varepsilon, y) = 0.$$

The inexact smoothing Newton method can be described as follows.

Algorithm 2: An inexact smoothing Newton method.

Step 0. Choose $r \in (0,1)$. Let $\hat{\varepsilon} \in (0,\infty)$ and $\eta \in (0,1)$ be such that $\delta := \sqrt{2}\max\{r\hat{\varepsilon}, \eta\} < 1$. Choose constants $\ell \in (0,1), \sigma \in (0,1/2), \tau \in (0,1)$, and $\hat{\tau} \in [1,\infty)$. Let $\varepsilon^0 := \hat{\varepsilon}$ and $y^0 \in \Re^{m+s}$ be an arbitrary starting point. Set $k := 0$.

Step 1. If $E(\varepsilon^k, y^k) = 0$, then stop. Otherwise, compute

$$\varsigma_k := r\min\{1, \|E(\varepsilon^k, y^k)\|^2\} \quad \text{and} \quad \eta_k := \min\{\tau, \hat{\tau}\|E(\varepsilon^k, y^k)\|\}.$$

Step 2. Solve the following equation

$$E(\varepsilon^k, y^k) + E'(\varepsilon^k, y^k)\begin{bmatrix} \Delta\varepsilon^k \\ \Delta y^k \end{bmatrix} = \begin{bmatrix} \varsigma_k \hat{\varepsilon} \\ 0 \end{bmatrix} \tag{80}$$

approximately such that

$$\|R_k\| \le \min\{\eta_k \|G(\varepsilon^k, y^k) + G'_\varepsilon(\varepsilon^k, y^k)\Delta\varepsilon^k\|, \eta\|E(\varepsilon^k, y^k)\|\}, \tag{81}$$

where $\Delta\varepsilon^k := -\varepsilon^k + \varsigma_k\hat{\varepsilon}$ and $R_k := G(\varepsilon^k, y^k) + G'(\varepsilon^k, y^k)\begin{bmatrix} \Delta\varepsilon^k \\ \Delta y^k \end{bmatrix}$.

Step 3. Let m_k be the smallest nonnegative integer m satisfying

$$\|E(\varepsilon^k + \ell^m \Delta\varepsilon^k, y^k + \ell^m \Delta y^k)\|^2 \le \Big[1 - 2\sigma(1-\delta)\ell^m\Big]\|E(\varepsilon^k, y^k)\|^2.$$

Set $(\varepsilon^{k+1}, y^{k+1}) = (\varepsilon^k + \ell^{m_k}\Delta\varepsilon^k, y^k + \ell^{m_k}\Delta y^k)$.

Step 4. Replace k by $k+1$ and go to **Step 1**.

Let

$$\mathcal{N} := \{(\varepsilon, y) \in \Re \times \Re^{m+s} \mid \varepsilon \ge \varsigma(\varepsilon, y)\hat{\varepsilon}\},$$

where $\varsigma(\varepsilon, y) = r\min\{1, \|E(\varepsilon, y)\|^2\}$. From [23, Theorem 4.1 & Theorem 3.6], we have the following convergence results for the inexact smoothing Newton method. For more details on the inexact smoothing Newton method, see [23].

Theorem 4.1. *Algorithm 2 is well defined and generates an infinite sequence* $\{(\varepsilon^k, y^k)\} \in \mathcal{N}$ *with the properties that any accumulation point* $(\bar{\varepsilon}, \bar{y})$ *of* $\{(\varepsilon^k, y^k)\}$ *is a solution of* $E(\varepsilon, y) = 0$ *and* $\lim_{k\to\infty}\|E(\varepsilon^k, y^k)\|^2 = 0$. *Additionally, if the Slater condition (47) holds, then* $\{(\varepsilon^k, y^k)\}$ *is bounded.*

Theorem 4.2. *Let* $(\bar{\varepsilon}, \bar{y})$ *be an accumulation point of the infinite sequence* $\{(\varepsilon^k, y^k)\}$ *generated by Algorithm 2. Suppose that* E *is strongly semismooth at* $(\bar{\varepsilon}, \bar{y})$ *and that all* $\mathcal{V} \in \partial E(\bar{\varepsilon}, \bar{y})$ *are nonsingular. Then the whole sequence* $\{(\varepsilon^k, y^k)\}$ *converges to* $(\bar{\varepsilon}, \bar{y})$ *quadratically, i.e.,*

$$\|(\varepsilon^{k+1} - \bar{\varepsilon}, y^{k+1} - \bar{y})\| = O(\|(\varepsilon^k - \bar{\varepsilon}, y^k - \bar{y})\|^2).$$

Suppose that the Slater condition (47) holds. Let $(\bar{\varepsilon}, \bar{y})$ be an accumulation point of the sequence $\{(\varepsilon^k, y^k)\}$ generated by Algorithm 2. Then, we know that $\bar{\varepsilon} = 0$ and $F(\bar{y}) = 0$, which means that $\bar{y} = (\bar{\zeta}; \bar{\xi}) \in \mathcal{K}$ is an optimal solution to the inner subproblem (72). Let $\overline{X} := D_{\rho\sigma}(W(\bar{y}; X))$. Then $(\bar{\zeta}, \overline{X})$ is the unique optimal solution to the problem (56).

For the quadratic convergence of Algorithm 2, we need the concept of constraint nondegeneracy. For a given closed set $K \subseteq \mathcal{X}$, we let $T_K(x)$ be the tangent cone of K at $x \in K$ as in convex analysis [45]. The largest linear space contained in $T_K(x)$ is denoted by $\mathrm{lin}(T_K(x))$, which is equal to $(-T_K(x)) \cap T_K(x)$. Define $g : \Re^{p\times q} \to \Re$ by $g(X) = \|X\|_*$. Let $K_{p,q}$ be the epigraph of g, i.e.,

$$K_{p,q} \;:=\; \mathrm{epi}(g) = \{(X,t) \in \Re^{p\times q} \times \Re \mid g(X) \le t\},$$

which is a closed convex cone. Let $\widehat{\mathcal{B}} := (\mathcal{B},\, 0)$. Then (1) can be rewritten in the following form:

$$\min\Big\{\frac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho\, t + \langle C, X\rangle \;:\; \widehat{\mathcal{B}}(X,t) \in d + \mathcal{Q},\ (X,t) \in K_{p,q}\Big\}. \tag{82}$$

It is easy to see that $\overline{X}$ is an optimal solution for (1) if and only if $(\overline{X}, \bar{t})$ is an optimal solution to (82) with $\bar{t} = \|\overline{X}\|_*$. Let $\mathcal{I}$ be the identity map from $\Re^{p\times q} \times \Re$ to $\Re^{p\times q} \times \Re$. Then the constraint nondegeneracy condition is said to hold at $(\overline{X}, \bar{t})$ if

$$\begin{pmatrix} \widehat{\mathcal{B}} \\ \mathcal{I} \end{pmatrix} (\Re^{p\times q} \times \Re) + \begin{pmatrix} \mathrm{lin}\big(T_{\mathcal{Q}}(\widehat{\mathcal{B}}(\overline{X}, \bar{t}) - d)\big) \\ \mathrm{lin}(T_{K_{p,q}}(\overline{X}, \bar{t})) \end{pmatrix} = \begin{pmatrix} \Re^s \\ \Re^{p\times q} \times \Re \end{pmatrix}. \tag{83}$$

Note that $\mathrm{lin}(T_{\mathcal{Q}}(\widehat{\mathcal{B}}(\overline{X}, \bar{t}) - d)) = \mathrm{lin}(T_{\mathcal{Q}}(\mathcal{B}(\overline{X}) - d))$. Let $\mathcal{E}(\overline{X})$ be the index set of active constraints at $\overline{X}$:

$$\mathcal{E}(\overline{X}) := \{i \mid \langle \mathcal{B}_i, \overline{X}\rangle = d_i,\ i = s_1 + 1, \ldots, s\},$$

and $l := |\mathcal{E}(\overline{X})|$. Without loss of generality, we assume that

$$\mathcal{E}(\overline{X}) := \{s_1 + 1, \ldots, s_1 + l\}.$$

Define $\widetilde{\mathcal{B}} : \Re^{p\times q} \to \Re^{s_1 + l}$ by

$$\widetilde{\mathcal{B}}(X) := \big[\langle \mathcal{B}_1, X\rangle, \ldots, \langle \mathcal{B}_{s_1+l}, X\rangle\big]^T, \quad X \in \Re^{p\times q}.$$

Let $\overline{\mathcal{B}} = (\widetilde{\mathcal{B}},\, 0)$. Since $\mathrm{lin}(T_{\mathcal{Q}}(\mathcal{B}(X) - d))$ can be computed directly as follows

$$\mathrm{lin}(T_{\mathcal{Q}}(\mathcal{B}(\overline{X}) - d)) = \{h \in \Re^s \mid h_i = 0, i \in \{1, \ldots, s_1\} \cup \mathcal{E}(\overline{X})\},$$

(83) can be reduced to

$$\begin{pmatrix} \overline{\mathcal{B}} \\ \mathcal{I} \end{pmatrix} (\Re^{p\times q} \times \Re) + \begin{pmatrix} \{0\}^{s_1+l} \\ \mathrm{lin}(T_{K_{p,q}}(\overline{X}, \bar{t})) \end{pmatrix} = \begin{pmatrix} \Re^{s_1+l} \\ \Re^{p\times q} \times \Re \end{pmatrix},$$

which is equivalent to

$$\overline{\mathcal{B}}\Big(\mathrm{lin}(T_{K_{p,q}}(\overline{X}, \bar{t}))\Big) = \Re^{s_1+l}. \tag{84}$$

Next, we shall characterize the linear space $\text{lin}(T_{K_{p,q}}(\overline{X}, g(\overline{X})))$. Let $W(\bar{y}; X)$ admit the SVD as in (10). Decompose the index set $\alpha = \{1, \ldots, p\}$ into the following three subsets:

$$\alpha_1 := \{i \,|\, \sigma_i(W) > \rho\sigma,\ i \in \alpha\},\ \alpha_2 := \{i \,|\, \sigma_i(W) = \rho\sigma,\ i \in \alpha\},\ \alpha_3 := \{i \,|\, \sigma_i(W) < \rho\sigma,\ i \in \alpha\}.$$

Then $U = [U_{\alpha_1}\ U_{\alpha_2}\ U_{\alpha_3}], V = [V_{\alpha_1}\ V_{\alpha_2}\ V_{\alpha_3}\ V_2]$, and $\overline{X} = D_{\rho\sigma}(W(\bar{y}; X))$ is of rank $|\alpha_1|$. For any $H \in \Re^{p\times q}$, by the results of Watson [55, Theorem 1], we have

$$g'(\overline{X}; H) = \begin{cases} \|H\|_* & \text{if } |\alpha_1| = 0, \\ \langle UV_1^T, H\rangle & \text{if } |\alpha_1| = p, \\ \langle U_{\alpha_1}V_{\alpha_1}^T, H\rangle + \|[U_{\alpha_2}\ U_{\alpha_3}]^T H[V_{\alpha_2}\ V_{\alpha_3}\ V_2]\|_* & \text{if } 0 < |\alpha_1| < p. \end{cases}$$

From [13, Proposition 2.3.6 & Theorem 2.4.9], we have

$$T_{K_{p,q}}(\overline{X}, g(\overline{X})) = \text{epi}(g'(\overline{X}; \cdot)),$$

from which we can readily get

$$T_{K_{p,q}}(\overline{X}, g(\overline{X})) = \{(H, t) \in \Re^{p\times q} \times \Re \mid \langle U_{\alpha_1}V_{\alpha_1}^T, H\rangle + \|[U_{\alpha_2}\ U_{\alpha_3}]^T H[V_{\alpha_2}\ V_{\alpha_3}\ V_2]\|_* \le t\}.$$

Thus its linearity space is as follows:

$$\text{lin}(T_{K_{p,q}}(\overline{X}, g(\overline{X}))) = \{(H, t) \in \Re^{p\times q} \times \Re \,|\, [U_{\alpha_2}\ U_{\alpha_3}]^T H[V_{\alpha_2}\ V_{\alpha_3}\ V_2] = 0,\ t = \langle U_{\alpha_1}V_{\alpha_1}^T, H\rangle\}.$$

Let

$$\mathcal{T}(\overline{X}) := \{H \in \Re^{p\times q} \mid [U_{\alpha_2}\ U_{\alpha_3}]^T H[V_{\alpha_2}\ V_{\alpha_3}\ V_2] = 0\},$$

which is a subspace of $\Re^{p\times q}$. The orthogonal complement of $\mathcal{T}(\overline{X})$ is given by

$$\mathcal{T}(\overline{X})^\perp = \{H \in \Re^{p\times q} \mid U_{\alpha_1}^T H = 0,\ HV_{\alpha_1} = 0\}.$$

Since $\overline{\mathcal{B}} = (\widetilde{\mathcal{B}}, 0)$, the constraint nondegeneracy condition (84) can be further reduced to

$$\widetilde{\mathcal{B}}(\mathcal{T}(\overline{X})) = \Re^{s_1+l}. \tag{85}$$

Lemma 4.1. *Let $W(\bar{y}; X) = X - \sigma(C - \widehat{\mathcal{A}}^*\bar{y})$ admit the SVD as in (10). Then the constraint nondegeneracy condition (85) holds at $\overline{X} = D_{\rho\sigma}(W(\bar{y}; X))$ if and only if for any $h \in \Re^{s_1+l}$,*

$$U_{\alpha_1}^T(\widetilde{\mathcal{B}}^*h) = 0 \quad \text{and } (\widetilde{\mathcal{B}}^*h)V_{\alpha_1} = 0 \implies h = 0. \tag{86}$$

Proof. "$\Longrightarrow$" Suppose $h \in \Re^{s_1+l}$ satisfies $U_{\alpha_1}^T(\widetilde{\mathcal{B}}^*h) = 0$ and $(\widetilde{\mathcal{B}}^*h)V_{\alpha_1} = 0$. Since the constraint nondegeneracy condition (85) holds at $\overline{X} = D_{\rho\sigma}(W(\bar{y}; X))$, there exist $Z \in \mathcal{T}(\overline{X})$ such that $h = \widetilde{\mathcal{B}}(Z)$. Let $\bar{\alpha}_1 = \alpha_2 \cup \alpha_3$. Then we have

$$\begin{aligned} &\langle h, h\rangle = \langle h, \widetilde{\mathcal{B}}(Z)\rangle = \langle \widetilde{\mathcal{B}}^*h, Z\rangle = \langle U^T(\widetilde{\mathcal{B}}^*h)V, U^TZV\rangle \\ &= \left\langle \begin{bmatrix} 0 & 0 & 0 \\ 0 & U_{\bar{\alpha}_1}^T(\widetilde{\mathcal{B}}^*h)V_{\bar{\alpha}_1} & U_{\bar{\alpha}_1}^T(\widetilde{\mathcal{B}}^*h)V_2 \end{bmatrix}, \begin{bmatrix} U_{\alpha_1}^TZV_{\alpha_1} & U_{\alpha_1}^TZV_{\bar{\alpha}_1} & U_{\alpha_1}^TZV_2 \\ U_{\bar{\alpha}_1}^TZV_{\alpha_1} & 0 & 0 \end{bmatrix} \right\rangle = 0, \end{aligned}$$

which means $h = 0$.

"$\Longleftarrow$" If the constraint nondegeneracy condition (85) does not hold at $\overline{X}$, then there exists a non-zero $h \in [\widetilde{\mathcal{B}}(\mathcal{T}(\overline{X}))]^{\perp}$. And we have

$$0 = \langle h, \widetilde{\mathcal{B}}(Z)\rangle = \langle \widetilde{\mathcal{B}}^* h, Z\rangle \quad \forall\ Z \in \mathcal{T}(\overline{X}),$$

which means $\widetilde{\mathcal{B}}^* h \in \mathcal{T}(\overline{X})^{\perp}$. Thus, $U_{\alpha_1}^T(\widetilde{\mathcal{B}}^* h) = 0,\ (\widetilde{\mathcal{B}}^* h)V_{\alpha_1} = 0$, from which we must have $h = 0$. This contradiction shows that the constraint nondegeneracy condition (85) holds at $\overline{X}$. □

Lemma 4.2. *Let $\widetilde{\mathcal{A}} = (\mathcal{A}; \widetilde{\mathcal{B}})$ and $\widetilde{\mathcal{A}}^* = (\mathcal{A}^*,\ \widetilde{\mathcal{B}}^*)$ be the adjoint of $\widetilde{\mathcal{A}}$. Let $\overline{D}_{\rho\sigma} : \Re \times \Re^{p\times q} \to \Re^{p\times q}$ be defined by (27). Assume that the constraint nondegeneracy condition (85) holds at $\overline{X}$. Then for any $\mathcal{V} \in \partial \overline{D}_{\rho\sigma}(0, W(\bar{y}; X))$, we have*

$$\langle h, \widetilde{T}h + \sigma\widetilde{\mathcal{A}}\mathcal{V}(0, \widetilde{\mathcal{A}}^* h)\rangle > 0 \quad \forall\ 0 \neq h \in \Re^{m+s_1+l}, \tag{87}$$

where $\widetilde{T} = [I_m, 0; 0, 0]$ is a matrix of size $m + s_1 + l$.

Proof. For any $0 \neq h = (h_1; h_2) \in \Re^{m+s_1+l}$, where $h_1 \in \Re^m$ and $h_2 \in \Re^{s_1+l}$, we have

$$\langle h, \widetilde{T}h + \sigma\widetilde{\mathcal{A}}\mathcal{V}(0, \widetilde{\mathcal{A}}^* h)\rangle = \|h_1\|^2 + \sigma\langle h, \widetilde{\mathcal{A}}\mathcal{V}(0, \widetilde{\mathcal{A}}^* h)\rangle = \|h_1\|^2 + \sigma\langle \widetilde{\mathcal{A}}^* h, \mathcal{V}(0, \widetilde{\mathcal{A}}^* h)\rangle.$$

If $h_1 \neq 0$, since $\langle \widetilde{\mathcal{A}}^* h, \mathcal{V}(0, \widetilde{\mathcal{A}}^* h)\rangle \geq 0$, we have

$$\langle h, \widetilde{T}h + \sigma\widetilde{\mathcal{A}}\mathcal{V}(0, \widetilde{\mathcal{A}}^* h)\rangle > 0.$$

In the following proof, we assume $h_1 = 0$. Consider $0 \neq h = (0; h_2) \in \Re^{m+s_1+l}$, we have $\widetilde{\mathcal{A}}^* h = \widetilde{\mathcal{B}}^* h_2$ and $\langle h, \widetilde{T}h + \sigma\widetilde{\mathcal{A}}\mathcal{V}(0, \widetilde{\mathcal{A}}^* h)\rangle = \sigma\langle h_2, \widetilde{\mathcal{B}}\mathcal{V}(0, \widetilde{\mathcal{B}}^* h_2)\rangle \geq 0$. Suppose that there exists $0 \neq h_2 \in \Re^{s_1+l}$ such that

$$\langle h_2, \widetilde{\mathcal{B}}\mathcal{V}(0, \widetilde{\mathcal{B}}^* h_2)\rangle = 0.$$

Let $H = \widetilde{\mathcal{B}}^* h_2, H_1 = U^T H V_1, H_2 = U^T H V_2$, and $H_1^s = \frac{1}{2}(H_1 + H_1^T),\ H_1^a = \frac{1}{2}(H_1 - H_1^T)$. Then we have

$$0 = \langle H, \mathcal{V}(0, H)\rangle = \frac{1}{2}\langle \Xi(H),\ \Xi(\mathcal{V}(0, H))\rangle = \frac{1}{2}\Big\langle Q^T\Xi(H)Q,\ Q^T\Xi(\mathcal{V}(0, H))Q\Big\rangle.$$

From (41), (42) and Proposition 2.3, we know that there exists $\mathcal{V}_{|\alpha_2|} \in \partial(\overline{D}_{\rho\sigma})_{|\alpha_2|}(0, \rho\sigma I_{|\alpha_2|})$ such that

$$Q^T\Xi(\mathcal{V}(0, H))Q = Q^T \begin{bmatrix} 0 & \mathcal{V}(0, H) \\ (\mathcal{V}(0, H))^T & 0 \end{bmatrix} Q = \begin{bmatrix} M_{\alpha\alpha} & M_{\alpha\gamma} & M_{\alpha\beta} \\ M_{\alpha\gamma}^T & M_{\gamma\gamma} & M_{\gamma\beta} \\ M_{\alpha\beta}^T & M_{\gamma\beta}^T & M_{\beta\beta} \end{bmatrix},$$

where

$$M_{\alpha\alpha} = \begin{bmatrix} (H_1^s)_{\alpha_1\alpha_1} & (H_1^s)_{\alpha_1\alpha_2} & \Omega_{\alpha_1\alpha_3} \circ (H_1^s)_{\alpha_1\alpha_3} \\ (H_1^s)_{\alpha_1\alpha_2}^T & \mathcal{V}_{|\alpha_2|}(0, (H_1^s)_{\alpha_2\alpha_2}) & 0 \\ \Omega_{\alpha_1\alpha_3}^T \circ (H_1^s)_{\alpha_1\alpha_3}^T & 0 & 0 \end{bmatrix},$$

$$M_{\alpha\gamma} = \Gamma_{\alpha\gamma} \circ (-H_1^a),\ M_{\alpha\beta} = \Gamma_{\alpha\beta} \circ (\frac{1}{\sqrt{2}} H_2),\ M_{\gamma\gamma} = -M_{\alpha\alpha},\ M_{\gamma\beta} = M_{\alpha\beta},\ M_{\beta\beta} = 0,$$

and $\Omega_{\alpha_1\alpha_3}, \Gamma_{\alpha\gamma}, \Gamma_{\alpha\beta}$ have the forms as in (36), (37), (38), respectively. Since $Q^T\Xi(H)Q$ has the form in (30), we have

$$\langle H, \mathcal{V}(0, H)\rangle = \langle H_1^s, M_{\alpha\alpha}\rangle + \langle H_1^a, \Gamma_{\alpha\gamma} \circ H_1^a\rangle + \langle H_2, \Gamma_{\alpha\beta} \circ H_2\rangle.$$

Since $\langle (H_1^s)_{\alpha_2\alpha_2}, \mathcal{V}_{|\alpha_2|}(0, (H_1^s)_{\alpha_2\alpha_2})\rangle \geq 0$, we obtain from $\langle H, \mathcal{V}(0, H)\rangle = 0$ that

$$(H_1^s)_{\alpha_1\alpha} = 0,\ (H_1^s)_{\alpha\alpha_1} = 0,\ (H_1^a)_{\alpha_1\alpha} = 0,\ (H_1^a)_{\alpha\alpha_1} = 0,\ (H_2)_{\alpha_1\bar{\beta}} = 0,$$

where $\bar{\beta} = \{1, \ldots, q-p\}$. Since $H_1 = H_1^s + H_1^a$, we have that $(H_1)_{\alpha_1\alpha} = 0$ and $(H_1)_{\alpha\alpha_1} = 0$. Now $H_1 = [U_{\alpha_1}\, U_{\alpha_2}\, U_{\alpha_3}]^T H [V_{\alpha_1}\, V_{\alpha_2}\, V_{\alpha_3}]$ and $H_2 = [U_{\alpha_1}\, U_{\alpha_2}\, U_{\alpha_3}]^T H V_2$, we obtain that

$$U_{\alpha_1}^T H V_1 = 0, \quad U_{\alpha_1}^T H V_2 = 0, \quad U^T H V_{\alpha_1} = 0.$$

Since both U and $V = [V_1\ V_2]$ are orthogonal matrices, we have $U_{\alpha_1}^T H = 0$, $HV_{\alpha_1} = 0$, which means that

$$U_{\alpha_1}^T (\widetilde{\mathcal{B}}^* h_2) = 0 \quad \text{and } (\widetilde{\mathcal{B}}^* h_2) V_{\alpha_1} = 0.$$

Because the constraint nondegeneracy condition (85) holds at $\overline{X}$, we obtain from Lemma 4.1 that $h_2 = 0$, which contradicts the assumption that $h_2 \neq 0$. This contradiction shows that for any $\mathcal{V} \in \partial \overline{D}_{\rho\sigma}(0, W(\bar{y}; X))$, (87) holds. □

Proposition 4.2. *Let $\Upsilon : \Re \times \Re^{m+s} \to \Re^{m+s}$ be defined by (77). Assume that the constraint nondegeneracy condition (85) holds at $\overline{X}$. Then for any $\mathcal{W} \in \partial\Upsilon(0, \bar{y})$, $\mathcal{W}$ is a P-matrix, i.e.,*

$$\max_i h_i(\mathcal{W}(0, h))_i > 0 \quad \forall\ 0 \neq h \in \Re^{m+s}. \tag{88}$$

Proof. Let $\mathcal{W} \in \partial\Upsilon(0, \bar{y})$. Suppose that there exists $0 \neq h \in \Re^{m+s}$ such that (88) does not hold, i.e.,

$$\max_i h_i(\mathcal{W}(0, h))_i \leq 0. \tag{89}$$

Then from part (iv) of Proposition 4.1, we know that there exist $\mathcal{D} \in \partial\pi(0, \bar{z})$ and $\mathcal{V} \in \partial \overline{D}_{\rho\sigma}(0, W(\bar{y}; X))$ such that

$$\mathcal{W}(0, h) = h - \mathcal{D}\Big(0, h - (Th + \sigma\widehat{\mathcal{A}}\mathcal{V}(0, \widehat{\mathcal{A}}^* h))\Big) = h - \mathcal{D}(0, h) + \mathcal{D}(0, Th + \sigma\widehat{\mathcal{A}}\mathcal{V}(0, \widehat{\mathcal{A}}^* h)),$$

where $\bar{z} = \bar{y} - (T\bar{y} + \widehat{\mathcal{A}D}_{\rho\sigma}(0, W(\bar{y}; X)) - \hat{b})$. By simple calculations, we can find a non-negative vector $d \in \Re^{m+s}$ satisfying

$$d_i = \begin{cases} 1 & \text{if } 1 \leq i \leq m + s_1, \\ \in [0, 1] & \text{if } m + s_1 + 1 \leq i \leq m + s_1 + l, \\ 0 & \text{if } m + s_1 + l + 1 \leq i \leq m + s, \end{cases}$$

such that for any $y \in \Re^{m+s}$,

$$(\mathcal{D}(0, y))_i = d_i y_i,\ i = 1, \ldots, m + s.$$

Thus we have $h_i(\mathcal{W}(0,h))_i = h_i\left[h_i - d_ih_i + d_i\Big(Th + \sigma\widehat{\mathcal{A}}\mathcal{V}(0,\widehat{\mathcal{A}}^*h)\Big)_i\right]$, $i = 1,\ldots,m+s$. This, together with (89), implies that

$$\begin{cases} h_i(Th + \sigma\widehat{\mathcal{A}}\mathcal{V}(0,\widehat{\mathcal{A}}^*h))_i \le 0 & \text{if } 1 \le i \le m+s_1, \\ h_i(Th + \sigma\widehat{\mathcal{A}}\mathcal{V}(0,\widehat{\mathcal{A}}^*h))_i \le 0 \text{ or } h_i = 0 & \text{if } m+s_1+1 \le i \le m+s_1+l, \\ h_i = 0 & \text{if } m+s_1+l+1 \le i \le m+s. \end{cases}$$

Hence, $\langle h, Th + \sigma\widehat{\mathcal{A}}\mathcal{V}(0,\widehat{\mathcal{A}}^*h\rangle = \langle \tilde{h}, \widetilde{T}\tilde{h} + \sigma\widetilde{\mathcal{A}}\mathcal{V}(0,\widetilde{\mathcal{A}}^*\tilde{h}\rangle \le 0$, where $0 \ne \tilde{h} \in \Re^{m+s_1+l}$ is defined by $\tilde{h}_i = h_i, i = 1,\ldots,m+s_1+l$. However, the above inequality contradicts (87) in Lemma 4.2. Hence, we have that (88) holds. □

Theorem 4.3. *Let $(\bar{\varepsilon},\bar{y})$ be an accumulation point of the infinite sequence $\{(\varepsilon^k,y^k)\}$ generated by Algorithm 2. Assume that the constraint nondegeneracy condition (85) holds at $\overline{X}$. Then the whole sequence $\{(\varepsilon^k,y^k)\}$ converges to $(\bar{\varepsilon},\bar{y})$ quadratically, i.e.,*

$$\|(\varepsilon^{k+1}-\bar{\varepsilon}, y^{k+1}-\bar{y})\| = O(\|(\varepsilon^k-\bar{\varepsilon}, y^k-\bar{y})\|^2).$$

Proof. To prove the quadratic convergence of $\{(\varepsilon^k,y^k)\}$, by Theorem 4.2, it is enough to show that E is strongly semismooth at $(\bar{\varepsilon},\bar{y})$ and all $\mathcal{V} \in \partial E(\bar{\varepsilon},\bar{y})$ are nonsingular. The strong semismoothness of E at $(\bar{\varepsilon},\bar{y})$ follows from part (iii) of Proposition 4.1 and the fact that the modulus function $|\cdot|$ is strongly semismooth everywhere on $\Re$.

Next, we show the nonsingularity of all elements in $\partial E(\bar{\varepsilon},\bar{y})$. For any $\mathcal{V} \in \partial E(\bar{\varepsilon},\bar{y})$, from Proposition 4.2 and the definition of E, we have that for any $0 \ne h \in \Re^{m+s+1}$, $\max_i h_i(\mathcal{V}d)_i > 0$, which implies that $\mathcal{V}$ is a P-matrix, and thus nonsingular [14, Theorem 3.3.4]. □

4.1 Efficient implementation of the inexact smoothing Newton method

When applying Algorithm 2 to solve the inner subproblem (72), the most expensive step is in solving the linear system (80). In our numerical implementation, we first obtain $\Delta\varepsilon^k = -\varepsilon^k + \varsigma_k\hat{\varepsilon}$, and then apply the BiCGStab iterative solver of Van der Vost [53] to the following linear system

$$G'_y(\varepsilon^k,y^k)\Delta y^k = -G(\varepsilon^k,y^k) - G'_\varepsilon(\varepsilon^k,y^k)\Delta\varepsilon^k \tag{90}$$

to obtain a Δy^k satisfying condition (81). For convenience, we suppress the superscript k in our subsequent analysis. By noting that $G(\varepsilon,y)$ and $\Upsilon(\varepsilon,y)$ are defined by (79) and (77), respectively, we have that

$$G'_y(\varepsilon,y)\Delta y = (1+\kappa\varepsilon)\Delta y + \pi'_z(\varepsilon,z)\Big(T\Delta y + \sigma\widehat{\mathcal{A}}(\overline{D}_{\rho\sigma})'_W(\varepsilon,W)\widehat{\mathcal{A}}^*\Delta y - \Delta y\Big), \tag{91}$$

where $z := y - (Ty + \widehat{\mathcal{A}}\overline{D}_{\rho\sigma}(\varepsilon,W) - \hat{b})$ and $W := X - \sigma(C - \widehat{\mathcal{A}}^*y)$. Let W have the SVD as in (10). Then, by (33), we have

$$(\overline{D}_{\rho\sigma})'_W(\varepsilon,W)(\widehat{\mathcal{A}}^*\Delta y) = U(\Lambda_{\alpha\alpha}\circ H_1^s + \Lambda_{\alpha\gamma}\circ H_1^a)V_1^T + U(\Lambda_{\alpha\beta}\circ H_2)V_2^T, \tag{92}$$

where $\Lambda_{\alpha\alpha}, \Lambda_{\alpha\gamma}$ and $\Lambda_{\alpha\beta}$ are given by (28), $H_1 = U^T(\widehat{\mathcal{A}}^*\Delta y)V_1$, and $H_2 = U^T(\widehat{\mathcal{A}}^*\Delta y)V_2$. When implementing the BiCGStab iterative method, one needs to repeatedly compute the matrix-vector multiplication $G'_y(\varepsilon, y)\Delta y$. From (92), it seems that a full SVD of W is required. But as we shall explain next, it is not necessary to compute V_2 explicitly.

For a problem where p is moderate but q is large, computing the full SVD would incur huge memory space since the matrix $V \in \Re^{q\times q}$ is large and dense. To over this difficulty, we first compute the economical form of the SVD of W, which is given by

$$W = U\Sigma V_1^T.$$

Then we construct V_2 via the QR factorization of V_1 with

$$V_1 = QR,$$

where $Q \in \Re^{q\times q}$ is orthogonal and $R \in \Re^{q\times p}$ is upper triangular. Decompose Q as $Q = [Q_1 \;\; Q_2]$, where $Q_1 \in R^{q\times p}$ and $Q_2 \in \Re^{q\times(q-p)}$. From [26, Theorem 5.2.1], we know

$$\text{range}(Q_2) = \text{range}(V_1)^\perp = \text{range}(V_2),$$

where range(Q_2) is the range space of Q_2. Since Q_2 has orthonormal columns which are orthogonal to those of V_1, Q_2 can be used in place of V_2. In our numerical implementation, Householder transformations are used to compute the QR factorization. Note that instead of storing the full Householder matrices, we only need to store the Householder vectors so as to compute the matrix-vector product involving V_2.

To achieve fast convergence for the BiCGStab method, we introduce an easy-to-compute diagonal preconditioner for the linear system (90). Since both $\pi'_z(\varepsilon, z)$ and T are diagonal matrices, we know from (91) that it is enough to find a good diagonal approximation of $\widehat{\mathcal{A}}(\overline{D}_{\rho\sigma})'_W(\varepsilon, W)\widehat{\mathcal{A}}^*$. Let

$$M := \widehat{\mathbf{A}}\mathbf{S}\widehat{\mathbf{A}}^T,$$

where $\widehat{\mathbf{A}}$ and $\mathbf{S}$ denote the matrix representation of the linear map $\widehat{\mathcal{A}}$ and $(\overline{D}_{\rho\sigma})'_W(\varepsilon, W)$ with respect to the standard bases in $\Re^{p\times q}$ and $\Re^{m+s}$, respectively. Let the standard basis in $\Re^{p\times q}$ be $\{E^{ij} \in \Re^{p\times q} : 1 \le i \le p, 1 \le j \le q\}$, where for each E^{ij}, its (i,j)-th entry is one and all the others are zero. Then the diagonal element of $\mathbf{S}$ with respect to the standard basis E^{ij} is given by

$$\mathbf{S}_{(i,j),(i,j)} = ((U \circ U)\widetilde{\Lambda}(V \circ V)^T)_{ij} + \frac{1}{2}\langle H_1^{ij} \circ (H_1^{ij})^T, \Lambda_{\alpha\alpha} - \Lambda_{\alpha\gamma}\rangle,$$

where $\widetilde{\Lambda} := [\frac{1}{2}(\Lambda_{\alpha\alpha} + \Lambda_{\alpha\gamma}), \Lambda_{\alpha\beta}]$ and $H_1^{ij} = U^T E^{ij} V_1$. Based on the above expression, the total cost of computing all the diagonal elements of $\mathbf{S}$ is equal to $2(p+q)pq + 3p^3q$ flops, which is too expensive if $p^2 \gg p+q$. Fortunately, the first term

$$\mathbf{d}_{(ij)} = ((U \circ U)\widetilde{\Lambda}(V \circ V)^T)_{ij}$$

is usually a very good approximation of $\mathbf{S}_{(i,j),(i,j)}$, and the cost of computing all the elements $\mathbf{d}_{(ij)}$, for $1 \le i \le p, 1 \le j \le q$, is $2(p+q)pq$ flops. Thus we propose the following diagonal preconditioner for the coefficient matrix $G'_y(\varepsilon, y)$:

$$M_G := (1 + \kappa\varepsilon)I + \pi'_z(\varepsilon, z)\Big(T + \sigma\text{diag}(\widehat{\mathbf{A}}\text{diag}(\mathbf{d})\widehat{\mathbf{A}}^T) - I\Big).$$

5 Numerical experiments

In this section, we report some numerical results to demonstrate the efficiency of our smoothing-Newton partial PPA.

In our numerical implementation, we apply the well-known alternating direction method of multipliers (ADMM) proposed by Gabay and Mercier [22], and Glowinski and Marrocco [25] for generating a good starting point for our PPA. To use the ADMM, we introduce two auxiliary variables Y and v, and consider the following equivalent form of (1):

$$\min_{X\in\Re^{p\times q},Y\in\Re^{p\times q},v\in\mathcal{Q}} \left\{\frac{1}{2}\|\mathcal{A}(X)-b\|^2+\rho\|Y\|_*+\langle C,X\rangle \;:\; Y=X,\,\mathcal{B}(X)-v=d\right\}. \tag{93}$$

Consider the following augmented Lagrangian function for the problem (93):

$$\begin{aligned} L_\beta(X,Y,v;Z,\lambda) &= \frac{1}{2}\|\mathcal{A}(X)-b\|^2+\rho\|Y\|_*+\langle C,X\rangle+\langle Z,X-Y\rangle+\langle \lambda,d-\mathcal{B}(X)+v\rangle \\ &+ \frac{\beta}{2}\|X-Y\|^2+\frac{\beta}{2}\|d-\mathcal{B}(X)+v\|^2, \end{aligned}$$

where $Z\in\Re^{p\times q}$ and $\lambda\in\Re^s$ are the Lagrangian multipliers for the linear equality constraints and $\beta>0$ is a penalty parameter. Given a starting point $(X^0,Y^0,v^0,Z^0,\lambda^0)$, the ADMM generates new iterates according to the following procedure:

$$\begin{aligned} X^{k+1} &:= \operatorname{argmin}_{X\in\Re^{p\times q}} L_\beta(X,Y^k,v^k;Z^k,\lambda^k), \\ (Y^{k+1},v^{k+1}) &:= \operatorname{argmin}_{Y\in\Re^{p\times q},\,v\in\mathcal{Q}} L_\beta(X^{k+1},Y,v;Z^k,\lambda^k), \\ &= \Big(D_{\rho/\beta}(X^{k+1}+\beta^{-1}Z^k),\,\Pi_{\mathcal{Q}}(\mathcal{B}(X^{k+1})-d-\beta^{-1}\lambda^k)\Big), \\ Z^{k+1} &:= Z^k+\gamma\beta(X^{k+1}-Y^{k+1}), \quad \lambda^{k+1}:=\lambda^k+\gamma\beta(d-\mathcal{B}(X^{k+1})+v^{k+1}), \end{aligned}$$

where $\gamma\in(0,(1+\sqrt{5})/2)$ is a given constant. Note that the theoretical convergence for the above procedure is guaranteed; see [18, Thm. 8]. During our numerical implementation, we observe that the performance of the ADMM is very sensitive to the choice of the penalty parameter β.

In order to measure the infeasibilities of the primal problem (46), we define two linear operators $\mathcal{B}_e:\Re^{p\times q}\to\Re^{s_1}$ and $\overline{\mathcal{B}}_e:\Re^{p\times q}\to\Re^{s_2}$, respectively, as follows:

$$\begin{cases} (\mathcal{B}_e(X))_i := \langle \mathcal{B}_i, X\rangle, \text{ for } i=1,\ldots,s_1, \\ (\overline{\mathcal{B}}_e(X))_i := \langle \mathcal{B}_i, X\rangle, \text{ for } i=s_1+1,\ldots,s. \end{cases}$$

Let $d=(d_{s_1};d_{s_2})$ where $d_{s_1}\in\Re^{s_1}$ and $d_{s_2}\in\Re^{s_2}$. We measure the infeasibilities and optimality for the primal problem (46) and the dual problem (50) as follows:

$$\begin{aligned} R_P &= \frac{\|(b-\zeta-\mathcal{A}(X);d_{s_1}-\mathcal{B}_e(X);\max(0,d_{s_2}-\overline{\mathcal{B}}_e(X)))\|}{1+\|\hat{b}\|}, \\ R_D &= \frac{\|C-\widehat{\mathcal{A}}^*y-Z\|}{1+\|\widehat{\mathcal{A}}^*\|}, \quad \text{relgap} = \frac{f_\rho(\zeta,X)-g_\rho(\zeta,\xi)}{1+|f_\rho(\zeta,X)|+|g_\rho(\zeta,\xi)|}, \end{aligned}$$

where $y = (\zeta;\xi), Z = (D_{\rho\sigma}(W) - W)/\sigma$ with $W = X - \sigma(C - \widehat{\mathcal{A}}^* y)$, and $f_\rho(\zeta, X)$ and $g_\rho(\zeta,\xi)$ are the objective functions of (46) and (50), respectively. The infeasibility of the condition $\|Z\|_2 \leq \rho$ is not checked since it is satisfied up to machine precision throughout the algorithm. In our numerical experiments, we stop the partial PPA when

$$\max\{R_P, R_D\} \leq 10^{-6} \quad \text{and} \quad |\text{relgap}| \leq 10^{-5}.$$

We choose the initial iterate $X^0 = 0, y^0 = 0$, and $\sigma_0 = 1$. The parameter ρ in (1) is set to be $\rho = 10^{-3}\|\mathcal{A}^* b\|_2$ if the data is not contaminated by noise; otherwise, the parameter ρ is set to be $\rho = 5 \times 10^{-3}\|\mathcal{A}^* b\|_2$.

5.1 Example 1

We consider the nearest matrix approximation problem which was discussed by Golub, Hoffman and Stewart in [27], where the classic Eckart-Young[17]-Mirsky[39] theorem was extended to obtain the nearest lower-rank approximation while certain specified columns of the matrix are fixed. The Eckart-Young-Mirsky theorem has the drawback that the approximation generally differs from the original matrix in all its entries. Thus it is not suitable for applications where some columns of the original matrix must be fixed. For example, in statistics the regression matrix for the multiple regression model with a constant term has a column of all ones, and this column should not be perturbed.

For each triplet (p, q, r), where r is the predetermined rank, we generate a random matrix $M \in \Re^{p\times q}$ of rank r by setting $M = M_1 M_2^T$ where both $M_1 \in \Re^{p\times r}$ and $M_2 \in \Re^{q\times r}$ have i.i.d. standard uniform entries in $(0, 1)$. As observed entries in practice are rarely exact, we corrupt the entries of M by Gaussian noises to simulate the situation where the observed data may be noisy as follows. First we generate a random matrix $N \in \Re^{p\times q}$ with i.i.d Gaussian entries. Then we assume that the observed data is given by $\widetilde{M} = M + \tau N\|M\|/\|N\|$, where τ is the noise factor. In our numerical experiments, we set $\tau = 0.1$. We assume that the first column of M should be fixed, and consider the following minimization problem:

$$\min_{X\in\Re^{p\times q}} \Big\{\frac{1}{2}\|X - \widetilde{M}\|^2 + \rho\|X\|_* \,:\, Xe_1 \,=\, Me_1,\, X \geq 0\Big\}, \tag{94}$$

where e_1 is the first column of the $q\times q$ identity matrix. Here we impose the extra constraint $X \geq 0$ since the original matrix M is nonnegative. Note that the approximation derived in [27] generally is not nonnegative.

For each (p, q, r) and τ, we generate 5 random instances. In Table 1, we report the total number of constraints $(m + s)$ in (46), the average number of outer iterations, the average total number of inner iterations, the average number of BiCGStab steps taken to solve (90), the average infeasibilities, the average relative gap, the average relative mean square error MSE $:= \|X - M\|/\|M\|$ (where M is the original matrix), the mean value of the rank (#sv) of X, and the average CPU time taken (in the format hours:minutes:seconds). We may observe from the table that the partial PPA is very efficient for solving (94). For the problem where p is moderate but q is large, e.g., $p = 100$ and $q = 20000$, we only compute the economical form of the SVD and use the technique introduced in section 4.1 to compute

V_2 via the QR factorization of V_1. It takes about 3.5 minutes to solve the last instance to achieve the tolerance of 10^{-6} while the MSE is reasonably small.

$p \times q$	r	$m+s$	it.|itsub|bicg	R_p | R_D | relgap	MSE	#sv	time
500×500	10	500500	5.0 | 13.0 | 4.0	5.57e-7 | 7.29e-7 | 3.61e-7	3.33e-2	169 (10)	38
500×500	50	500500	3.0 | 8.0 | 3.3	1.46e-7 | 4.58e-7 | 1.30e-7	4.01e-2	177 (NA)	26
500×500	100	500500	3.0 | 7.8 | 3.4	3.63e-7 | 6.68e-7 | 1.88e-7	3.72e-2	177 (NA)	26
1000×1000	10	2001000	7.0 | 16.4 | 4.6	1.03e-7 | 5.33e-7 | -3.15e-6	2.09e-2	121 (10)	3:14
1000×1000	50	2001000	9.0 | 16.2 | 3.0	6.13e-9 | 1.90e-8 | -3.75e-6	3.31e-2	138 (NA)	2:27
1000×1000	100	2001000	9.0 | 15.8 | 2.7	9.45e-9 | 1.92e-8 | -3.82e-6	3.10e-2	143 (NA)	2:20
1500×1500	10	4501500	9.0 | 19.0 | 4.4	3.47e-8 | 4.01e-8 | 9.14e-6	1.85e-2	22 (10)	9:18
1500×1500	50	4501500	9.0 | 16.2 | 3.2	1.06e-8 | 2.48e-8 | -3.71e-6	3.26e-2	54 (50)	6:54
1500×1500	100	4501500	8.0 | 15.2 | 3.2	1.14e-8 | 1.52e-8 | -4.50e-6	3.19e-2	67 (NA)	6:41
100×5000	10	1000100	10.0 | 12.8 | 1.5	2.67e-8 | 4.10e-8 | 3.85e-6	5.72e-2	100 (10)	46
100×10000	10	2000100	10.0 | 12.4 | 1.4	2.00e-8 | 4.09e-8 | 4.05e-6	5.70e-2	100 (10)	1:40
100×20000	10	4000100	10.4 | 13.0 | 1.4	1.89e-8 | 4.13e-8 | 3.90e-6	5.70e-2	100 (10)	3:32

Table 1: Numerical performance of the partial PPA on (94).

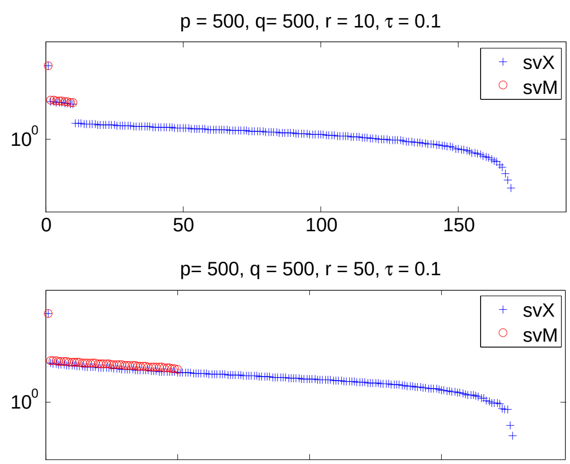

Figure 1: Distribution of singular values of X and M.

In the numerical experiments, we observe that when the generated matrix M is of small rank, e.g., $r = 10$, the singular values of the computed solution X are separated into two clusters with the first cluster having much larger mean value than that of the second cluster (see, e.g., Figure 1). We may view the number of singular values in the first cluster as a good estimate of the rank of the true solution, while the smaller positive singular values in the second cluster may be attributed to the presence of noise in the given data. When the matrix M is of high rank, e.g., $r = 50$, the singular values of X are usually not well separated into two clusters (see, e.g., Figure 1), excluding the largest singular value. In

Table 1, when the singular values of X are well separated into two clusters, we also report the number of singular values in the first cluster in parenthesis next to #sv. In the table, "NA" means that the singular values of X are not well separated into two clusters.

5.2 Example 2

In [32], Lin proposed the Latent Markov Analysis (LMA) approach for finding the reduced rank approximations of transition matrices. The LMA is applied to clustering based on pairwise similarities such that the inferred cluster relationships can be described probabilistically by the reduced-rank transition matrix. In [5], Benczúr, Csalogány and Sarlós considered the problem of finding the low rank approximation of the transition matrix for computing the personalized PageRank, which describes the backlink-based page quality around user-selected pages.

In this example, we evaluate the performance of our partial PPA for finding the nearest transition matrix of lower rank. Consider the set of n web pages as a directed graph whose nodes are the web pages and whose edges are all the links between pages. Let $\deg(i)$ be the outdegreee of the page i, i.e., the number of pages which can be reached by a direct link from page i. Note that all the self-referential links are excluded. Let $P \in \Re^{n\times n}$ be the matrix which describes the transition probability between pages i and j, where $P_{ij} = 1/\deg(i)$ if $\deg(i) > 0$ and there is a link form i to j. For some page i having no outlink (dangling pages), we assume $P_{ij} = 1/n$ for $j = 1, \dots, n$, i.e., the user will make a random choice with uniform distribution $1/n$. Since the matrix P for the web graph generally is reducible, P may have several eigenvalues on the unit circle, which could cause convergence problems to the power method for computing the PageRank [41]. The standard way of ensuring irreducibility is that we replace P by the matrix

$$P_c = cM + (1-c)ev^T,$$

where $c \in (0,1)$, $e \in \Re^n$ is a vector of all ones, and $v \in \Re^n$ is a vector such that $v \geq 0$ and $e^T v = 1$. We generate a random matrix $N \in \Re^{n\times n}$ with i.i.d Gaussian entries. Then we assume that the observed data is given by $\widetilde{P}_c = P_c + \tau N\|P_c\|/\|N\|$, where τ is the noise factor. In our numerical experiments, we set $\tau = 0.1$, $c = 0.85$ which is a typical value used by Google, and $v_i = 1/n \;\forall\, i$. The minimization problem which we solve is as follows:

$$\min_{X\in\Re^{n\times n}} \left\{ \frac{1}{2}\|X - \widetilde{P}_c\|^2 + \rho\|X\|_* \,:\, Xe = e,\ X \geq 0 \right\}. \tag{95}$$

We use the data `Harvard500.mat` generated by Cleve Moler's MATLAB program `surfer.m`[1] to evaluate the performance of our algorithm. We also use `surfer.m` to generate three adjacency graphs of a portion of web pages starting at the root page "`http://www.nus.edu.sg`". We also apply our algorithm to the data sets[2] collected by Tsaparas on querying the Google search engine about four topics: "automobile industries", "computational complexity", "computational geometry", and "randomized algorithms". Table 2 reports the average numerical results of PPA for solving (95) over 5 runs, where r denotes the rank of P_c. We can

[1] Available at `http://www.mathworks.com/moler/ncmfilelist.html`

[2] Available at: `http://www.cs.toronto.edu/~tsap/experiments/datasets /index.html`

observe from the table that the partial PPA is very efficient for solving (95) when applied to the real web graph data sets.

Problem	n	r	$m+s$	it.|itsub|bicg	R_p | R_D | relgap	MSE	#sv	time
Harvard500	500	218	500500	6.0 | 14.6 | 7.8	7.58e-8 | 4.92e-9 | -7.48e-6	5.87e-2	366	1:01
NUS500	500	225	500500	6.2 | 12.4 | 5.9	4.51e-8 | 1.60e-9 | -5.22e-6	5.70e-2	382	47
NUS1000	1000	466	2001000	5.4 | 14.2 | 7.7	3.35e-7 | 4.67e-9 | -6.46e-6	5.62e-2	658	5:19
NUS1500	1500	807	4501500	5.0 | 15.0 | 8.8	3.34e-7 | 5.68e-9 | -7.00e-6	6.35e-2	957	17:21
RandomAlg	742	216	1101870	7.0 | 16.0 | 7.7	3.37e-7 | 3.19e-9 | -7.02e-6	4.48e-2	631	2:48
Complexity	884	255	1563796	7.0 | 16.2 | 7.7	6.75e-8 | 1.75e-9 | -4.65e-6	4.77e-2	712	4:22
Automobile	1196	206	2862028	6.0 | 16.4 | 8.7	2.02e-7 | 5.91e-9 | -8.80e-6	4.05e-2	844	10:14
Geometry	1226	416	3007378	7.0 | 17.2 | 8.0	7.22e-8 | 1.99e-9 | -4.18e-6	4.67e-2	1018	11:01

Table 2: Numerical performance of the partial PPA on (95).

5.3 Example 3

We consider the problem of finding a low rank doubly stochastic matrix with a prescribed entry. A matrix $M \in \Re^{n\times n}$ is called doubly stochastic if it is nonnegative and all its row and column sums are equal to one. This problem arises from numerical simulation of large circuit networks. In order to reduce the complexity of simulating the whole system, Padé approximation via a Krylov subspace method, such as the Lanczos algorithm, is used to generate a low order approximation to the linear system matrix describing the large network [3]. The tridiagonal matrix $M \in \Re^{n\times n}$ produced by the Lanczos algorithm is generally not doubly stochastic. But if the original matrix is doubly stochastic, then we need to find a low rank approximation of M, which is doubly stochastic and matches the maximal moments. In our numerical experiments, we will not restrict the matrix M to be tridiagonal.

For each pair (n, r), we generate a positive matrix $\overline{M} \in \Re^{n\times n}$ with rank r by the same method as in Example 1. Then we use the Sinkhorn-Knopp algorithm [48] to find two positive definite diagonal matrices $D_1, D_2 \in \Re^{n\times n}$ such that $M = D_1\overline{M}D_2$ is a doubly stochastic matrix of rank r. We sample a subset $\mathcal{E}$ of m entries of M uniformly at random, and generate a random matrix $N_{\mathcal{E}} \in \Re^{p\times q}$ with sparsity pattern $\mathcal{E}$ and i.i.d standard Gaussian random entries. Then we assume that the observed data is given by $\widetilde{M}_{\mathcal{E}} = M_{\mathcal{E}} + \tau N_{\mathcal{E}}\|M_{\mathcal{E}}\|/\|N_{\mathcal{E}}\|$, where τ is the noise factor. The problem for matching the first moment of M can be stated as follows:

$$\min_{X\in\Re^{n\times n}} \left\{\frac{1}{2}\|X_{\mathcal{E}} - \widetilde{M}_{\mathcal{E}}\|^2 + \rho\|X\|_* \,:\, Xe = e,\, X^T e = e,\, X_{11} = M_{11},\, X \geq 0\right\}. \tag{96}$$

In our numerical experiments, we set $\tau = 0, 0.1$, and the number of sampled entries to be $m = 10dr$, where $dr = r(2n - r)$ is the degree of freedom in an $n \times n$ matrix of rank r. In Table 3, we report the average numerical results for solving (96) on randomly generated matrices over 5 runs, where m is the average number of sampled entries, and $m + s$ is the average number of total constraints in (46). For problems with noise, if the singular values of X are well separated into two clusters, we also report the number of singular values in the first cluster in parenthesis next to #sv as in Example 1. We can observe from the table that the partial PPA can solve (96) very efficiently for all the instances.

n/τ	r	m	$m+s$	it.|itsub|bicg	R_p | R_D | relgap	MSE	#sv	time
500	10	99148	350148	7.0 | 15.4 | 3.3	5.71e-7 | 6.88e-8 | -4.30e-6	3.53e-3	10	26
0	50	250000	501000	6.0 | 8.2 | 1.5	2.03e-7 | 8.45e-8 | -3.45e-6	7.07e-3	50	09
	100	250000	501000	5.0 | 7.0 | 1.4	1.02e-7 | 1.50e-7 | -3.62e-6	1.01e-2	100	08
1000	10	199034	1201034	9.0 | 20.0 | 4.0	6.80e-7 | 5.66e-8 | -8.17e-6	4.07e-3	10	2:56
0	50	974915	1976915	6.0 | 12.0 | 2.7	2.88e-7 | 4.68e-8 | -3.69e-6	7.11e-3	50	1:24
	100	1000000	2002000	5.0 | 7.0 | 1.4	3.63e-8 | 7.41e-8 | -3.52e-6	1.01e-2	100	39
1500	10	299194	2552194	10.0 | 23.0 | 4.0	5.81e-7 | 3.94e-8 | -8.79e-6	4.41e-3	10	9:29
0	50	1474481	3727481	7.0 | 13.8 | 2.6	1.41e-7 | 4.40e-8 | -4.70e-6	7.54e-3	50	4:12
	100	2250000	4503000	5.0 | 7.0 | 1.4	1.34e-8 | 4.92e-8 | -3.50e-6	1.01e-2	100	2:02
500	10	99148	350148	7.0 | 16.0 | 3.2	1.97e-7 | 1.93e-7 | -6.27e-6	5.42e-2	174 (10)	26
0.1	50	250000	501000	5.0 | 9.2 | 2.0	1.65e-7 | 2.31e-7 | -8.58e-6	3.97e-2	177 (NA)	12
	100	250000	501000	5.0 | 9.0 | 2.1	1.11e-7 | 1.83e-7 | -5.37e-6	3.65e-2	177 (NA)	12
1000	10	199034	1201034	8.0 | 18.8 | 3.6	1.45e-7 | 9.18e-8 | -9.31e-6	5.50e-2	234 (10)	2:41
0.1	50	974915	1976915	5.0 | 10.0 | 2.7	7.25e-7 | 7.91e-8 | -3.93e-6	3.30e-2	145 (NA)	1:13
	100	1000000	2002000	3.0 | 6.6 | 2.1	4.43e-7 | 3.32e-7 | -7.58e-6	3.07e-2	143 (NA)	45
1500	10	299194	2552194	9.0 | 22.2 | 3.9	1.69e-7 | 3.84e-8 | -5.68e-6	5.49e-2	275 (10)	8:56
0.1	50	1474481	3727481	5.0 | 11.0 | 2.7	4.76e-7 | 1.11e-7 | -6.87e-6	3.41e-2	194 (NA)	3:36
	100	2250000	4503000	2.0 | 5.2 | 3.1	2.11e-7 | 2.71e-7 | -3.26e-6	3.19e-2	68 (NA)	1:55

Table 3: Numerical performance of the partial PPA on (96). In the table, $m = 10\text{dof}$ and $\text{dof} = r(2n-r)$.

We may also consider the following generalized version of (96), where we want to find a low rank doubly stochastic matrix with k prescribed entries of M:

$$\min_{X\in\Re^{n\times n}} \left\{ \frac{1}{2}\|X_{\mathcal{E}} - \widetilde{M}_{\mathcal{E}}\|^2 + \rho\|X\|_* \mid Xe = e, X^T e = e, X_{i_t,j_t} = M_{i_t,j_t}, 1 \le t \le k, X \ge 0 \right\}, \tag{97}$$

where $(i_1, j_1), \ldots, (i_k, j_k)$ are distinct pairs. In our numerical experiments, we set $k = \lceil 10^{-3} n^2 \rceil$, which is the number of prescribed entries selected uniformly at random. Table 4 presents the average numerical results for solving (97) on randomly generated matrices over 5 runs.

n/τ	r	m	$m+s$	it.|itsub|bicg	R_p | R_D | relgap	MSE	#sv	time
500	10	99148	350148	7.0 | 18.2 | 6.4	5.62e-7 | 6.89e-8 | -4.27e-6	3.47e-3	10	39
0	50	250000	501000	6.0 | 10.8 | 3.0	3.71e-7 | 8.39e-8 | -3.39e-6	7.05e-3	50	15
	100	250000	501000	5.0 | 11.0 | 2.9	4.78e-7 | 1.49e-7 | -3.64e-6	1.00e-2	100	14
1000	10	199034	1201034	9.0 | 21.0 | 5.3	6.56e-7 | 5.78e-8 | -8.35e-6	3.94e-3	10	3:32
0	50	974915	1976915	6.0 | 12.4 | 4.3	3.06e-7 | 4.65e-8 | -3.64e-6	7.06e-3	50	1:47
	100	1000000	2002000	5.0 | 12.0 | 3.4	3.76e-7 | 7.31e-8 | -3.47e-6	1.00e-2	100	1:27
1500	10	299194	2552194	10.0 | 25.8 | 6.3	6.68e-7 | 3.94e-8 | -8.85e-6	4.20e-3	11	12:54
0	50	1474481	3727481	6.8 | 16.4 | 4.9	3.53e-7 | 5.39e-8 | -5.64e-6	7.50e-3	50	6:36
	100	2250000	4503000	5.0 | 12.0 | 4.3	3.39e-7 | 4.82e-8 | -3.42e-6	1.00e-2	100	4:41

500	10	99148	350148	7.0 | 16.0 | 3.6	2.10e-7 | 1.92e-7 | -6.26e-6	5.41e-2	174 (10)	28
0.1	50	250000	501000	5.0 | 11.0 | 2.9	5.43e-7 | 2.29e-7 | -8.51e-6	3.97e-2	177 (NA)	17
	100	250000	501000	5.0 | 11.0 | 3.0	4.18e-7 | 1.81e-7 | -5.35e-6	3.65e-2	177 (NA)	17
1000	10	199034	1201034	8.0 | 19.0 | 4.1	1.61e-7 | 9.14e-8 | -9.28e-6	5.47e-2	234 (10)	2:58
0.1	50	974915	1976915	5.0 | 13.2 | 4.6	6.18e-7 | 7.67e-8 | -3.90e-6	3.29e-2	151 (NA)	2:06
	100	1000000	2002000	3.0 | 11.2 | 5.5	1.17e-7 | 3.18e-7 | -7.13e-6	3.06e-2	151 (NA)	1:57
1500	10	299194	2552194	9.0 | 22.0 | 4.5	1.35e-7 | 3.81e-8 | -5.64e-6	5.45e-2	276 (10)	9:43
0.1	50	1474481	3727481	5.0 | 13.0 | 4.6	6.26e-7 | 1.13e-7 | -6.75e-6	3.39e-2	203 (NA)	5:23
	100	2250000	4503000	2.0 | 10.2 | 8.1	2.07e-7 | 3.50e-7 | -6.19e-6	3.11e-2	119 (NA)	5:55

Table 4: Same as Table 3 but for the problem (97). In the table, $m = 10\text{dof}$ and $\text{dof} = r(2n - r)$.

5.4 Example 4

We consider the problem of finding a low rank nonnegative approximation which preserves the left and right principal eigenvectors of a square positive matrix. This problem was considered by Ho and Dooren in [29]. Let $M \in \Re^{n\times n}$ be a positive matrix. By the Perron-Frobenius theorem, M has a simple positive eigenvalue λ with the largest magnitude. Moreover, there exist two positive eigenvectors $v, w \in \Re^n$ such that $Mv = \lambda v$ and $M^T w = \lambda w$. As suggested by Bonacich [8], the principal eigenvector can be used to measure network centrality, where the i-th component of the eigenvector gives the centrality of the i-th node in the network. For example, the Google's PageRank [41] is a variant of the eigenvector centrality for ranking web pages.

For each pair (n, r), we generate a positive matrix $M \in \Re^{n\times n}$ of rank r by the same method as in Example 1. We sample a subset $\mathcal{E}$ of m entries of M that are possibly corrupted by Gaussian noise as in Example 3. Given the largest positive eigenvalue λ and the corresponding left and right eigenvectors v and w of M, the problem of finding a low rank approximation of M while preserving the left and right eigenvectors can be stated as follows:

$$\min_{X\in\Re^{n\times n}} \Big\{\frac{1}{2}\|X_{\mathcal{E}} - \widetilde{M}_{\mathcal{E}}\|^2 + \rho\|X\|_* \;:\; Xv = \lambda v, X^T w = \lambda w, X \geq 0\Big\}. \tag{98}$$

n/τ	r	m	$m+s$	it.|itsub|bicg	R_p | R_D | relgap	MSE	#sv	time
500	10	99157	350157	6.0 | 14.8 | 2.5	1.27e-7 | 1.45e-7 | -7.75e-6	3.54e-3	10	24
0	50	250000	501000	3.0 | 7.0 | 2.0	3.95e-7 | 5.25e-7 | 3.36e-6	6.97e-3	50	09
	100	250000	501000	3.0 | 7.0 | 2.0	2.41e-7 | 5.46e-7 | -1.68e-6	1.01e-2	100	11
1000	10	199029	1201029	7.0 | 17.2 | 2.6	2.15e-7 | 7.20e-8 | 5.32e-6	2.93e-3	10	2:12
0	50	974912	1976912	2.4 | 7.2 | 2.3	7.32e-8 | 6.53e-7 | -6.72e-6	7.39e-3	50	56
	100	1000000	2002000	2.0 | 7.0 | 2.1	2.13e-7 | 8.02e-7 | -9.34e-7	1.03e-2	100	52
1500	10	299187	2552187	8.0 | 22.4 | 2.7	5.26e-7 | 3.90e-8 | 4.89e-6	2.86e-3	10	8:23
0	50	1474471	3727471	3.0 | 9.0 | 2.2	6.12e-7 | 8.03e-8 | -5.41e-6	7.54e-3	50	3:51
	100	2250000	4503000	2.0 | 7.0 | 2.1	1.38e-7 | 4.92e-7 | 2.85e-7	1.02e-2	100	2:41

500	10	99157	350157	2.0	5.8	2.2	1.85e-7	4.88e-7	-4.58e-6	5.38e-2	170 (10)	16
0.1	50	250000	501000	1.6	5.6	2.1	4.68e-7	8.07e-9	-4.63e-7	3.94e-2	177 (NA)	18
	100	250000	501000	1.8	6.2	2.1	3.35e-7	9.18e-9	-2.35e-7	3.64e-2	176 (NA)	17
1000	10	199029	1201029	2.0	5.2	1.9	6.13e-7	2.54e-7	-2.08e-6	5.28e-2	230 (10)	1:20
0.1	50	974912	1976912	2.0	6.8	2.4	9.95e-8	1.61e-8	-5.18e-8	3.27e-2	145 (NA)	1:18
	100	1000000	2002000	2.0	6.0	2.2	9.21e-7	1.73e-7	-2.64e-6	3.04e-2	142 (NA)	1:13
1500	10	299187	2552187	2.0	5.0	1.8	4.56e-7	1.83e-7	2.16e-6	5.22e-2	278 (10)	3:53
0.1	50	1474471	3727471	2.0	5.6	2.4	3.95e-7	2.93e-8	1.75e-7	3.35e-2	192 (NA)	3:36
	100	2250000	4503000	2.0	7.4	2.2	6.33e-8	4.31e-8	-5.30e-7	3.14e-2	67 (NA)	3:40

Table 5: Numerical performance of the partial PPA on (98). In the table, $m = 10dr$ and $dr = r(2n - r)$.

Table 5 reports the average numerical results of the partial PPA for solving (98) over 5 runs. Again, we observe that our partial PPA is very efficient in solving the problem (98).

6 Conclusion

In this paper, we introduced a partial PPA for solving nuclear norm regularized matrix least squares problems with equality and inequality constraints, and presented global and local convergence results based on the classical results for a general partial PPA. The inner subproblems, reformulated as a system of semismooth equations, were solved by an inexact smoothing Newton method, which is proved to be quadratically convergent under a constraint nondegeneracy condition, which we also characterized. Extensive numerical experiments conducted on a variety of large scale nuclear norm regularized matrix least squares problems demonstrated that our proposed algorithm is very efficient and robust.

We view our current work as providing an important tool for solving the very difficult problem of structured approximation with a prescribed rank. Although it is popular to regularize a structured approximation problem by the nuclear norm to promote a low rank solution, it cannot always deliver a solution with a prescribed rank. Thus it would be desirable to design an algorithm (along the line discussed in [24]) to find the best approximation with a prescribed rank efficiently while preserving certain desired structures.

References

[1] A.Y. Alfakih, A. Khandani, H. Wolkowicz, *Solving Euclidean distance matrix completion problems via semidefinite programming*, Computational Optimization and Applications 12 (1999), 13–30.

[2] B. Ames and S. Vavasis, *Nuclear norm minimization for the planted clique and biclique problems*, Mathematical Programming, 129 (2011), pp. 69–89.

[3] Z. Bai and R. Freund, *A Partial Padé-via-Lanczos Method for Reduced-Order Modeling*, Linear Algebra and its Applications 332-334 (2001), 139–164.

[4] M. Bee, *Estimating rating transition probabilites with missing data*, Statistical Methods and Applications 14 (2005), 127–141.

[5] A.A. Benczúr, K. Csalogány, and T. Sarlós, *On the feasibility of low-rank approximation for personalized PageRank*, Special interest tracks and posters of the 14th international conference on World Wide Web, ACM (2005), 972–973.

[6] D. Bertsekas and P. Tseng, *Partial proximal minimization algorithms for convex programming*, SIAM J. on Optimization 4 (1994), 551–572.

[7] R. Bhatia, Matrix Analysis, Springer-Verlag, NewYork, 1997.

[8] P. Bonacich, *Factoring and weighting approaches to status scores and clique identification*, The J. of Mathematical Sociology 2 (1972), 113–120.

[9] J. F. Cai, E. J. Candès and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM J. on Optimization 20 (2010), 1956–1982.

[10] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics 9 (2009), 717–772.

[11] S. Chennubhotla, *Spectral methods for multi-scale feature extraction and data clustering*, PhD thesis, University of Toronto, 2004.

[12] M. Chu, R. Funderlic, and R. Plemmons, *Structured low rank approximation*, Linear algebra and its applications 366 (2003), 157–172.

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

[14] R.W. Cottle, J.-S. Pang, and R.E. Stone, The Linear Complementarity Problem, Aacdemic Press, Boston, 1992.

[15] V. Dolezal, *Monotone operators and applications in Control and Network Theory*, vol. 2, Elsevier Scientific Pub. Co.,(Amsterdam and New York and New York), 1979.

[16] B.C. Eaves, *On the basic theorem of complementarity*, Mathematical Programming 1 (1971), 68—75.

[17] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika 1 (1936), 211–218.

[18] J. Eckstein, aand D. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Mathematical Programming, 55 (1992), pp. 293–318.

[19] M. Fazel, Matrix Rank Minimization with Applications. PhD thesis, Stanford University, 2002.

[20] M. Fazel, H. Hindi, and S. Boyd, *A rank minimization heuristic with application to minimum order system approximation*. In Proceedings of the American Control Conference, 2001.

[21] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Mathematical Programming 76 (1997), 513–532.

[22] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Computers and Mathematics with Applications 2 (1976), 17–40.

[23] Y. Gao and D.F. Sun, *Calibrating least squares covariance matrix problems with equality and inequality constraints*, SIAM J. on Matrix Analysis and Applications 31 (2009), 1432–1457.

[24] Y. Gao and D.F. Sun, *A majorized penalty approach for calibrating rank constrained correlation matrix problems*, March 2010.

[25] R. Glowinski and A. Marrocco, *Sur l'approximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet nonlinéaires*, Revue Française d'Automatique, Informatique, Recherche Opérationnelle 2 (1975), 41–76.

[26] G.H. Golub and C.F. van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore, USA, Third Edition, 1996.

[27] G.H. Golub, A. Hoffman, and G.W. Stewart, *A generalization of the Eckart-Young-Mirsky matrix approximation theorem*, Linear Algebra and its applications 88 (1987), 317–327.

[28] C. Ha, *A generalization of the proximal point algorithm*, SIAM J. on Control and Optimization 28 (1990), 503–512.

[29] N.D. Ho, and P. Van Dooren, *Non-negative matrix factorization with fixed row and column sums*, Linear Algebra and its Applications 429 (2008), 1020–1025.

[30] S. Ibaraki and M. Fukushima, *Partial proximal method of multipliers for convex programming problems*, J. of the Operations Research Society of Japan-Keiei Kagaku, 39 (1996), 213–229.

[31] K.F. Jiang, *Algorithms for large scale nuclear norm minimization and convex quadratic semidefinite programming problems*, PhD thesis, National University of Singapore, August 2011.

[32] Juan K. Lin, *Reduced rank approximations of transition matrices*, Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, 2003.

[33] Y.J. Liu, D.F. Sun, and K.C. Toh, *An implementable proximal point algorithmic framework for nuclear norm minimization*, Mathematical Programming, accepted, 2010.

[34] K. Löwner, *Über monotone matrixfunktionen*, Mathematische Zeitschrift 38 (1934), 177–216.

[35] Y. Luo, *A smoothing Newton-BICGStab method for least squares matrix nuclear norm problems*, Master thesis, National University of Singapore, January 2010.

[36] Ma, S., Goldfarb, D., and Chen, L., *Fixed point and bregman iterative methods for matrix rank minimization*, Mathematical Programming, 128 (2011), pp. 321–353.

[37] F. Meng, D. Sun, and G. Zhao, *Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization*, Mathematical Programming 104 (2005), 561–581.

[38] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. on Control and Optimization 15 (1977), 959–972.

[39] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, The Quarterly J. of Mathematics 11 (1960), 50–59.

[40] J.J. Moreau, *Proximite et dualite dans un espace hilbertien*, Bulletin de la Societe Mathematique de France 93 (1965), 273–299

[41] L. Page, S. Brin, R. Motwani, and T. Winograd, *The pagerank citation ranking: Bringing order to the web*, Technical report, Stanford Digital Library Technologies Project, 1999. Available at `http://dbpubs.stanford.edu/pub/1999-66`.

[42] T.K. Pong, P. Tseng, S. Ji, and J. Ye, *Trace Norm Regularization: Reformulations, Algorithms, and Multi-task Learning*, SIAM J. on Optimization 20 (2010), 3465–3489.

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

[44] B. Recht, M. Fazel and P.A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Review 52 (2010), 471–501.

[45] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, 1970.

[46] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. on Control and Optimization 14 (1976), 877–898.

[47] R.T. Rockafellar, *Augmented Lagrangains and applications of the proximal point algorithm in convex programming*, Mathematics of Operation Research 1 (1976), 97–116.

[48] R. Sinkhorn and P. Knopp, *Concerning nonnegative matrices and doubly stochastic matrices*, Pacific J. of Mathematics 21 (1967), 343–348.

[49] J.F. Sturm, *Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones*, Optimization Methods and Software 11 (1999), 625–653.

[50] D.F. Sun and J. Sun, *Semismooth matrix valued functions*, Mathematics of Operations Research 27 (2002), 150–169.

[51] K.C. Toh, M. J. Todd, and R. H. Tütüncü, SDPT3 — a Matlab software package for semidefinite programming, Optimization Methods and Software, 11 (1999), pp. 545–581.

[52] K.C. Toh and S.W. Yun,*An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems*, Pacific J. of Optimization 6 (2010), 615–640.

[53] H.A. Van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. on scientific and Statistical Computing 13 (1992), 631–644.

[54] J. Warga, *Fat homeomorphisms and unbounded derivate containers*, J. of Mathematical Analysis and Applications 81 (1981), 545—560.

[55] G.A. Watson, *Characterization of the subdifferential of some matrix norms*, Linear Algebra and its Applications 170 (1992), 33–45.

[56] K. Yosida, Functional Analysis, Springer Verlag, Berlin, 1964.

[57] E.H. Zarantonello, *Projections on Convex Sets in Hilbert Space and Spectral Theory I and II*, Contributions to Nonlinear Functional Analysis (E.H. Zarantonello, ed.), Academic Press, New York (1971), 237–424.

[58] X.Y. Zhao, D.F. Sun, and K.C. Toh, *A Newton-CG augmented Lagrangian method for semidefinite programming*, SIAM J. on Optimization 20 (2010), 1737–1765.