A Perturbation Inequality for Concave Functions of Singular Values and Its Applications in Low–Rank Matrix Recovery

Man–Chung Yue* Anthony Man–Cho So†

July 1, 2015

Abstract

In this paper, we establish the following perturbation result concerning the singular values of a matrix: Let $A, B \in \mathbb{R}^{m\times n}$ be given matrices, and let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function satisfying $f(0) = 0$. Then, we have

$$\sum_{i=1}^{\min\{m,n\}} \big|f(\sigma_i(A)) - f(\sigma_i(B))\big| \leq \sum_{i=1}^{\min\{m,n\}} f(\sigma_i(A-B)),$$

where $\sigma_i(\cdot)$ denotes the i–th largest singular value of a matrix. This answers an open question that is of interest to both the compressive sensing and linear algebra communities. In particular, by taking $f(\cdot) = (\cdot)^p$ for any $p \in (0, 1]$, we obtain a perturbation inequality for the so–called Schatten p–quasi–norm, which allows us to confirm the validity of a number of previously conjectured conditions for the recovery of low–rank matrices via the popular Schatten p–quasi–norm heuristic. We believe that our result will find further applications, especially in the study of low–rank matrix recovery.

Keywords: Singular value perturbation inequality; Schatten quasi–norm; Low–rank matrix recovery; Exact and robust recovery

1 Introduction

The problem of low–rank matrix recovery, with its many applications in computer vision [12, 20], trace regression [31, 23], network localization [19, 21], etc., has been attracting intense research interest in recent years. In a basic version of the problem, the goal is to reconstruct a low–rank matrix from a set of possibly noisy linear measurements. To achieve this, one immediate idea is to formulate the recovery problem as a rank minimization problem:

$$\begin{array}{ll} \text{minimize} & \text{rank}(X) \\ \text{subject to} & \|\mathcal{A}(X) - y\|_2 \leq \eta, \;\; X \in \mathbb{R}^{m\times n}, \end{array} \tag{1}$$

*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E–mail: mcyue@se.cuhk.edu.hk

†Department of Systems Engineering and Engineering Management, and, by courtesy, CUHK–BGI Innovation Institute of Trans–omics, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E–mail: manchoso@se.cuhk.edu.hk

where the linear measurement map $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$, the vector of measurements $y \in \mathbb{R}^l$, and the noise level $\eta \geq 0$ are given. However, Problem (1) is NP–hard in general, as it includes the NP–hard vector cardinality minimization problem [30] as a special case. Moreover, since the rank function is discontinuous, Problem (1) can be challenging from a computational point–of–view. To circumvent this intractability, a popular approach is to replace the objective of (1) with the so–called Schatten (quasi)–norm of X. Specifically, given a matrix $X \in \mathbb{R}^{m\times n}$ and a number $p \in (0, 1]$, let $\sigma_i(X)$ denote the i–th largest singular value of X and define the *Schatten p–quasi–norm* of X by

$$\|X\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(X) \right)^{1/p}.$$

One can then consider the following *Schatten p–quasi–norm heuristic* for low–rank matrix recovery:

$$\begin{array}{ll} \text{minimize} & \|X\|_p^p \\ \text{subject to} & \|\mathcal{A}(X) - y\|_2 \leq \eta, \;\; X \in \mathbb{R}^{m\times n}. \end{array} \tag{2}$$

Note that the function $X \mapsto \|X\|_p^p$ is continuous for each $p \in (0, 1]$. Thus, algorithmic techniques for continuous optimization can be used to tackle Problem (2). The Schatten quasi–norm heuristic is motivated by the observation that $\|X\|_p^p \to \text{rank}(X)$ as $p \searrow 0$. In particular, when $p = 1$, the function $X \mapsto \|X\|_1$ defines a norm—known as the *nuclear norm*—on the set of $m \times n$ matrices, and we obtain the well–known *nuclear norm heuristic* [13]. In this case, Problem (2) is a convex optimization problem that can be solved efficiently by various algorithms; see, e.g., [18] and the references therein. On the other hand, when $p \in (0, 1)$, the function $X \mapsto \|X\|_p$ only defines a quasi–norm. In this case, Problem (2) is a non–convex optimization problem and is NP–hard in general; cf. [16]. Nevertheless, a number of numerical algorithms implementing the Schatten p–quasi–norm heuristic (where $p \in (0, 1)$) have been developed (see, e.g., [28, 32, 21, 26] and the references therein), and they generally have better empirical recovery performance than the (convex) nuclear norm heuristic.

From a theoretical perspective, a natural and fundamental question concerning the aforementioned heuristics is about their recovery properties. Roughly speaking, this entails determining the conditions under which a given heuristic can recover, either exactly or approximately, a solution to Problem (1). A first study in this direction was done by Recht, Fazel, and Parrilo [35], who showed that techniques used to analyze the ℓ_1 heuristic for sparse vector recovery (see [40] for an overview and further pointers to the literature) can be extended to analyze the nuclear norm heuristic. Since then, recovery conditions based on the restricted isometry property (RIP) and various nullspace properties have been established for the nuclear norm heuristic; see, e.g., [33, 7, 6, 22] for some recent results. In fact, many recovery conditions for the nuclear norm heuristic can be derived in a rather simple fashion from their counterparts for the ℓ_1 heuristic by utilizing a perturbation inequality for the nuclear norm [33].

Compared with the nuclear norm heuristic, recovery properties of the Schatten p–quasi–norm heuristic are much less understood, even though the corresponding heuristic for sparse vector recovery, namely the ℓ_p heuristic with $p \in (0, 1)$, has been extensively studied; see, e.g., [42, 44, 34, 43] and the references therein. As first pointed out in [33] and later further elaborated in [25], the difficulty seems to center around the following question, which concerns the validity of certain perturbation inequality for the Schatten p–quasi–norm:

Question (Q) *Given a number $p \in (0,1)$ and matrices $A, B \in \mathbb{R}^{m\times n}$, does the inequality*

$$\sum_{i=1}^{\min\{m,n\}} |\sigma_i^p(A) - \sigma_i^p(B)| \leq \sum_{i=1}^{\min\{m,n\}} \sigma_i^p(A-B) \tag{3}$$

hold?

Indeed, assuming the validity of (3), one can establish a *necessary and sufficient* nullspace–based condition for the recovery of low–rank matrices via the Schatten p–quasi–norm heuristic [33]. This, coupled with the arguments in [33], would then allow one to translate various recovery conditions for the ℓ_p heuristic directly into those for the Schatten p–quasi–norm heuristic [33, 25]. Thus, there is a strong motivation to study Question (Q).

Although the authors of [33] are widely credited with formulating Question (Q) and pointing out its relevance in low–rank matrix recovery, the question itself has been studied long before the interest in low–rank matrix recovery takes shape. For instance, in 1988, Ando [1] showed, among other things, that the perturbation inequality (3) is valid when A, B are positive semidefinite. However, a complete answer to Question (Q) remains elusive. In a recent work [46], Zhang and Qiu claimed to have resolved Question (Q) in the affirmative by establishing a subadditivity inequality for singular values [46, Corollary 2.3]. However, as we shall explain in Section 2, there is a critical gap in the proof.[1] Thus, to the best of our knowledge, Question (Q) remains open; see also [3, Section 7].

In this paper, we show that the perturbation inequality (3) is indeed valid, thereby giving the first complete answer to Question (Q). In fact, we shall prove the following more general result:

Theorem 1 *Let $A, B \in \mathbb{R}^{m\times n}$ be given matrices. Suppose that $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a concave function satisfying $f(0) = 0$. Then, we have*

$$\sum_{i=1}^{\min\{m,n\}} \big|f(\sigma_i(A)) - f(\sigma_i(B))\big| \leq \sum_{i=1}^{\min\{m,n\}} f(\sigma_i(A-B)). \tag{4}$$

Since $x \mapsto |x|^p$ is concave on $\mathbb{R}_+$ for any $p \in (0,1]$, by taking $f(\cdot) = (\cdot)^p$ in (4), we immediately obtain (3). It is interesting to compare Theorem 1 with a result of Ando [1], which states that the perturbation inequality (4) is valid when $f : \mathbb{R}_+ \to \mathbb{R}_+$ is an operator concave function[2] satisfying $f(0) = 0$ and A, B are positive semidefinite. An essential ingredient in Ando's proof is the fact that every non–negative operator concave function on $\mathbb{R}_+$ admits an integral representation, which provides an explicit handle on the function. However, even though operator concavity implies concavity, the converse is not true in general. Moreover, the assumption that A and B are positive semidefinite is crucial to Ando's arguments. Thus, in order to prove Theorem 1, a substantially different approach is needed. Our proof, which is given in Section 3, is inspired in part by the work of Fiedler [14] and makes heavy use of matrix perturbation theory. In Section 4, we shall discuss some applications of the perturbation inequality (3) in the study of low–rank matrix recovery. Finally, we close with some concluding remarks in Section 5.

The following notations will be used throughout this paper. Let $\mathcal{S}^n$ (resp. $\mathcal{O}^n$) denote the set of $n \times n$ real symmetric (resp. orthogonal) matrices. For an arbitrary matrix $Z \in \mathbb{R}^{m\times n}$,

[1] This is also confirmed by the authors of [46] in a private correspondence.

[2] For the definition of operator concavity, see [4, Chapter V.1].

we use $\sigma(Z)$ and $\sigma_i(Z)$ to denote its vector of singular values and i–th largest singular value, respectively. Similarly, for an arbitrary symmetric matrix $Z \in \mathcal{S}^n$, we use $\lambda_i(Z)$ to denote its i–th largest eigenvalue. The spectral norm (i.e., the largest singular value) and Frobenius norm of Z are denoted by $\|Z\|$ and $\|Z\|_F$, respectively. Given a vector v, we use $\mathrm{Diag}(v)$ to denote the diagonal matrix with v on the diagonal. Similarly, given matrices $A_1, \ldots, A_l$, we use $\mathrm{BlkDiag}(A_1, \ldots, A_l)$ to denote the block diagonal matrix whose i–th diagonal block is A_i, for $i = 1, \ldots, l$. We say that $Z = O(\alpha)$ if $\|Z\|/\alpha$ is uniformly bounded as $\alpha \to 0$.

2 Gap in the Zhang–Qiu Proof

In this section, we review the main steps in Zhang and Qiu's proof of the perturbation inequality (4) and explain the gap in the proof. To set the stage, let us recall two classic perturbation inequalities:

(a) (Lidskii–Wielandt Eigenvalue Perturbation Inequality) Let $A, B \in \mathcal{S}^l$ be given. Then, for any $k \in \{1, \ldots, l\}$ and $i_1, \ldots, i_k \in \{1, \ldots, l\}$ satisfying $1 \leq i_1 < \cdots < i_k \leq l$,

$$\sum_{j=1}^{k} (\lambda_{i_j}(A) - \lambda_{i_j}(B)) \leq \sum_{i=1}^{k} \lambda_i(A - B); \tag{5}$$

see, e.g., [39, Chapter IV, Theorem 4.8].

(b) (Mirsky Singular Value Perturbation Inequality) Let $\bar{A}, \bar{B} \in \mathbb{R}^{m \times n}$ be given. Set $\bar{l} = \min\{m, n\}$. Then, for any $k \in \{1, \ldots, \bar{l}\}$ and $i_1, \ldots, i_k \in \{1, \ldots, \bar{l}\}$ satisfying $1 \leq i_1 < \cdots < i_k \leq \bar{l}$,

$$\sum_{j=1}^{k} \left|\sigma_{i_j}(\bar{A}) - \sigma_{i_j}(\bar{B})\right| \leq \sum_{i=1}^{k} \sigma_i(\bar{A} - \bar{B}); \tag{6}$$

see, e.g., [39, Chapter IV, Theorem 4.11].

Mirsky [29] observed that (6) is a simple consequence of (5), and his argument goes as follows. Let

$$A = \begin{bmatrix} \mathbf{0} & \bar{A} \\ \bar{A}^T & \mathbf{0} \end{bmatrix} \in \mathcal{S}^{m+n}, \quad B = \begin{bmatrix} \mathbf{0} & \bar{B} \\ \bar{B}^T & \mathbf{0} \end{bmatrix} \in \mathcal{S}^{m+n}, \tag{7}$$

and suppose without loss of generality that $m \leq n$. It is well–known (see Fact 1 below) that 0 is an eigenvalue of both A and B of multiplicity $n - m$, and the remaining eigenvalues of A and B are $\pm\sigma_1(\bar{A}), \ldots, \pm\sigma_m(\bar{A})$ and $\pm\sigma_1(\bar{B}), \ldots, \pm\sigma_m(\bar{B})$, respectively. Thus, we have

$$\{\lambda_i(A) - \lambda_i(B) : i = 1, \ldots, m+n\} = \left\{\pm\left|\sigma_i(\bar{A}) - \sigma_i(\bar{B})\right| : i = 1, \ldots, m\right\} \cup \{0\}.$$

In particular, by substituting (7) into (5), we obtain (6).

Motivated by the above argument, Zhang and Qiu first established a Lidskii–Wielandt–type singular value perturbation inequality by extending a matrix–valued triangle inequality of Bourin and Uchiyama [5] and invoking Horn's inequalities for characterizing the eigenvalues of sums of

Hermitian matrices [15]. Specifically, they showed that for any concave function $f : \mathbb{R}_+ \to \mathbb{R}_+$ and matrices $A, B \in \mathbb{R}^{m\times n}$, the inequality

$$\sum_{j=1}^{k} \big(f(\sigma_{i_j}(A)) - f(\sigma_{i_j}(B))\big) \leq \sum_{i=1}^{k} f(\sigma_i(A-B)) \tag{8}$$

holds for any $k \in \{1,\ldots,\bar{l}\}$ and $i_1,\ldots,i_k \in \{1,\ldots,\bar{l}\}$ satisfying $1 \leq i_1 < \cdots < i_k \leq \bar{l}$, where $\bar{l} = \min\{m,n\}$; cf. [46, Theorem 2.1]. Then, they claimed that the perturbation inequality (4) follows by applying Mirsky's argument above to (8); cf. [46, Corollary 2.3]. However, the reasoning in this last step is flawed. Indeed, the inequality (8) is concerned with *singular values*, while the inequality (5) is concerned with *eigenvalues*. In particular, for the matrices A, B given in (7), we only have

$$\{f(\sigma_i(A)) - f(\sigma_i(B)) : i = 1,\ldots,m+n\} = \big\{f(\sigma_i(\bar{A})) - f(\sigma_i(\bar{B})) : i = 1,\ldots,\bar{l}\big\} \cup \{0\},$$

and there is no guarantee that the set on the right–hand side (RHS) contains any element of the set

$$\big\{\big|f(\sigma_i(\bar{A})) - f(\sigma_i(\bar{B}))\big| : i = 1,\ldots,\bar{l}\big\}.$$

Hence, Mirsky's argument does not lead to the desired conclusion. In fact, we do not see a straightforward way of proving (4) using (8). The difficulty stems in part from the fact that f is always non–negative, while the eigenvalues in (5) can be negative. This suggests that (8) is fundamentally different from (5).

3 Proof of the Perturbation Inequality (4)

In this section, we give the first complete proof of the perturbation inequality (4). The proof can be divided into six steps.

Step 1: Reduction to the symmetric case.

A first observation concerning (4) is that we can restrict our attention to the case where both A and B are symmetric. To prove this, consider the linear operator $\Xi : \mathbb{R}^{m\times n} \to \mathcal{S}^{m+n}$ given by

$$\Xi(Z) = \begin{bmatrix} \mathbf{0} & Z \\ Z^T & \mathbf{0} \end{bmatrix}.$$

We shall make use of the following standard fact, which establishes a relationship between the singular value decomposition of an arbitrary matrix $Z \in \mathbb{R}^{m\times n}$ and the spectral decomposition of $\Xi(Z) \in \mathcal{S}^{m+n}$:

Fact 1 *(cf. [39, Chapter I, Theorem 4.2]) Let $Z \in \mathbb{R}^{m\times n}$ be a given matrix with $m \leq n$. Consider its singular value decomposition $Z = U \begin{bmatrix} \Sigma & \mathbf{0} \end{bmatrix} V^T$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma = Diag(\sigma_1(Z),\ldots,\sigma_m(Z)) \in \mathcal{S}^m$ is diagonal. Write $V = \begin{bmatrix} V^1 & V^2 \end{bmatrix}$, where $V^1 \in \mathbb{R}^{n\times m}$ and $V^2 \in \mathbb{R}^{n\times(n-m)}$. Then, the matrix $\Xi(Z)$ admits the spectral decomposition*

$$\Xi(Z) = W \begin{bmatrix} \Sigma & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} W^T,$$

where

$$W = \frac{1}{\sqrt{2}} \begin{bmatrix} U & U & \mathbf{0} \\ V^1 & -V^1 & \sqrt{2}\, V^2 \end{bmatrix}$$

is orthogonal. In particular, 0 *is an eigenvalue of* $\Xi(Z)$ *of multiplicity* $n-m$*, and the remaining eigenvalues of* $\Xi(Z)$ *are* $\pm\sigma_1(Z), \ldots, \pm\sigma_m(Z)$*.*

Fact 1 implies that the i–th largest singular value of $\Xi(Z)$ is given by

$$\sigma_i(\Xi(Z)) = \begin{cases} \sigma_{\lceil i/2 \rceil}(Z) & \text{for } i = 1, \ldots, 2m, \\ 0 & \text{for } i = 2m+1, \ldots, m+n. \end{cases} \tag{9}$$

This in turn implies the following result:

Proposition 1 *The inequality (4) holds for all matrices* $A, B \in \mathbb{R}^{m \times n}$ *iff it holds for all symmetric matrices* $A, B \in \mathcal{S}^l$*.*

Proof The "only if" part of the proposition is clear. Suppose then the inequality (4) holds for all symmetric matrices $A, B \in \mathcal{S}^l$. Consider arbitrary matrices $A, B \in \mathbb{R}^{m \times n}$, and without loss of generality, suppose that $m \leq n$. By assumption and the linearity of Ξ, we have

$$\sum_{i=1}^{m+n} \left| f(\sigma_i(\Xi(A))) - f(\sigma_i(\Xi(B))) \right| \leq \sum_{i=1}^{m+n} f(\sigma_i(\Xi(A-B))).$$

Together with (9), this implies that

$$\begin{aligned} 2\sum_{i=1}^{m} \left| f(\sigma_i(A)) - f(\sigma_i(B)) \right| &= \sum_{i=1}^{2m} \left| f(\sigma_i(\Xi(A))) - f(\sigma_i(\Xi(B))) \right| \\ &\leq \sum_{i=1}^{2m} f(\sigma_i(\Xi(A-B))) \\ &= 2\sum_{i=1}^{m} f(\sigma_i(A-B)). \end{aligned}$$

This completes the proof. □

In view of Proposition 1, we will focus on proving Theorem 1 for the case where A, B are symmetric. Our strategy is to first establish (4) for those functions f that, in addition to the assumptions in Theorem 1, satisfy a regularity condition called *well–behavedness* (the precise definition will be given shortly). Then, we show how the well–behavedness assumption can be removed using a limiting argument, thereby completing the proof of Theorem 1.

Step 2: Local behavior of a well–behaved f*.*

To introduce the notion of well–behavedness, let us first recall some basic facts from convex analysis. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function satisfying $f(0) = 0$. Then, for any $x_l, x_r, y_l, y_r \geq 0$ satisfying $x_l < x_r$, $y_l < y_r$, $x_l \leq y_l$, and $x_r \leq y_r$, we have

$$\frac{f(x_r) - f(x_l)}{x_r - x_l} \geq \frac{f(y_r) - f(x_l)}{y_r - x_l} \geq \frac{f(y_r) - f(y_l)}{y_r - y_l}; \tag{10}$$

cf. [37, Chapter 5, Lemma 16]. This implies that for each $x > 0$, the right–hand derivative of f at x, which is defined as

$$d_f(x) = \lim_{\tau \searrow 0} \frac{f(x+\tau) - f(x)}{\tau},$$

exists and is finite. Moreover, we have $f(y) \leq f(x) + d_f(x)(y-x)$ for any $y \geq 0$. Now, define the extension $\bar{d}_f : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ of $d_f : \mathbb{R}_{++} \to \mathbb{R}$ by

$$\bar{d}_f(x) = \begin{cases} d_f(x) & \text{for } x > 0, \\ \limsup_{t \searrow 0} d_f(t) & \text{for } x = 0. \end{cases}$$

Using (10), it can be easily verified that $\bar{d}_f(y) \geq \bar{d}_f(x)$ for all $x \geq y \geq 0$. We say that f is *well–behaved* if $\bar{d}_f(x) < +\infty$ for all $x \geq 0$. Note that for a well–behaved f, we have

$$f(y) \leq f(x) + \bar{d}_f(x)(y-x) \tag{11}$$

for all $x, y \geq 0$.

Consider an arbitrary symmetric matrix $M \in \mathcal{S}^n$. We say that $\pi = (\pi_1, \ldots, \pi_n)$ is a *spectrum–sorting permutation of* M if π is a permutation of $\{1, \ldots, n\}$ and $\sigma_i(M) = |\lambda_{\pi_i}(M)|$ for $i = 1, \ldots, n$. Note that there can be more than one spectrum–sorting permutation of M, as multiple eigenvalues can have the same magnitude. Now, given a spectrum–sorting permutation π of M, let $M = U\Lambda U^T$ be a spectral decomposition of M, where $\Lambda = \mathrm{Diag}(\lambda_{\pi_1}(M), \ldots, \lambda_{\pi_n}(M)) \in \mathcal{S}^n$. Furthermore, define $M_\pi = U\Lambda_\pi U^T$, where $\Lambda_\pi = \mathrm{Diag}(s_1, \ldots, s_n) \in \mathcal{S}^n$ and

$$s_i = \mathrm{sgn}(\lambda_{\pi_i}(M)) \cdot \bar{d}_f(\sigma_i(M)) \quad \text{for } i = 1, \ldots, n.$$

Our immediate objective is to prove the following theorem, which is the crux of our proof of the perturbation inequality (4):

Theorem 2 *Let $M, N \in \mathcal{S}^n$ be given symmetric matrices. Suppose that $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a well–behaved concave function satisfying $f(0) = 0$. Then, for any spectrum–sorting permutation π of M and any scalar $t > 0$,*

$$\sum_{i=1}^n f(\sigma_i(M + tN)) \leq \sum_{i=1}^n f(\sigma_i(M)) + t \cdot \mathrm{tr}\,(NM_\pi) + O(t^2).$$

Note that if $\mathrm{tr}(NM_\pi) < 0$, then for sufficiently small $t > 0$, the RHS of the above inequality is strictly smaller than $\sum_{i=1}^n f(\sigma_i(M))$. We shall make use of this simple observation later.

The proof of Theorem 2 relies on the following fact concerning the singular values of a perturbed symmetric matrix:

Fact 2 *(cf. [27, Section 5.1]) Let $M, N \in \mathcal{S}^n$ be given symmetric matrices. Suppose that M has $l+1$ distinct singular values for some $l \in \{0, 1 \ldots, n-1\}$, and that they are arranged as follows:*

$$\begin{array}{rcccl} & \sigma_{i_0}(M) & = \cdots = & \sigma_{i_1 - 1}(M) \\ > & \sigma_{i_1}(M) & = \cdots = & \sigma_{i_2 - 1}(M) \\ & & \vdots & \\ > & \sigma_{i_j}(M) & = \cdots = & \sigma_{i_{j+1} - 1}(M) \\ & & \vdots & \\ > & \sigma_{i_l}(M) & = \cdots = & \sigma_{i_{l+1} - 1}(M). \end{array} \tag{12}$$

Here, the indices $i_0, i_1, \ldots, i_l, i_{l+1} \in \{1, \ldots, n+1\}$ *satisfy* $1 = i_0 < i_1 < \cdots < i_l < i_{l+1} = n+1$. *Then, for any* $t > 0$, $j \in \{0, 1, \ldots, l\}$, *and* $i \in \{i_j, \ldots, i_{j+1} - 1\}$, *we have*

$$\sigma_i(M + tN) = \sigma_i(M) + t \cdot \lambda_{i-i_j+1}\left((Q^j)^T \Xi(N) Q^j\right) + O(t^2), \tag{13}$$

where Q^j *is a* $2n \times (i_{j+1} - i_j)$ *matrix whose columns are the eigenvectors associated with the* i_j*–th to the* $(i_{j+1} - 1)$*–st eigenvalue of* $\Xi(M)$.

Proof of Theorem 2 Using (11) and (13), we have

$$f(\sigma_i(M + tN)) \leq f(\sigma_i(M)) + t \cdot \bar{d}_f(\sigma_i(M)) \cdot \lambda_{i-i_j+1}\left((Q^j)^T \Xi(N) Q^j\right) + O(t^2) \tag{14}$$

for any $t > 0$, $j \in \{0, 1, \ldots, l\}$, and $i \in \{i_j, \ldots, i_{j+1} - 1\}$. Hence,

$$\begin{aligned}
\sum_{i=1}^{n} f(\sigma_i(M + tN)) &\leq \sum_{i=1}^{n} f(\sigma_i(M)) + t \sum_{j=0}^{l} \sum_{i=i_j}^{i_{j+1}-1} \bar{d}_f(\sigma_i(M)) \cdot \lambda_{i-i_j+1}\left((Q^j)^T \Xi(N) Q^j\right) + O(t^2) \\
&= \sum_{i=1}^{n} f(\sigma_i(M)) + t \sum_{j=0}^{l} \bar{d}_f(\sigma_{i_j}(M)) \cdot \mathrm{tr}\left((Q^j)^T \Xi(N) Q^j\right) + O(t^2),
\end{aligned}$$

where the last equality follows from (12). Now, fix a spectrum–sorting permutation π of M. Let $M = U \Sigma V^T$ be a singular value decomposition of M, where $\Sigma = \mathrm{Diag}(\sigma_1(M), \ldots, \sigma_n(M)) \in \mathcal{S}^n$. Here, we take u_i to be the eigenvector corresponding to the eigenvalue $\lambda_{\pi_i}(M)$ and $v_i = \mathrm{sgn}(\lambda_{\pi_i}(M)) u_i$, where u_i (resp. v_i) is the i–th column of U (resp. V), for $i = 1, \ldots, n$. Then, by Fact 1, the matrix Q^j can be put into the form

$$Q^j = \frac{1}{\sqrt{2}} \begin{bmatrix} U^j \\ V^j \end{bmatrix},$$

where U^j (resp. V^j) is the $n \times (i_{j+1} - i_j)$ matrix formed by the i_j–th to $(i_{j+1} - 1)$–st columns of U (resp. V), for $j = 0, 1, \ldots, l$. Upon letting

$$D(M) = \mathrm{BlkDiag}\left(\bar{d}_f(\sigma_{i_0}(M)) I_{i_1 - i_0}, \ldots, \bar{d}_f(\sigma_{i_l}(M)) I_{i_{l+1} - i_l}\right) \in \mathbb{R}^{n \times n}$$

and noting, because of (12), that $D(M) = \mathrm{Diag}(\bar{d}_f(\sigma_1(Z)), \ldots, \bar{d}_f(\sigma_n(Z)))$, we compute

$$\begin{aligned}
\sum_{j=0}^{l} \bar{d}_f(\sigma_{i_j}(M)) \cdot \mathrm{tr}\left((Q^j)^T \Xi(N) Q^j\right) &= \sum_{j=0}^{l} \bar{d}_f(\sigma_{i_j}(M)) \cdot \mathrm{tr}\left((V^j)^T N (U^j)\right) \\
&= \mathrm{tr}\left(NUD(M)V^T\right) \\
&= \mathrm{tr}(NM_\pi).
\end{aligned}$$

This completes the proof. □

In the sequel, we fix $A, B \in \mathcal{S}^n$ and let $A = U_A \Sigma_A U_A^T$ and $B = U_B \Sigma_B U_B^T$ be spectral decompositions of A and B, respectively.

Step 3: Lower bounding the RHS of (4) via an optimization problem.

Consider a function f that satisfies the assumptions in Theorem 2. Let $V = U_A^T U_B \in \mathcal{O}^n$. Then, we can lower bound the RHS of (4) as follows:

$$\sum_{i=1}^{n} f(\sigma_i(A-B)) = \sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - V\Sigma_B V^T\right)\right) \geq \inf_{Q\in\mathcal{O}^n} \sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - Q\Sigma_B Q^T\right)\right). \tag{15}$$

We claim that the minimum value above can be attained. This follows from the compactness of $\mathcal{O}^n$ and the following result:

Proposition 2 *For each $i \in \{1,\ldots,n\}$, the function $f(\sigma_i(\cdot))$ is continuous on $\mathcal{S}^n$.*

Proof Let $i \in \{1,\ldots,n\}$ be fixed. By [39, Chapter IV, Theorem 4.11], $\sigma_i(\cdot)$ is 1–Lipschitz continuous. Moreover, since $f(\cdot)$ is concave on $\mathbb{R}_+$, it is continuous on $\mathbb{R}_{++}$ [38, Lemma 2.70]. Thus, $f(\sigma_i(\cdot))$ is continuous at all $Z \in \mathcal{S}^n$ satisfying $\sigma_i(Z) > 0$. Now, let $Z \in \mathcal{S}^n$ be such that $\sigma_i(Z) = 0$. Then, using (11) and the fact that $f(0) = 0$, we have

$$|f(\sigma_i(Y)) - f(\sigma_i(Z))| \leq \left|\bar{d}_f(0)\right| \cdot |\sigma_i(Y) - \sigma_i(Z)|$$

for all $Y \in \mathcal{S}^n$. This, together with the 1–Lipschitz continuity of $\sigma_i(\cdot)$, implies that $f(\sigma_i(\cdot))$ is continuous at all $Z \in \mathcal{S}^n$ satisfying $\sigma_i(Z) = 0$ as well. □

Hence, let $Q_0 \in \mathcal{O}^n$ be the orthogonal matrix that attains the minimum value in (15). Clearly, in order to establish the perturbation inequality (4) for f, it suffices to show that

$$\sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - Q_0\Sigma_B Q_0^T\right)\right) \geq \sum_{i=1}^{n} \left|f(\sigma_i(A)) - f(\sigma_i(B))\right|.$$

Towards that end, we need the following result, which concerns the structure of the minimizer Q_0:

Theorem 3 *Let $\bar{B} = Q_0\Sigma_B Q_0^T \in \mathcal{S}^n$ and $C = \Sigma_A - \bar{B} \in \mathcal{S}^n$. Then, for any spectrum–sorting permutation π of C, $\bar{B}$ and C_π commute.*

Proof Suppose that $\bar{B}$ and C_π do not commute for some spectrum–sorting permutation π of C. Set $D = C_\pi\bar{B} - \bar{B}C_\pi \neq \mathbf{0}$. It is easy to verify that D is skew–symmetric, i.e., $D = -D^T$. Hence, we have $V(t) = \exp(tD) \in \mathcal{O}^n$ for all $t \in \mathbb{R}$. Since f is well–behaved, we compute

$$\begin{aligned}
\sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - V(t)\bar{B}V(t)^T\right)\right) &= \sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - (I+tD)\bar{B}(I-tD) + O(t^2)\right)\right) \\
&\leq \sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - \bar{B} + t(\bar{B}D - D\bar{B})\right)\right) \\
&\quad + \sum_{i=1}^{n} \left[\bar{d}_f\left(\sigma_i\left(\Sigma_A - \bar{B} + t(\bar{B}D - D\bar{B})\right)\right) \cdot O(t^2)\right] \quad (16) \\
&\leq \sum_{i=1}^{n} f(\sigma_i(C)) + t \cdot \mathrm{tr}\left((\bar{B}D - D\bar{B})C_\pi\right) + O(t^2), \quad (17)
\end{aligned}$$

where (16) follows from (11) and the 1–Lipschitz continuity of $\sigma_i(\cdot)$ for all $i \in \{1,\ldots,n\}$, while (17) follows from Theorem 2 and the fact that

$$\bar{d}_f\left(\sigma_i\left(\Sigma_A - \bar{B} + t(\bar{B}D - D\bar{B})\right)\right) \leq \bar{d}_f(0) < +\infty$$

for all $t \in \mathbb{R}$ and $i \in \{1,\ldots,n\}$. Using the identity $\mathrm{tr}(XY^T) = \mathrm{tr}(Y^TX)$, which is valid for arbitrary matrices of the same dimensions, we have

$$\mathrm{tr}\left((\bar{B}D - D\bar{B})C_\pi\right) = \mathrm{tr}\left(-DD^T\right) = -\|D\|_F^2 < 0. \tag{18}$$

It follows from (17) and (18) that for sufficiently small $t > 0$,

$$\sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - V(t)\bar{B}V(t)^T\right)\right) < \sum_{i=1}^{n} f(\sigma_i(C)),$$

which contradicts the minimality of Q_0. Hence, we have $D = \mathbf{0}$, or equivalently, $\bar{B}$ and C_π commute. □

In view of Theorem 3, the definition of C_π, and the fact that two symmetric matrices commute iff they are simultaneously diagonalizable, it is tempting to claim that $\bar{B}$ and C also commute. This conclusion, which is equivalent to $\Sigma_A\bar{B} = \bar{B}\Sigma_A$, would be more useful, as it reveals the relationship that the minimizer Q_0 imposes on the eigenvalues of A and B. Unfortunately, the claim is not true in general, for there may exist a set of eigenvectors of C_π that is not a set of eigenvectors of C. To illustrate this possibility, consider the following example:

Example 1 *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be given by $f(x) = x$. Then, by definition, we have $\bar{d}_f(x) = 1$ for all $x \geq 0$. Now, let*

$$\bar{B} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}, \quad C = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

It is easy to verify that $\lambda_1(C) = 3$ and $\lambda_2(C) = 1$. Thus, for any spectrum–sorting permutation π of C, we have $C_\pi = I$, which clearly commutes with $\bar{B}$. However, we have

$$\bar{B}C = \begin{bmatrix} -1 & 5 \\ 2 & 5 \end{bmatrix} \neq \begin{bmatrix} -1 & 2 \\ 5 & 5 \end{bmatrix} = C\bar{B}.$$

The above example shows that $\bar{B}$ and C need not commute when $\lambda_{\pi_i}(C) \neq \lambda_{\pi_j}(C)$ for some spectrum–sorting permutation π of C and $i, j \in \{1,\ldots,n\}$, but the corresponding eigenvalues of C_π, namely, $s_i = \mathrm{sgn}(\lambda_{\pi_i}(C)) \cdot \bar{d}_f(|\lambda_{\pi_i}(C)|)$ and $s_j = \mathrm{sgn}(\lambda_{\pi_j}(C)) \cdot \bar{d}_f(|\lambda_{\pi_j}(C)|)$, are equal. To circumvent this difficulty, we will first focus on the case where the function f satisfies the assumptions in Theorem 2 and has *strictly decreasing slope*; i.e., $\bar{d}_f$ is strictly decreasing on $\mathbb{R}_+$ (recall that the concavity of f on $\mathbb{R}_+$ implies that $\bar{d}_f$ is non–increasing on $\mathbb{R}_+$). The assumption on $\bar{d}_f$ is sufficient to guarantee that $s_i = \mathrm{sgn}(\lambda_{\pi_i}(C)) \cdot \bar{d}_f(|\lambda_{\pi_i}(C)|)$ and $s_j = \mathrm{sgn}(\lambda_{\pi_j}(C)) \cdot \bar{d}_f(|\lambda_{\pi_j}(C)|)$ are distinct whenever $\lambda_{\pi_i}(C)$ and $\lambda_{\pi_j}(C)$ are. Consequently, C and C_π have the same sets of eigenvectors, which, together with Theorem 3, implies that $\bar{B}$ and C commute. After establishing the perturbation inequality (4) for this case, we will show how the assumption on $\bar{d}_f$ can be removed using a limiting argument.

Step 4: Establishing the perturbation inequality (4) under well–behavedness and strictly decreasing slope assumptions.

Throughout this step, let f be a function that satisfies the assumptions in Theorem 2 and has strictly decreasing slope; i.e., $\bar{d}_f$ is strictly decreasing on $\mathbb{R}_+$. We have already seen from the discussion following the proof of Theorem 3 that $\bar{B}$ and C commute. Using the definition of C, this implies that $\Sigma_A \bar{B} = \bar{B}\Sigma_A$; i.e., Σ_A and $\bar{B}$ commute.

Now, suppose that A has distinct eigenvalues. Since Σ_A and $\bar{B}$ commute and Σ_A is diagonal, it is straightforward to show that $\bar{B}$ must also be diagonal. In particular, we can write $\bar{B} = \mathrm{Diag}(\lambda_{\theta_1}(B), \ldots, \lambda_{\theta_n}(B))$ for some permutation $\theta = (\theta_1, \ldots, \theta_n)$ of $\{1, \ldots, n\}$. Using this and [3, Proposition 1], we obtain

$$\begin{aligned}
\sum_{i=1}^{n} f(\sigma_i(A-B)) &\geq \sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - Q_0 \Sigma_B Q_0^T\right)\right) \\
&= \sum_{i=1}^{n} f\left(\sigma_i\left(\Sigma_A - \mathrm{Diag}(\lambda_{\theta_1}(B), \ldots, \lambda_{\theta_n}(B))\right)\right) \\
&\geq \sum_{i=1}^{n} \left|f(\sigma_i(A)) - f(\sigma_i(B))\right|; \qquad (19)
\end{aligned}$$

i.e., the perturbation inequality (4) holds in this case. On the other hand, suppose that A has repeated eigenvalues. By considering a sequence $\{A^l\}_{l=1}^{\infty}$ of matrices in $\mathcal{S}^n$ with distinct eigenvalues such that $A^l \to A$ and using (19), we have

$$\sum_{i=1}^{n} f(\sigma_i(A^l - B)) \geq \sum_{i=1}^{n} \left|f(\sigma_i(A^l)) - f(\sigma_i(B))\right|$$

for $l = 1, 2, \ldots$. This, together with Proposition 2, implies that the perturbation inequality (4) holds in this case as well.

Step 5: Relaxing the strictly decreasing slope assumption.

Now, suppose that f satisfies only the assumptions in Theorem 2. Our strategy is to approximate f using a sequence of functions that not only satisfy the assumptions in Theorem 2 but also have strictly decreasing slope. The perturbation inequality (4) for f would then follow from the result in Step 4 and a limiting argument. To begin, for each $\delta > 0$, define $\tilde{f}_\delta : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\tilde{f}_\delta(x) = f(x) - \delta^2 \exp(-x/\delta) + \delta^2.$$

The functions $\{\tilde{f}_\delta\}_{\delta>0}$ have the following properties:

Proposition 3 *The following hold:*

(a) For each $\delta > 0$, $\tilde{f}_\delta$ is concave on $\mathbb{R}_+$ and $\tilde{f}_\delta(0) = 0$.

(b) For each $\delta > 0$, we have $\bar{d}_{\tilde{f}_\delta}(x) = \bar{d}_f(x) + \delta \exp(-x/\delta)$ for all $x \geq 0$. In particular, $\tilde{f}_\delta$ is well–behaved and $\bar{d}_{\tilde{f}_\delta}$ is strictly decreasing on $\mathbb{R}_+$ for each $\delta > 0$.

(c) For each $x \geq 0$, we have $\tilde{f}_\delta(x) \to f(x)$ as $\delta \searrow 0$.

Proof

(a) Consider a fixed $\delta > 0$. By direct substitution, we have $\tilde{f}_\delta(0) = 0$. Moreover, it is easy to verify that $x \mapsto -\delta^2 \exp(-x/\delta) + \delta^2$ is concave on $\mathbb{R}_+$. It follows that $\tilde{f}_\delta$, which is the sum of two concave functions on $\mathbb{R}_+$, is concave on $\mathbb{R}_+$.

(b) Consider a fixed $\delta > 0$. By definition, for any $x > 0$, we have

$$\begin{aligned} \bar{d}_{\tilde{f}_\delta}(x) &= d_{\tilde{f}_\delta}(x) \\ &= \lim_{\tau \searrow 0} \left(\frac{f(x+\tau) - f(x)}{\tau} + \frac{\delta^2(\exp(-x/\delta) - \exp(-(x+\tau)/\delta))}{\tau} \right) \\ &= d_f(x) + \delta \exp(-x/\delta). \end{aligned}$$

Moreover,

$$\bar{d}_{\tilde{f}_\delta}(0) = \limsup_{t \searrow 0} d_{\tilde{f}_\delta}(t) = \limsup_{t \searrow 0} \left(d_f(t) + \delta \exp(-t/\delta) \right) = \bar{d}_f(0) + \delta.$$

It follows that $\bar{d}_{\tilde{f}_\delta}(x) = \bar{d}_f(x) + \delta \exp(-x/\delta)$ for all $x \geq 0$. In particular, since f is well–behaved and $\delta \exp(-x/\delta) \leq \delta$ for all $x \geq 0$, we conclude that $\tilde{f}_\delta$ is well–behaved.

Finally, since $\bar{d}_f$ is non–increasing on $\mathbb{R}_+$ and $x \mapsto \delta \exp(-x/\delta)$ is strictly decreasing on $\mathbb{R}_+$, we conclude that $\bar{d}_{\tilde{f}_\delta}$ is strictly decreasing on $\mathbb{R}_+$.

(c) For each $x \geq 0$, we have $\tilde{f}_\delta(x) - f(x) = -\delta^2 \exp(-x/\delta) + \delta^2$. It follows that $\tilde{f}_\delta(x) \to f(x)$ as $\delta \searrow 0$.

□

Proposition 3(a,b) implies that for each $\delta > 0$, the function $\tilde{f}_\delta$ satisfies the assumptions in Theorem 2 and has strictly decreasing slope. Hence, the result in Step 4 implies that for each $\delta > 0$, we have

$$\sum_{i=1}^n \tilde{f}_\delta(\sigma_i(A - B)) \geq \sum_{i=1}^n \left| \tilde{f}_\delta(\sigma_i(A)) - \tilde{f}_\delta(\sigma_i(B)) \right|.$$

By taking $\delta \searrow 0$ and applying Proposition 3(c), we conclude that the perturbation inequality (4) holds for f as well.

Step 6: Relaxing the well–behavedness assumption and completing the proof of Theorem 1.

To complete the proof of Theorem 1, it remains to relax the well–behavedness assumption. Towards that end, let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function satisfying $f(0) = 0$. For each $\delta > 0$, define $f_\delta : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$f_\delta(x) = \min \left\{ \frac{f(\delta)}{\delta} x, f(x) \right\}.$$

The following result shows that the functions $\{f_\delta\}_{\delta > 0}$ satisfy the assumptions in Theorem 2 and converge pointwise to f:

Proposition 4 *The following hold:*

(a) For each $\delta > 0$, f_δ is concave on $\mathbb{R}_+$ and $f_\delta(0) = 0$.

(b) For each $\delta > 0$, we have

$$\bar{d}_{f_\delta}(x) = \begin{cases} \bar{d}_f(x) & \text{for } x \geq \delta, \\ \dfrac{f(\delta)}{\delta} & \text{for } 0 \leq x < \delta. \end{cases}$$

In particular, f_δ is well–behaved.

(c) For each $x \geq 0$, we have $f_\delta(x) \to f(x)$ as $\delta \searrow 0$.

Proof

(a) Consider a fixed $\delta > 0$. By direct substitution, we have $f_\delta(0) = 0$. Moreover, since f_δ is the pointwise minimum of two concave functions on $\mathbb{R}_+$, it is concave on $\mathbb{R}_+$.

(b) Consider a fixed $\delta > 0$. The concavity of f on $\mathbb{R}_+$ and the inequalities in (10) imply that

$$f_\delta(x) = \begin{cases} f(x) & \text{for } x \geq \delta, \\ \dfrac{f(\delta)}{\delta} x & \text{for } 0 \leq x \leq \delta. \end{cases} \tag{20}$$

Thus, for $0 < x < \delta$, we have

$$\bar{d}_{f_\delta}(x) = d_{f_\delta}(x) = \lim_{\tau \searrow 0} \frac{1}{\tau} \left(\frac{f(\delta)}{\delta}((x+\tau) - x) \right) = \frac{f(\delta)}{\delta}.$$

On the other hand, for $x \geq \delta$, we have

$$\bar{d}_{f_\delta}(x) = d_{f_\delta}(x) = \lim_{\tau \searrow 0} \frac{f(x+\tau) - f(x)}{\tau} = d_f(x).$$

It follows that

$$\bar{d}_{f_\delta}(0) = \limsup_{t \searrow 0} d_{f_\delta}(t) = \frac{f(\delta)}{\delta}.$$

Since $\delta > 0$ is fixed, we have $\bar{d}_{f_\delta}(0) < +\infty$, which implies that f_δ is well–behaved.

(c) Clearly, we have $f_\delta(0) = f(0) = 0$ for all $\delta > 0$. Hence, $f_\delta(0) \to f(0)$ as $\delta \searrow 0$. Now, let $x > 0$ be fixed. Using (20), we have $f_\delta(x) = f(x)$ for all $\delta \in (0, x)$. It follows that $f_\delta(x) \to f(x)$ as $\delta \searrow 0$, as desired.

□

Proposition 4(a,b) and the result in Step 5 imply that for each $\delta > 0$, we have

$$\sum_{i=1}^{n} f_\delta(\sigma_i(A - B)) \geq \sum_{i=1}^{n} \big| f_\delta(\sigma_i(A)) - f_\delta(\sigma_i(B)) \big|.$$

Thus, by taking $\delta \searrow 0$ and applying Proposition 4(c), we conclude that the perturbation inequality (4) holds for f, which completes the proof of Theorem 1.

4 Applications in Low–Rank Matrix Recovery

As pointed out in [33], one important consequence of the perturbation inequality (3) is that it connects the sufficient conditions for the recovery of low–rank matrices via the Schatten p–quasi–norm heuristic to those for the recovery of sparse vectors via the ℓ_p heuristic. For completeness' sake, let us briefly elaborate on the connection here.

For a given number $p \in (0,1]$ and integer $k \geq 1$, let $\mathscr{S}_k^p$ be the set of $s \times t$ matrices (where $t \geq k$) such that whenever $A \in \mathscr{S}_k^p$, every vector $\bar{x} \in \mathbb{R}^t$ with $\|\bar{x}\|_0 = |\{i : \bar{x}_i \neq 0\}| \leq k$ and $y = A\bar{x} \in \mathbb{R}^s$ can be exactly recovered by solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|x\|_p^p \\ \text{subject to} & Ax = y. \end{array} \tag{21}$$

We then have the following theorem:

Theorem 4 *(cf. [33, Theorem 1]) Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ be a given linear operator with $m \leq n$. Suppose that $\mathcal{A}$ possesses the following property for some number $p \in (0,1]$ and integer $k \geq 1$:*

PROPERTY (E). For any orthogonal matrices $U \in \mathcal{O}^m$ and $V \in \mathcal{O}^n$, the matrix $A_{U,V} \in \mathbb{R}^{l\times m}$ induced by the linear map $x \mapsto \mathcal{A}\left(U \left[\begin{array}{cc} \mathrm{Diag}(x) & \mathbf{0} \end{array}\right] V^T\right)$ belongs to $\mathscr{S}_k^p$.

Then, every matrix $\bar{X} \in \mathbb{R}^{m\times n}$ with rank$(\bar{X}) \leq k$ and $y = \mathcal{A}(\bar{X}) \in \mathbb{R}^l$ can be exactly recovered by solving Problem (2) with $\eta = 0$.

The proof of Theorem 4 relies on the following two results, the latter of which is established using the perturbation inequality (3):

Fact 3 *(cf. [17]) Let $A \in \mathbb{R}^{s\times t}$ be a given matrix, $p \in (0,1]$ be a given number, and $k \geq 1$ be a given integer. Then, we have $A \in \mathscr{S}_k^p$ iff*

$$\sum_{i=1}^{k} |z_i^{\downarrow}|^p < \sum_{i=k+1}^{t} |z_i^{\downarrow}|^p \quad \textit{for all } z \in \mathcal{N}(A)\backslash\{\mathbf{0}\},$$

where $z^{\downarrow} \in \mathbb{R}^t$ is the vector whose i–th entry is the i–th largest (in absolute value) entry of z, and $\mathcal{N}(A) = \{z \in \mathbb{R}^t : Az = \mathbf{0}\}$ is the nullspace of A.

Proposition 5 *Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ be a given linear operator with $m \leq n$, $p \in (0,1]$ be a given number, and $k \geq 1$ be a given integer. Then, every matrix $\bar{X} \in \mathbb{R}^{m\times n}$ with rank$(\bar{X}) \leq k$ and $y = \mathcal{A}(\bar{X}) \in \mathbb{R}^l$ can be exactly recovered by solving Problem (2) with $\eta = 0$ iff*

$$\sum_{i=1}^{k} \sigma_i^p(Z) < \sum_{i=k+1}^{m} \sigma_i^p(Z) \tag{22}$$

holds for all $Z \in \mathcal{N}(\mathcal{A})\backslash\{\mathbf{0}\}$.

Proof Suppose that (22) holds for all $Z \in \mathcal{N}(\mathcal{A})\backslash\{\mathbf{0}\}$. Let $\bar{X}, \bar{X}' \in \mathbb{R}^{m\times n}$ be such that $\mathrm{rank}(\bar{X}) \leq k$ and $\mathcal{A}(\bar{X}) = \mathcal{A}(\bar{X}') = y$. Clearly, we have $\bar{Z} = \bar{X}' - \bar{X} \in \mathcal{N}(\mathcal{A})$. If $\bar{Z} \neq \mathbf{0}$,

or equivalently, if $\bar{X}' \neq \bar{X}$, then by taking $f(\cdot) = (\cdot)^p$ in Theorem 1 and using the fact that $\mathrm{rank}(\bar{X}) \leq k$, we obtain

$$\begin{aligned}
\sum_{i=1}^{m} \sigma_i^p(\bar{X}+\bar{Z}) &\geq \sum_{i=1}^{m} \left|\sigma_i^p(\bar{X}) - \sigma_i^p(\bar{Z})\right| \\
&= \sum_{i=1}^{k} \left|\sigma_i^p(\bar{X}) - \sigma_i^p(\bar{Z})\right| + \sum_{i=k+1}^{m} \left|\sigma_i^p(\bar{X}) - \sigma_i^p(\bar{Z})\right| \\
&\geq \sum_{i=1}^{k} \sigma_i^p(\bar{X}) - \sum_{i=1}^{k} \sigma_i^p(\bar{Z}) + \sum_{i=k+1}^{m} \sigma_i^p(\bar{Z}) \\
&> \sum_{i=1}^{m} \sigma_i^p(\bar{X}).
\end{aligned}$$

Since $\bar{X}' = \bar{X} + \bar{Z}$ is arbitrary, this shows that $\bar{X}$ is the unique optimal solution to Problem (2) when $\eta = 0$.

Conversely, suppose there exists a $\bar{Z} \in \mathcal{N}(\mathcal{A})\backslash\{\mathbf{0}\}$ such that $\sum_{i=1}^{k} \sigma_i^p(\bar{Z}) \geq \sum_{i=k+1}^{m} \sigma_i^p(\bar{Z})$. Let $\bar{Z} = U \left[\begin{array}{cc} \mathrm{Diag}(\sigma(\bar{Z})) & \mathbf{0} \end{array}\right] V^T$ be its singular value decomposition, and define

$$\bar{X} = -U \left[\begin{array}{cc} \Sigma_1^k(\bar{Z}) & \mathbf{0} \end{array}\right] V^T, \quad \bar{X}' = U \left[\begin{array}{cc} \Sigma_{k+1}^m(\bar{Z}) & \mathbf{0} \end{array}\right] V^T,$$

where

$$\begin{aligned}
\Sigma_1^k(\bar{Z}) &= \mathrm{Diag}\left(\sigma_1(\bar{Z}), \ldots, \sigma_k(\bar{Z}), 0, \ldots, 0\right) \in \mathcal{S}^m, \\
\Sigma_{k+1}^m(\bar{Z}) &= \mathrm{Diag}\left(0, \ldots, 0, \sigma_{k+1}(\bar{Z}), \ldots, \sigma_m(\bar{Z})\right) \in \mathcal{S}^m.
\end{aligned}$$

Clearly, we have $\mathrm{rank}(\bar{X}) \leq k$. Moreover, since $\mathcal{A}(\bar{X}' - \bar{X}) = \mathcal{A}(\bar{Z}) = \mathbf{0}$, we have $\mathcal{A}(\bar{X}) = \mathcal{A}(\bar{X}')$. Now, using the definition of $\bar{Z}$, we compute

$$\|\bar{X}\|_p^p = \sum_{i=1}^{k} \sigma_i^p(\bar{Z}) \geq \sum_{i=k+1}^{m} \sigma_i^p(\bar{Z}) = \|\bar{X}'\|_p^p.$$

This shows that $\bar{X}$ is not the unique optimal solution to Problem (2) when $\eta = 0$ and $y = \mathcal{A}(\bar{X})$. □

Proof of Theorem 4 Consider an arbitrary matrix $Z \in \mathcal{N}(\mathcal{A})\backslash\{\mathbf{0}\}$ with singular value decomposition $Z = U \left[\begin{array}{cc} \mathrm{Diag}(\sigma(Z)) & \mathbf{0} \end{array}\right] V^T$. We have $\mathbf{0} = \mathcal{A}(Z) = A_{U,V}(\sigma(Z))$. Hence, Property (E) and Fact 3 imply that Z satisfies (22). The desired conclusion now follows from Proposition 5. □

By invoking existing results in the literature and applying Theorem 4, exact recovery properties of the Schatten p–quasi–norm heuristic can be deduced in a rather straightforward manner. As an illustration, let us establish two recovery conditions based on notions of restricted isometry for the Schatten p–quasi–norm heuristic. We begin with the following simple observation:

Proposition 6 *Let m, n, r be integers such that $r \leq m \leq n$. Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ be a given linear operator.*

(a) Suppose there exists a constant $\alpha_r \in (0,1)$ such that

$$(1-\alpha_r)\|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1+\alpha_r)\|X\|_F^2$$

for all $X \in \mathbb{R}^{m\times n}$ with rank$(X) \leq r$. Then, for any $U \in \mathcal{O}^m$ and $V \in \mathcal{O}^n$, the matrix $A_{U,V} \in \mathbb{R}^{l\times m}$ satisfies

$$(1-\alpha_r)\|x\|_2^2 \leq \|A_{U,V}(x)\|_2^2 \leq (1+\alpha_r)\|x\|_2^2 \tag{23}$$

for all $x \in \mathbb{R}^m$ with $\|x\|_0 \leq r$.

(b) Let $p \in (0,1]$ be given. Suppose there exists a constant $\beta_{p,r} \in (0,1)$ such that

$$(1-\beta_{p,r})\|X\|_F^p \leq \|\mathcal{A}(X)\|_p^p \leq (1+\beta_{p,r})\|X\|_F^p$$

for all $X \in \mathbb{R}^{m\times n}$ with rank$(X) \leq r$. Then, for any $U \in \mathcal{O}^m$ and $V \in \mathcal{O}^n$, the matrix $A_{U,V} \in \mathbb{R}^{l\times m}$ satisfies

$$(1-\beta_{p,r})\|x\|_2^p \leq \|A_{U,V}(x)\|_p^p \leq (1+\beta_{p,r})\|x\|_2^p \tag{24}$$

for all $x \in \mathbb{R}^m$ with $\|x\|_0 \leq r$.

Proof Let $x \in \mathbb{R}^m$ be such that $\|x\|_0 \leq r$. For any $U \in \mathcal{O}^m$ and $V \in \mathcal{O}^n$, the matrix $X = U \left[\begin{array}{cc} \mathrm{Diag}(x) & \mathbf{0} \end{array}\right] V^T \in \mathbb{R}^{m\times n}$ has rank at most r. Moreover, we have $\|X\|_F = \|x\|_2$, $\|\mathcal{A}(X)\|_2 = \|A_{U,V}(x)\|_2$ and $\|\mathcal{A}(X)\|_p = \|A_{U,V}(x)\|_p$. This completes the proof. □

Condition (23) (resp. (24)) implies that for any orthogonal matrices $U \in \mathcal{O}^m$ and $V \in \mathcal{O}^n$, the matrix $A_{U,V} \in \mathbb{R}^{l\times m}$ satisfies the *restricted isometry property (RIP) of order r* [9] (resp. *restricted p–isometry property (p–RIP) of order r* [11]) with constant at most α_r (resp. $\beta_{p,r}$). Hence, we shall say that a linear operator $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ satisfies the RIP of order r (resp. p–RIP of order r) with constant at most α_r (resp. $\beta_{p,r}$) if it satisfies the hypothesis of Proposition 6(a) (resp. Proposition 6(b)). Now, the results in [11, 44], together with Theorem 4, imply the following recovery conditions:

Theorem 5 *Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ be a given linear operator with $m \leq n$. Furthermore, let $p \in (0,1)$ be given.*

(a) (cf. [44]; see also [43]) Let $k \geq 1$ be an integer such that $2k \leq m$. Suppose that $\mathcal{A}$ satisfies the RIP of order $2k$ with constant at most α_{2k}, and that $p < \min\{1, 1.0873 \times (1-\alpha_{2k})\}$. Then, every matrix $\bar{X} \in \mathbb{R}^{m\times n}$ with rank$(\bar{X}) \leq k$ and $y = \mathcal{A}(\bar{X}) \in \mathbb{R}^l$ can be exactly recovered by solving Problem (2) with $\eta = 0$.

(b) (cf. [11, Theorem 2.4]) Given an integer $k \geq 1$ and a real number $b > 1$, let $a = \lceil b^{2/(2-p)}k \rceil / k$. Suppose that $\mathcal{A}$ satisfies the p–RIP of order $(a+1)k$ with constant at most $\beta_{p,(a+1)k}$, and that $\beta_{p,ak} + b\beta_{p,(a+1)k} < b - 1$. Then, every matrix $\bar{X} \in \mathbb{R}^{m\times n}$ with rank$(\bar{X}) \leq k$ and $y = \mathcal{A}(\bar{X}) \in \mathbb{R}^l$ can be exactly recovered by solving Problem (2) with $\eta = 0$.

It is worth noting that the recovery conditions in Theorem 5 improve upon those in [24, 45], which are obtained by analyzing the Schatten p–quasi–norm heuristic directly. This shows that it could be advantageous to first establish recovery conditions for the ℓ_p heuristic (which is often easier to do) and then translate them into recovery conditions for the Schatten p–quasi–norm heuristic using Theorem 4.

So far our discussion has focused on the case of noiseless recovery (i.e., $\eta = 0$). For the case of noisy recovery (i.e., $\eta > 0$), we do not know whether a result similar to Theorem 4 holds (cf. [33, Theorem 1]). Nevertheless, the perturbation inequality (3) can still facilitate a direct analysis of the Schatten p–quasi–norm heuristic in this case. For instance, using arguments similar to those in [25], we can prove the following robust recovery result:

Theorem 6 *Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$, where $m \leq n$, be a given linear operator that satisfies the RIP of order $2k$ with constant at most α_{2k} for some integer $k \geq 1$. Furthermore, let $\bar{X} \in \mathbb{R}^{m\times n}$ be a matrix with rank$(\bar{X}) \leq k$. Set*

$$\bar{p} = \ln\left(1 + \frac{1}{2k}\right) \Bigg/ \ln\left(\sqrt{\frac{1+\alpha_{2k}}{1-\alpha_{2k}}} \cdot \frac{2m}{\sqrt{k}}\right) \in (0, 1). \tag{25}$$

Given $p \in (0, \bar{p}]$, $\eta > 0$, and $y = \mathcal{A}(\bar{X}) + z \in \mathbb{R}^l$ for some $z \in \mathbb{R}^l$ satisfying $\|z\|_2 \leq \eta$, let $X^* \in \mathbb{R}^{m\times n}$ be an optimal solution to Problem (2). Then,

$$\|X^* - \bar{X}\|_p < \left(1 - \frac{1}{2\sqrt{k}}\right)^{-1} \frac{2(2k)^{1/p}\eta}{\sqrt{k(1-\alpha_{2k})}}.$$

Proof To ease notation, let us assume that m is a multiple of k. The case where m is not a multiple of k can be handled in the same manner. Define $Z = X^* - \bar{X} \in \mathbb{R}^{m\times n}$ and let $Z = U \left[\begin{array}{cc} \mathrm{Diag}(\sigma(Z)) & \mathbf{0} \end{array}\right] V^T$ be its singular value decomposition. Furthermore, for $i = 1, \ldots, m/k$, let

$$\Sigma_i = \mathrm{Diag}\left(\sigma_{(i-1)k+1}(Z), \ldots, \sigma_{ik}(Z)\right) \in \mathcal{S}^k.$$

Then, we have $Z = Z_1 + \cdots + Z_{m/k}$, where $Z_i = U_i \Sigma_i V_i^T$ and U_i (resp. V_i) is the $m\times k$ (resp. $n\times k$) matrix formed by the $((i-1)k+1)$–st to (ik)–th columns of U (resp. V).

Since $\|Z\|_p^p = \|Z_1\|_p^p + \cdots + \|Z_{m/k}\|_p^p$, in order to prove Theorem 6, it suffices to control $\|Z_i\|_p^p$ for $i = 1, \ldots, m/k$. Towards that end, recall that for any $x \in \mathbb{R}^h$ and $p \in (0, 1]$, we have the inequality

$$\|x\|_p^p \leq h^{1-p/2}\|x\|_2^p. \tag{26}$$

Thus, it suffices to consider $\|Z_i\|_F^p$ for $i = 1, \ldots, m/k$, which are more amenable to analysis. To begin, let us derive a bound on $\|Z_1\|_F$ using the RIP of $\mathcal{A}$.

Proposition 7 *The following holds:*

$$\|Z_1\|_F \leq \frac{2\eta}{\sqrt{1-\alpha_{2k}}} + \sqrt{\frac{1+\alpha_{2k}}{1-\alpha_{2k}}} \cdot \sum_{i=3}^{m/k} \|Z_i\|_F.$$

To prove Proposition 7, we need the following result:

Fact 4 *(cf. [8, Lemma 3.3]) Let $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ be a given linear operator and $X, Y \in \mathbb{R}^{m\times n}$ be given matrices. Suppose that rank$(X) \le r_X$ and rank$(Y) \le r_Y$ for some integers $r_X, r_Y \ge 1$, and that tr$(X^TY) = 0$. Furthermore, suppose that $\mathcal{A}$ satisfies the RIP of order $r_X + r_Y$ with constant at most $\alpha_{r_X+r_Y}$. Then,*

$$\left|\mathcal{A}(X)^T\mathcal{A}(Y)\right| \le \alpha_{r_X+r_Y}\|X\|_F\|Y\|_F.$$

Proof of Proposition 7 We compute

$$\begin{aligned}
(1-\alpha_{2k})\left(\|Z_1\|_F^2 + \|Z_2\|_F^2\right) &= (1-\alpha_{2k})\|Z_1+Z_2\|_F^2 && (27)\\
&\le \|\mathcal{A}(Z_1+Z_2)\|_2^2 && (28)\\
&= \left\|\mathcal{A}(Z) - \sum_{i=3}^{m/k}\mathcal{A}(Z_i)\right\|_2^2 \\
&\le \left(2\eta + \left\|\sum_{i=3}^{m/k}\mathcal{A}(Z_i)\right\|_2\right)^2, && (29)
\end{aligned}$$

where (27) follows from the fact that $\mathrm{tr}(Z_1^TZ_2) = 0$; (28) follows from the assumption on $\mathcal{A}$ and the fact that $\mathrm{rank}(Z_1+Z_2) \le 2k$; (29) follows from the fact that $\|\mathcal{A}(Z)\|_2 \le \|\mathcal{A}(X^*)-y\|_2+\|z\|_2 \le 2\eta$. In particular, we have

$$\|Z_1\|_F \le \frac{2\eta}{\sqrt{1-\alpha_{2k}}} + \frac{1}{\sqrt{1-\alpha_{2k}}}\left\|\sum_{i=3}^{m/k}\mathcal{A}(Z_i)\right\|_2. \qquad (30)$$

Now, since $\mathrm{tr}(Z_i^TZ_j) = 0$ for all $i, j \in \{1, \ldots, m/k\}$ with $i \ne j$, we have

$$\begin{aligned}
\left\|\sum_{i=3}^{m/k}\mathcal{A}(Z_i)\right\|_2^2 &= \sum_{i=3}^{m/k}\|\mathcal{A}(Z_i)\|_2^2 + \sum_{3\le i\ne j\le m/k}\mathcal{A}(Z_i)^T\mathcal{A}(Z_j)\\
&\le (1+\alpha_{2k})\sum_{i=3}^{m/k}\|Z_i\|_F^2 + \alpha_{2k}\sum_{3\le i\ne j\le m/k}\|Z_i\|_F\|Z_j\|_F && (31)\\
&\le \sum_{i=3}^{m/k}\|Z_i\|_F^2 + \alpha_{2k}\left(\sum_{i=3}^{m/k}\|Z_i\|_F\right)^2\\
&\le (1+\alpha_{2k})\left(\sum_{i=3}^{m/k}\|Z_i\|_F\right)^2,
\end{aligned}$$

where (31) follows from the fact that $\mathrm{rank}(Z_i) \le k$ for $i = 1, \ldots, m/k$, the assumption on $\mathcal{A}$, and Fact 4. Upon combining the above inequality with (30), the proof is completed. □

To proceed, suppose that $\sigma_{2k+1}(Z) = 0$. Then, we have $Z_i = \mathbf{0}$ for $i = 3, \ldots, m/k$. Hence, Proposition 7 yields $\|Z_1\|_F \leq 2\eta/\sqrt{1-\alpha_{2k}}$. Using (26), we compute

$$\|Z_1\|_p^p \leq k^{1-p/2}\|Z_1\|_F^p \leq k^{1-p/2}\left(\frac{2\eta}{\sqrt{1-\alpha_{2k}}}\right)^p.$$

This, together with the fact that $\|Z\|_p^p = \|Z_1\|_p^p + \|Z_2\|_p^p \leq 2\|Z_1\|_p^p$, implies that

$$\|X^* - \bar{X}\|_p = \|Z\|_p \leq 2^{1/p}\|Z_1\|_p \leq \frac{2(2k)^{1/p}\eta}{\sqrt{k(1-\alpha_{2k})}}. \tag{32}$$

On the other hand, suppose that $\sigma_{2k+1}(Z) > 0$. Then, we have $Z_3 \neq \mathbf{0}$. Nevertheless, it can be shown that $\sum_{i=3}^{m/k}\|Z_i\|_F$ is not too large.

Proposition 8 *The following holds:*

$$\sum_{i=3}^{m/k}\|Z_i\|_F < \frac{m\|Z_1+Z_2+Z_3\|_p}{\sqrt{k}(2k+1)^{1/p}}.$$

Proof of Proposition 8 On one hand, we have

$$\sum_{i=3}^{m/k}\|Z_i\|_F \leq \left(\frac{m}{k}-2\right)\|Z_3\|_F \leq \left(\frac{m}{k}-2\right)\sqrt{k}\,\sigma_{2k+1}(Z) < \frac{m}{\sqrt{k}}\sigma_{2k+1}(Z).$$

On the other hand, we have

$$\begin{aligned}
\|Z_1+Z_2+Z_3\|_p &= \sigma_{2k+1}(Z)\left(\sum_{i=1}^{3k}\left(\frac{\sigma_i(Z)}{\sigma_{2k+1}(Z)}\right)^p\right)^{1/p} \\
&\geq \sigma_{2k+1}(Z)\left(\sum_{i=1}^{2k+1}\left(\frac{\sigma_i(Z)}{\sigma_{2k+1}(Z)}\right)^p\right)^{1/p} \\
&\geq (2k+1)^{1/p}\sigma_{2k+1}(Z).
\end{aligned}$$

The desired result then follows by combining the above inequalities. □

Now, by Propositions 7 and 8 and (26), we have

$$\begin{aligned}
\|Z_1\|_p &\leq \frac{k^{1/p}}{\sqrt{k}}\|Z_1\|_F \\
&\leq \frac{k^{1/p}}{\sqrt{k}}\left(\frac{2\eta}{\sqrt{1-\alpha_{2k}}} + \sqrt{\frac{1+\alpha_{2k}}{1-\alpha_{2k}}}\cdot\sum_{i=3}^{m/k}\|Z_i\|_F\right) \\
&< \frac{k^{1/p}}{\sqrt{k}}\left(\frac{2\eta}{\sqrt{1-\alpha_{2k}}} + \sqrt{\frac{1+\alpha_{2k}}{1-\alpha_{2k}}}\cdot\frac{m\|Z_1+Z_2+Z_3\|_p}{\sqrt{k}(2k+1)^{1/p}}\right).
\end{aligned} \tag{33}$$

It is easy to verify that for all $p \in (0, \bar{p}]$, where $\bar{p}$ is defined in (25), we have

$$\sqrt{\frac{1+\alpha_{2k}}{1-\alpha_{2k}}} \cdot \frac{m}{\sqrt{k}(2k+1)^{1/p}} \leq \frac{1}{2(2k)^{1/p}}. \tag{34}$$

Moreover, similar to the proof of Proposition 5, we compute

$$\begin{aligned}
\sum_{i=1}^{m} \sigma_i^p(\bar{X}) &\geq \sum_{i=1}^{m} \sigma_i^p(X^*) && (35)\\
&= \sum_{i=1}^{m} \sigma_i^p(\bar{X} + Z) \\
&\geq \sum_{i=1}^{m} \left|\sigma_i^p(\bar{X}) - \sigma_i^p(Z)\right| && (36)\\
&\geq \sum_{i=1}^{k} \sigma_i^p(\bar{X}) - \sum_{i=1}^{k} \sigma_i^p(Z) + \sum_{i=k+1}^{m} \sigma_i^p(Z), && (37)
\end{aligned}$$

where (35) follows from the fact that $\bar{X}$ and X^* are feasible and optimal for Problem (2), respectively; (36) follows from the perturbation inequality (3); (37) follows from the fact that $\mathrm{rank}(\bar{X}) \leq k$. This yields

$$\|Z_1\|_p^p \geq \sum_{i=2}^{m/k} \|Z_i\|_p^p. \tag{38}$$

Upon substituting (34) and (38) into (33), we obtain

$$\begin{aligned}
\|Z_1\|_p &< \frac{k^{1/p}}{\sqrt{k}} \left(\frac{2\eta}{\sqrt{1-\alpha_{2k}}} + \frac{\|Z_1 + Z_2 + Z_3\|_p}{2(2k)^{1/p}} \right) \\
&\leq \frac{k^{1/p}}{\sqrt{k}} \left[\frac{2\eta}{\sqrt{1-\alpha_{2k}}} + \frac{1}{2(2k)^{1/p}} \left(\|Z_1\|_p^p + \sum_{i=2}^{m/k} \|Z_i\|_p^p \right)^{1/p} \right] \\
&\leq \frac{k^{1/p}}{\sqrt{k}} \left(\frac{2\eta}{\sqrt{1-\alpha_{2k}}} + \frac{\|Z_1\|_p}{2k^{1/p}} \right),
\end{aligned}$$

which is equivalent to

$$\|Z_1\|_p < \left(1 - \frac{1}{2\sqrt{k}}\right)^{-1} \frac{2k^{1/p}\eta}{\sqrt{k(1-\alpha_{2k})}}. \tag{39}$$

Since $\|Z\|_p^p = \|Z_1\|_p^p + \cdots + \|Z_{m/k}\|_p^p$, it follows from (38) and (39) that

$$\|X^* - \bar{X}\|_p = \|Z\|_p \leq 2^{1/p}\|Z_1\|_p < \left(1 - \frac{1}{2\sqrt{k}}\right)^{-1} \frac{2(2k)^{1/p}\eta}{\sqrt{k(1-\alpha_{2k})}}. \tag{40}$$

Upon comparing (32) and (40), the desired result follows. □

5 Conclusion

In this paper, we established the perturbation inequality (4) concerning concave functions of the singular values of a matrix. Such an inequality has proven to be fundamental in understanding the recovery properties of the Schatten p–quasi–norm heuristic. Thus, a natural future direction is to find other applications of (4) in the study of low–rank matrix recovery. In particular, it would be interesting to extend Theorem 4 to the noisy recovery setting. Another direction is to consider random linear operators $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ and determine the number of measurements l needed to guarantee the recovery of a low–rank matrix via the Schatten p–quasi–norm heuristic; cf. related results for the nuclear norm heuristic in [36, 10]. Lastly, it would be interesting to prove or disprove the following generalization of (4), which has already attracted some attention in the linear algebra community:

Conjecture 1 *([3, Conjecture 6]) Let $A, B \in \mathbb{R}^{m\times n}$ be given matrices. Suppose that $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a concave function satisfying $f(0) = 0$. Then, for any $k \in \{1, \ldots, \min\{m, n\}\}$,*

$$\sum_{i=1}^{k} \left| f(\sigma_i(A)) - f(\sigma_i(B)) \right| \leq \sum_{i=1}^{k} f(\sigma_i(A - B)).$$

An immediate idea for proving Conjecture 1 would be to adapt our proof in Section 3. As an intermediate step, one would need to establish a bound similar to that in Theorem 2 on $\sum_{i=1}^{k} f(\sigma_i(M + tN))$, where $M, N \in \mathcal{S}^n$, $k \in \{1, \ldots, n\}$, and $t > 0$ are arbitrary. However, it is not clear how to generalize the proof of Theorem 2 to achieve this. Indeed, in the notation of that proof, if there exists an index $j \in \{0, 1, \ldots, l\}$ such that $i_j \leq k < i_{j+1} - 1$, then as we sum the inequality (14) over $i = 1, \ldots, k$, we will encounter the sum of the $k - i_j + 1$ largest eigenvalues of $(Q^j)^T \Xi(N) Q^j$, which does not seem to admit a simple analytical form. This should be contrasted with the sum of *all* eigenvalues of $(Q^j)^T \Xi(N) Q^j$, which can be expressed simply as $\mathrm{tr}\left((Q^j)^T \Xi(N) Q^j\right)$. Thus, we believe that some new ideas will be needed to settle Conjecture 1 in the affirmative.

Remark: After the submission of our manuscript, Audenaert [2] has resolved Conjecture 1 in the affirmative. His proof employs the so–called Thompson–Freede inequalities [41] for the eigenvalues of sums of Hermitian matrices and the proof technique is quite different from ours. We refer the reader to [2] for details.

Acknowledgements

We thank the two anonymous referees for their insightful comments. We would also like to thank Liang Chen and Defeng Sun for pointing out an inaccuracy in an earlier version of Theorem 3 and for suggesting Example 1. This research is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project CUHK 416413 and in part by a gift grant from Microsoft Research Asia.

References

[1] T. Ando. Comparison of Norms $|||f(A) - f(B)|||$ and $|||f(|A - B|)|||$. *Mathematische Zeitschrift*, 197(3):403–409, 1988.

[2] K. M. R. Audenaert. A Generalisation of Mirsky's Singular Value Inequalities. Manuscript, available at `http://arxiv.org/abs/1410.4941`, 2014.

[3] K. M. R. Audenaert and F. Kittaneh. Problems and Conjectures in Matrix and Operator Inequalities. Manuscript, available at `http://arxiv.org/abs/1201.5232`, 2012.

[4] R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer–Verlag New York, Inc., New York, 1997.

[5] J.-C. Bourin and M. Uchiyama. A Matrix Subadditivity Inequality for $f(A+B)$ and $f(A)+f(B)$. *Linear Algebra and Its Applications*, 423(2–3):512–518, 2007.

[6] T. T. Cai and A. Zhang. Sharp RIP Bound for Sparse Signal and Low–Rank Matrix Recovery. *Applied and Computational Harmonic Analysis*, 35(1):74–93, 2013.

[7] E. Candès and B. Recht. Simple Bounds for Recovering Low–Complexity Models. *Mathematical Programming, Series A*, 141(1–2):577–589, 2013.

[8] E. J. Candès and Y. Plan. Tight Oracle Inequalities for Low–Rank Matrix Recovery from a Minimal Number of Noisy Random Measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[9] E. J. Candès and T. Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[10] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[11] R. Chartrand and V. Staneva. Restricted Isometry Properties and Nonconvex Compressive Sensing. *Inverse Problems*, 24(3):Article 035020, 2008.

[12] P. Chen and D. Suter. Recovering the Missing Components in a Large Noisy Low–Rank Matrix: Application to SFM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1051–1063, 2004.

[13] M. Fazel, H. Hindi, and S. P. Boyd. A Rank Minimization Heuristic with Application to Minimum Order System Approximation. In *Proceedings of the 2001 American Control Conference*, pages 4734–4739, 2001.

[14] M. Fiedler. Bounds for the Determinant of the Sum of Hermitian Matrices. *Proceedings of the American Mathematical Society*, 30(1):27–31, 1971.

[15] W. Fulton. Eigenvalues, Invariant Factors, Highest Weights, and Schubert Calculus. *Bulletin (New Series) of the American Mathematical Society*, 37(3):209–249, 2000.

[16] D. Ge, X. Jiang, and Y. Ye. A Note on the Complexity of L_p Minimization. *Mathematical Programming, Series B*, 129(2):285–299, 2011.

[17] R. Gribonval and M. Nielsen. Sparse Representations in Unions of Bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

[18] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and Accurate Matrix Completion via Truncated Nuclear Norm Regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2013.

[19] A. Javanmard and A. Montanari. Localization from Incomplete Noisy Distance Measurements. *Foundations of Computational Mathematics*, 13(3):297–345, 2013.

[20] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust Video Denoising Using Low Rank Matrix Completion. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 1791–1798, 2010.

[21] S. Ji, K.-F. Sze, Z. Zhou, A. M.-C. So, and Y. Ye. Beyond Convex Relaxation: A Polynomial–Time Non–Convex Optimization Approach to Network Localization. In *Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM 2013)*, pages 2499–2507, 2013.

[22] A. Juditsky, F. K. Karzan, and A. Nemirovski. On a Unified View of Nullspace–Type Conditions for Recoveries Associated with General Sparsity Structures. *Linear Algebra and Its Applications*, 441:124–151, 2014.

[23] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear–Norm Penalization and Optimal Rates for Noisy Low–Rank Matrix Completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[24] L. Kong and N. Xiu. Exact Low–Rank Matrix Recovery via Nonconvex Schatten p–Minimization. *Asia–Pacific Journal of Operational Research*, 30(3), 2013.

[25] M.-J. Lai, S. Li, L. Y. Liu, and H. Wang. Two Results on the Schatten p–Quasi–Norm Minimization for Low–Rank Matrix Recovery. Manuscript, 2012.

[26] M.-J. Lai, Y. Xu, and W. Yin. Improved Iteratively Reweighted Least Squares for Unconstrained Smoothed ℓ_q Minimization. *SIAM Journal on Numerical Analysis*, 51(2):927–957, 2013.

[27] A. S. Lewis and H. S. Sendov. Nonsmooth Analysis of Singular Values. Part II: Applications. *Set–Valued Analysis*, 13(3):243–264, 2005.

[28] G. Marjanovic and V. Solo. On l_q Optimization and Matrix Completion. *IEEE Transactions on Signal Processing*, 60(11):5714–5724, 2012.

[29] L. Mirsky. Symmetric Gauge Functions and Unitarily Invariant Norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[30] B. K. Natarajan. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[31] S. Negahban and M. J. Wainwright. Estimation of (Near) Low–Rank Matrices with Noise and High–Dimensional Scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

[32] F. Nie, H. Huang, and C. Ding. Low–Rank Matrix Recovery via Efficient Schatten p–Norm Minimization. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12)*, pages 655–661, 2012.

[33] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A Simplified Approach to Recovery Conditions for Low Rank Matrices. In *Proceedings of the 2011 IEEE International Symposium on Information Theory (ISIT 2011)*, pages 2318–2322, 2011.

[34] Z. Pan and C. Zhang. Relaxed Sparse Eigenvalue Conditions for Sparse Estimation via Non–Convex Regularized Regression. *Pattern Recognition*, 48(1):231–243, 2015.

[35] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed Minimum–Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010.

[36] B. Recht, W. Xu, and B. Hassibi. Null Space Conditions and Thresholds for Rank Minimization. *Mathematical Programming, Series B*, 127(1):175–202, 2011.

[37] H. L. Royden. *Real Analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[38] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, Princeton, New Jersey, 2006.

[39] G. W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, Boston, 1990.

[40] T. Strohmer. Measure What Should be Measured: Progress and Challenges in Compressive Sensing. *IEEE Signal Processing Letters*, 19(12):887–893, 2012.

[41] R. C. Thompson and L. J. Freede. On the Eigenvalues of Sums of Hermitian Matrices. *Linear Algebra and Its Applications*, 4(4):369–376, 1971.

[42] M. Wang, W. Xu, and A. Tang. On the Performance of Sparse Recovery via ℓ_p–Minimization $(0 \leq p \leq 1)$. *IEEE Transactions on Information Theory*, 57(11):7255–7278, 2011.

[43] J. Wen, D. Li, and F. Zhu. Stable Recovery of Sparse Signals via l_p–Minimization. *Applied and Computational Harmonic Analysis*, 38(1):161–176, 2015.

[44] R. Wu and D.-R. Chen. The Improved Bounds of Restricted Isometry Constant for Recovery via ℓ_p–Minimization. *IEEE Transactions on Information Theory*, 59(9):6142–6147, 2013.

[45] M. Zhang, Z.-H. Huang, and Y. Zhang. Restricted p–Isometry Properties of Nonconvex Matrix Recovery. *IEEE Transactions on Information Theory*, 59(7):4316–4323, 2013.

[46] Y. Zhang and L. Qiu. From Subadditive Inequalities of Singular Values to Triangle Inequalities of Canonical Angles. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1606–1620, 2010.