Optimal Exploitation of Subspace Prior Information in Matrix Sensing

Sajad Daei, Farzad Haddadi, Arash Amini

Abstract—Low-rank matrix recovery is the problem of reconstructing a low-rank matrix from a few linear measurements. In many applications such as collaborative filtering, the famous Netflix prize problem and seismic data interpolation, there exists some prior information about the column and row spaces of the true low rank matrix. In this paper, we exploit this prior information by proposing a weighted optimization problem where its objective function promotes both rank and prior subspace information. Using the recent results in conic integral geometry, we obtain the unique optimal weights that minimize the required number of measurements. As simulation results confirm, the proposed convex program with optimal weights substantially needs fewer measurements than the regular nuclear norm minimization.

Index Terms—Matrix sensing, subspace prior information, conic integral geometry.

I. INTRODUCTION

Low-rank matrix recovery (also known as matrix sensing) has come up in a great number of applications in recent years. For example, Netflix prize problem [1], [2], collaborative filtering [3], seismic data interpolation [4, 5], system identification [6], sensor network localization [7]. Mathematically, it can be stated as recovering a low rank matrix $X \in \mathbb{R}^{n_1 \times n_2}$ with rank $r \ll \min\{n_1, n_2\}$ from a few linear measurements of the form $y = A(X)$ where $A : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^m$ is a linear operator. A hypothetical approach is the following optimization problem:

$$\begin{align*}
\min_{Z \in \mathbb{R}^{n_1 \times n_2}} & \quad \text{rank}(Z) \\
\text{s.t.} & \quad y = A(Z),
\end{align*}$$

However, this problem is NP-hard and is computationally intractable. A common alternative is to relax the objective function into the closest convex function. In fact, since rank is the number of nonzero elements of the singular value vector, its convex relaxation amounts to $\ell_1$ norm of this vector known as the nuclear norm of the matrix. Then, one may solve the following convex problem:

$$\begin{align*}
P_{\text{nuc}} & : \min_{Z \in \mathbb{R}^{n_1 \times n_2}} \|Z\|_* \\
\text{s.t.} & \quad y = A(Z),
\end{align*}$$

where $\|\cdot\|_*$ computes the sum of singular values. The problem also has a very special (and important) case that seeks to complete $X$ from a few observed entries known as matrix completion:
directions than true directions\footnote{In fact, inaccurate directions would be penalized more.} leading to a reduction in the required number of measurements. Now, let us go back to the matrix world. Consider a matrix $X$ that lives in a union of row and column subspaces denoted by $T$. Suppose that we are given a subspace $\tilde{T}$ that is slightly angled from $T$. Can we hope for a reduction in the required number of measurements with more penalizing orthogonal complement of $\tilde{T}$? Are the parallels still strong?

**B. Notation**

Throughout the page, scalars are denoted by lowercase letters, vectors by lowercase boldface letters, and matrices by uppercase boldface letters. The $i$th element of the vector $x$ is given either by $x_i$ or $x_i, (\cdot)^\dagger$ denotes pseudo inverse operator. $I_n$ is the identity matrix of size $n \times n$. The complement of an event $\mathcal{E}$ is shown by $\bar{\mathcal{E}}$. The nullspace of linear operators is denoted by $\null()$. For a matrix $A$, the operator norm is defined as $\|A\|_{p \to q} = \sup_{\|x\|_p \leq 1} \|Ax\|_q$. The unit ball and unit sphere are shown by $B^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ and $\mathbb{B}^n := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, respectively. Also, we have $\mathbb{B}_e^{n \times n} := \{Z \in \mathbb{R}^{n \times n} : \|Z\|_F \leq \epsilon\}$ which refers to the $\epsilon$-ball of matrices. Consider a matrix $X \in \mathbb{R}^{n \times n}$ with reduced SVD form $X = U_n \Sigma_{n \times r} V_n^H$. Define $\mathcal{U} := \text{span}(U)$ and $\mathcal{V} := \text{span}(V)$. We denote the matrix $UV^H$ by the notation $\text{sgn}(X)$. Also define the support of $X$ by the linear subspace $T = \{Z \in \mathbb{R}^n : Z = P_T ZP_V + P_{T^\perp} ZP_{V^\perp} + P_{T^\perp} ZP_{V^\perp}\} := \text{supp}(X)$, (4) where $P_T := UU^H$ and $P_V := VV^H$ are unique orthogonal projections onto $\mathcal{U}$ and $\mathcal{V}$, respectively. $P_T(Z)$ and $P_{T^\perp}(Z)$ are the projection of matrix $Z$ onto the linear subspace $T$ and $T^\perp$, respectively and are defined as $P_T(Z) := P_T ZP_V + P_{T^\perp} ZP_{V^\perp} + P_{T^\perp} ZP_{V^\perp}$, $P_{T^\perp}(Z) := P_{T^\perp} ZP_{V^\perp}$. Also, we denote projection onto a cone $\mathcal{C}$ with the same notation but different definition. Namely, $P_{\mathcal{C}}(X) := \text{arg min}_{Z \in \mathcal{C}} \|Z - X\|_F$. (5) The polar of a cone $\mathcal{C}$ is defined as $C^o = \{z : \langle z, x \rangle \leq 0 \ \forall x \in \mathcal{C}\}$. $x \in [a, b]^n$ for a vector $x \in \mathbb{R}^n$ means that $a \leq x_i \leq b$, $i = 1, \ldots, n$. Also, by $x \in [a, b]^n$, we mean $a < x_i < b$, $i = 1, \ldots, n$diag$(x)$ is a diagonal matrix with its main diagonal determined by the elements of $x$. For a function $f : \mathbb{R}^{p \times n} \rightarrow \mathbb{R}$, $f^*$ means the adjoint of the function $f$. $\sigma(A) \in \mathbb{R}^n$ denotes the singular values of $A$ sorted non-increasingly. $(a)_+$, $a \vee b$ and $a \wedge b$ denote max$(a, 0)$, max$(a, b)$ and min$(a, b)$. $\langle A, B \rangle_F = tr(AB^H)$ denotes Frobenius inner product of two conforming matrices $A$ and $B$.

**C. Contributions**

In this work, we propose a new approach for optimal exploitation of prior subspace information leading to a considerable reduction in the required number of measurements.

Consider a rank $r$ matrix $X \in \mathbb{R}^{n \times n}$ with column and row subspaces $\mathcal{U}$ and $\mathcal{V}$. Assume that we are given two subspaces $\mathcal{U}$ and $\mathcal{V}$, each with dimension $r' \geq r$, that are angled from $\mathcal{U}$ and $\mathcal{V}$. Let $\theta_u \in [0, 90^\circ]$ and $\theta_v \in [0, 90^\circ]$ be the known principle angles that $\mathcal{U}$ and $\mathcal{V}$ form with $\mathcal{U}$ and $\mathcal{V}$, respectively. We implicitly take these prior subspace information into account by proposing the following optimization problem.

$$P_{w, \text{nuc}} : \min_{Z \in \mathbb{R}^{n \times n}_2} \|h_w(Z)\|_*$$

s.t. $y = A(Z)$ (6)

where,

$$h_w(Z) = w_1 P_{U} ZP_{V} + w_2 P_{U} ZP_{V^\perp} + w_3 P_{U^\perp} ZP_{V} + w_4 P_{U^\perp} ZP_{V^\perp} = \frac{1}{w_3} \left( w_1 P_{U} + w_3 P_{U^\perp} \right) Z \left( w_3 P_{V} + w_4 P_{V^\perp} \right)$$

$w := [w_1, w_2, w_3]^T, \ w_4 := 1/w_1$ (7)

The weights $w_1$ and $w_3$ reflect our uncertainty in the prior column space information. The same argument holds for $w_3$ and $w_4$ in the prior row space information. In this work, we obtain the unique weights that minimize the required number of measurements. These weights are optimal since they minimize the number of measurements that $P_{w, \text{nuc}}$ needs for exact recovery of $X$. To find optimal weights, we exploit the concept of statistical dimension in conic integral geometry. Statistical dimension specifies the boundary of success and failure of $P_{w, \text{nuc}}$. To be precise, we obtain upper and lower bounds with asymptotically vanishing distance for the statistical dimension of a certain convex cone and thereby calculate a threshold $m_0(w, \theta_u, \theta_v)$ for the minimum required number of measurements. Then, we solve the optimization problem

$$\min_{w \in \mathbb{R}^3_+} m_0(w, \theta_u, \theta_v)$$

(8) to reach the optimal weight $w^*$. For completeness and better highlight our contributions, we summarize the novelties of our work below.

1. **Proposing a new optimization model for matrix sensing:** We propose a new convex optimization problem in (6) that promotes both rank and subspace information. The benefits of this model is that by suitably tuning the weights, it always outperforms $P_{\text{nuc}}$ whether the accuracy of subspace prior information is reliable or unreliable. The more probable that $\mathcal{U} \approx \mathcal{U}$, less penalty is assigned to $w_1$ than $w_3$. The same argument also holds for $\mathcal{V}$. If the subspace prior information is at the boundary of reliability and unreliability (i.e. $\theta_u(i) = \theta_v(i) = 45^\circ \forall i = 1, \ldots, r$), then by setting $w_1 = w_2 = w_3 = 1$, $P_{w, \text{nuc}}$ reduces to $P_{\text{nuc}}$.  

2. **Obtaining an upper-bound for the required sample complexity of $P_{W, \theta_u, \theta_v}$:** We obtain a closed-form relation for the sufficient number of measurements that $P_{w, \theta_u, \theta_v}$ needs for successful recovery denoted by $m_{w, \theta_u, \theta_v}$. This bound depends on the weights $w$ and the principal angles $\theta_u, \theta_v$. By setting $w_1 = w_2 = w_3 = 1$, the bound reaches the required sample complexity of $P_{\text{nuc}}$. 


3) Obtaining an error bound estimate for $\hat{\mathbf{m}}_{w, \theta, \theta}$: We prove that the sufficient number of measurements i.e. $\hat{\mathbf{m}}_{w, \theta, \theta}$ is also necessary for successful recovery. To be more precise, we show that $\hat{\mathbf{m}}_{w, \theta, \theta}$ differs from the minimum required number of measurements up to an asymptotically constant term.

4) Proposing a new strategy for finding optimal weights: In the proposed model (6), we obtain the weights $\mathbf{w}$ that minimizes the required sample complexity for exact recovery. If one takes the sample complexity as optimality criterion, then, these weights are optimal. Also, we show that, they are unique up to a positive scaling. Moreover, we propose a simple algorithm (called Optweights) that efficiently returns the unique optimal weights.

5) Obtaining closed-form relations for support $(\mathbf{h}(\mathbf{X}))$ and $\text{sgn}(h_w(X))$: We find that the spaces and sign of $\mathbf{h}(\mathbf{X})$ (i.e. $\text{supp}(h_w(X))$ and $\text{sgn}(h_w(X))$, respectively) are rotated versions of $\mathbf{I}_r$ spaces. To be precise, it holds that

\[
\mathcal{P}_F(Z) = \mathcal{Q}_L\mathcal{P}_{\text{supp}(t)}(\mathcal{I}_r)(\mathcal{Q}_L^H Z \mathcal{Q}_R)\mathcal{Q}_R^H,
\]

\[
\mathcal{P}_{F^\perp}(Z) = \mathcal{Q}_L\mathcal{P}_{\text{supp}(t)}(\mathcal{I}_r^c)(\mathcal{Q}_L^H Z \mathcal{Q}_R)\mathcal{Q}_R^H,
\]

\[
\text{sgn}(h_w(X)) = \mathcal{Q}_L\mathcal{P}_{\text{supp}(t)}(\mathcal{I}_n)\mathcal{Q}_R^H,
\]

for any arbitrary $Z \in \mathbb{R}^{n \times n}$, where $Q_L \in \mathbb{R}^{n \times n}$ and $Q_R \in \mathbb{R}^{n \times n}$ are some orthonormal bases of $\mathbb{R}^{n \times n}$ which explicitly depends on the weights $\mathbf{w}$ and the principal angles $\theta_\alpha, \theta_\beta$.

6) Obtaining the limiting behavior of spectral functions: For any function $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ and any random matrix $\mathbf{G} \in \mathbb{R}^{n_1 \times n_2}$ with i.i.d. Gaussian ensemble, we obtain a closed-form relation for the limiting behavior of $E[f(\sigma(G))]$.

D. Intuition and what we can expect

There is a nice intuition in selecting good weights in $\mathbf{P}_{w, \text{nuc}}$. In fact, if the subspace $\mathbf{U}$ lie close to $\mathbf{U}$, it is better to penalize the orthogonal complement $\mathbf{U}^\perp$ more, i.e. associate larger weight. Conversely, if the subspace $\mathbf{U}$ is very far from $\mathbf{U}$ (close to $\mathbf{U}^\perp$), it is better to associate larger weight to $\mathbf{U}$. Intuitively, we expect from our model, to a great extent, satisfies the following statements.

- When the principle angles $\{\theta(i)\}_{i=1}^r$ between $\mathbf{U}$ and $\mathbf{U}$, with dimensions $r$ and $r'$ respectively, are all small (close to $0^\circ$), we predict that $\mathbf{w}$ is large and $w_1$ small. Further, if $r' \approx r$, it is expected that the required number of measurements for $\mathbf{P}_{w, \text{nuc}}$ approach the optimal number of measurements i.e. $r^2$.
- When the principle angles $\{\theta(i)\}_{i=1}^r$ are all large (close to $90^\circ$), we predict a large value for $w_1$ and a small value for $w_3$. Again, we predict a reduced number of required measurements for $\mathbf{P}_{w, \text{nuc}}$.
- When $\{\theta(i)\}_{i=1}^r$ are all around $45^\circ$, we predict no improvement in the required number of measurements for $\mathbf{P}_{w, \text{nuc}}$. In fact, we believe that the number of required measurements for $\mathbf{P}_{w, \text{nuc}}$ is almost the same as $\mathbf{P}_{\text{nuc}}$ needs.

- When the angles $\{\theta(i)\}_{i=1}^r$ are such close to $0^\circ$ as they are close to $90^\circ$, again, no improvement is predicted. For instance, consider $\theta = [25^\circ, 45^\circ, 75^\circ]^T$ for $r = 3$.

In this case, $\mathbf{U}$ is as close to $\mathbf{U}$ as to $\mathbf{U}^\perp$. So, it is indistinguishable which subspace to penalize more.

Similar statements also hold for row space prior information $\mathbf{V}$ and $\mathbf{V}^\perp$.

E. Applications

The application of subspace prior information in matrix sensing is very broad; we only listed some of them below.

- The Netflix problem [1]. Some users (represented by the rows of the Netflix matrix) rate a few movies (Columns of the matrix). The problem then is to complete this matrix in order to predict users’ ratings on unobserved movies. It might occur that one has some extra information about the column space of the matrix. For example, consider a case that only the factors scenario, music and actors have main effects on users’ rating and the attribute of each factor is somehow available for some users or that some users share the same attributes. As another example, consider movies that has been evaluated before release by some professional reviewers [2]. This provides us with a higher dimensional subspace [3] that has known principle angles with column subspace of the true matrix. Row subspace prior information is also imaginable. Consider for example two persons that have known common interests in the importance of each factor in ratings.
- Subspace tracking. Here, the aim is to recover a subspace $\mathbf{U}$ from only few vectors inside it. Consider a case that the intended subspace is estimated in previous stages. This provides one with a subspace $\mathbf{U}$ that is slightly deviated from $\mathbf{U}$.
- Dynamic sensor network localization [4]. Consider a network of low-power dynamic sensors scattered in an area forming a distance matrix [5]. The aim is to find the current position given only a few distance measurements. If some sensors know their previous positions, it gives some prior information about row and column spaces of the current distance matrix and might be used to reduce the required number of measurements in the current state.
- Time-varying channel estimation in FDD [6] massive MIMO [7] [12]. Some single antenna users send a few training sequences to a multi antenna base station. The aim is to estimate the users’ channel matrix that is correlated and thus low-rank in a coherence time-bandwidth. Moreover, consider a case that one has some additional information from channel estimates in previous coherence blocks. By taking singular value decomposition (SVD) of the channel matrix, one has access to the row and column subspaces. Also, it is possible to estimate the

\footnotesize
3 This case is very common in Rotten Tomato review aggression or in film festivals.
4 The context that we consider in this work.
5 This matrix is formed of distances between sensors.
6 Frequency Division Multiplexing
7 Multiple Input Multiple Output
angles between previous and current subspaces by using the statistics of previous data. This can help much in reducing the number of training sequences and thus the overhead of downlink pilot signaling.

F. Roadmap

The paper is organized as follows. A more clear definition of principal angles between subspaces besides a few concepts from convex geometry are reviewed in Section II. Section III is dedicated to obtaining bounds for the required number of measurements in $P_{\text{nuc}}$ and $P_{\text{w,nuc}}$. Section IV is about our strategy of finding optimal weights. In Section V, we present some numerical experiments which validate our theory. In Section VI, we shall describe related works in Section VII. Section VIII is devoted to important lemmas that frequently used in our analysis. Lastly, the paper is concluded in Section VIII.

II. PRELIMINARIES

A. Principal angles between subspaces

Consider two subspaces $U$ and $W$ of an Euclidean vector space $\mathbb{R}^n$ with $\dim(U) := r \leq \dim(W) := r'$. There exist $r$ non increasingly sorted angles $\theta := [\theta(1), ..., \theta(r)]^T \in [0^\circ, 90^\circ]^r$ called the principal angles, the least one is obtained by:

$$\theta(r) := \min \left\{ \cos^{-1} \left( \frac{\langle u, w \rangle}{\|u\|_2 \|w\|_2} \right) : u \in U, w \in W \right\}$$

$$= \angle(u_r, w_r)$$

(9)

The $i$th one ($i < r$) is given by:

$$\theta(i) := \min \left\{ \cos^{-1} \left( \frac{\langle u, w \rangle}{\|u\|_2 \|w\|_2} \right) : u \in U, w \in W \right\}$$

$$, u \perp u_j, w \perp w_j \forall j \in \{i + 1, ..., r\}$$

$$= \angle(u_i, w_i)$$

(10)

The set of principal angles is denoted by $\Theta = \{\theta(1), ..., \theta(r)\}$. Moreover, each subspace $U, W$ is spanned by a set of linearly independent vectors. In fact, there exist orthonormal bases $U := [u_1, ..., u_r] \in \mathbb{R}^{n \times r}$ and $V := \{w_1, ..., w_{r'}, w_{r'+1}, ..., w_{r'}\} \in \mathbb{R}^{n \times r'}$ for subspaces $U$ and $W$, respectively. Also,

$$U = \text{span}(U)$$

$$W = \text{span}(W)$$

$$U^*W = [\cos(\theta) \ 0_{r \times r'-r}]$$

where

$$\cos(\theta) := \text{diag}(\{\cos(\theta(1)), \cos(\theta(2)), ..., \cos(\theta(1))\}) \in \mathbb{R}^{r \times r}$$

(12)

In the following, basic concepts of convex geometry are reviewed.

B. Descent Cones

The descent cone $\mathcal{D}(f, x)$ at a point $x \in \mathbb{R}^n$ consists of the set of directions that do not increase $f$ and is given by:

$$\mathcal{D}(f, x) = \bigcup_{t \geq 0} \{z \in \mathbb{R}^n : f(x + tz) \leq f(x)\}$$

(13)

Fig. 1. Principal angles and vectors in a three-dimensional Euclidean space. In this Figure, the subspaces $W$ and $W^+$ form principal angles $\theta = [\theta_1, \theta_2]$ and $\alpha$ with the subspace $U$. $u_1$, $w_1$ and $w_2$ are the corresponding principal vectors. Also, $\theta_2 = 0, u_2$ and $w_2$ (which are not depicted) are in the line $U \cap W$ in the same direction.

C. Statistical Dimension

Definition 1. Statistical Dimension [14]: Let $C \subseteq \mathbb{R}^n$ be a convex closed cone. Statistical dimension of $C$ is defined as:

$$\delta(C) := E[\|P_C(g)\|_2^2] = \text{Edist}^2(g, C)$$

(15)

where, $g$ is i.i.d. standard normal vector and $P_C(x)$ is the projection of $x \in \mathbb{R}^n$ onto the set $C$ defined as: $P_C(x) = \arg \min \{z \in C : \|z - x\|_2\}$.

Statistical dimension extends the concept of linear subspaces to convex cones. Intuitively, it measures the size of a cone. Furthermore,

$$\delta(\mathcal{D}(f, x)) := E \inf_{t \geq 0} \inf_{z \in \mathcal{D}(f, x)} \|g - tz\|_2^2$$

(16)

determines the precise number of measurements corresponding to transition from failure to success in $P_f$.

D. Optimality Condition

In the following, we characterize when $P_f$ succeeds in the noise-free case.

Proposition 1. [15] Proposition 2.1]: Optimality condition: Let $f$ be a proper convex function. The vector $x \in \mathbb{R}^n$ is the
unique optimal point of $\mathcal{P}_f$ if and only if $\mathcal{D}(f, x) \cap \text{null}(A) = \{0\}$.

The next theorem determines the number of measurements needed for successful recovery of $\mathcal{P}_f$ for any proper convex function $f$.

**Theorem 1.** ([14] Theorem 2) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function and $x \in \mathbb{R}^n$ a fixed sparse vector. Suppose that $m$ independent Gaussian linear measurements of $x$ are observed via the affine constraint $y = Ax \in \mathbb{R}^m$. Then for a given tolerance $\eta \in [0, 1]$ if

$$m \geq \delta(\mathcal{D}(f, x)) + \sqrt{8 \log(\frac{4}{\eta})} n$$

we have:

$$\mathbb{P}(\mathcal{D}(f, x) \cap \text{null}(A) = \{0\}) \geq 1 - \eta.$$

Besides, if

$$m \leq \delta(\mathcal{D}(f, x)) - \sqrt{8 \log(\frac{4}{\eta})} n$$

then,

$$\mathbb{P}(\mathcal{D}(f, x) \cap \text{null}(A) = \{0\}) \leq \eta.$$

Also in [14], the following error bound for the statistical dimension is provided:

**Theorem 2.** ([14] Theorem 4.3) For any $x \in \mathbb{R}^n \setminus \{0\}$:

$$0 \leq \inf_{t \geq 0} \mathbb{E} \text{dist}^2(g, \tilde{t} \hat{f}(x)) - \delta(\mathcal{D}(f, x)) \leq \frac{2 \sup_{x \in \hat{f}(x)} \|s\|_2}{f(\frac{x}{\|x\|^2})}.$$

### III. The Measurement Threshold for Successful Recovery

Fix a probability of failure $\eta \in [0, 1]$. Denote the normalized number of measurements that $\mathcal{P}_{\text{nuc}}$ and $\mathcal{P}_{\text{w, nuc}}$ need for exact recovery of a matrix $X \in \mathbb{R}^{n \times n}$ by

$$m_{\text{nuc}} := \frac{\delta(\mathcal{D}(\|\cdot\|_*, X))}{n^2},$$

$$m_{\text{w, nuc}} := \frac{\delta(\mathcal{D}(\|h_\text{w}(-)\|_*, X))}{n^2}.$$

respectively. In [14] Proposition 4.7, an upper-bound for $m_{\text{nuc}}$ is provided. To facilitate the calculations, we obtain an upper-bound for $m_{\text{nuc}}$ in harmony with our strategy of finding optimal weights in this work. The proposed upper-bound asymptotically equals the upper-bound in [14] Proposition 4.7.

**Proposition 2.** Consider a matrix $X \in \mathbb{R}^{n \times n}$ with rank $r$. Suppose that $r, r', n \to \infty$ with limiting ratios $\sigma_1 := \frac{r}{n}$ and $\sigma_2 := \frac{r'}{n}$ with $r' \geq r$. Then,

$$m_{\text{nuc}} \to \tilde{m}_{\text{nuc}}$$

for

$$\tilde{m}_{\text{nuc}} := \inf_{t \geq 0} \Psi(\sigma_1, \sigma_2)$$

with

$$\Psi(\sigma_1, \sigma_2) = 3\sigma_1^2 + 2\sigma_1^2 \int_{0}^{1} (u - \frac{t}{\sqrt{r'}})^2 \frac{\sqrt{4 - u^2}}{\pi} du + 2\sigma_1 (1 - \sigma_1 - \sigma_2) \int_{b_2}^{n_{\text{us}}} (u - \frac{t}{\sqrt{r'} + (n - r')} \vee (n - r'))^2 \frac{\sqrt{4 - u^2}}{\pi} du + 2\sigma_1 (1 - \sigma_1 - \sigma_2) \int_{b_2}^{n_{\text{us}}} (\frac{t}{\sqrt{r'}})^2 \frac{\sqrt{4 - u^2}}{\pi} du + (1 - \sigma_1 - \sigma_2)^2 \int_{0}^{t} (u - \frac{t}{\sqrt{r'}})^2 \frac{\sqrt{4 - u^2}}{\pi} du$$

where

$$s_1 = \frac{r \wedge (r' - r)}{r \vee (r' - r)}$$

$$s_2 = \frac{r \wedge (n - r - r')}{r \vee (n - r - r')}$$

$$s_3 = \frac{(r' - r) \wedge (n - r - r')}{(r' - r) \vee (n - r - r')}$$

$$b_1 = 1 - \sqrt{s_1}$$

$$b_2 = 1 - \sqrt{s_2}$$

$$b_3 = 1 - \sqrt{s_3}$$

$$u_{b_1} = 1 + \sqrt{s_1}$$

$$u_{b_2} = 1 + \sqrt{s_2}$$

$$u_{b_3} = 1 + \sqrt{s_3}.$$

### Proof.

*Discussion.* In [14] Equation 4.8 an upper-bound is derived for $m_{\text{nuc}}$. Here, we compare our bound i.e. $\tilde{m}_{\text{nuc}}$ with theirs. Define the difference between $\tilde{m}_{\text{nuc}}$ and the upper-bound in [14] Equation 4.8 by $E_1$. From Figure 3 it seems that the
error $E_1$ is negligible when $r'$ is not far from $r$ which is common in practice. Moreover, since the upper-bound \[ \text{Equation 4.8} \] describes $m_{nuc}$ well (the error is at most $\frac{n}{\sqrt{n}}$), regarding Figure 3 one can infer that $\hat{m}_{nuc}$ also approximate $m_{nuc}$ suitably up to an asymptotically vanishing error term.

In what follows, we obtain an upper-bound for $m_{w, \theta_u, \theta_v}$. This bound helps us to find the optimal weights later. The strategy of providing this bound is, to some extent, similar to Proposition 2. However, the derivation is more elaborate; in fact, this bound, unlike $\hat{m}_{nuc}$, depends on the principal angles $\theta_u, \theta_v \in [0, 90)^\circ$ and the weights $w$ making it more involved.

**Proposition 3.** Consider a rank $r$ matrix $X \in \mathbb{R}^{n \times n}$ with column and row subspaces $\mathcal{U}$ and $\mathcal{V}$, respectively. Also, assume that we are given the subspaces $\mathcal{U} \subseteq \mathbb{R}^n$ and $\mathcal{V} \subseteq \mathbb{R}^n$ with dimension $r' \geq r$ that have known principal angles $\theta_u \in [0^\circ, 90^\circ]^\circ$ and $\theta_v \in [0^\circ, 90^\circ]^\circ$ with $\mathcal{U}$ and $\mathcal{V}$, respectively. Then,

$$m_{w, \theta_u, \theta_v} \leq \hat{m}_{w, \theta_u, \theta_v}$$

for

$$\hat{m}_{w, \theta_u, \theta_v} := \inf_{i > 0} \frac{\Psi_i(w, \theta_u, \theta_v)}{n^2}$$

with

$$\Psi_i(w, \theta_u, \theta_v) = 3r^2 + (tw_3)^2 \sum_{i=1}^{r} \cos^2(\theta_u(i)) \cos^2(\theta_v(i)) +$$

$$\sum_{i=1}^{r} \cos^2(\theta_u(i)) \sin^2(\theta_u(i)) + (tw_3)^2 \sum_{i=1}^{r} \sin^2(\theta_u(i))$$

$$\cos^2(\theta_v(i)) + (tw_4)^2 \sum_{i=1}^{r} \sin^2(\theta_v(i)) \sin^2(\theta_v(i)) +$$

$$\left(\frac{w_4}{w_3} - 1\right)^2 (tw_1 + tw_2)^2 \sum_{i=1}^{r} \left\{ \frac{w_1^2 \cos^2(\theta_u(i)) + w_3^2 \sin^2(\theta_u(i))}{w_1^2 \cos^2(\theta_v(i)) + w_3^2 \sin^2(\theta_v(i))} \right\}$$

$$+ \left(\frac{w_4}{w_2} - 1\right)^2 (tw_1 + tw_3)^2 \sum_{i=1}^{r} \left\{ \frac{w_1^2 \cos^2(\theta_v(i)) + w_2^2 \sin^2(\theta_v(i))}{w_1^2 \cos^2(\theta_u(i)) + w_2^2 \sin^2(\theta_u(i))} \right\}$$

$$+ \left\{ \frac{w_1^2 \cos^2(\theta_u(i)) + w_2^2 \sin^2(\theta_u(i))}{w_1^2 \cos^2(\theta_v(i)) + w_2^2 \sin^2(\theta_v(i))} \right\}$$

$$\sum_{i=1}^{r} \int_{0}^{\theta_1} \left( u - \frac{\sigma_i(E_{22})}{\sqrt{r}} \right)^2 \frac{\sqrt{4 - u^2}}{\pi} \, du +$$

$$\left( tw_1(tw_3) \sigma_1(C_{L}^{-1}) \sigma_1(C_{R}^{-1}) \sigma_2 \right)^2 \frac{\sqrt{4 - u^2}}{\pi} \, du +$$

$$\left( tw_2(tw_3) \sigma_1(C_{L}^{-1}) \sigma_2 \right)^2 \frac{\sqrt{4 - u^2}}{\pi} \, du +$$

$$\left( tw_3(tw_4) \sigma_1(C_{L}^{-1}) \sigma_2 \right)^2 \frac{\sqrt{4 - u^2}}{\pi} \, du +$$

$$\left( tw_2(tw_3) \sigma_1(C_{L}^{-1}) \sigma_2 \right)^2 \frac{\sqrt{4 - u^2}}{\pi} \, du +$$

$$\left( tw_3(tw_4) \sigma_1(C_{L}^{-1}) \sigma_2 \right)^2 \frac{\sqrt{4 - u^2}}{\pi} \, du +$$

and

$$\frac{\sqrt{4 - u^2}}{\pi} \, du +$$

$$\frac{\sqrt{4 - u^2}}{\pi} \, du +$$

where

$$E_{22} = \left( tw_4^2 - tw_3^2 \right)^2 \left( tw_2^2 + tw_1^2 \right)^2$$

$$\left( tw_1^2 \cos^2(\theta_u) \cos^2(\theta_v) + tw_2^2 \cos^2(\theta_u) \sin^2(\theta_v) +$$

$$tw_3^2 \sin^2(\theta_u) \cos^2(\theta_v) + tw_4^2 \sin^2(\theta_u) \sin^2(\theta_v) \right)$$

$$\sin^2(\theta_u) \cos(\theta_u) \sin(\theta_v) \cos(\theta_v)$$

$$C_L := \left( tw_1^2 \cos^2(\theta_u) + tw_3^2 \sin^2(\theta_u) \right)^{\frac{1}{2}}$$

$$C_R := \left( tw_2^2 \cos^2(\theta_v) + tw_4^2 \sin^2(\theta_v) \right)^{\frac{1}{2}}.$$
The function $J$ satisfies the following error bound.

In essence, $\hat{m}_{w, \theta_u, \theta_v}$ coincides with $m_{\text{nuc}}$ when we set $w = 1 \in \mathbb{R}_+$. The question that arises is

- Is $\hat{m}_{w, \theta_u, \theta_v}$ a good description of $m_{w, \theta_u, \theta_v}$?

In the following Lemma, we provide a positive answer to this question. In fact, we demonstrate that the proposed upper-bound in Proposition 1 is asymptotically tight.

**Lemma 1.** The required number of measurements that $P_{w, \text{nuc}}$ with parameters $\theta_u = [\theta_u(1), \ldots, \theta_u(r)]^T$ and $\theta_v = [\theta_v(1), \ldots, \theta_v(r)]^T$, needs for exact recovery of $X \in \mathbb{R}^{n \times n}$ satisfies the following error bound.

$$\hat{m}_{w, \theta_u, \theta_v} - \frac{2}{\sqrt{n} \sqrt{T_c}} \leq m_{w, \theta_u, \theta_v} \leq \hat{m}_{w, \theta_u, \theta_v}$$

(27)

where,

$$c = \min\{\sin(\theta_u(1)), \cos(\theta_v(r))\}, \min\{\sin(\theta_v(1)), \cos(\theta_v(r))\}.$$ 

Proof. See Appendix E.

It is worth mentioning that the error term is independent of $w$ and is constant.

#### IV. HOW TO FIND OPTIMAL WEIGHTS

In this section, we propose the strategy of finding the unique optimal weights. First, we present a general Lemma about the function $\delta(D[|h_w(X)|_a, X])$. Actually, this Lemma states that this function (ignoring the infimum on $t \geq 0$ in the definition of statistical dimension) is strictly convex with respect to $w \in \mathbb{R}_+^3$. This Lemma helps us later in proving the uniqueness of optimal weights.

**Lemma 2.** Assume $C := \bar{c} \cdot \|h_w(X)\|$ does not contain the origin. Also, denote $G \in \mathbb{R}^{n \times n}$ a random matrix with i.i.d standard normal entries. Consider the function

$$J(v) := \text{E}d^2(G, h_w(C)) := \text{E}[J_G(v)]$$

with $v := [v_1, v_2, v_3]$. Further, it attains its minimum in the set $\left[0, n \left(1 + \frac{(n^2 + 1)^\frac{1}{2}}{\sqrt{n} + 1 - n}\right)^\frac{1}{2}\right].$

(28)

Proof. See Appendix F.

Now, we introduce our strategy of finding the unique optimal weights. Consider the error bound in Lemma 1. By taking infimum from both sides, it holds that

$$\inf_{w \in \mathbb{R}_+^3} \hat{m}_{w, \theta_u, \theta_v} - \frac{2}{\sqrt{n} \sqrt{T_c}} \leq \inf_{w \in \mathbb{R}_+^3} m_{w, \theta_u, \theta_v} \leq \hat{m}_{w, \theta_u, \theta_v}.$$

(29)

$m_{w, \theta_u, \theta_v}$ is surrounded by the same upper and lower-bounds up to an asymptotically vanishing constant term. We minimize this expression so as to reach the optimal weights $w^* = [w_1^*, w_2^*, w_3^*]^T$ via

$$w^* := \arg \min_{w \in \mathbb{R}_+^3} \hat{m}_{w, \theta_u, \theta_v} = \arg \min_{w \in \mathbb{R}_+^3} \inf_{t \geq 0} \Psi_t(w, \theta_u, \theta_v) = \arg \min_{w \in \mathbb{R}_+^3} J(v).$$

(30)

The reason to name these weights, optimal, lies in that they asymptotically (as $n \rightarrow \infty$) minimize the required number of measurements in $P_{w, \text{nuc}}$. Note that in the second equality of (30), we converted two variables $w$ and $t$ into a single vector variable $v = t w \in \mathbb{R}_+^3$. This is since $w$ in $\Psi_t(w, \theta_u, \theta_v)$ of (31) always appears along with the scalar $t$ (namely in the form of $tw$). Therefore, by finding $v^*$ (last term in (30)), we can reach the optimal weights $w^*$ up to a positive scaling factor. As a matter of fact, by the aid of Lemma 2, $v^*$ is unique and lies in the set $\left[0, n \left(1 + \frac{(n^2 + 1)^\frac{1}{2}}{\sqrt{n} + 1 - n}\right)^\frac{1}{2}\right]$. Hence, $w^*$ is unique up to a positive scaling factor. Note that this scaling factor is not the case since it is effectless on $P_{w, \text{nuc}}$. To obtain $w^*$ in (30), we propose a simple algorithm in Algorithm 1 called Optweights. In Optweights, we solve the convex optimization problem

$$[w_1^*, w_2^*, w_3^*] = \text{argmin}_{w_1 \geq 0, \ w_2 \geq 0, \ w_3 \geq 0} \hat{m}_{w, \theta_u, \theta_v}$$

(31)

to reach the triple $[w_1^*, w_2^*, w_3^*]$.

Qualitatively spoken, Algorithm 1 is based on Alternating Minimization (AM) approach. AM method is used to solve multivariate unconstrained optimization problems. The idea is based on optimizing each coordinate, individually. The advantages of our proposed algorithm are

- Each iteration is cheap.
- Unlike the gradient-based algorithms, it needs no step-size tuning.
- It is simple to implement.

In essence, Optweights (Algorithm 1) converts the multivariate optimization problem into some with scalar variables. For solving scalar optimization problems in Optweights (i.e. Step 10 in Algorithm 1), we use Golden Section Search (GSS) method (Algorithm 2) which tries to narrow the range of values (a and b in Algorithm 2) inside which the minimum is known to exist.

#### V. NUMERICAL EXPERIMENTS

In this section, we present the result of some computer experiments designed to evaluate the effect of optimal weighting strategy in matrix sensing given some prior subspace information. Note that the optimal weights are obtained using Algorithm 1. First, we construct a matrix

$$X = U_{n \times r} \Sigma_{r \times r} V_{n \times r}^H$$

(32)

with $n = 10$, $r = 3$. Then, we construct two subspaces $\tilde{U}$ and $\tilde{V}$ with dimension $r' \geq r$, that have known principal angles $\theta_u \in [0, 90^\circ]$ and $\theta_v \in [0, 90^\circ]$ with column and row subspaces of the ground truth matrix $X$ i.e. $U = \text{span}(U)$
and $\mathbf{V} = \text{span}(\mathbf{V})$, respectively. Note that, the bases $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are chosen such that

$$\hat{\mathbf{U}}^H \hat{\mathbf{U}} = \begin{bmatrix} \cos(\theta_a) & 0_{r \times (r'-r)} \\ \end{bmatrix} \text{ and } \hat{\mathbf{V}}^H \hat{\mathbf{V}} = \begin{bmatrix} \cos(\theta_v) & 0_{r \times (r'-r)} \\ \end{bmatrix}.$$  

Next, we compute the optimal weights $\mathbf{w}^*$ by Algorithm 1 namely by the function Optweights. We compare $P_{\text{nucl}}$ with $P_{\text{w*},\text{nucl}}$ for different $\theta_a$ and $\theta_v$. Our assessment criterion is probability of success over 50 Monte Carlo trials. A trial is declared successful if

$$\|\mathbf{X} - \hat{\mathbf{X}}\|_F / \|\mathbf{X}\|_F < 10^{-2},$$

where $\hat{\mathbf{X}}$ is the solution of optimization problems provided by CVX MATLAB package [16]. Below, we investigate different cases of principal angles.

In Figure 4 we tested some cases of optimal prior subspace information in which $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are slightly diverged from $\mathbf{U}$ and $\mathbf{V}$. Also, we set the deviation level of column and row subspaces roughly the same. From Figures 4(a)-4(e), it is observed that the required sample complexity of $P_{\text{w*},\text{nucl}}$ reaches the optimal number of measurements i.e. $r^2$. Besides, its sample complexity is far from that in $P_{\text{nucl}}$. In Figure 5 $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$, are close to $\mathbf{U}^\perp$ and $\mathbf{V}^\perp$, respectively. Figures 5(a)-5(c) show that even when $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are very far from $\mathbf{U}$ and $\mathbf{V}$, respectively, the reduction of sample complexity is possible. It is worth mentioning that one can also hope to reach the optimal number of measurements when there exists a subspace with dimension $r' = n - r$ very close to $\mathbf{U}^\perp$. This case can be observed in Figure 5(c).

In Figure 6 we tested a scenario where the principal angles are not so small but less than 45°. One can see from Figures 6(a)-6(e) that as much as the principal angles get less, more reduction is achievable in the required sample complexity.

In Figure 7 optimal weighting strategy is investigated when there exists weak prior subspace information about the column and row space of $\mathbf{X}$. What we mean by weak prior is a case that $\mathbf{U}$ and $\mathbf{V}$ are almost as close to $\mathbf{U}$ and $\mathbf{V}$ as they are to $\mathbf{U}^\perp$ and $\mathbf{V}^\perp$. In these cases, (See Figures 7(a)-7(d)) the sample complexity of our algorithm approaches the one in $P_{\text{nucl}}$.

In the last experiment shown in Figure 8 we consider the case where accuracies of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are different. From Figures 8(a)-8(d), it is observed that a huge sample complexity reduction is feasible when either prior column or row subspace information is close to the respective subspaces of the ground truth matrix.

VI. RELATED WORKS AND KEY DIFFERENCES

In [3], a non-uniform sampling distribution is considered for a Netflix data set and is shown that a properly weighted trace norm of the form

$$\|\mathbf{X}\|_{tr} := \|\text{diag}(\sqrt{\mathbf{p}}) \mathbf{X} \text{diag}(\sqrt{\mathbf{q}})\|_F,$$

works well where $p(i), i = 1, \ldots, n$ and $q(j), j = 1, \ldots, n$ are the probability of observing row $i$ and column $j$ of the matrix.

In [17], a non-uniform sampling scheme is considered in which authors proposed a generalized nuclear norm which penalizes the directions in the vector space of $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$...
The principal angles are $\theta$ as in [5]. They showed that the isometry constant for the calculated optimal weights are equal to $w* = 18.6213$. [b] $r = 3, n = 10$. The principal angles and optimal weights are equal to $\theta_u = [0.1858, 0.1426, 0.0742]^T, \theta_v = [0.205, 0.1374, 0.0878]^T, w_1^u = 1.1487 \times 10^{-4}, w_2^u = 0.0366, w_3^u = 0.0398, w_4^u = 12.6870$. [c] $r' = r = 3, n = 10$. The principal angles are $\theta_u = [0.2636, 0.1592, 0.0281]^T, \theta_v = [0.3212, 0.1438, 0.0470]^T$ and the optimal weights are equal to $w_1^u = 0.013, w_2^u = 0.4596, w_3^u = 0.4917, w_4^u = 17.3836$. [d] $r' = r = 5, n = 20$. The principal angles are $\theta_u = [0.3236, 0.2660, 0.2465, 0.2104, 0.135]^T, \theta_v = [0.2836, 0.2667, 0.2512, 0.1917, 0.1703]^T$ and the calculated optimal weights are equal to $w_1^u = 0.0008, w_2^u = 0.1305, w_3^u = 0.1232, w_4^u = 19.5313$. [e] $r' = r = 5, n = 10$. The principal angles are $\theta_u = [0.0295, 0.024, 0.0156, 0.0147, 0.0108]^T, \theta_v = [0.2996, 0.2635, 0.2346, 0.1656, 0.1470]^T$ and the calculated optimal weights are equal to $w_1^u = 0.0001, w_2^u = 0.0357, w_3^u = 0.0977, w_4^u = 26.6213$.

non-uniformly; namely, allocates more weights on certain directions less than other directions.

In [5], the authors heuristically proposed the following optimization problem to exploit prior subspace information.

$$\min_{Z \in \mathbb{R}^{n \times n}} \| (\lambda P_{\hat{U}} + P_{\hat{V}}) Z (\rho P_{\hat{U}} + P_{\hat{V}}) \|_F$$

s.t. $\| y - A(Z) \|_F \leq \epsilon$ (35)

where $\hat{U}$ and $\hat{V}$ with dimension $r$ are the estimates of column and row subspaces of the rank $r$ ground truth matrix $X \in \mathbb{R}^{n \times n}$. However, they did not answer how to explicitly find $\lambda$ and $\rho$.

In [18], the authors investigated the same objective function as in [5]. They showed that the isometry constant for $A(\cdot)$ can be more conservative and thus, the required bound for robust recovery can be lowered provided that prior subspace information is good $(\theta_u(1), \theta_v(1) < 45^\circ)$. Only in case $\theta_u(1) = \theta_v(1) = \theta$, they suggested to choose $\lambda = \rho = \sqrt{\tan^2(\theta) + \tan^2(\theta) - \tan^2(\theta)}$ so as to maximize the RIP bound. There are some key differences between our work and [18] which are listed below:

- They assume that the subspace estimate and the ground truth subspace are of the same dimension $r$. This assumption fails to occur in practical scenarios in some certain settings for example in Netflix problem where a higher dimensional subspace estimate is available to the practitioner (see Subsection 1.2 for more discussion).
- In our work, we consider a generalized case where high dimensional row and column subspaces angles from the row and column subspaces of interest.
- The meaning of optimal in that work differs from ours in that their weights maximize the RIP constants while ours minimize the required sample complexity.
- [18] only considers the effect of largest principal angles in performance bounds while in fact all principal angles directly affects the performance bounds.
- The measurement bound in [18] depends on $A(\cdot)$ while our bound is independent of the sampling operator.
- There is a wide range of principal angles ($45^\circ \leq \theta \leq 90^\circ$) in which no improvement is predicted in [18], inevitably reaching the performance bound of $P_{\text{nuc}}$. The only exception that our algorithm reaches the performance bound of $P_{\text{nuc}}$ is $\{\theta_u(i), \theta_v(i)\}_{i=1}^r = 45^\circ$. For instance, $\theta_u(i) = \theta_v(i) \approx 90^\circ i = 1, \ldots, r$ is considered to be a weak prior subspace information while it is excellent in our work leading to a huge sample complexity reduc-
Also, when $\theta_u \leq 45^\circ$, unlike ours, their bound is not optimal in the sense of sample complexity. Overall, our proposed method acts much better in terms of required sample complexity.

VII. USEFUL LEMMAS

This section provides necessary mathematical tools for Sections III and IV.

A. Constructing a basis for $\mathbb{R}^{n \times n}$

In this section, we find a special basis for $\mathbb{R}^{n \times n}$ that simplifies the analysis on required number of measurements of $\mathbb{P}_{w, \text{nuc}}$. The following lemma precisely states this.

**Lemma 3.** Consider a rank $r$ matrix $X \in \mathbb{R}^{n \times n}$ with column and row subspaces $\mathcal{U}$ and $\mathcal{V}$, respectively. Also, assume that we are given the subspaces $\mathcal{U} \subseteq \mathbb{R}^n$ and $\mathcal{V} \subseteq \mathbb{R}^n$ with dimension $r' \geq r$ that have known principal angles $\theta_u \in [0^\circ, 90^\circ]^r$ and $\theta_v \in [0^\circ, 90^\circ]^r$ with $\mathcal{U}$ and $\mathcal{V}$, respectively. Then, there exist bases $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{n \times r'}$, $\tilde{U} \in \mathbb{R}^{n \times r'}$, $\tilde{V} \in \mathbb{R}^{n \times r'}$

such that

\[ U = \text{span}(U), \quad \tilde{U} = \text{span}(\tilde{U}), \quad V = \text{span}(V), \quad \tilde{V} = \text{span}(\tilde{V}). \]

\[ \tilde{U} = B_L \begin{bmatrix} \cos(\theta_u) & 0 \\ -\sin(\theta_u) & 0 \\ 0 & -I_{r'-r} \\ 0 & 0 \end{bmatrix}, \quad \tilde{V} = B_R \begin{bmatrix} \cos(\theta_v) & 0 \\ -\sin(\theta_v) & 0 \\ 0 & -I_{r'-r} \\ 0 & 0 \end{bmatrix} \]

(37)

where

\[ U' = -P_{\mathcal{U}} \tilde{U} \begin{bmatrix} \sin^{-1}(\theta_u) \\ 0_{r' \times r} \end{bmatrix}, \quad V' = -P_{\mathcal{V}} \tilde{V} \begin{bmatrix} \sin^{-1}(\theta_v) \\ 0_{r' \times r} \end{bmatrix} \]

(38)

and, $\cos(\theta_u)$ is defined as

\[ \cos(\theta_u) := \text{diag}\left[ \cos(\theta_u(r)), \cos(\theta_u(r - 1)), ..., \cos(\theta_u(1)) \right] \in \mathbb{R}^{n \times n} \]

(39)

\[ \cos(\theta_v) := \text{diag}\left[ \cos(\theta_v(r)), \cos(\theta_v(r - 1)), ..., \cos(\theta_v(1)) \right] \in \mathbb{R}^{n \times n} \]

(40)

Lemma (3) allows us to find supp($h_w(X)$) which is later helpful. Below, we state a lemma that includes this, along with a crucial decomposition of $h_w(Z)$ for an arbitrary matrix $Z \in \mathbb{R}^{n \times n}$.

**Lemma 4.** Consider a matrix $X \in \mathbb{R}^{n \times n}$ with column and row spaces $\mathcal{U}$ and $\mathcal{V}$, respectively. Then, $h_w(Z)$ in (7) with the convention $w_k := \frac{w_k w_{2k}}{w_1}$ for an arbitrary matrix $Z \in \mathbb{R}^{n \times n}$ is decomposed as:

\[ h_w(Z) = \left( \frac{1}{w_3} \right) B_L O_L L^H \hat{Z} B_R R^H O_R^H B_R^H \]

(41)
where, $B_L \in \mathbb{R}^{n \times n}$ and $B_R \in \mathbb{R}^{n \times n}$ are defined in Lemma 5.

Also,

$$O_L := \begin{bmatrix}
(w_1 \cos^2(\theta_u) + w_3 \sin^2(\theta_u))(C_L)^{-1} & 0 & 0 \\
(w_3 - w_1) \sin(\theta_u) \cos(\theta_u)(C_L)^{-1} & 0 & 0 \\
0 & 0 & I_{r' - r'}
\end{bmatrix}$$

$$O_R := \begin{bmatrix}
(w_3 \cos^2(\theta_v) + w_4 \sin^2(\theta_v))(C_R)^{-1} & 0 & 0 \\
(w_4 - w_3) \sin(\theta_v) \cos(\theta_v)(C_R)^{-1} & 0 & 0 \\
0 & 0 & I_{n - r'}
\end{bmatrix}$$

where

$$C_L = (w_1 \cos^2(\theta_u) + w_3 \sin^2(\theta_u))^2$$

$$C_R = (w_3 \cos^2(\theta_v) + w_4 \sin^2(\theta_v))^2$$

$$L_{12} = (w_3^2 - w_1^2) \sin(\theta_u) \cos(\theta_v)(C_L)^{-1}$$

$$R_{12} = (w_3^2 - w_4^2) \sin(\theta_v) \cos(\theta_v)(C_R)^{-1}$$

$B_L, B_R, O_L$ and $O_R$ are orthonormal bases. Also, $L$ and $R$ are upper triangular matrices.

**Lemma 5.** Let $X = U_{n \times n} \Sigma_{r \times r} V_{n \times r}^H$ be the reduced SVD form of $X \in \mathbb{R}^{n \times n}$. Then, the unsorted SVD of $h_w(X)$ is obtained as

$$h_w(X) = B_L O_L \begin{bmatrix}
\frac{1}{\sqrt{3}} & C_L \Sigma C_R & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix} O_H^H B_H^T.$$
The principal angles are

Then,

\[ \text{Corollary 1.} \quad \text{Let } \hat{T} := \text{supp}(h_w(X)) \text{ and } T_1 := \text{supp}(I_r). \text{ Then, } \text{sgn}(h_w(X)) \text{ and } \mathcal{P}_{\hat{T}_1}(Z) \text{ for an arbitrary matrix } Z \in \mathbb{R}^{n \times n} \text{ are obtained by} \]

\[ \text{sgn}(h_w(X)) = B_L O_L \mathcal{P}_{T_1} (I_s) O_R^H B_R^H \quad (51) \]

\[ \mathcal{P}_{\hat{T}_1}(Z) = B_L O_L \mathcal{P}_{\hat{T}_1} (O_L^H B_L^H Z B_R O_R) O_R^H B_R^H. \quad (52) \]

Proof. See Appendix I.

B. Spectral Analysis of Large Random Matrices

In this part, we intend to specify the behavior of singular values of large i.i.d random Gaussian matrices e.g. \( G \in \mathbb{R}^{n_1 \times n_2} \). First, we state a well-known fact that specifies the limiting behavior of eigenvalues of random matrices to the Marchenko Pastur law \[19\] (20 Theorem 3.6). Here, we approximate the distribution of singular values of a random i.i.d standard normal matrix by a version of Marchenko–Pastur Law \[19\]. The proof uses a change of variable to match the argument for singular values which does not much differ from \[20\] Theorem 3.6] and thus we omitted the uninteresting details of this change.

**Fact 1.** Let \( G \in \mathbb{R}^{n_1 \times n_2} \) be a matrix with i.i.d standard normal distribution with \( s, u_b, l_b \) defined in Lemma 6. Then, the probability density function (pdf) of \( \sigma(G) \) is given by:

\[ f(x) = \frac{\sqrt{(u_b^2 - x^2)(x^2 - l_b^2)}}{\pi u s} \quad \forall x \in [l_b, u_b] \quad (53) \]

One can see from Figure 10 that the empirical density estimate of singular values of a random matrix \( G \) with Gaussian ensemble (shown with bars) harmonizes with the obtained bound in Fact 1 (shown with dashed line).

In the following lemma, we obtain the limiting behavior of \( \text{E}f(\sigma(G)) \) for a random matrix \( G \in \mathbb{R}^{n_1 \times n_2} \) and a function \( f : \mathbb{R} \rightarrow \mathbb{R} \). Here, for simplicity, we only consider \( f(x) = x^2 \). However, the result holds for any function of this kind.

**Lemma 6.** Consider a random matrix \( G \in \mathbb{R}^{n_1 \times n_2} \) whose elements are drawn from i.i.d standard normal distribution. Suppose \( n_1, n_2 \rightarrow \infty \) and \( s := \frac{n_1}{n_2} \in (0, 1] \). Then, we have:

\[ S := \text{E} \sum_{i=1}^{n_1} \left( \frac{\sigma_i(G)}{\sqrt{n_2}} - \lambda_i \right)^2 \rightarrow \sum_{i=1}^{n_1} \int_{l_b}^{u_b} (u - \lambda_i)^2 \frac{\sqrt{(u_b^2 - u^2)(u^2 - l_b^2)}}{\pi u s} du := S_{\text{ap}} \quad (54) \]

where

\[ u_b = 1 + \sqrt{s}, \quad l_b = 1 - \sqrt{s}, \quad (55) \]

Proof. See Appendix I. In Figure 9 we compare \( S_{\text{ap}} \) with \( S \) in Lemma 6 for a randomly generated vector \( \lambda := [\lambda_1, \ldots, \lambda_{n_1}]^T \) (the dotted label is for \( S \) and the dashed label for \( S_{\text{ap}} \)). It is observed that \( S_{\text{ap}} \) approximates \( S \) well and their difference vanishes when \( n_1 \) and \( n_2 \) are large. Note that in Figure 9 the value
The optimal weights are equal to \( \theta \). The outcome of our work and obtained the unique optimal weights that minimize the angles by introducing a new weighted optimization problem row spaces of the ground truth matrix. We exploited these subspaces form some known angles with the column and prior knowledge in matrix sensing. We assumed that two given of \( S \) is computed by replacing the expectation by empirical mean with \( 10^4 \) realization of \( G \).

VIII. CONCLUSION

In this work, we presented a new approach for exploiting prior knowledge in matrix sensing. We assumed that two given subspaces form some known angles with the column and row spaces of the ground truth matrix. We exploited these angles by introducing a new weighted optimization problem and obtained the unique optimal weights that minimize the required number of measurements. The outcome of our work is to use considerably less measurements in seismic data interpolation [5], FDD massive MIMO [12], Dynamic sensor network localization [7], collaborative filtering [3]. Netflix problem [1] and subspace tracking.

REFERENCES

[1] A. SIGKDD, “Netflix,” in Proceedings of kdd cup and workshop, 2007.
[2] N. Srebro, “Learning with matrix factorizations phd thesis,” 2004.
[3] N. Srebro and R. R. Salakhutdinov, “Collaborative filtering in a non-uniform world: Learning with the weighted trace norm,” in Advances in Neural Information Processing Systems, pp. 2056–2064, 2010.
A. Proof of Lemma 3

Proof. Assume \( \hat{U} \in \mathbb{R}^{n \times r} \) and \( \hat{U}^\perp \in \mathbb{R}^{n \times (n-r')} \) be some orthonormal bases for \( \tilde{U} \) and \( \tilde{U}^\perp \), respectively. Also, since \( \overset{\sim}{U} \) and \( \overset{\sim}{U}^\perp \) are uniquely characterized by their respective projection matrices i.e. \( P_{\hat{U}} \in \mathbb{R}^{n \times n} \) and \( P_{\hat{U}^\perp} \in \mathbb{R}^{n \times n} \), without loss of generality, assume that

\[
U^H \hat{U} = \begin{bmatrix}
\cos(\theta_u) & 0 \\
\sin(\theta_u) & 0
\end{bmatrix}
\]

where \( \hat{U}_1 \in \mathbb{R}^{n \times r} \), \( \hat{U}_2 \in \mathbb{R}^{n \times (n-r')} \), \( \hat{U}^\perp_1 \in \mathbb{R}^{n \times r} \), and \( \hat{U}^\perp_2 \in \mathbb{R}^{n \times (n-r')} \) are orthonormal bases forming the subspaces \( \hat{U}_1 \subseteq \tilde{U} \), \( \hat{U}_2 \subseteq \tilde{U}^\perp \), \( \hat{U}^\perp_1 \subseteq \tilde{U}^\perp \), and \( \hat{U}^\perp_2 \subseteq \tilde{U}^\perp \), respectively. (otherwise one could redefine \( \hat{U}, \hat{U}^\perp \) and \( U \) by taking SVD of \( U^H \hat{U} \) and \( U^H \hat{U}^\perp \). Since rotation in \( \hat{U} \) and \( \hat{U}^\perp \) does not affect \( P_{\hat{U}} \) and \( P_{\hat{U}^\perp} \).

The column space of any matrix in \( \mathbb{R}^{n \times n} \) can be decomposed into the spaces \( U, U^\perp \cap \tilde{U}_1, U^\perp \cap \tilde{U}_2 \) and \( U^\perp \cap \tilde{U}^\perp \), where for the last three, we construct an orthonormal basis as follows

\[
U'_1 := -P_{U^\perp} \tilde{U}_1 \sin^{-1}(\theta_u) \in \mathbb{R}^{n \times r} \\
U'_2 := -P_{U^\perp} \tilde{U}_2 \in \mathbb{R}^{n \times r'} \\
U'' := -P_{U^\perp} \tilde{U}^\perp_2 \in \mathbb{R}^{n \times (n-r')} \\
\]

such that

\[
B_L := \begin{bmatrix}
U'_{n \times r} & U''_{n \times (n-r')}
\end{bmatrix}
\]

forms an orthonormal basis for column span of any matrix in \( \mathbb{R}^{n \times n} \). Similar to the above statements, there exist orthonormal bases

\[
V'_1 := -P_{V' \perp} \tilde{V}_1 \sin^{-1}(\theta_v) \in \mathbb{R}^{n \times r} \\
V'_2 := -P_{V' \perp} \tilde{V}_2 \in \mathbb{R}^{n \times r'} \\
V'' := -P_{V' \perp} \tilde{V}^\perp_2 \in \mathbb{R}^{n \times (n-r')} \\
\]

such that

\[
B_R := \begin{bmatrix}
V'_{n \times r} & V''_{n \times (n-r')}
\end{bmatrix}
\]

forms an orthonormal basis for the row space of any arbitrary matrix in \( \mathbb{R}^{n \times n} \). Lastly, it is easy to verify that the matrices \( \hat{U} \) and \( \hat{V} \) can be represented in the bases \( B_L \) and \( B_R \) as follows.

\[
\hat{U} = B_L \begin{bmatrix}
\cos(\theta_u) & 0 \\
-\sin(\theta_u) & 0 \\
0 & -I_{r'} \\
0 & 0
\end{bmatrix} \in \mathbb{R}^{n \times r'} \\
\hat{V} = B_R \begin{bmatrix}
\cos(\theta_v) & 0 \\
-\sin(\theta_v) & 0 \\
0 & 0 \\
0 & -I_{r'}
\end{bmatrix} \in \mathbb{R}^{n \times r'}
\]

B. Proof of Lemma 4

Proof. Define

\[
w_4 := \frac{w_2 w_3}{w_1}.
\]

\( h_w(Z) \) in (1) can be reformulated as

\[
h_w(Z) = \frac{1}{w_3} \left( w_1 P_{\hat{U}} + w_3 P_{\hat{U}^\perp} \right) Z \left( w_3 P_{\hat{V}} + w_4 P_{\hat{V}^\perp} \right)
\]

(58)
We start our derivation by (37) and (38) to find $P_{\tilde{U}}$, $P_{\tilde{V}}$ which are the essential components of $h_w(Z)$. By (37) and (38), it is simply holds that

$$P_{\tilde{U}} := \tilde{U}\tilde{U}^H = \begin{bmatrix} 
\cos^2(\theta_u) & -\sin(\theta_u)\cos(\theta_u) & 0_{r\times(r'-r)} & 0 \\
-\sin(\theta_u)\cos(\theta_u) & \sin^2(\theta_u) & 0 & 0 \\
0 & 0 & I_{r'-r} & 0 \\
0 & 0 & 0 & I_{n-r-\hat{r}'-r} 
\end{bmatrix}$$

Also, we have:

$$P_{\tilde{U}^+} := I - P_{\tilde{U}} = B_L \begin{bmatrix} 
\sin^2(\theta_u) & \sin(\theta_u)\cos(\theta_u) & 0_{r\times(r'-r)} & 0 \\
\sin(\theta_u)\cos(\theta_u) & \cos^2(\theta_u) & 0 & 0 \\
0 & 0 & 0_{n-r-\hat{r}'-r} & 0 \\
0 & 0 & 0 & I_{n-r-\hat{r}'-r} 
\end{bmatrix} B_L^H$$

where $O_L$ is an orthonormal basis and $L$ is an upper-triangular matrix. We rewritten (61) as

$$w_1P_{\tilde{U}} + w_3P_{\tilde{U}^+} = B_L O_L L^H = B_L L^H O_L^H B_L^H,$$  \hspace{1cm} (63)

where the last equality is since $O_L L = L^H O_L^H$. With a similar $B_{\tilde{V}}^H$ approach on the row space of $X$, one may write

$$w_3P_{\tilde{V}} + w_4P_{\tilde{V}^+} = B_R O_R R^H = B_R R^H O_R^H B_R^H$$  \hspace{1cm} (64)

where

$$O_R := \begin{bmatrix} 
(w_3 \cos^2(\theta_v) + w_4 \sin^2(\theta_v))(C_R)^{-1} \\
(w_4 - w_3) \sin(\theta_v) \cos(\theta_v)(C_R)^{-1} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & I_{r'-r} & 0 \\
0 & 0 & 0 & I_{n-r-\hat{r}'-r} 
\end{bmatrix}$$

is an orthonormal basis of $\mathbb{R}^n$ and

$$R := \begin{bmatrix} 
C_R & R_{12} & 0 & 0 \\
0 & w_3 w_4 \Delta_R^{-1} & 0 & 0 \\
0 & 0 & w_3 I_{r'-r} & 0 \\
0 & 0 & 0 & w_4 I_{n-r-\hat{r}'-r} 
\end{bmatrix}$$  \hspace{1cm} (65)

is a triangular matrix. Lastly, $h_w(Z)$ for an arbitrary $Z \in \mathbb{R}^{n \times n}$ (the relation (58)) may be written as (41).

C. Proof of Lemma 5

Proof. Since

$$B_L^H X B_R = \begin{bmatrix} 
\Sigma_{r\times r} & 0_{r \times (n-r)} \\
0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} 
\end{bmatrix},$$  \hspace{1cm} (67)

it follows that

$$L B_L^H X B_R R_{\tilde{R}}^H = \begin{bmatrix} 
C_L \Sigma C_R & 0_{r \times (n-r)} \\
0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} 
\end{bmatrix}$$

As $C_L \Sigma C_R$ is a diagonal matrix, one may deduce from (41) that

$$h_w(X) = B_L O_L \begin{bmatrix} 
\Sigma_{n-r\times r} & 0_{n-r \times (n-r)} \\
0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} 
\end{bmatrix} O_R^H B_R^H$$

(69)

provides an unsroted SVD form for $h_w(X)$.  \hspace{1cm}
D. Proof of Proposition 2

Proof. Let us define below notations before proving the result.

\[ T_1 := \left\{ Z \in \mathbb{R}^{n \times n}, \ Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \right\} \]

\[ T_1^\perp := \left\{ Z \in \mathbb{R}^{n \times n}, \ Z = \begin{bmatrix} 0_{r \times r} & 0_{r \times r - r'} \\ 0_{r' \times r} & 0_{r' \times r} \\ 0_{r \times r} & 0_{r \times r} \\ 0_{r' \times r} & 0_{r' \times r} \end{bmatrix} \right\} \]

\[ T_{11} := \left\{ Z \in \mathbb{R}^{n \times n}, \ Z = \begin{bmatrix} Z_{11} & Z_{12} & Z_{13} \\ Z_{21} & Z_{22} & Z_{23} \\ Z_{31} & Z_{32} & Z_{33} \end{bmatrix} \right\} \]

\[ t Z_{43} \frac{\partial}{\partial t} + E \left( \inf_{i=1}^{n} \sigma_i(G_{43}) - t \sigma_i(Z_{43}) \right) \]

\[ \inf_{t > 0} \left\{ 3 r^2 + t^2 r + E \sum_{i=1}^{r} \sigma_i(G_{22}) - t \sigma_i(Z_{22}) \right\} \]

\[ + E \left( \inf_{i=1}^{n} \sigma_i(G_{23}) - t \sigma_i(Z_{23}) \right) + E \sum_{i=1}^{r} \sigma_i(G_{32}) - t \sigma_i(Z_{32}) \]

\[ \left( \inf_{t \geq 0} \left\{ 3 r^2 + t^2 r + E \sum_{i=1}^{r} \sigma_i(G_{22}) - t \sigma_i(Z_{22}) \right\} \right) \]

\[ \left( \inf_{t \geq 0} \left\{ 3 r^2 + t^2 r + E \sum_{i=1}^{r} \sigma_i(G_{23}) - t \sigma_i(Z_{23}) \right\} \right) \]

\[ t Z_{43} \frac{\partial}{\partial t} + E \left( \inf_{i=1}^{n} \sigma_i(G_{41}) - t \sigma_i(Z_{41}) \right) \]

\[ \left( \inf_{t \geq 0} \left\{ 3 r^2 + t^2 r + E \sum_{i=1}^{r} \sigma_i(G_{22}) - t \sigma_i(Z_{22}) \right\} \right) \]

\[ + E \left( \inf_{i=1}^{n} \sigma_i(G_{23}) - t \sigma_i(Z_{23}) \right) + E \sum_{i=1}^{r} \sigma_i(G_{32}) - t \sigma_i(Z_{32}) \]

\[ \left( \inf_{t \geq 0} \left\{ 3 r^2 + t^2 r + E \sum_{i=1}^{r} \sigma_i(G_{22}) - t \sigma_i(Z_{22}) \right\} \right) \]

\[ + E \left( \inf_{i=1}^{n} \sigma_i(G_{23}) - t \sigma_i(Z_{23}) \right) + E \sum_{i=1}^{r} \sigma_i(G_{32}) - t \sigma_i(Z_{32}) \]
Before proving the result, let define some notations. 

\[ 2(r' - r)(n - r - r') \int_{t/3}^{u_{h_3}} \left( u - \frac{t}{\sqrt{(n - r') \sqrt{(n - r - r')}} \right)^2 \left( u^2 + (t' - r)^2 \right)^2 \left( u^2 + (n - r - r') \right)^2 \]

\[ \frac{1}{\pi u_\delta_3} \left( u - \frac{t}{\sqrt{(n - r') \sqrt{(n - r - r')}} \right)^2 \left( u^2 + (t' - r)^2 \right)^2 \left( u^2 + (n - r - r') \right)^2 \]

\[ \frac{1}{\pi} \left( u^2 + (n - r) \right)^2 \left( u - \frac{t}{\sqrt{(n - r')}} \right)^2 \]

In (72), (I) follows from the fact that the infimum of an affine function is concave and Jensen’s inequality. In (II), we decomposed the term in Frobenius norm into \( T \) and \( T^\perp \) and used the relation

\[ \mathcal{P}_T \cdot \mathcal{P}_T^2 : \| \mathcal{P}_T \cdot (Z) \|_{2 \rightarrow 2} \leq 1 \]

In (III), we benefited from the facts

\[ \mathcal{P}_T \cdot \mathcal{P}_T = BL \cdot \mathcal{P}_T (B^T H \cdot AB) \cdot B^T H \cdot \mathcal{P}_T \cdot \mathcal{P}_T \]

for an arbitrary matrix \( A \in \mathbb{R}^{n \times n} \). In (III), we also used the rotational invariance of spectral norm. In (IV), since \( B_L \) and \( B_R \) each have orthonormal columns, the entries of \( B^T H \cdot GB \) are i.i.d standard Gaussian which, without loss of generality, we denote by \( G \) again. For simplicity, we replace \( B^T H \cdot GB \) by \( Z \). We also used the rotational invariance of Frobenius and spectral norms. In (V), we only used the argument that the entries of \( G_{11}, G_{12} \) and \( G_{13} \) have i.i.d standard normal distribution. In (VI), we decompose the space \( T^\perp \) into the spaces \( T_{ij} \) for \( i, j = \{2, 3, 4\} \). Also, we used the relation

\[ \| \mathcal{P}_T \cdot (Z) \|_{2 \rightarrow 2} \leq \sum_{i,j=2}^4 \| \mathcal{P}_T \cdot (Z) \|_{2 \rightarrow 2} \leq \sum_{i,j=2}^4 \alpha_{ij} \leq 1, \]

(74) where the first inequality is due to the triangle inequality of spectral norm. The second is due to the definition of \( \alpha_{ij} \) in (71). In fact,

\[ \{ Z : \| \mathcal{P}_T \cdot (Z) \|_{2 \rightarrow 2} \leq 1 \} \subseteq \{ Z : \| \mathcal{P}_T \cdot (Z) \|_{2 \rightarrow 2} \leq \alpha_{ij} \forall i, j = \{2, 3, 4\} \}. \]

(75) (VII) follows from Hoffman Wielandt theorem [27] Corollary 7.3.5. (VIII) is the consequence of the below argument. Namely:

\[ \sqrt{(g - z)^2} = \sqrt{|g - a|^2}, \]

for arbitrary scalar \( g \) and positive \( a \). For (IX), we invoke Lemma [6].

E. Proof of Proposition [8]

Proof. Before proving the result, let define some notations.

\[ \hat{T} := \text{supp}(h_w(X)) \]

\[ w := [w_1, w_2, w_3]^T \]

\[ w_1 := \frac{w_2 w_3}{w_1}. \]

(77)
We decompose the resulting matrix into smaller matrices

\[
C_L = \left( (tw_1)^2 \cos^2(\theta_u) + (tw_3)^2 \sin^2(\theta_u) \right)^{1/2}
\]

\[
C_R = \left( (tw_3)^2 \cos^2(\theta_v) + (tw_1)^2 \sin^2(\theta_v) \right)^{1/2}
\]

(79)

The argument (78) follows from a few reasons where we explain each in the following.

In (I), we used the chain rule lemma of subdifferential [13 Theorem 23.9]. In (II), we used the fact that \( h_{w_1} = h_{w_2} \). Also, the decomposition (41) is used. (III) is the consequence of following definition of the set \( \partial \|_{\ast} (h_w(X)) \):

\[
\partial \|_{\ast} (h_w(X)) := \left\{ \text{sgn}(h_w(X)) + \mathcal{P}_{\partial \|_{\ast}}(W) : \| \mathcal{P}_{\partial \|_{\ast}}(W) \|_{2-\alpha} \leq \epsilon \right\}.
\]

(80)

(IV) is since \( B_L, O_L, B_R, O_R \) are orthonormal bases and the rotational invariance of Frobenius norm. Also, we used the fact that \( O_h^1 B_L^1 G B_R O_R \) has the same distribution as \( G \). So, for simplicity, we set \( G \) instead of \( O_h^1 B_L^1 G B_R O_R \). (V) is the result of (51) and (52) where comes from Lemma (5) and replacing \( O_h^1 B_L^1 G B_R O_R \) with \( Z \). (VI) comes from the relations

\[
\frac{t}{w_3} L^H \mathcal{P}_{T_1}(I_n) R = \begin{bmatrix}
0_{r \times r} & L_{12} C_R & 0_{r \times (n-2r)} \\
0_{(r'-r) \times r} & \frac{t}{w_3} C_R L_{12} & 0_{r \times (n-2r)} \\
0_{(n-r'-r) \times r} & \frac{t}{w_3} C_R C_R L_{12} & 0_{r \times (n-2r)}
\end{bmatrix},
\]

(81)

\[
\frac{t}{w_3} L^H \mathcal{P}_{T_1}^T(Z) R = \begin{bmatrix}
0_{r \times r} & 0_{r \times (n-2r)} \\
0_{(r'-r) \times r} & \frac{t}{w_3} w_1 w_3 x_4 Z_{23} C_{22}^{-1} & 0_{r \times (n-2r)} \\
0_{(n-r'-r) \times r} & \frac{t}{w_3} w_1 w_3 x_4 Z_{23} C_{22}^{-1} & 0_{r \times (n-2r)}
\end{bmatrix}.
\]

(82)

Also, (81) and (82) can be more simplified as follows.

\[
\frac{t}{w_3} L^H \mathcal{P}_{T_1}(I_n) R = \begin{bmatrix}
E_{11} & E_{12} & 0_{r \times (n-2r)} \\
E_{21} & E_{22} & 0_{r \times (n-2r)} \\
0_{r \times r} & 0_{r \times (n-2r)}
\end{bmatrix}
\]

(83)

\[
\frac{t}{w_3} L^H \mathcal{P}_{T_1}^T(Z) R = \begin{bmatrix}
0_{r \times r} & 0_{r \times (n-2r)} \\
0_{(r'-r) \times r} & \frac{t}{w_3} (tw_1)(tw_3) C_{R}^{-1} Z_{23} C_{22}^{-1} & 0_{r \times (n-2r)} \\
0_{(n-r'-r) \times r} & \frac{t}{w_3} (tw_1)(tw_3) C_{R}^{-1} Z_{23} C_{22}^{-1} & 0_{r \times (n-2r)}
\end{bmatrix}
\]

We decompose the resulting matrix into smaller matrices which are disjointly supported on \( \{T_{ij}\}_{i,j=1}^4 \). Note that, unlike (46) and (47), \( C_L \) and \( C_R \) are redefined here so as to include \( t \). We proceed (78) by writing:

\[
\Psi_t(w, \theta_u, \theta_v) = r^2 + \| E_{11} \|_F^2 + r^2 + \| E_{12} \|_F^2 + r^2 + \| E_{21} \|_F^2 + \| E_{22} \|_F^2 + \inf_{|r| \leq \alpha_{22}} \| \mathcal{P}_{T_1}(G) - E_{22} - (tw_1)(tw_3)(tw_4) \|_F^2
\]

(84)
\[
\inf_{\sigma_i(Z_{44}) \leq 1} (\sigma_i(G_{44}) - (tw_4)\sigma_i(Z_{44}))^2 + E \sum_{i=1}^{r \wedge (n-r-r')} \inf_{\sigma_i(Z_{44}) \leq 1} (\sigma_i(G_{43}) - (tw_3)\sigma_i(Z_{43}))^2 + E \sum_{i=1}^{r \wedge (n-r-r')} \inf_{\sigma_i(Z_{43}) \leq 1} (\sigma_i(G_{42}) - (tw_3)\sigma_i(Z_{42}))^2 \frac{C_{11}\sigma_i(C_{R})}{2} \geq \inf_{\sigma_i(Z_{41}) \leq 1} (\sigma_i(G_{41}) - (tw_4)\sigma_i(Z_{41}))^2
\]

(85)

In (85), we proceeded one step further by simplifying \( \|E_{11}\|_{F}, \|E_{12}\|_{F}, \|E_{21}\|_{F}, \|E_{22}\|_{F} \). Some beneficial facts are employed in (85). First, in line six, we used the relation

\[
\inf_{\|P_{T_{22}}(Z)\|_{2 \to 2} \leq \alpha_{22}} \|P_{T_{22}}(G) - E_{22} - (tw_1)(tw_3)(tw_4)\|_{F}^2 \frac{C_{11}\sigma_i(C_{R})}{2} \geq \inf_{\sigma_i(Z_{22}) \leq \alpha_{22}} (\sigma_i(G_{22}) - \sigma_i(E_{22}))^2 \frac{C_{11}\sigma_i(C_{R})}{2} \geq \inf_{\sigma_i(Z_{22}) \leq \alpha_{22}} (\sigma_i(G_{22}) - \sigma_i(E_{22}))^2 \frac{C_{11}\sigma_i(C_{R})}{2}
\]

(86)

where in above equations, the first equality (I) is due to Hoffman–Wielandt Theorem [22 Corollary 7.3.5]. The second (II) is because of the relations

\[
\sum_{i=1}^{r} \left( \sigma_i(G_{22}) - \sigma_i(E_{22}) \right) \geq \sum_{i=1}^{r} \left( \sigma_i(G_{22}) - \sigma_i(E_{22}) \right) \geq (tw_1)(tw_3)(tw_4) \sigma_i(C_{L}^{-1}) \sigma_i(Z_{22}) \sigma_i(C_{R}^{-1}) \frac{C_{11}\sigma_i(C_{R})}{2}
\]

(87)

[22 Lemma 3.3.8] and \( f(x) = x^2 \) is an increasing convex function on \([0, \infty]\). In (87), we benefit from [22 Theorem 3.3.14 a] and [22 Problem 3]. In line eight of (85), we used

\[
\inf_{\|P_{T_{23}}(Z)\|_{2 \to 2} \leq \alpha_{23}} \|P_{T_{23}}(G) - (tw_3)C_{L}^{-1}P_{T_{23}}(Z)\|_{F}^2 = \frac{C_{11}\sigma_i(C_{R})}{2} \geq \inf_{\sigma_i(Z_{23}) \leq 1} \left( \sigma_i(G_{23}) - (tw_1)(tw_3)\sigma_i(C_{L}^{-1}Z_{23}) \right) \frac{C_{11}\sigma_i(C_{R})}{2} \geq \inf_{\sigma_i(Z_{23}) \leq 1} \left( \sigma_i(G_{23}) - (tw_1)(tw_3)\sigma_i(C_{L}^{-1}Z_{23}) \right) \frac{C_{11}\sigma_i(C_{R})}{2}
\]

(88)

Lemma 3.3.8], convexity, besides monotonicity, of \( f(x) = x^2 \) in the interval \([0, \infty]\] and

\[
\sum_{i=1}^{r \wedge (n-r-r')} \left( \sigma_i(G_{23}) - (tw_1)(tw_3)\sigma_i(C_{L}^{-1}Z_{23}) \right) \frac{C_{11}\sigma_i(C_{R})}{2} \geq \sum_{i=1}^{r \wedge (n-r-r')} \left( \sigma_i(G_{23}) - (tw_1)(tw_3)\sigma_i(C_{L}^{-1})\sigma_i(Z_{23}) \right)
\]

(89)

which follows from [22 Theorem 3.14 a]. In the next lines of (85), we used a similar strategy.

\[
\Psi_1(w, \theta_u, \theta_v) = 3r^2 + w_2^2 \sum_{i=1}^{r} \cos^2(\theta_u(i)) \cos^2(\theta_v(i)) + w_2^2 \sum_{i=1}^{r} \cos^2(\theta_u(i)) \sin^2(\theta_v(i)) + w_2^3 \sum_{i=1}^{r} \sin^2(\theta_u(i)) \cos^2(\theta_v(i))
\]

(90)

In (90), we benefited from the relation (76). We proceed to prove by writing,

\[
\Psi_1(w, \theta_u, \theta_v) = 3r^2 + w_2^2 \sum_{i=1}^{r} \cos^2(\theta_u(i)) \cos^2(\theta_v(i)) + w_2^2 \sum_{i=1}^{r} \cos^2(\theta_u(i)) \sin^2(\theta_v(i)) + w_2^3 \sum_{i=1}^{r} \sin^2(\theta_u(i)) \cos^2(\theta_v(i)) + \ldots
\]

(91)
\[ w_2^2 \sum_{i=1}^{r} \sin^2(\theta_u(i)) \sin^2(\theta_v(i)) + \left( \frac{w_1}{w_3} - 1 \right)^2 (w_1 + w_2)^2 \sum_{i=1}^{r} \left( \frac{w_1}{w_3} \cos^2(\theta_u(i)) + \frac{w_2}{w_3} \sin^2(\theta_u(i)) \right) \sin^2(\theta_v(i)) \cos^2(\theta_v(i)) + \left( \frac{w_3}{w_1} - 1 \right)^2 (w_1 + w_3)^2 \sum_{i=1}^{r} \left( \frac{w_1}{w_3} \cos^2(\theta_u(i)) + \frac{w_2}{w_3} \sin^2(\theta_u(i)) \right) \sin^2(\theta_v(i)) \cos^2(\theta_v(i)) \]
\[ + \left( \frac{w_3}{w_1} \right)^2 (w_1 + w_3)^2 \sum_{i=1}^{r} \left( \frac{w_1}{w_3} \cos^2(\theta_u(i)) + \frac{w_2}{w_3} \sin^2(\theta_u(i)) \right) \sin^2(\theta_v(i)) \cos^2(\theta_v(i)) \]
where

\[
\begin{align*}
\beta_1 &= \sum_{i=1}^{r} \cos^2(\theta_u(i)) \cos^2(\theta_v(i)) \\
\beta_2 &= \sum_{i=1}^{r} \cos^2(\theta_u(i)) \sin^2(\theta_v(i)) \\
\beta_3 &= \sum_{i=1}^{r} \sin^2(\theta_u(i)) \cos^2(\theta_v(i)) \\
\beta_4 &= \sum_{i=1}^{r} \sin^2(\theta_u(i)) \cos^2(\theta_v(i))
\end{align*}
\] (98)

In (I), we used the definition of nuclear norm. In (II), \( \text{sgn}(h_w(X)) \) is obtained from [51]. Also, we used the fact that \( h_w = h_w^* \). In (III), we used [41] and the facts \( O_L L = L^H O_L^H \) and \( O_R R = R^H O_R^H \). (IV) is the consequence of the fact that

\[
X = B_L \left[ \Sigma \begin{array}{c} 0_{r \times n-r} \\ 0_{(n-r) \times (n-r)} \end{array} \right] B_R^H,
\] (99)

and only \( \mathcal{P}_{F_1}(L^H R) = C_L C_R \) contribute to the Frobenius inner product. (V) follows from Cauchy Schwarz inequality and the rotational invariance of Frobenius norm. By considering [13], [41], the fact that

\[
\varepsilon \|h_w(\cdot)\|_F(X) = \left( \frac{1}{w^3} \right) B_L O_L L^H \varepsilon \cdot \|u(h_w(X))\|_F - \|B_R R^H O_R^H B_R^H\|_F \] (100)

and [51], \( D(\|h_w(\cdot)\|_F, X) \) does not depend on the singular values of \( X \), i.e. \( \Sigma \). So, a matrix

\[
Z = U_{n \times r} C_L C_R V_{n \times r}^H
\] (101)

can be chosen to have equality in (V). Lastly, (VI) follows from the definitions [46] and [47] and some simplification. Therefore, by (93) and (97), the error bound reads:

\[
\frac{2\sqrt{n} \sqrt{w_1^2 + w_2^2 + w_3^2 + w_4^2}}{\sqrt{w_1^2 \beta_1 + w_2^2 \beta_2 + w_3^2 \beta_3 + w_4^2 \beta_4}} \leq 2\sqrt{n} \frac{1}{\min(\beta_1, \beta_2, \beta_3, \beta_4)}
\] (102)

Since \( \cos(\theta_u) \) and \( \sin(\theta_u) \) are arranged in decreasing and increasing order, respectively, it holds that

\[
\min\{\beta_1, \beta_2, \beta_3, \beta_4\} \\ r \min\{\cos(\theta_u(r)), \sin(\theta_u(1))\} \min\{\cos(\theta_v(r)), \sin(\theta_v(1))\}
\] (103)

and hence the result follows.  

\[\Box\]

\[G. \ \text{Proof of Lemma 2}\]

\[\text{Proof. Continuity in bounded points. For continuity, it must be shown that sufficiently small changes in \( v \) result in arbitrary small changes in \( J(v) \). Let} \ v_1, v_2 \in \mathbb{R}_+^d. \ \text{By definition of} \ J_G, \ \text{it holds that}\]

\[
J_G(v_1) - J_G(v_2) = \|G - \mathcal{P}_{h_{v_1}}(C)\|_F^2 - \|G - \mathcal{P}_{h_{v_2}}(C)\|_F^2 = 2 \left[ \mathcal{P}_{h_{v_1}}(C) \right] \left( \mathcal{P}_{h_{v_1}}(C) - \mathcal{P}_{h_{v_2}}(C) \right) + \left( \|\mathcal{P}_{h_{v_1}}(C)\|_F - \|\mathcal{P}_{h_{v_2}}(C)\|_F \right) \left( \|\mathcal{P}_{h_{v_1}}(C)\|_F + \|\mathcal{P}_{h_{v_2}}(C)\|_F \right)
\]

(104)

Since

\[
\|\mathcal{P}_{h_{v}}(C)\|_F \leq \sup_{Z \in C} \|h_w(Z)\|_F \leq \sqrt{n} \max \left\{ \|v\|_{\infty}, \frac{v(2)v(3)}{v(1)} \right\}
\]

(105)

(due to [23]) and

\[
\|\mathcal{P}_{h_{v_1}}(C)\|_F - \|\mathcal{P}_{h_{v_2}}(C)\|_F \leq \sup_{Z \in C} \left( \|h_{v_1}(Z)\|_F - \|h_{v_2}(Z)\|_F \right) \leq \sup_{Z \in C} \left( \|h_{v_1}(Z) - h_{v_2}(Z)\|_F = \sup_{Z \in C} \|h_{v_1} - h_{v_2}(Z)\|_F \leq \sqrt{n} \max \left\{ \|v_1 - v_2\|_{\infty}, \frac{(v_1(2) - v_2(2))(v_1(3) - v_2(3))}{(v_1(1) - v_2(1))} \right\} \right)
\]

(106)

(see [23]), we have

\[
\|J_G(v_1) - J_G(v_2)\| \leq \left( 2\|G\|_F \sqrt{n} + n \max \left\{ \|v_1\|_{\infty}, \frac{v_1(2)v_1(3)}{v_1(1)} \right\} \right) \max \left\{ \|v_1 - v_2\|_{\infty}, \frac{(v_1(2) - v_2(2))(v_1(3) - v_2(3))}{(v_1(1) - v_2(1))} \right\}
\]

(107)

As a consequence, we obtain

\[
\|J_G(v_1) - J_G(v_2)\| \to 0 \ \text{as} \ v_1 \to v_2.
\]

(108)

Since \( \|v\|_{\infty} \) is bounded, continuity holds.

\[\text{Convexity. Let} \ v_1, v_2 \in \mathbb{R}_+^d \ \text{and} \ \theta \in [0, 1]. \ \text{Then, for all}\]

\[
\forall \epsilon, \tilde{\epsilon} > 0, \exists Z, \tilde{Z} \in C \ \text{such that}: \]

\[
\|G - h_{v_1}(Z)\|_F \leq \text{dist}(G, h_{v_1}(C)) + \epsilon \]

\[
\|G - h_{v_2}(Z)\|_F \leq \text{dist}(G, h_{v_2}(C)) + \tilde{\epsilon}
\]

Since otherwise, we have:

\[
\forall Z, \tilde{Z} \in C: \]

\[
\|G - h_{v_1}(Z)\|_F > \text{dist}(G, h_{v_1}(C)) + \epsilon \]

\[
\|G - h_{v_2}(Z)\|_F > \text{dist}(G, h_{v_2}(C)) + \tilde{\epsilon}
\]

By taking the infimum over \( Z, \tilde{Z} \in C \), we reach a contradiction. Below, we proceed to prove convexity of \( \text{dist}(G, h_{v}(C)). \)

\[
\text{dist}(G, h_{\theta v_1} + (1 - \theta) v_2(Z)) \]

\[
\text{inf}_{Z \in C} \|G - h_{\theta v_1} + (1 - \theta) v_2(Z)\|_F \leq \text{dist}(G, h_{v_1}(C)) + (1 - \theta) \text{dist}(G, h_{v_2}(C)) + \epsilon + \tilde{\epsilon}
\]

(109)

Since this holds for any \( \epsilon, \tilde{\epsilon}, \text{dist}(G, h_{v}(C)) \) is a convex function. As the square of a non-negative convex function is convex, \( J_G(v) \) is a convex function. Finally, the function \( J(v) \) is the average of convex functions, hence is convex. In (109),
the equality (I) comes from the definition of “dist”. (II) uses the argument
\[ \forall Z_1, Z_2 \in C \exists Z \in C \text{ such that } \theta h_{v_1}(Z_1) + (1 - \theta)h_{v_2}(Z_2) = (\theta h_{v_1} + (1 - \theta) h_{v_2})(Z) \] (110)
In fact, the left and right hand side of (110) have the same value on \( \tilde{T} := \text{supp}(h_v(X)) \). To more clarify this fact, when \( Z_1, Z_2, Z \in \tilde{T} \), both the right and left hand side of (110), takes the same value
\[ (\theta h_{v_1} + (1 - \theta) h_{v_2})(\text{sgn}(h_v(X))) \] (111)
To verify (110), it remains to prove
\[ \theta h_{v_1}(P_{\tilde{T}}(Z_1)) + (1 - \theta) h_{v_2}(P_{\tilde{T}}(Z_2)) = \theta h_{v_1} + (1 - \theta) h_{v_2})(P_{\tilde{T}}(Z)) \] (112)
To prove the above equality, we argue by contradiction. Suppose that the above “=” turns to “\neq” for all \( Z_1, Z_2, Z \). By setting \( Z_1 = Z_2 = Z = I_n \), we reach a contradiction.

Strict convexity. We prove strict convexity by contradiction. If \( J(v) \) were not strictly convex, there would be vectors \( v_1, v_2 \in \mathbb{R}^n_+ \) such that
\[ \mathbb{E}\left[J_G(\theta v_1 + (1 - \theta) v_2)\right] = \mathbb{E}\left[\theta J_G(v_1) + (1 - \theta) J_G(v_2)\right] \] (113)
For each \( G \) in (113), the left-hand side is smaller than or equal to the right-hand side. Therefore, in (113), \( J_G(\theta v_1 + (1 - \theta) v_2) \) and \( \theta J_G(v_1) + (1 - \theta) J_G(v_2) \) are almost surely equal (except at a measure zero set) with respect to Gaussian measure. Moreover, it holds that
\[ J_0(\theta v_1 + (1 - \theta) v_2) = \text{dist}^2(0, h_{\theta v_1 + (1 - \theta) v_2}(C)) \] (I)
\[ \inf_{Z_1, Z_2 \in C} \|h_{v_1}(Z_1) + (1 - \theta) h_{v_2}(Z_2)\|_F^2 \leq \theta \inf_{Z \in C} \|h_{v_1}(Z_1)\|_F^2 + (1 - \theta) \inf_{Z \in C} \|h_{v_2}(Z_2)\|_F^2 \] (II)
\[ \Rightarrow \theta J_0(v_1) + (1 - \theta) J_0(v_2) \] (III)
where the inequality (I) follows from (110). (II) stems from the strict convexity of \( \|\cdot\|_F^2 \). (III) is due to the definition of \( J_0 \). From (110), it can be deduced that the set \( h_v(C) \) is a convex set. Since the distance to a convex set, e.g. \( E \subseteq \mathbb{R}^{n \times n} \) (i.e. \( \text{dist}(G, E) \)) is a 1-Lipschitz function, namely
\[ \|\text{dist}(G_1, E) - \text{dist}(G_2, E)\| \leq \|G_1 - G_2\|_F \quad \forall G_1, G_2 \in \mathbb{R}^{n \times n}, \] (115)
and continuous with respect to \( G \), \( J_G(v) \) is continuous with respect to \( G \). So, there exist an open ball around \( G = 0 \) in \( \mathbb{R}^{n \times n} \) that we may write the following relation for some \( \epsilon > 0 \)
\[ \exists U \in \mathbb{R}^{n \times n} : \quad J_U(\theta v_1 + (1 - \theta) v_2) < \theta J_U(v_1) + (1 - \theta) J_U(v_2). \] (116)
Since \( \mathbb{R}^{n \times n}_+ \) is a measure zero set, the above statement contradicts (113). Hence, we have strict convexity. Continuity besides convexity of \( J \) implies that \( J(v) \) is convex on the whole domain \( v \in \mathbb{R}^3_+ \).

Attainment of the minimum. Suppose that \( v_{\min} := \min\{v_1, v_2, v_3, \frac{v_2 v_3}{v_1}\} > \|G\|_F. \) Then, we may write:
\[ \text{dist}(G, h_v(C)) = \inf_{Z \in C} \|G - h_v(Z)\|_F \] (I)
\[ \inf_{Z \in C} \|h_v(Z)\|_F - \|G\|_F \] (II)
\[ \inf_{Z \in C} \|LB^H_L Z B^H R^H - \|G\|_F \| \geq v_{\min} - \|G\|_F \geq 0, \] (117)
where in [117] the inequality (I) comes from triangle inequality of Frobenius norm. The equality (II) is the result of the decomposition provided in Lemma 4 and the rotational invariance of Frobenius norm. Lastly, (III) is obtained by combining the facts
\[ \|ABC\|_F \geq \frac{\|B\|_F}{\|A^{-1}\|_{2 \to 2} \|C^{-1}\|_{2 \to 2}} \]
for any nonsingular and conforming matrices \( A, B, C \in \mathbb{R}^{n \times n} \),
\[ \|Z\|_F \geq 1 \quad \forall Z \in C \]
and (95), (96). By squaring (117), we reach
\[ J_G(v) \geq \left( v_{\min} - \|B\|_F \right)^2 : \text{ when } v_{\min} > \|G\|_F. \] (118)
Using the relation \( E\|G\|_F \geq \frac{n}{\sqrt{n^2 - 1}} \) (25 Proposition 8.1) and Marcov’s inequality, we obtain:
\[ \mathbb{P}\left(\|G\|_F \leq n\right) \geq 1 - \frac{n}{\sqrt{n^2 - 1}}. \] (119)
Then, it holds that
\[ J(v)^{(I)} \geq \mathbb{E}\left[J_G(v)\right] \|G\|_F \leq n\mathbb{P}\left(\|G\|_F \leq n\right) + \mathbb{E}\left[J_G(v)\right] \|G\|_F > n\mathbb{P}\left(\|G\|_F > n\right) \] (II)
\[ \geq \mathbb{E}\left[J_G(v)\right] \|G\|_F \leq n\mathbb{P}\left(\|G\|_F \leq n\right) \] (III)
\[ \left(1 - \frac{n}{\sqrt{n^2 - 1}}\right) \mathbb{E}\left[\left( v_{\min} - \|G\|_F \right)^2 \|G\|_F \leq n\right] \] (IV)
\[ \left(1 - \frac{n}{\sqrt{n^2 - 1}}\right) \left( v_{\min} - n \right)^2 \] (120)
where in [120], (I) stems from total probability theorem. (II) is since \( J_G(v) \) is positive. (III) follows from (118) and (119). Lastly, (IV) is because \( (v_{\min} - n)^2 \) provides a lower-bound for the expression in the brackets inside the expectation.
From (120), one can infer that when
\[ v_{\min} > n\left(1 + \frac{\sqrt{n^2 - 1}}{\sqrt{n^2} - 1} - n\right) \] (121)
we have that \( J(v) > J(0) = n^2 \). Thus any minimizer of \( J \) must be in the set \( \left[0, n\left(1 + \frac{\sqrt{n^2 - 1}}{\sqrt{n^2} - 1} - n\right)^3\right] \).
Minimum is not the origin. Assume that \( w_i = \lambda \forall i = 1, \ldots, 3 \). Then, one may write \( J(w) \) in \((28)\) as

\[
J(w) = \text{Edist}^2(G, t\lambda e \cdot \|X \|_2^2) \leq (1)
\]

\[
\inf_{t \geq 0} \inf_{Z \in \mathcal{C} \in \|h_w(X) \|_2} \|G - t\lambda Z \|_F^2 \leq (II)
\]

\[
E \inf_{Z \in \mathcal{C} \in \|h_w(X) \|_2} \|G - \lambda Z \|_F^2 = (III)
\]

\[
E \sum_{i=1}^n \inf_{\sigma_i \leq 1} \left( \sigma_i(G) - \lambda \sigma_i(Z) \right)^2 = (IV)
\]

\[
E \sum_{i=1}^n (\sigma_i(G) - \lambda)^2 = (V)
\]

\[
v^2 \frac{1}{n} (u - \frac{\lambda}{\sqrt{n}})^2 \frac{\sqrt{4-u^2}}{\pi} \, du = (VI)
\]

\[
n^2 \varphi(\frac{\lambda}{\sqrt{n}}) < n^2 = J(0)
\]

where \( \varphi(\alpha) := \frac{-26\alpha + \alpha^2 + 24(1 + \alpha^2) \cos^{-1}(\frac{\alpha}{2})}{12\pi} \).

In \((122)\), \((I)\) is because the infimum of an affine function is concave and Jensen’s inequality. In \((II)\), we set \( t = 1 \). \((III)\) is because of Hoffman–Wielandt Theorem \([21, \text{Corollary 7.3.5}]\). \((IV)\) follows from \([76]\). \((V)\) is the result of Lemma \([6]\) \((VI)\) is since

\[
\int_0^2 (u - \alpha)^2 \frac{\sqrt{4-u^2}}{\pi} \, du =
\]

\[
\begin{cases}
\frac{3\pi - 16\alpha + 3\pi\alpha^2}{3\pi} & \alpha \leq 0 \\
\frac{-26\alpha + \alpha^2 + 24(1 + \alpha^2) \cos^{-1}(\frac{\alpha}{2})}{12\pi} & 0 \leq \alpha \leq 2 \\
0 & \alpha > 2
\end{cases}
\]

(VII) comes from \( \varphi(\alpha) \) is a decreasing function and for sufficiently small \( \lambda > 0 \) is less than 1. So, this completes the proof. ■

H. Proof of Corollary \([7]\)

Proof. In relations \([51]\) and \([52]\), using MATLAB matrix notation, \( sgn(h_w(X)) \) is obtained by \([50]\) and

\[
sgn(h_w(X)) = B_L O_L(:,1:r) O_R(:,1:r) H B_R^H,
\]

while \([52]\) is the result of the expression \( P_{T_r} (Z) = B_L O_L(:,r+1:n) O_L(:,r+1:n) H B_L^H Z B_R \)

\[
O_R(:,r+1:n) O_R(:,r+1:n) H B_R^H = B_L O_L P_{T_r}(I_n) O_L^H B_L^H Z B_R O_R P_{T_r}(I_n) O_R^H B_R^H = B_L O_L P_{T_r}(I_n) O_L^H B_L^H Z B_R O_R O_R^H B_R^H
\]

\( (124) \)

I. Proof of Lemma \([6]\)

Proof. The expression \( S_{ap} \) in Lemma \([6]\) is the result of the fact that the distribution of the singular values of a Gaussian matrix tends to the Marčenko–Pastur law \([20, \text{Theorem 3.6}]\) with probability one. In fact, this result is a more generalized statement of the argument in \([14, \text{Fact D.1}]\). ■