Maximum entropy low-rank matrix recovery

Simon Mak, Yao Xie

***Abstract*—We propose a novel, information-theoretic method, called `MaxEnt`, for efficient data requisition for low-rank matrix recovery. This proposed method has important applications to a wide range of problems in image processing, text document indexing and system identification, and is particularly effective when the desired matrix $\mathbf{X}$ is high-dimensional, and measurements from $\mathbf{X}$ are expensive to obtain. Fundamental to this design approach is the so-called maximum entropy principle, which states that the measurement masks which maximize the entropy of observations also maximize the information gain on the unknown matrix $\mathbf{X}$. Coupled with a low-rank stochastic model for $\mathbf{X}$, such a principle (i) reveals several insightful connections between information-theoretic sampling, compressive sensing and coding theory, and (ii) yields efficient algorithms for constructing initial and adaptive masks for recovering $\mathbf{X}$, which significantly outperforms random measurements. We illustrate the effectiveness of `MaxEnt` using several simulation experiments, and demonstrate its usefulness in two real-world applications on image recovery and text document indexing.**

I. Introduction

Low-rank matrices play a fundamental role in solving a wide range of statistical, engineering and machine learning problems. For many such problems, however, the low-rank matrix $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$ cannot be directly or fully observed as data. Instead, the observations $\mathbf{y} \in \mathbb{R}^n$ are typically obtained as $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\epsilon}$, where $\mathcal{A} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ is a linear measurement operator, and $\boldsymbol{\epsilon}$ is a vector of measurement noise. The goal then is to recover the underlying matrix $\mathbf{X}$ from observations $\mathbf{y}$, a problem known as *matrix recovery*. There are two key challenges for such a recovery. First, the low-rank matrix $\mathbf{X}$ is typically high-dimensional in practice, meaning m_1 or m_2 can grow quite large. For example, in the well-known Netflix prize problem [1], the matrix $\mathbf{X}$ of user-movie preferences in the database can span millions of users and thousands of movies. Second, in many applications in the physical or biological sciences, the cost of obtaining measurements from $\mathbf{X}$ can also be quite expensive. One such example is in genetic studies, where costly, time-intensive experiments are needed to gain information on the matrix $\mathbf{X}$ of gene-disease expression levels [2]. In light of these two challenges, we propose in this paper a novel, information-theoretic method for *designing* the measurement operator $\mathcal{A}$, so that more information can be extracted and a better recovery can be achieved on the unknown matrix $\mathbf{X}$ compared to random measurements.

Given the increasing prevalence of low-rank modeling in scientific and engineering problems [3], the proposed methodology has important applications to a broad spectrum of important problems. We briefly review several such problems which motivated this work:

- *Image processing*: For many imaging systems, the measurements obtained are of the form $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is the pixel matrix for the image of interest, and $\mathcal{A}$ corresponds to the measurement masks used to observe this image. Imaging systems of this form arise in many real-world applications, including single-pixel cameras [4], compressive hyperspectral imaging [5], X-ray imaging [6] and spacecraft imaging [7]. Particularly for the latter two applications, the image measurements $\mathbf{y}$ can be expensive to generate. For such applications, the proposed method can be used to *design* the measurement masks, so that one can maximize image recovery performance given a certain budget constraint.
- *Text document indexing*: A key challenge encountered in data mining is the sheer size of the database at hand (e.g., the large number of documents in text analysis), which greatly restricts the use of standard data analytic techniques. For large text databases, one way of tackling this problem is to compress (or *index*) the large database into a smaller representative summary. This compression is typically performed by randomly projecting the large database onto a smaller subspace (see [8], [9] for details); in other words, the summary data $\mathbf{y}$ is generated as $\mathbf{y} = \mathcal{A}(\mathbf{X})$, where $\mathcal{A}$ is a random linear projection operator. For this application, the proposed method can be used to *design* the projection operator $\mathcal{A}$ so that a good compression of the database $\mathbf{X}$ can be achieved.
- *System identification*: Many engineering processes can be well-approximated by the following simple discrete-time, linear time-invariant model:

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t),$$

  where $\mathbf{x}(t)$ denotes the system states, $\mathbf{u}(t)$ denotes the control inputs, and $\mathbf{y}(t)$ denotes the observed outputs, all at time t. Given data on the system of interest, [10] showed that the parameter matrices (i.e., $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$) can be estimated using a low-rank matrix recovery formulation. For such applications, however, the system data can be expensive to collect, since measurements must be taken when the system is running. The proposed method can be used to *design* the measurement procedure, so that a maximum amount of information can be gained on the system given a fixed operating budget.

Over the past decade, there has been a rapidly growing body of literature on the topic of low-rank *matrix recovery*, focusing largely on the theoretical properties of this recovery via convex programming. This includes the seminal works of [11] and

Simon Mak (e-mail: smak6@gatech.edu) and Yao Xie (e-mail: yao.xie@isye.gatech.edu) are with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

This work was partially supported by NSF grants CCF-1442635, CMMI-1538746, and an NSF CAREER Award CCF-1650913.

[12], who investigated the theoretical conditions necessary for a successful recovery of $\mathbf{X}$ using nuclear-norm minimization, as well as numerous subsequent works (e.g., [13], [14]) which improved upon such conditions. A related problem, called *matrix completion* (where matrix entries are directly observed), has received special attention; this includes the pioneering papers [15], [16], [17], [18], as well as many subsequent works. However, nearly all of the literature on matrix recovery (or matrix completion) assumes the measurement masks (or the missing entries) are sampled at random, since this allows for easy implementation and more amenable theoretical analysis. For the specific case of matrix completion, there has been some recent work on developing an informed (or *designed*) strategy for sampling matrix entries [19], [20], [21]. To the best of our knowledge, no one has tackled the design problem for the more general setting of matrix recovery from a maximum entropy design perspective; this is the aim of the current paper.

We propose here a novel information-theoretic framework for designing the measurement masks in the linear operator $\mathcal{A}$, with the desired goal being an improved recovery of the low-rank matrix $\mathbf{X}$ compared to randomly sampled masks. The contributions of our work are three-fold. First, we derive a design principle called the *maximum entropy principle* for the matrix recovery problem; this principle states that *the measurement operator $\mathcal{A}$ which maximizes the entropy of observations $\mathbf{y}$ corresponds to the operator which maximizes information gain on the underlying matrix $\mathbf{X}$*. Such a principle yields a simple design criterion involving only the *observations* $\mathbf{y}$, which serves as a proxy for more complicated criteria involving the low-rank *matrix* $\mathbf{X}$ (the matrix $\mathbf{X}$ is typically much more high-dimensional than the observations $\mathbf{y}$, see, e.g., [12]). Next, adopting a so-called singular matrix-variate Gaussian stochastic model on $\mathbf{X}$, we reveal several interesting and valuable insights between the maximum entropy design principle and well-known concepts from compressive sensing and coding theory (such as coherence property). Using such insights, we then develop a novel algorithm called `MaxEnt` for efficiently designing initial and sequential measurement masks in $\mathcal{A}$. Finally, we demonstrate the effectiveness of this proposed design strategy using several numerical simulations, and two real-world applications on image processing and text document indexing.

There is a large body of work on information-theoretic design (e.g., for compressive sensing) and its many real-world applications (see, e.g., [22]), and we would be remiss if we did not mention important developments in this field. This includes the seminal paper [23], who showed the profound fact that, for linear vector Gaussian channels, the gradient of the mutual information is related to the minimum mean-squared error matrix for parameter estimation. Such a result is further developed by [24], [25] and [26] for designing measurement matrices in compressive sensing and phase retrieval. The key difference between the above approaches and our work is that the former aims to directly maximize the mutual information between signal (i.e., $\mathbf{X}$) and measurements (i.e., $\mathbf{y}$), whereas our work examines a dual (but equivalent) problem of maximizing the entropy of observations $\mathbf{y}$, using the maximum entropy design principle. As we show in the paper, this dual view sometimes provides the advantage of efficient initial and adaptive constructions of measurement masks via code design, which then allows for effective matrix recovery in high-dimensions. A related maximum entropy approach was also employed in [27] for developing a general minimax approach to supervised learning. Our work on matrix recovery can also be seen as a novel extension of the information-theoretic matrix completion work [21]; by exploring general measurement masks (rather than entrywise measurements), this paper reveals new insights on mask construction and code design, and provides an efficient and closed-form method for constructing masks.

This paper is organized as follows. Section 2 outlines the matrix recovery framework and the proposed model specification. Section 3 introduces the maximum entropy principle, and demonstrates how such a principle can be applied for designing the measurement masks in $\mathcal{A}$. Sections 4 and 5 reveal several insights on initial and adaptive design, including a useful connection between maximum entropy masks and optimal subspace packings. Section 6 details a design algorithm called `MaxEnt`, which can efficiently construct initial and adaptive masks to maximize information gain on $\mathbf{X}$. Section 7 demonstrates the effectiveness of `MaxEnt` using several simulation experiments, and its usefulness in solving real-world problems on image recovery and text document indexing. Section 8 concludes with thoughts on future work.

II. Model Framework for Matrix Recovery

We begin by first outlining the matrix recovery problem, then reviewing the singular matrix-variate Gaussian model [21]. This model will serve as a versatile probabilistic model for the low-rank matrices of interest throughout the paper.

A. Problem set-up

Let $\mathbf{X} = (X_{i,j}) \in \mathbb{R}^{m_1 \times m_2}$ be the low-rank matrix of interest. Suppose $\mathbf{X}$ is observed through a set of masks $\{\mathbf{A}_i\}_{i=1}^n \subseteq \mathbb{R}^{m_1 \times m_2}$, with the resulting samples then corrupted by independent and identically distributed (i.i.d.) Gaussian noise. The resulting noisy observations, $\{y_i\}_{i=1}^n$, then follow the model:

$$y_i = \mu_i + \epsilon_i,\ \mu_i = \langle \mathbf{A}_i, \mathbf{X} \rangle_F,\ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \eta^2),\ i = 1, \cdots, n, \tag{1}$$

where $\langle \mathbf{A}, \mathbf{X} \rangle_F = \mathrm{tr}(\mathbf{A}^T \mathbf{X})$ is the Frobenius inner-product. We assume all masks satisfy the unit power constraint $\|\mathbf{A}_i\|_F^2 \leq 1$, as is typical in matrix sensing problems [3]. With $\mathbf{y} = (y_i)_{i=1}^n$ and $\boldsymbol{\epsilon} = (\epsilon_i)_{i=1}^n$, (1) can then be written in vector form:

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathcal{A} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ is the linear measurement operator returning $\mathcal{A}(\mathbf{X}) = (\mu_i)_{i=1}^n$, the mean vector of observations from $\mathbf{X}$.

With this in hand, the desired goal of designing masks for matrix recovery can be made more precise. We consider here the following two-step design approach, commonly employed in the field of (statistical) experimental design [28]. First, the *initial* masks $\{\mathbf{A}_i\}_{i=1}^n$ are designed so that a maximum amount

of initial information can be gained on $\mathbf{X}$, given no prior knowledge on the unknown matrix. Next, using this initial information, the *sequential* masks $\mathbf{A}_{n+1}, \mathbf{A}_{n+2}, \cdots$ are then designed to adaptively maximize subsequent information gain on $\mathbf{X}$. The singular matrix-variate Gaussian distribution, presented below, provides an appealing framework for developing this two-step information-theoretic design methodology.

B. Model specification

1) The singular matrix-variate Gaussian distribution: Suppose $\mathbf{X}$ is normalized with zero mean, and consider the following model for $\mathbf{X}$:

Definition 1 (Singular matrix-variate Gaussian (SMG); Definition 2.4.1, [29])**.** *Let $\mathbf{Z} \in \mathbb{R}^{m_1 \times m_2}$ be a random matrix with entries $Z_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \cdots, m_1$ and $j = 1, \cdots, m_2$. The random matrix $\mathbf{X}$ has a* singular matrix-variate Gaussian *distribution if $\mathbf{X} \overset{d}{=} \mathcal{P}_{\mathcal{U}} \mathbf{Z} \mathcal{P}_{\mathcal{V}}$ for some choice of projection matrices $\mathcal{P}_{\mathcal{U}} = \mathbf{U}\mathbf{U}^T$ and $\mathcal{P}_{\mathcal{V}} = \mathbf{V}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{m_1 \times R}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V} \in \mathbb{R}^{m_2 \times R}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $R < m_1 \wedge m_2$.*[1] *We will denote this as $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_{\mathcal{U}}, \mathcal{P}_{\mathcal{V}}, \sigma^2, R)$.*

From a simulation point-of-view, a realization from the SMG distribution is obtained by first (i) simulating a random matrix $\mathbf{Z}$ with each entry following an i.i.d. $\mathcal{N}(0, \sigma^2)$ distribution, then (ii) performing a left and right projection of $\mathbf{Z}$ using the projection matrices $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$. Here, $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ can be seen as projection operators which map vectors from $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$ onto $\mathcal{U}$ and $\mathcal{V}$, the R-dimensional linear subspaces spanned by the orthonormal columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. After this left-right projection of $\mathbf{Z}$, one can then show that the resulting matrix $\mathbf{X} = \mathcal{P}_{\mathcal{U}} \mathbf{Z} \mathcal{P}_{\mathcal{V}}$ has rank $R < m_1 \wedge m_2$, with its row and column spaces lying in $\mathcal{U}$ and $\mathcal{V}$, respectively. For R sufficiently small, the SMG distribution provides a flexible framework for modeling the low-rank structure of $\mathbf{X}$.

Similar to its application for the design problem in matrix completion [21], the parametrization of the SMG model using projection matrices yields several appealing features for mask design in the broader problem of matrix recovery. First, such a parametrization encodes valuable information on the subspaces of $\mathbf{X}$. Recall that each projection operator $\mathcal{P}_{\mathcal{W}} \in \mathbb{R}^{m \times m}$ of rank R corresponds to a unique R-plane $\mathcal{W}$ (i.e., an R-dimensional linear subspace) in $\mathbb{R}^m$, so $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ parametrize important information on the row space $\mathcal{U} \in \mathcal{G}_{R,m_1-R}$ and the column space $\mathcal{V} \in \mathcal{G}_{R,m_2-R}$, where $\mathcal{G}_{R,m-R}$ is the *Grassmann manifold* consisting of all R-planes in $\mathbb{R}^m$. This parametrization can then be used to derive insightful connections between initial mask design and optimal Grassmann packings (see Section IV). Second, using the fact that the projection of Gaussian random vector is still Gaussian-distributed, we show later in this section that, conditional on observations $\mathbf{y}$, the observations from subsequent masks will also be Gaussian-distributed. This property is key for deriving a closed-form sequential mask design scheme which adaptively maximizes information gain on $\mathbf{X}$ (see Section V).

[1]Here, $m_1 \wedge m_2 := \min(m_1, m_2)$ and $m_1 \vee m_2 := \max(m_1, m_2)$.

C. Connection to nuclear-norm recovery

For most practical scenarios, there is little-to-no prior knowledge on either the rank of $\mathbf{X}$ or its subspaces. In such a scenario, a Bayesian approach [30] would be to assign non-informative prior distributions to the model parameters R, $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$, by assuming all possible ranks and subspaces are equally likely to occur. Adopting these non-informative priors, the following lemma reveals an insightful expression for the maximum-a-posteriori (MAP) estimator of $\mathbf{X}$:

Lemma 1 (MAP estimator)**.** *Assume η^2 and σ^2 are fixed, and suppose the subspaces for $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ are uniformly distributed over the Grassmann manifolds $\mathcal{G}_{R,m_1-R}$ and $\mathcal{G}_{R,m_2-R}$, with matrix rank R uniformly sampled from $\{1, \cdots, m_1 \wedge m_2\}$. Conditional on $\mathbf{y}$, the* MAP estimator *for $\mathbf{X}$ becomes:*

$$\widetilde{\mathbf{X}} \in \underset{\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}}{\text{Argmin}} \left[\frac{\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2}{\eta^2} + \log(2\pi\sigma^2)\text{rank}^2(\mathbf{X}) + \frac{\|\mathbf{X}\|_F^2}{\sigma^2} \right], \tag{3}$$

where $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{i,j}^2}$ is the Frobenius norm of $\mathbf{X}$.

Consider now the following approximation of the formulation in (3). First, viewing the rank penalty term $\log(2\pi\sigma^2)\text{rank}^2(\mathbf{X})$ as a Lagrange multiplier, such a term can be replaced by the constraint $\text{rank}(\mathbf{X}) \leq \sqrt{\xi}$. Next, changing this constraint into its Lagrangian form, and relaxing the rank function $\text{rank}(\mathbf{X})$ into the nuclear norm $\|\mathbf{X}\|_*$, which is the tightest convex relaxation [11], the formulation in (3) becomes:

$$\widetilde{\mathbf{X}} = \underset{\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}}{\text{Argmin}} \left[\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2 + \lambda \left\{ \alpha \|\mathbf{X}\|_* + (1-\alpha)\|\mathbf{X}\|_F^2 \right\} \right], \tag{4}$$

for an appropriate choice of $\lambda > 0$ and $\alpha \in (0, 1)$. The problem in (4) can then be viewed as an "elastic net" [31] formulation for low-rank matrix recovery.

The formulation in (4) yields an interesting link between the MAP estimator $\widetilde{\mathbf{X}}$ and existing recovery methods. Setting $\alpha = 1$, (3) reduces to the nuclear-norm formulation:

$$\widehat{\mathbf{X}} = \underset{\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}}{\text{Argmin}} \left[\|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2 + \lambda \|\mathbf{X}\|_* \right], \tag{5}$$

which is widely used for the recovery of low-rank matrices [32], [33]. Such a connection allows us to adapt state-of-the-art algorithms for solving (5) to efficiently guide the sequential mask design procedure, as detailed in Section VI.

D. Useful model properties

We now introduce several properties of the SMG model which will prove useful later for mask design. These properties concern the joint distribution of measurements, both prior to and conditional on obtaining observations from $\mathbf{X}$.

1) Joint distribution prior to observations: Let y_i and y_j be observations from (1) using masks $\mathbf{A}_i$ and $\mathbf{A}_j$, respectively. The following lemma gives a closed-form joint distribution for y_i and y_j, prior to observing any data on $\mathbf{X}$:

Lemma 2 (Unconditional joint distribution)**.** *Suppose $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_{\mathcal{U}}, \mathcal{P}_{\mathcal{V}}, \sigma^2, R)$, with $\mathcal{P}_{\mathcal{U}}$, $\mathcal{P}_{\mathcal{V}}$, σ^2 and R fixed. Then,*

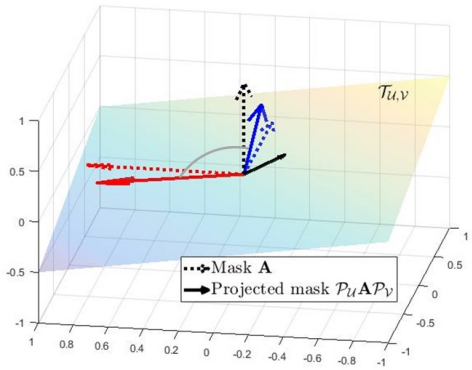

Figure 1. A visualization of three masks (dotted vectors) on the Frobenius inner-product space $\langle\cdot,\cdot\rangle_F$, and its projections (solid vectors) onto subspace $\mathcal{T}_{\mathcal{U},\mathcal{V}}$ (corresponding to the row and column spaces of $\mathbf{X}$).

for $i \neq j$:

$$\begin{bmatrix} y_i \\ y_j \end{bmatrix} \sim \mathcal{N}\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} \xi_{i,i} & \xi_{i,j} \\ \xi_{i,j} & \xi_{j,j} \end{pmatrix} \right\}, \tag{6}$$

where:

$$\begin{aligned} \xi_{i,i} &= \text{Var}(y_i) = \sigma^2 \|\mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}\|_F^2 + \eta^2, \\ \xi_{i,j} &= \text{Cov}(y_i, y_j) = \sigma^2 \langle \mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}}\mathbf{A}_j\mathcal{P}_{\mathcal{V}} \rangle_F. \end{aligned} \tag{7}$$

These closed-form variance and covariance expressions can nicely be interpreted as the similarities between the measurement masks and the subspaces of $\mathbf{X}$ (i.e., the row space $\mathcal{U}$ and column space $\mathcal{V}$). Consider first the variance expression for $\text{Var}(y_i)$, which (ignoring the measurement noise term η^2) is proportional to $\|\mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}\|_F^2$. Viewed geometrically, the variance of an observation from mask $\mathbf{A}_i$ is proportional to the norm of $\mathbf{A}_i$, *after* accounting for the *similarity* of such a mask with the subspaces of $\mathbf{X}$ via a left-right projection by $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$. Consider next the covariance between y_i and y_j, which is proportional to the inner-product $\langle \mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}}\mathbf{A}_j\mathcal{P}_{\mathcal{V}} \rangle_F$. This can be interpreted as the *angle* between the two measurement masks $\mathbf{A}_i$ and $\mathbf{A}_j$, *after* accounting for their similarities with the subspaces of $\mathbf{X}$.

Figure 1 provides a visualization of this geometric interpretation. Here, the shaded subspace $\mathcal{T}_{\mathcal{U},\mathcal{V}}$ represents the projected subspace from a left-right projection by $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ (further details on $\mathcal{T}_{\mathcal{U},\mathcal{V}}$ in Lemma 11), the three dotted vectors represent three measurement masks in the Frobenius inner-product space $\langle\cdot,\cdot\rangle_F$, and the three solid vectors represent the projection of these masks onto $\mathcal{T}_{\mathcal{U},\mathcal{V}}$. The variance term $\text{Var}(y_i) \propto \|\mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}\|_F^2$ (again, ignoring η^2) can be viewed as the *length* of the projected vector (solid) for measurement mask $\mathbf{A}_i$. Comparing the three projected vectors in Figure 1, we see that the mask with greater similarity to $\mathcal{U}$ and $\mathcal{V}$ (the red vector) results in a longer projected vector, meaning observations from such a mask have greater variability. Likewise, the covariance term $\text{Cov}(y_i, y_j) \propto \langle \mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}}\mathbf{A}_j\mathcal{P}_{\mathcal{V}} \rangle_F$ can be viewed as the *inner-product* between the projected vectors for two measurement masks $\mathbf{A}_i$ and $\mathbf{A}_j$. A larger inner-product (or smaller angle) between these two projected vectors indicates a higher correlation between the observations from masks $\mathbf{A}_i$ and $\mathbf{A}_j$. As we demonstrate later, the goal of designing masks which *maximize* information on $\mathbf{X}$ can approximately be viewed as finding masks which *maximize* the angles between their projected vectors. This intuitive notion of "angle maximization" will reappear in Section IV in the context of initial mask design.

2) Joint distribution conditional on observations: Now, suppose the observations $\mathbf{y}$ have been observed from (1), and let y_{n+1} and y_{n+2} be new observations from masks $\mathbf{A}_{n+1}$ and $\mathbf{A}_{n+2}$. Using the conditional property of the Gaussian distribution, the following lemma provides a closed-form joint distribution for y_{n+1} and y_{n+2} conditional on $\mathbf{y}$:

Lemma 3 (Conditional joint distribution)**.** *Let $\mathbf{y}$ be observations obtained from masks $\mathbf{A}_{1:n} = [\mathbf{A}_1\ \mathbf{A}_2\ \cdots\ \mathbf{A}_n]$, and let y_{n+1} and y_{n+2} be new observations from masks $\mathbf{A}_{n+1}$ and $\mathbf{A}_{n+2}$. Assuming $\mathbf{X} \sim \mathcal{SMG}(\mathcal{P}_{\mathcal{U}}, \mathcal{P}_{\mathcal{V}}, \sigma^2, R)$, with $\mathcal{P}_{\mathcal{U}}$, $\mathcal{P}_{\mathcal{V}}$, σ^2 and R fixed, we have:*

$$\begin{bmatrix} y_{n+1} \\ y_{n+2} \end{bmatrix} \Big| \mathbf{y} \sim \mathcal{N}\left\{ \begin{bmatrix} \mathbb{E}(y_{n+1}|\mathbf{y}) \\ \mathbb{E}(y_{n+2}|\mathbf{y}) \end{bmatrix}, \begin{pmatrix} \xi_{n+1,n+1}|\mathbf{y} & \xi_{n+1,n+2}|\mathbf{y} \\ \xi_{n+1,n+2}|\mathbf{y} & \xi_{n+2,n+2}|\mathbf{y} \end{pmatrix} \right\}, \tag{8}$$

where, with:

$$\begin{aligned} \mathbf{R}_n(\mathbf{A}_{1:n}) &:= [\langle \mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}}\mathbf{A}_j\mathcal{P}_{\mathcal{V}} \rangle_F]_{i,j=1}^n \in \mathbb{R}^{n\times n}, \\ \mathbf{r}_n(\mathbf{A}) &:= [\langle \mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}}\mathbf{A}\mathcal{P}_{\mathcal{V}} \rangle_F]_{i=1}^n \in \mathbb{R}^n, \end{aligned} \tag{9}$$

and $\gamma^2 := \eta^2/\sigma^2$, we have:

$$\begin{aligned} \mathbb{E}(y_i|\mathbf{y}) &= \mathbf{r}_n^T(\mathbf{A}_i)[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{y}, \\ \xi_{i,i}|\mathbf{y} &:= \text{Var}(y_i|\mathbf{y}) \\ &= \text{Var}(y_i) - \sigma^2\mathbf{r}_n^T(\mathbf{A}_i)[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{r}_n(\mathbf{A}_i), \\ \xi_{i,j}|\mathbf{y} &:= \text{Cov}(y_i, y_j|\mathbf{y}) \\ &= \text{Cov}(y_i, y_j) - \sigma^2\mathbf{r}_n^T(\mathbf{A}_i)[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{r}_n(\mathbf{A}_j). \end{aligned} \tag{10}$$

These closed-form conditional variance and covariance expressions also enjoy intuitive interpretations. In particular, the *conditional* variance of a new observation, $\text{Var}(y_{n+1}|\mathbf{y})$, can be decomposed as the *unconditional* variance $\text{Var}(y_{n+1})$, minus a *reduction* term which quantifies how correlated the new mask $\mathbf{A}_{n+1}$ is to the observed masks $\mathbf{A}_{1:n}$ after accounting for the subspaces of $\mathbf{X}$. Similarly, the *conditional* covariance between two new observations, $\text{Cov}(y_{n+1}, y_{n+2}|\mathbf{y})$, can be decomposed as the *unconditional* covariance $\text{Cov}(y_{n+1}, y_{n+2})$, minus a *reduction* term quantifying how correlated the new masks $\mathbf{A}_{n+1}$ and $\mathbf{A}_{n+2}$ are to the observed masks $\mathbf{A}_{1:n}$ after a left-right projection. These expressions provide a new way of quantifying the correlation between two measurement masks, one which takes into account the desired low-rank matrix $\mathbf{X}$, as well as *prior* observation masks. The expressions in (10) will again reappear when deriving the sequential mask design procedure in Section V.

III. An Information-Theoretic View on Mask Design

Next, we present an information-theoretic mask design approach for the matrix recovery problem. We first outline the design principle of maximum entropy, then illustrate how such a principle can be used to effectively design measurement masks which maximize information gain on $\mathbf{X}$.

A. The design principle of maximum entropy

The principle of maximum entropy was first introduced in [34] and further developed in [35] in the context of (statistical) experimental design for spatio-temporal modeling, although the origins of information-theoretic experimental design date back much further to the seminal works of [36] and [37]. Simply put, this principle states that, under certain regularity assumptions on an observation model with unknown model parameters, *a design scheme which maximizes the entropy of collected data is a design scheme which maximizes information gain on model parameters*. In other words, to find a design scheme which maximizes information on unknown model parameters, one can equivalently find the design which maximizes the entropy of the collected samples. As described in [35], the maximum entropy principle offers two important advantages for design. First, it allows for *efficient* design construction, since the entropy expression for observed data is typically simple to optimize. Second, the trade-off between sample entropy and model information often reveals useful insights. For the matrix recovery problem, we will demonstrate throughout the paper how such advantages play a role in both theory and algorithm.

To formally present the maximum entropy principle, we require the following definitions. Let (X, Y) be a pair of random variables (r.v.s) with marginal densities $(f_X(x), f_Y(y))$ and joint density $f_{X,Y}(x, y)$. Following [38], the *entropy* of X is defined as $\mathrm{H}(X) = \mathbb{E}[-\log f_X(X)]$, with larger values indicating greater uncertainty for r.v. X. Similarly, the *joint entropy* of (X, Y) is defined as $\mathrm{H}(X, Y) = \mathbb{E}[-\log f_{X,Y}(X, Y)]$, and the *conditional entropy* of Y given X is defined as the entropy of the conditional r.v. $Y|X$, denoted by $\mathrm{H}(Y|X)$. The following chain rule (Theorem 2.2.1 in [38]) provides a link between the joint entropy $\mathrm{H}(X, Y)$ with the conditional entropy $\mathrm{H}(Y|X)$:

$$\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y|X). \tag{11}$$

This identity will serve as the basis for deriving the maximum entropy principle below.

With this in hand, consider now the matrix recovery problem. Here, the parameter-of-interest is the unknown low-rank matrix $\mathbf{X}$, the design scheme is the choice of measurement masks $\mathbf{A}_{1:n} = [\mathbf{A}_1\ \mathbf{A}_2\ \cdots\ \mathbf{A}_n]$, and the collected data are the observations $\mathbf{y}$ taken from such masks. Using the chain rule in (11), we get the following decomposition:

$$\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y}, \mathbf{X}) = \mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y}) + \mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y}), \tag{12}$$

where the subscript for $\mathbf{A}_{1:n}$ indicates the measurements were collected from masks $\mathbf{A}_{1:n}$. The first term in (12) is the joint entropy of the observations $\mathbf{y}$ and the matrix $\mathbf{X}$, the middle term $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y})$ is the entropy of observations $\mathbf{y}$ from masks $\mathbf{A}_{1:n}$, and the last term $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$ corresponds to the conditional entropy of matrix $\mathbf{X}$ after observing $\mathbf{y}$. The goal here is to design the masks $\mathbf{A}_{1:n}$ which *minimize* the conditional entropy $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$, thereby *maximizing* the amount of information gained on $\mathbf{X}$ after observing $\mathbf{y}$.

The maximum entropy principle can then be derived as follows. Applying the chain rule again to the joint entropy $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y}, \mathbf{X})$ on the left-hand side of (12), we get:

$$\begin{aligned}
\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y}, \mathbf{X}) &= \mathrm{H}(\mathbf{X}) + \mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y}|\mathbf{X}) && \text{(by (11))} \\
&= \mathrm{H}(\mathbf{X}) + \mathrm{H}_{\mathbf{A}_{1:n}}(\mathcal{A}(\mathbf{X}) + \boldsymbol{\epsilon}|\mathbf{X}) && \text{(by (2))} \\
&= \mathrm{H}(\mathbf{X}) + \mathrm{H}(\boldsymbol{\epsilon}|\mathbf{X}) && (\mathcal{A}(\mathbf{X}) \text{ is fixed given } \mathbf{X}) \\
&= \mathrm{H}(\mathbf{X}) + \mathrm{H}(\boldsymbol{\epsilon}). && (\boldsymbol{\epsilon} \text{ and } \mathbf{X} \text{ are independent})
\end{aligned}$$

The key observation here is that the final quantity $\mathrm{H}(\mathbf{X}) + \mathrm{H}(\boldsymbol{\epsilon})$ – the sum of entropies for matrix $\mathbf{X}$ and measurement noise $\boldsymbol{\epsilon}$ – *does not depend on the choice of masks* $\mathbf{A}_{1:n}$. In view of the decomposition in (12), it follows that the left-hand side of (12) also does not depend on $\mathbf{A}_{1:n}$, and hence the masks $\mathbf{A}_{1:n}$ which *minimize* $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$ in (12) are the same masks that *maximize* $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y})$ as well. This is precisely the maximum entropy principle for matrix recovery – *a mask design which maximizes the entropy of observations* $\mathbf{y}$ *in turn yields maximum information gain on* $\mathbf{X}$. Computationally, such a principle allows us to employ for our setting the simpler entropy expression $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y})$ as an efficient proxy for the desired entropy quantity $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$, which is much more complicated and difficult to minimize in high-dimensions.

B. Maximum entropy masks

Consider now the entropy of observations $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y})$, which (with a slight abuse of notation) we abbreviate as $\mathrm{H}(\mathbf{A}_{1:n})$. For fixed $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$, a straight-forward extension of Lemma 2 gives the following closed-form expression for $\mathrm{H}(\mathbf{A}_{1:n})$:

$$\mathrm{H}(\mathbf{A}_{1:n}) := \mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y}) = \det\{\sigma^2 \mathbf{R}_n(\mathbf{A}_{1:n}) + \eta^2 \mathbf{I}\}, \tag{13}$$

where $\mathbf{R}_n(\mathbf{A}_{1:n}) = [\langle \mathcal{P}_{\mathcal{U}} \mathbf{A}_i \mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}} \mathbf{A}_j \mathcal{P}_{\mathcal{V}} \rangle_F]_{i,j=1}^n$ is the matrix of inner-products in (9). The masks which maximize $\mathrm{H}(\mathbf{A}_{1:n})$ can then be defined as follows:

Definition 2 (Maximum entropy masks)**.** *For fixed* $\mathcal{P}_{\mathcal{U}}$ *and* $\mathcal{P}_{\mathcal{V}}$*, the* maximum entropy masks $\mathbf{A}^*_{1:n}$ *satisfy:*

$$\mathbf{A}^*_{1:n} := \operatorname*{Argmax}_{\substack{\mathbf{A}_1, \cdots, \mathbf{A}_n \in \mathbb{R}^{m_1 \times m_2} \\ \|\mathbf{A}_i\|_F^2 \leq 1}} \mathrm{H}(\mathbf{A}_{1:n}). \tag{14}$$

By the maximum entropy principle, the maximum entropy masks $\mathbf{A}^*_{1:n}$ can be interpreted as the mask design which *maximizes* the information gained on $\mathbf{X}$. As such, this mask design enjoys two appealing features for matrix recovery. First, the maximum entropy formulation in (14) yields several insightful connections to compressive sensing and coding theory. Such connections are not only of theoretical interest, but can be further exploited for developing efficient, closed-form mask construction algorithms. Second, by maximizing information gain on $\mathbf{X}$ (or equivalently, minimizing the conditional entropy $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$), these masks can yield considerable reductions for recovery error in a mean-squared sense, which is ultimately the desired goal. Indeed, the relation between conditional entropy $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$ and expected recovery error $\mathbb{E}[\|\mathbf{X} - \tilde{\mathbf{X}}\|_F^2|\mathbf{y}]$, $\tilde{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\mathbf{y}]$ can be expressed via the following lower bound (Equation 27 in [39]):

$$\mathbb{E}[\|\mathbf{X} - \tilde{\mathbf{X}}\|_F^2|\mathbf{y}] \geq \frac{1}{2\pi e} \exp\left\{2\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})\right\}. \tag{15}$$

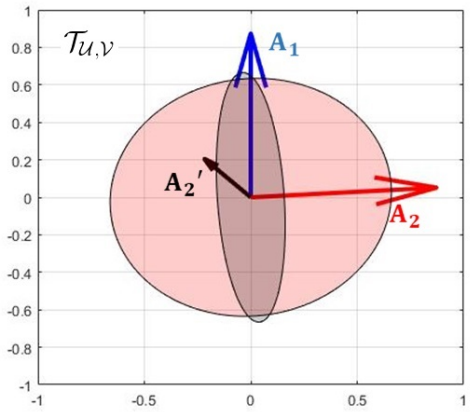

Figure 2. The covariance matrices for the red and blue masks (red ellipse), and for the black and blue masks (black ellipse) from Figure 1.

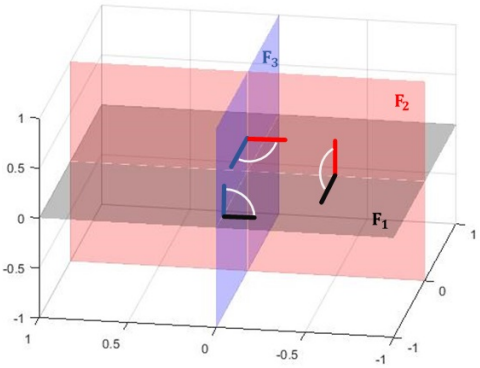

Figure 3. A visualization of the optimal Grassmann packing problem for $n = 3$ frames, each in $R = 2$ dimensions. White arcs denote principal angles between any two of the three subspaces.

The maximum entropy masks $\mathbf{A}^*_{1:n}$, which minimize the conditional entropy $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$, can in turn yield lower expected recovery error as a result of this bound. As mentioned above, the precise advantage in working with $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$ is that it allows us to manipulate the simpler observation entropy term $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{y})$, whereas the error term $\mathbb{E}[\|\mathbf{X} - \tilde{\mathbf{X}}\|_F^2|\mathbf{y}]$ is much more cumbersome to work with for mask construction, particularly in high-dimensions.

The maximum entropy masks also yield a natural geometric interpretation. The crux of this interpretation lies in the fact that the entropy $\mathrm{H}(\mathbf{A}_{1:n})$ in (13) can be viewed as the *volume* of the covariance matrix formed by the projected masks $\{\mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}\}_{i=1}^n$ on $\mathcal{T}_{\mathcal{U},\mathcal{V}}$. Figure 2 visualizes this using the previous example in Figure 1. Here, the red ellipse corresponds to the covariance matrix for the red and blue projected vectors, and the black ellipse corresponds to the covariance matrix for the black and blue projected vectors. Note that the volume of the red ellipse is larger than that for the black ellipse, which suggests that by sampling the red and blue masks, the resulting observations have greater entropy than that obtained from sampling the black and blue masks. By the maximum entropy principle, the former observations yield more information on $\mathbf{X}$ than the latter observations. The goal is then to design measurement masks which maximize the volume of their covariance matrix, after accounting for its similarity with the subspaces of $\mathbf{X}$.

IV. Insights on initial mask design

With this in hand, we consider first the problem of designing *initial* masks which maximize information gain on $\mathbf{X}$, given no prior knowledge or observations on the underlying matrix. Using the maximum entropy principle, we show how this initial design problem is intricately related to the problem of optimal subspace packing, which will prove useful for developing an effective construction of these initial masks.

A. Block coherence of frames

Recall that an R-frame in $\mathbb{R}^m$ is a matrix $\mathbf{F} \in \mathbb{R}^{m\times R}$ with orthonormal columns, i.e., $\mathbf{F}^T\mathbf{F} = \mathbf{I}$. The worst-case and average block coherence between two frames can then be defined as:

Definition 3 (Worst-case and avg. block coherence; [40])**.** *Let $\mathbf{F}_{1:n} = [\mathbf{F}_1\ \mathbf{F}_2\ \cdots\ \mathbf{F}_n] \in \mathbb{R}^{m\times nR}$ be a collection of R-frames in $\mathbb{R}^m$, where $m \geq 2R$, and let $\|\cdot\|_2$ be the spectral norm. The* worst-case block coherence *of $\mathbf{F}_{1:n}$ is defined as:*

$$\mu(\mathbf{F}_{1:n}) := \max_{i\neq j} \|\mathbf{F}_i^T\mathbf{F}_j\|_2, \tag{16}$$

and the average block coherence *of $\mathbf{F}_{1:n}$ is defined as:*

$$a(\mathbf{F}_{1:n}) := \frac{1}{n-1}\max_i \Big\|\sum_{j:j\neq i}\mathbf{F}_i^T\mathbf{F}_j\Big\|_2. \tag{17}$$

The worst-case block coherence $\mu(\mathbf{F}_{1:n})$ and average block coherence $a(\mathbf{F}_{1:n})$ play a role in quantifying the error performance of block compressive sensing methods [41]; the lower the coherence, the easier recovery becomes. The minimization of these coherence metrics has an appealing geometric connection to the problem of optimal *Grassmann packings* – the packing of R-dimensional subspaces in $\mathbb{R}^m$. In particular, it is shown [42] that the packing which maximizes the minimum principal angles between two subspaces, corresponds to frames $\mathbf{F}_{1:n}$ which minimize the worst-case block coherence $\mu(\mathbf{F}_{1:n})$. Figure 3 visualizes this for $m = 3$ frames, each in $R = 2$ dimensions. One sees that the principal angles between any two of the three subspaces are maximized, so the corresponding frames minimize the worst-case coherence $\mu(\mathbf{F}_{1:n})$. Similarly, the frames minimizing average-case coherence $a(\mathbf{F}_{1:n})$ can be interpreted as a packing which minimizes some averaged form of the principal angles between subspaces.

B. Initial masks and block coherence

With this in hand, we can now establish an interesting connection between maximum entropy masks and the block coherence of its corresponding frames, under the assumption of no prior knowledge on $\mathbf{X}$. We begin by providing a lower bound on the desired entropy criterion $\mathrm{H}(\mathbf{A}_{1:n})$:

Lemma 4 (Lower bound on entropy)**.** *Suppose $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ are fixed. The entropy $\mathrm{H}(\mathbf{A}_{1:n})$ can be lower bounded as:*

$$\mathrm{H}(\mathbf{A}_{1:n}) \geq \min_i \left\{ \mathrm{Var}(y_i) - \sigma^2 \left\langle \mathcal{P}_{\mathcal{U}} \mathbf{A}_i \mathcal{P}_{\mathcal{V}}, \sum_{j:j\neq i} \mathcal{P}_{\mathcal{U}} \mathbf{A}_j \mathcal{P}_{\mathcal{V}} \right\rangle_F \right\}^n, \tag{18}$$

where, following Lemma 2, $\mathrm{Var}(y_i) = \sigma^2 \|\mathcal{P}_{\mathcal{U}} \mathbf{A}_i \mathcal{P}_{\mathcal{V}}\|_F^2 + \eta^2$.

Using the lower bound in Lemma 4 as a proxy for the entropy $\mathrm{H}(\mathbf{A}_{1:n})$, the maximum entropy masks in (14) can be approximated as:

$$\underset{\substack{\mathbf{A}_1,\cdots,\mathbf{A}_n \\ \|\mathbf{A}_i\|_F^2 \leq 1}}{\mathrm{Argmax}} \min_i \left\{ \mathrm{Var}(y_i) - \sigma^2 \left\langle \mathcal{P}_{\mathcal{U}} \mathbf{A}_i \mathcal{P}_{\mathcal{V}}, \sum_{j:j\neq i} \mathcal{P}_{\mathcal{U}} \mathbf{A}_j \mathcal{P}_{\mathcal{V}} \right\rangle_F \right\}. \tag{19}$$

In the absense of prior knowledge on $\mathbf{X}$, we make two assumptions which reflect this lack of knowledge. First, following Section II-C, we assume **(A1)** independent and uniform prior distributions for $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ over the Grassmann manifolds $\mathcal{G}_{R,m_1-R}$ and $\mathcal{G}_{R,m_2-R}$; this reflects the belief that all subspaces in $\mathbf{X}$ are equally likely to occur. Second, letting $\mathbf{A}_i = \mathbf{R}_i \mathbf{\Lambda}_i \mathbf{S}_i^T$ be the singular-value decomposition (SVD) of mask $\mathbf{A}_i$, it makes sense that all subspaces should be weighted equally for the initial masks $\mathbf{A}_{1:n}$ given no prior information on $\mathbf{X}$. We therefore assume **(A2)** the weight matrices follow the scaled identity matrix $\mathbf{\Lambda}_i = R^{-1/2}\mathbf{I}$, where the scaling factor $R^{-1/2}$ ensures the power constraint is satisfied.

With this in hand, consider first the variance term $\mathrm{Var}(y_i) = \sigma^2 \|\mathcal{P}_{\mathcal{U}} \mathbf{A}_i \mathcal{P}_{\mathcal{V}}\|_F^2 + \eta^2$, which depends on the mask $\mathbf{A}_i$ as well as the matrices $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$. Under assumptions **(A1)** and **(A2)**, it is easy to see that the expected variance $\mathbb{E}_{\mathcal{P}_{\mathcal{U}},\mathcal{P}_{\mathcal{V}}} \mathrm{Var}(y_i)$ is constant for any choice of $\mathbf{A}_i$, $i = 1, \cdots, n$, since the uniform distributions on $\mathcal{G}_{R,m_1-R}$ and $\mathcal{G}_{R,m_2-R}$ are rotationally invariant. The optimization in (19) then reduces to only the inner-product term:

$$\underset{\substack{\mathbf{A}_1,\cdots,\mathbf{A}_n \\ \|\mathbf{A}_i\|_F^2 \leq 1}}{\mathrm{Argmin}} \max_i \left\langle \mathcal{P}_{\mathcal{U}} \mathbf{A}_i \mathcal{P}_{\mathcal{V}}, \sum_{j:j\neq i} \mathcal{P}_{\mathcal{U}} \mathbf{A}_j \mathcal{P}_{\mathcal{V}} \right\rangle_F. \tag{20}$$

Under the mask structure in **(A2)**, the lemma below provides an upper bound for the inner-product term in (20):

Lemma 5 (Upper bound on inner-product)**.** *Assume the initial masks* $\mathbf{A}_{1:n}$ *follow* **(A2)**. *Then, for fixed* i*:*

$$\left\langle \mathcal{P}_{\mathcal{U}} \mathbf{A}_i \mathcal{P}_{\mathcal{V}}, \sum_{j:j\neq i} \mathcal{P}_{\mathcal{U}} \mathbf{A}_j \mathcal{P}_{\mathcal{V}} \right\rangle_F \leq \frac{n-1}{2R} \left\{ \max_{j:j\neq i} \|(\mathcal{P}_{\mathcal{U}} \mathbf{R}_i)^T (\mathcal{P}_{\mathcal{U}} \mathbf{R}_j)\|_2^2 + \max_{j:j\neq i} \|(\mathcal{P}_{\mathcal{V}} \mathbf{S}_j)^T (\mathcal{P}_{\mathcal{V}} \mathbf{S}_i)\|_2^2 \right\}. \tag{21}$$

Using the above upper bound as a proxy for the inner-product criterion in (20), and assuming **(A1)** (uniform priors on $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$), the row frames $\mathbf{R}_{1:n}$ and $\mathbf{S}_{1:n}$ for initial masks $\mathbf{A}_{1:n}$ should therefore jointly minimize:

$$\max_{i\neq j} \|\mathbf{R}_i^T \mathbf{R}_j\|_2 \quad \text{and} \quad \max_{i\neq j} \|\mathbf{S}_i^T \mathbf{S}_j\|_2. \tag{22}$$

In other words, given no prior information on $\mathbf{X}$, the initial masks $\mathbf{A}_{1:n}$ which *maximize* the entropy criterion $\mathrm{H}(\mathbf{A}_{1:n})$ should have row and column frames with *low worst-case block coherences*.

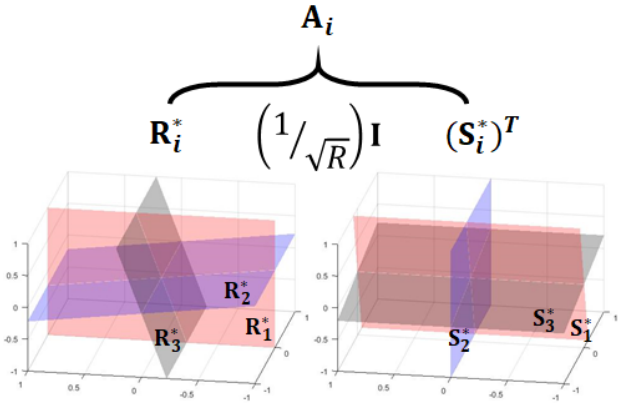

Figure 4. A visualization of the mask construction in (23), highlighting the link between maximum entropy masks and the Grassmann packing problem.

This suggests the following construction of $\mathbf{A}_{1:n}$:

$$\mathbf{A}_i = \mathbf{R}_i^* \mathbf{\Lambda}_i (\mathbf{S}_i^*)^T = \frac{1}{\sqrt{R}} \mathbf{R}_i^* (\mathbf{S}_i^*)^T, \tag{23}$$

where $\mathbf{R}_{1:n}^* = [\mathbf{R}_1^* \ \cdots \ \mathbf{R}_n^*]$ and $\mathbf{S}_{1:n}^* = [\mathbf{S}_1^* \ \cdots \ \mathbf{S}_n^*]$ are R-frames in $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$ which minimize their corresponding worst-case block coherences (or average block coherence, as a close approximation). Using the above reasoning, initial masks constructed this way can yield near-maximal entropy for the resulting observations; by the maximum entropy principle, this in turn maximizes the initial information gained on matrix $\mathbf{X}$.

The initial mask construction in (23) also admits an intuitive geometric interpretation in terms of Grassmann packings. Figure 4 visualizes this construction for $n = 3$ frames with $R = 2$ and $m_1 = m_2 = 3$. By restricting the row frames $\mathbf{R}_{1:n}^*$ and column frames $\mathbf{S}_{1:n}^*$ to have low block coherence, the corresponding row and column spaces for these initial masks are well spread-out in terms of their principal angles. These packed frames are then combined one-by-one via matrix multiplication to form the initial measurement masks. In the absense of prior information on $\mathbf{X}$, these initial masks $\mathbf{A}_{1:n}$, with *well-packed* row and column frames, can provide *near-maximal information* on $\mathbf{X}$. This link to the packing problem allows us to adapt state-of-the-art algorithms for optimal Grassmann packings to construct the optimal frames $\mathbf{R}_{1:n}^*$ and $\mathbf{S}_{1:n}^*$ in (23); this is detailed in Section VI.

V. Insights on Sequential Mask Design

Consider now the case where samples $\mathbf{y}$ have been observed using initial masks $\mathbf{A}_{1:n}$, and suppose informed estimates can be obtained on the projection matrices $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ via such samples (further details in Section VI). From (14), the sequential problem of finding the next mask which maximizes observational entropy can be formulated as the following

optimization problem:

$$\mathbf{A}^*_{n+1} := \underset{\mathbf{A}\in\mathbb{R}^{m_1\times m_2},\ \|\mathbf{A}\|_F^2\leq 1}{\text{Argmax}} \ \mathrm{H}(\mathbf{A}_{1:n},\mathbf{A}) = \underset{\mathbf{A}\in\mathbb{R}^{m_1\times m_2},\ \|\mathbf{A}\|_F^2\leq 1}{\text{Argmax}} \ \det\{\sigma^2\mathbf{R}_{n+1}([\mathbf{A}_{1:n}\ \ \mathbf{A}])+\eta^2\mathbf{I}\}, \tag{24}$$

where $\mathrm{H}(\mathbf{A}_{1:n},\mathbf{A})$ is shorthand for the joint entropy of observations from initial masks $\mathbf{A}_{1:n}$ and new mask $\mathbf{A}$, and $\mathbf{R}_{n+1}([\mathbf{A}_{1:n}\ \ \mathbf{A}])$ is the correlation matrix in (9) corresponding to $\mathbf{A}_{1:n}$ and $\mathbf{A}$. This corresponds to a greedy approach for optimizing measurement masks.

Using the Schur complement [43], (24) can be further simplified as follows:

Lemma 6 (Joint entropy simplification)**.** *The* joint entropy *in* (24) *can be written as:*

$$\mathrm{H}(\mathbf{A}_{1:n},\mathbf{A}) = \sigma^2\mathrm{H}(\mathbf{A}_{1:n})\left\{\|\mathcal{P}_{\mathcal{U}}\mathbf{A}\mathcal{P}_{\mathcal{V}}\|_F^2 + \gamma^2 - \mathbf{r}_n^T(\mathbf{A})(\mathbf{R}_n(\mathbf{A}_{1:n})+\gamma^2\mathbf{I})^{-1}\mathbf{r}_n(\mathbf{A})\right\}. \tag{25}$$

Using this lemma, the optimization in (24) then becomes:

$$\underset{\|\mathbf{A}\|_F^2\leq 1}{\text{Argmax}} \left\{\|\mathcal{P}_{\mathcal{U}}\mathbf{A}\mathcal{P}_{\mathcal{V}}\|_F^2 - \mathbf{r}_n^T(\mathbf{A})[\mathbf{R}_n(\mathbf{A}_{1:n})+\gamma^2\mathbf{I}]^{-1}\mathbf{r}_n(\mathbf{A})\right\}. \tag{26}$$

The sequential design criterion in (26) can be interpreted as subspace matching in the following sense. First, by maximizing the first term $\|\mathcal{P}_{\mathcal{U}}\mathbf{A}\mathcal{P}_{\mathcal{V}}\|_F^2$, we ensure that the projection of the new mask $\mathbf{A}$ has large norm (when projected onto the true subspaces of $\mathbf{X}$). This then encourages the *matching* of the subspaces for the new mask $\mathbf{A}$ to the true subspaces for $\mathbf{X}$ (see, e.g., [44], [45], [46]). Second, by minimizing the second term $\mathbf{r}_n^T(\mathbf{A})[\mathbf{R}_n(\mathbf{A}_{1:n})+\gamma^2\mathbf{I}]^{-1}\mathbf{r}_n(\mathbf{A})$, we force the new mask $\mathbf{A}$ to investigate *different* subspaces which are previously unexplored by the observed masks $\mathbf{A}_{1:n}$. The sequential criterion (26) can then be seen as balancing exploration versus exploitation.

Comparing (26) with the conditional expressions in (10) from Lemma 3, this sequential problem can equivalently be viewed as finding a new mask $\mathbf{A}$ which maximizes the *conditional variance* of an observation, given observed masks $\mathbf{A}_{1:n}$. From Section II-D1, the first term encourages masks yielding observations with large *variances*; from Section II-D2, the second term then aims to minimize the *correlation* between a new observation from $\mathbf{A}$ and the observed sample $\mathbf{y}$, after accounting for its similarity with $\mathbf{X}$. Again, the sequential criterion in (26) offers a balance between these two contrasting criteria.

The criterion in (26) also enjoys a nice visualization using the earlier illustration of maximum entropy (see Figure 2). Recall that maximum entropy masks can be viewed as vectors which maximize the volume of their covariance matrix after projection onto $\mathcal{T}_{\mathcal{U},\mathcal{V}}$. Suppose the black and blue vectors in Figure 2 were taken as initial measurement masks; one then aims to select the next mask which maximizes the volume of the covariance matrix ellipse. Figure 5 shows these two initial masks using black vectors, along with the projection

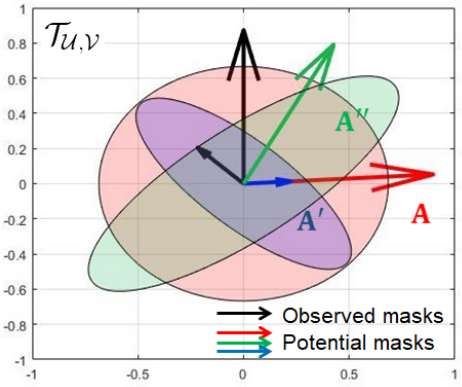

Figure 5. Visualizing three potential new masks $\mathbf{A}$, $\mathbf{A}'$ and $\mathbf{A}''$ (red, blue and green vectors, respectively) and their corresponding covariance matrices, given two observed masks (black vectors).

of three potential mask choices $\mathbf{A}$, $\mathbf{A}'$ and $\mathbf{A}''$ onto $\mathcal{T}_{\mathcal{U},\mathcal{V}}$ using red, blue and green vectors. Comparing the potential masks $\mathbf{A}$ (red) and $\mathbf{A}'$ (blue), we see that both $\mathbf{A}$ and $\mathbf{A}'$ have the same correlation with observed masks, but $\mathbf{A}$ has greater projected length. The resulting covariance matrix volume is therefore larger for $\mathbf{A}$ than for $\mathbf{A}'$, meaning $\mathbf{A}$ is a better sequential mask for recovering $\mathbf{X}$. Likewise, comparing $\mathbf{A}$ (red) and $\mathbf{A}''$ (green), we see that both masks have the same projected length, but $\mathbf{A}$ has smaller correlations with observed masks. The covariance matrix volume is therefore larger for $\mathbf{A}$ than for $\mathbf{A}''$, meaning $\mathbf{A}$ is again a better sequential mask for recovering $\mathbf{X}$. The mask $\mathbf{A}$ therefore satisfies the two-fold objective imposed by the two terms in (26). This intuition will again arise when we derive a closed-form solution for the sequential optimization in (26).

VI. MaxEnt – Constructing maximum entropy masks

We now employ the results from previous sections to derive an efficient algorithm, called `MaxEnt`, for constructing initial and sequential measurement masks.

A. Initial mask construction

Consider first the problem of constructing the initial masks $\mathbf{A}_{1:n} = [\mathbf{A}_1\ \mathbf{A}_2\ \cdots\ \mathbf{A}_n] \in \mathbb{R}^{m_1\times m_2}$ for preliminary learning on $\mathbf{X}$. Given the inherent link between maximum entropy masks and optimal packings of Grassmann manifolds (see Section IV), we extend here two Grassmann packing algorithms from [42] to construct initial masks for `MaxEnt`. Both constructions use the same form as in (23), but employ different approaches for packing the row and column frames $\mathbf{R}^*_{1:n}$ and $\mathbf{S}^*_{1:n}$. The first method, called the *flipping construction* (`ini.flip`), can be used for all choices of matrix dimensions $m_1\times m_2$ and any initial guess of the matrix rank R_{ini}. The second method, called the *Kerdock-Kronecker construction* (`ini.kk`), provides higher-quality masks than the flipping construction, but can only be used for specific matrix dimensions $m_1\times m_2$, initial matrix rank R_{ini} and initial sample size n.

Algorithm 1 $\texttt{flip}(\mathbf{R}_{1:n})$ – Flipping algorithm

- $\mathbf{R}_1^* \leftarrow \mathbf{R}_1$, $\mathbf{F}_1 \leftarrow \mathbf{R}_1$
- **For** $k = 1, \cdots, n-1$:
  - **if** $\|\mathbf{F}_k + \mathbf{R}_{k+1}\|_2 \leq \|\mathbf{F}_k - \mathbf{R}_{k+1}\|_2$ **then** $\mathbf{R}_{k+1}^* \leftarrow \mathbf{R}_{k+1}$ **else** $\mathbf{R}_{k+1}^* \leftarrow -\mathbf{R}_{k+1}$
  - $\mathbf{F}_{k+1} = \mathbf{F}_k + \mathbf{R}_{k+1}^*$
- **return** $\mathbf{R}_{1:n}^* = [\mathbf{R}_1^* \, \mathbf{R}_2^* \cdots \mathbf{R}_n^*]$

Algorithm 2 $\texttt{ini.flip}(m_1, m_2, n, R_{ini})$ – Flipping construction of initial masks

- Generate uniformly random frames $\mathbf{R}_{1:n}$ and $\mathbf{S}_{1:n}$
- $\mathbf{R}_{1:n}^* \leftarrow \texttt{flip}(\mathbf{R}_{1:n})$, $\mathbf{S}_{1:n}^* \leftarrow \texttt{flip}(\mathbf{S}_{1:n})$
- $\mathbf{A}_{1:n} \leftarrow [\mathbf{A}_1 \, \mathbf{A}_2 \cdots \mathbf{A}_n]$, where $\mathbf{A}_i \leftarrow R_{ini}^{-1/2} \mathbf{R}_i^* (\mathbf{S}_i^*)^T$
- **return** $\mathbf{A}_{1:n}$

1) Flipping construction: The flipping construction relies on the flipping algorithm `flip` (Algorithm 1 in [42]) to generate the row frames $\mathbf{R}_{1:n}^*$ and column frames $\mathbf{S}_{1:n}^*$ in (23). Given any set of frames $\mathbf{R}_{1:n}$, `flip` can be viewed as a post-processing step which reduces the average block coherence of $\mathbf{R}_{1:n}$, while retaining low worst-case block coherence. In words, `flip` iteratively performs random flips of each frame in $\mathbf{R}_{1:n}$, and accept such flips only if it results in a reduction in average coherence. Using the row and column frames $\mathbf{R}_{1:n}^*$ and $\mathbf{S}_{1:n}^*$ returned from `flip`, the proposed flipping construction `ini.flip` then incorporates these frames within (23) to form the initial masks $\mathbf{A}_{1:n}$. Algorithms 1 and 2 outline the details for the flipping algorithm `flip` and the flipping mask construction `ini.flip`.

The following lemma from [42] provides an upper bound guarantee for the average block coherence of frames returned by `flip`:

Lemma 7 (Avg. block coherence of `flip`; Lemma 3.4 in [42])**.** *The row and column frames $\mathbf{R}_{1:n}^*$ and $\mathbf{S}_{1:n}^*$ returned by* `flip` *satisfy $a(\mathbf{R}_{1:n}^*) = a(\mathbf{S}_{1:n}^*) = (\sqrt{n}+1)/(n-1)$.*

This average block coherence guarantee for `flip` can be shown to improve upon that for uniformly random frames (see Section 4 of [42]). Given the link between information and block coherence in Section IV-B (and using average coherence as a proxy for worst-case coherence), such a lemma shows that the initial mask construction using `ini.flip` yields more information on $\mathbf{X}$, compared to uniformly sampled masks.

One advantage of `ini.flip` over the Kerdock-Kronecker construction `ini.kk` (introduced in the next section) is that it can be used for any choice of matrix dimension or rank. However, when m_1, m_2 and R_{ini} satisfy certain conditions, the latter can provide better recovery performance.

2) Kerdock-Kronecker (KK) construction: The Kerdock-Kronecker (KK) construction, `ini.kk`, uses the form in (23), with row and column frames $\mathbf{R}_{1:n}^*$ and $\mathbf{S}_{1:n}^*$ following the Kronecker form:

$$\mathbf{R}_{1:n}^* = \mathbf{K}_{\mathbf{R}} \otimes \mathbf{Q}_{\mathbf{R}}, \quad \mathbf{S}_{1:n}^* = \mathbf{K}_{\mathbf{S}} \otimes \mathbf{Q}_{\mathbf{S}}. \tag{27}$$

Algorithm 3 $\texttt{ini.kk}(m_1, m_2, n, R_{ini})$ – Kerdock-Kronecker construction of initial masks

- Generate Kerdock frames for $\mathbf{K}_{\mathbf{R}} \in \mathbb{R}^{m_1/R_{ini} \times n}$ and $\mathbf{K}_{\mathbf{S}} \in \mathbb{R}^{m_2/R_{ini} \times n}$
- Generate uniformly random unitary matrices $\mathbf{Q}_{\mathbf{R}} \in \mathbb{R}^{R_{ini} \times R_{ini}}$ and $\mathbf{Q}_{\mathbf{S}} \in \mathbb{R}^{R_{ini} \times R_{ini}}$
- $\mathbf{R}_{1:n}^* \leftarrow \mathbf{K}_{\mathbf{R}} \otimes \mathbf{Q}_{\mathbf{R}}$, $\mathbf{S}_{1:n}^* \leftarrow \mathbf{K}_{\mathbf{S}} \otimes \mathbf{Q}_{\mathbf{S}}$
- $\mathbf{R}_{1:n}^* \leftarrow \texttt{flip}(\mathbf{R}_{1:n}^*)$, $\mathbf{S}_{1:n}^* \leftarrow \texttt{flip}(\mathbf{S}_{1:n}^*)$
- $\mathbf{A}_{1:n} \leftarrow [\mathbf{A}_1 \, \mathbf{A}_2 \cdots \mathbf{A}_n]$, where $\mathbf{A}_i \leftarrow R_{ini}^{-1/2} \mathbf{R}_i^* (\mathbf{S}_i^*)^T$
- **return** $\mathbf{A}_{1:n}$

Here, $\mathbf{K}_{\mathbf{R}} \in \mathbb{R}^{(m_1/R_{ini}) \times n}$ and $\mathbf{K}_{\mathbf{S}} \in \mathbb{R}^{(m_2/R_{ini}) \times n}$ are taken from the Kerdock family of frames [47], and $\mathbf{Q}_{\mathbf{R}}$ and $\mathbf{Q}_{\mathbf{S}} \in \mathbb{R}^{R_{ini} \times R_{ini}}$ are independent and uniformly distributed unitary matrices. One can prove that such a construction returns frames with low block coherence, both in the average-case and the worst-case:

Lemma 8 (Worst-case and avg. block coherence of KK; Thm. 2.10 and Table 1, [42])**.** *The frames $\mathbf{R}_{1:n}^*$ and $\mathbf{S}_{1:n}^*$ in (27) satisfy (a) $\mu(\mathbf{R}_{1:n}^*) = 1/\sqrt{m_1}$ and $\mu(\mathbf{S}_{1:n}^*) = 1/\sqrt{m_2}$, and (b) $a(\mathbf{R}_{1:n}^*) = a(\mathbf{S}_{1:n}^*) = 1/(n-1)$.*

In particular, the worst-case block coherences in Lemma 8 nearly achieve the universal lower bound for block coherences provided in Theorem 3.6 of [48] (it achieves this bound asymptotically, as $n \to \infty$). From the discussion in Section IV, this suggests the initial masks constructed using `ini.kk` are near-optimal for extracting information from $\mathbf{X}$. In our implementation of `ini.kk`, we use the flipping algorithm `flip` as a post-processing step to further improve average block coherence. Algorithm 3 summarizes the steps for `ini.kk`.

One restriction of `ini.kk` is that the Kerdock frames $\mathbf{K}_{\mathbf{R}}$ and $\mathbf{K}_{\mathbf{S}}$ can be only generated for certain choices of m_1, m_2, R_{ini} and n, specifically, when the matrix dimensions satisfy $m_1 = (2^{k_1+1})R_{ini}$ and $m_2 = (2^{k_2+1})R_{ini}$ for some odd positive integers k_1 and k_2, with initial sample size $n \leq \min(2^{2k_1+2}, 2^{2k_2+2})$ [49]. These conditions are summarized in Table I. Despite such limitations, `ini.kk` enjoys two important advantages. First, the initial masks from `ini.kk` provide improved block coherence (and thereby better information-theoretic properties) than that for `ini.flip`: (i) the worst-case block coherences for its subspaces nearly achieve the universal coherence lower bound in [48], and (ii) its average block coherences are much lower than that for `ini.flip` (see Lemmas 7 and 8). `ini.kk` should therefore be used whenever the conditions in Table I are satisfied (see, e.g., the image recovery applications in Section VII-B). Second, Kerdock frames enjoy a nice packing property, which allows more blocks to be packed into a frame while retaining low block coherence [42]. This packing property provides users with the option of increasing initial sample size n by setting the initial rank guess R_{ini} to be as small as possible, all the while ensuring the resulting masks enjoy good recovery performance for matrices with higher rank.

Table I
RESTRICTIONS ON MATRIX DIMENSIONS $m_1 \times m_2$ AND INITIAL SAMPLE SIZE n FOR THE FLIPPING CONSTRUCTION `INI.FLIP` AND THE KERDOCK-KRONECKER CONSTRUCTION `INI.KK`.

	`ini.flip`	`ini.kk`
m_1	Any	$(2^{k_1+1})R_{ini}$, for some odd integer k_1
m_2	Any	$(2^{k_2+1})R_{ini}$, for some odd integer k_2
n	Any	$n \leq \min(2^{2k_1+2}, 2^{2k_2+2})$

B. Sequential mask construction

Given initial measurements from masks $\mathbf{A}_{1:n}$ and informed estimates of $\mathcal{U}$, $\mathcal{V}$ and R from such samples, consider next the construction of sequential masks. The following theorem gives a closed-form expression for the sequential optimization in (26):

Theorem 9 (Sequential mask construction)**.** *For fixed row and column spaces $\mathcal{U}$ and $\mathcal{V}$, the optimal* sequential mask $\mathbf{A}^*_{n+1}$ *in* (26) *takes the form:*

$$\mathbf{A}^*_{n+1} = \mathbf{U}\mathbf{\Sigma}^*\mathbf{V}^T, \tag{28}$$

for any choice of $\mathbf{U} \in \mathcal{U}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V} \in \mathcal{V}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Here, $\mathrm{vec}(\mathbf{\Sigma}^*) \in \mathbb{R}^{R^2}$ is the unit eigenvector for the smallest eigenvalue of $\mathbf{D}^T[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{D}$, $\mathbf{D} = [\mathrm{vec}(\mathbf{\Sigma}_1)^T; \cdots; \mathrm{vec}(\mathbf{\Sigma}_n)^T] \in \mathbb{R}^{n \times R^2}$, with $\mathbf{\Sigma}_i = \mathbf{U}^T\mathbf{A}_i\mathbf{V}$.

This theorem can be explained as follows. Having sampled from masks $\mathbf{A}_{1:n}$, the optimal sequential mask is constructed using the row and column frames $\mathbf{U}$ and $\mathbf{V}$ from the underlying subspaces $\mathcal{U}$ and $\mathcal{V}$, along with a weight matrix $\mathbf{\Sigma}^*$. The weight matrix $\mathbf{\Sigma}^*$, which is obtained via a minimum eigenvector computation, can be seen as an *optimal allocation* of measurement power to the row and column frames $\mathbf{U}$ and $\mathbf{V}$, in such a way that the correlations between the new mask $\mathbf{A}^*_{n+1}$ and observed masks $\mathbf{A}_{1:n}$ are minimized (after accounting for the subspaces of $\mathbf{X}$). Equivalently, the power allocation in $\mathbf{\Sigma}^*$ can be interpreted as distributing measurement power to important subspaces of $\mathbf{X}$ which have yet to be explored by previous masks $\mathbf{A}_{1:n}$.

Computationally, the key appeal of Theorem 9 is that it provides a *closed-form* method for greedily maximizing information gain on $\mathbf{X}$. Compared to a numerical optimization of (26), which quickly becomes computationally infeasible for moderate choices of m_1 or m_2, the closed-form solution in (28) offers a much more efficient construction of sequential masks. It should be emphasized that this closed-form sequential mask construction is a direct result of the maximum entropy design principle, which allows us to optimize the simpler observation entropy term $\mathrm{H}(\mathbf{A}_{1:n})$ as a proxy for the more complicated entropy term $\mathrm{H}_{\mathbf{A}_{1:n}}(\mathbf{X}|\mathbf{y})$ or the error term in (15). The latter expressions, which involve the high-dimensional matrix $\mathbf{X}$, are too computationally cumbersome to manipulate, and would likely not yield the nice closed-form construction in (28).

The closed-form solution in (28) also yields an insightful geometric interpretation. For fixed row and column frames $\mathbf{U}$ and $\mathbf{V}$, consider the representation of the observed masks $\{\mathbf{A}_i\}_{i=1}^n$ by its power allocation matrices $\{\mathbf{\Sigma}_i\}_{i=1}^n$, where $\mathbf{\Sigma}_i = \mathbf{U}^T\mathbf{A}_i\mathbf{V}$ as in Theorem 9. Note that $\langle \mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}}\mathbf{A}_j\mathcal{P}_{\mathcal{V}} \rangle_F = \langle \mathbf{\Sigma}_i, \mathbf{\Sigma}_j \rangle_F$, so, similar to the interpretation in Section III-B, the goal of maximum entropy masks can be viewed as maximizing the covariance matrix *volume* for the vectors corresponding to these observed power matrices $\{\mathbf{\Sigma}_i\}_{i=1}^n$. Figure 6 visualizes these vectors and their covariance ellipse for $n = 4$ observed masks. The minimum eigenvector solution for $\mathbf{\Sigma}^*$ in (28) can be viewed as the *minor axis* – the axis with shortest length – of the covariance ellipse. Such a choice is quite intuitive, because adding the vector $\mathbf{\Sigma}^*$ (red arrow in Figure 6) along the minor axis maximizes the *volume gain* of the ellipse. The sequential mask yielding the greatest *information gain* can then be constructed by using the weight matrix $\mathbf{\Sigma}^*$ to optimally allocate *measurement power* to the row and column frames $\mathbf{U}$ and $\mathbf{V}$ (bottom of Figure 6). This interpretation is related to the "water-filling" allocation in information-theoretic frame design for compressive sensing (see [24], [38]), except instead of allocating more measurement power to *less noisy* channels, the optimal mask in (28) allocates more power to frames *more correlated* with the desired matrix $\mathbf{X}$ and *less correlated* with previously observed masks.

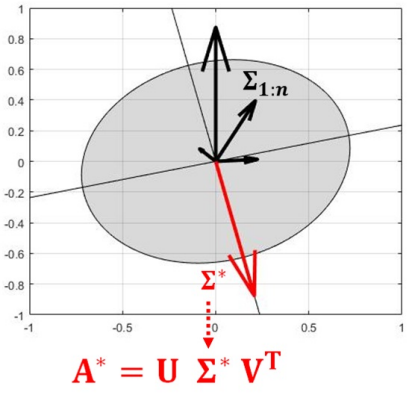

Figure 6. A visualization of the sequential mask in (28) as an optimal power allocation scheme from a minimum eigenvector problem.

Algorithm 4 `MaxEnt`$(m_1, m_2, n_{ini}, R_{ini}, n_{seq})$ – Information-theoretic matrix recovery

- **If** conditions in Table I satisfied
  then $\mathbf{A}_{1:n_{ini}} \leftarrow$ `ini.kk`$(m_1, m_2, n_{ini}, R_{ini})$
  else $\mathbf{A}_{1:n_{ini}} \leftarrow$ `ini.flip`$(m_1, m_2, n_{ini}, R_{ini})$
- Observe initial samples $y_i \leftarrow \langle \mathbf{A}_i, \mathbf{X} \rangle_F + \epsilon_i, i = 1, \cdots, n_{ini}$
- **For** $i = n_{ini} + 1, \cdots, n_{ini} + n_{seq}$:
  - Estimate $\hat{\mathbf{X}}$ via the nuclear-norm formulation (5)
  - Estimate row and column spaces $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ from `svd`$(\hat{\mathbf{X}})$
  - Construct next mask $\mathbf{A}_i$ from (28) using $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$
  - Observe new sample $y_i \leftarrow \langle \mathbf{A}_i, \mathbf{X} \rangle_F + \epsilon_i$
- **Return** recovered matrix $\hat{\mathbf{X}}$

C. Full algorithm

Algorithm 4 summarizes the full `MaxEnt` algorithm for information-theoretic matrix recovery. First, initial masks

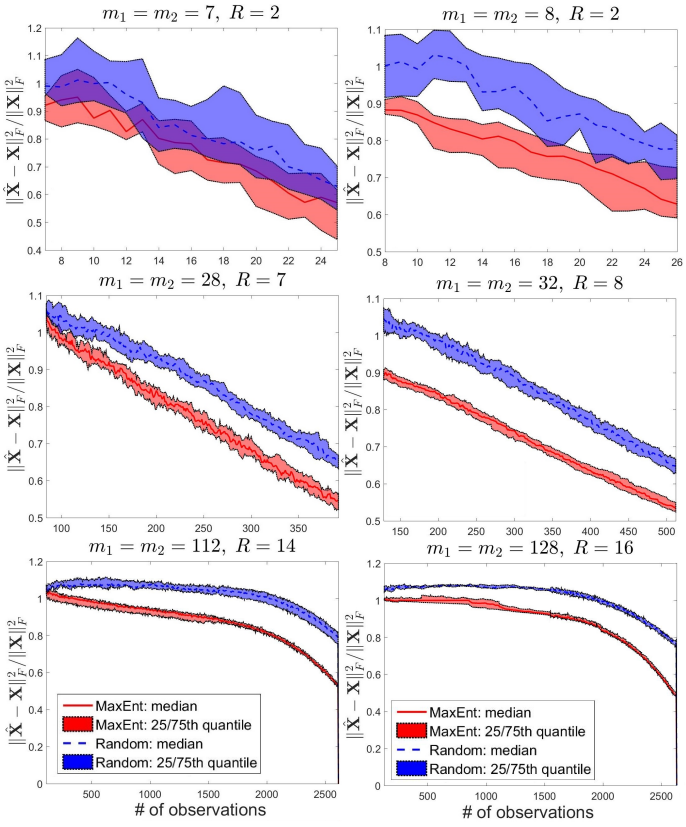

Figure 7. Normalized recovery errors for $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$ with rank R, using `MaxEnt` (initialized with `ini.flip` on left, with `ini.kk` on right) and random masks. Solid lines mark median error, and shaded bands mark 25-th/75-th error quantiles.

are generated via either `ini.kk` (whenever possible) or `ini.flip`. Next, using the observations from these initial masks, $\mathbf{X}$ is approximated using the nuclear-norm estimator $\hat{\mathbf{X}}$ in (5). In our implementation, $\hat{\mathbf{X}}$ is optimized via the Matlab solver CVX [50] for small matrices, and a straight-forward extension of the singular-value thresholding algorithm [51] for larger matrices. $\hat{\mathbf{X}}$ can then be used to provide estimates on the row and column spaces $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ via a SVD step. The next mask to observe is then constructed from the closed-form expression in (28), using subspace estimates $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$. This sequential procedure is repeated until a desired sample size is obtained, or a desired error tolerance achieved. For larger problems, the sequential procedure in `MaxEnt` can be sped up via batch sampling, by supplementing the optimal mask in (28) with (i) masks constructed using the smallest few eigenvectors for the power allocation in (28), or (ii) randomly sampled masks.

VII. Numerical Examples

A. Simulated examples

We now investigate the numerical recovery performance of `MaxEnt` for simulated instances of $\mathbf{X}$. Here, $\mathbf{X}$ is simulated using the singular matrix-variate Gaussian distribution $\mathcal{SMG}(\mathcal{P}_{\mathcal{U}}, \mathcal{P}_{\mathcal{V}}, \sigma^2, R)$, where $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{V}}$ are uniformly simulated, with variance parameters $\sigma^2 = 1$ and $\eta^2 = 10^{-4}$. Six simulation cases are conducted for different choices of matrix dimension and rank (m_1, m_2, R): the first three cases $(7, 7, 2)$, $(28, 28, 7)$ and $(112, 112, 14)$ investigate the performance of `MaxEnt` using the flipping construction `ini.flip`, while the next three cases $(2^3, 2^3, 2)$, $(2^5, 2^5, 8)$ and $(2^7, 2^7, 16)$ investigate the Kerdock-Kronecker construction `ini.kk` (where R_{ini} is set as 2 to exploit the packing property of Kerdock frames; see Section VI-A2). As for sample size, the first three cases employ an initial and total sample size $(n_{ini}, n_{ini} + n_{seq})$ of $(7, 25)$, $(84, 400)$ and $(112, 2600)$, while the sample sizes for next three cases are set as $(8, 26)$, $(128, 520)$ and $(128, 2600)$. For each case, we replicate the simulation procedure for ten trials to provide an estimate of error variability.

For each of the six cases, Figure 7 shows the normalized recovery errors $\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2/\|\mathbf{X}\|_F^2$ as a function of samples taken, where as before, $\hat{\mathbf{X}}$ is the nuclear-norm estimate in (5). To benchmark performance, the proposed method `MaxEnt` is compared with uniformly random measurement masks satisfying the unit power constraints. Consider first the *initial* recovery performance of `MaxEnt` (using `ini.flip` or `ini.kk`) and random masks. For the three cases on the left, the initial masks from `ini.flip` give noticeable improvements over random masks, both in terms of median error and error quantiles. For the three cases on the right, the initial masks from `ini.kk` yield more pronounced improvements over random masks; the 75-th error quantiles for the former are sizably smaller than the 25-th error quantiles for the latter. These results corroborate two earlier insights. First, the improved performance of `ini.flip` and `ini.kk` over random masks supports the link between maximum entropy masks and low block coherence of subspaces, as noted in Section IV-B. By ensuring a tight packing of row and column frames for initial masks, `MaxEnt` can yield greater information gain on $\mathbf{X}$ over random masks, as seen in Figure 7. Second, this confirms the improved performance of `ini.kk` to `ini.flip`, which is expected because the former has better packing properties than the latter (see Section VI-A).

Consider next the *sequential* performance of `MaxEnt`. From Figure 7, we see that the sequential recovery from `MaxEnt` is markedly better than that from random masks, at nearly all sample sizes. In particular, the error gap between `MaxEnt` and random masks appears to grow larger as more sequential observations are taken. This suggests the closed-form sequential scheme in Section VI-B is indeed effective; by designing masks which greedily maximize information on $\mathbf{X}$, `MaxEnt` provides an informed adaptive scheme which targets *important* row and column space features of $\mathbf{X}$. This growing error gap also hints at an improved theoretical error rate for `MaxEnt` over random masks; we look forward to exploring this further in a future work.

B. Real data examples

1) Image recovery: Next, we investigate the performance of `MaxEnt` for recovering (i) a 64×64-pixel peppers image[2] (hereby called 'peppers'), and (ii) an 80×80-pixel solar flare

[2] www.statemaster.com/encyclopedia/Standard-test-image.

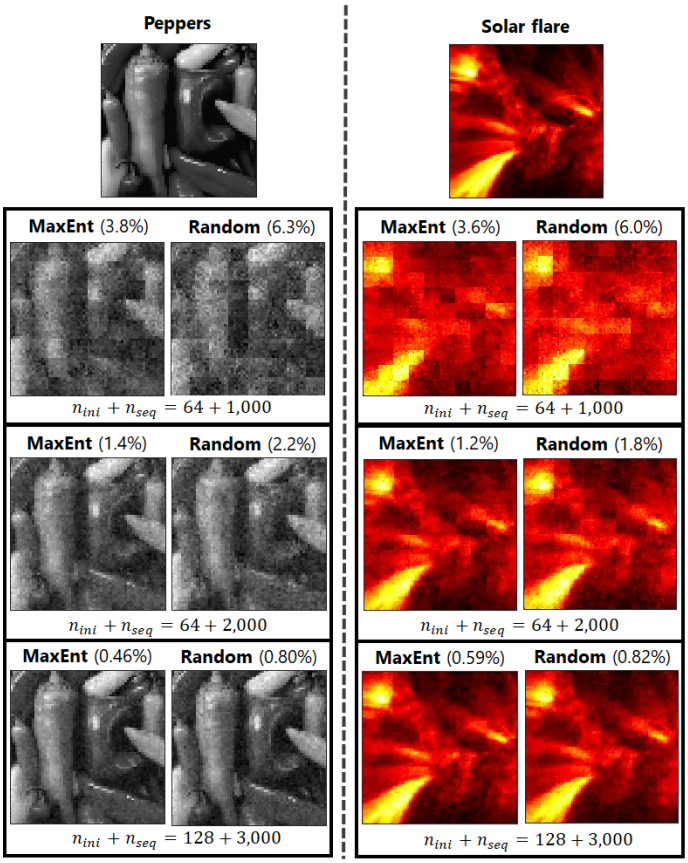

Figure 8. (Top) The original 'peppers' and 'flare' images, and (bottom) the recovered images using `MaxEnt` and random masks with $n_{ini} + n_{seq}$ measurements. Normalized errors are bracketed.

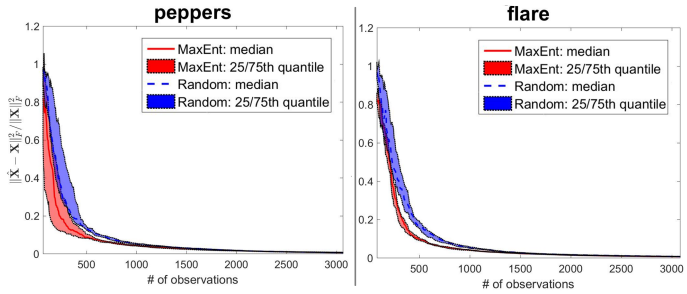

Figure 9. Normalized recovery errors using `MaxEnt` and random masks. Solid lines mark median errors, and shaded bands mark 25-th/75-th error quantiles.

image captured by the NASA SDO satellite (hereby called 'flare'; see [52] for details). These images are shown on the top of Figure 8. To generate $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$, we first break each image into 8×8 patches, then vectorize these patches and collect the resulting vectors into an $m_1 \times m_2 = 64 \times 64$ matrix for 'peppers', and an $m_1 \times m_2 = 64 \times 100$ matrix for 'flare' (details on this patching step can be found in [53]). Using this patched matrix $\mathbf{X}$ (which takes values from $\{0, \cdots, 255\}$), observations are then sampled with noise variance $\eta^2 = 1.0$. For 'peppers', $n_{ini} = 64$ initial measurement masks are constructed using `ini.kk` with $R_{ini} = 4$, with the remaining $n_{seq} = 3,000 - 64$ samples taken sequentially; for 'flare', $n_{ini} = 80$ initial measurement masks are constructed using `ini.flip` with $R_{ini} = 8$, with the remaining $n_{seq} = 3,000 - 80$ samples taken sequentially. This sampling procedure is replicated five times to give an estimate of error variability.

For these images, Figure 9 shows the normalized recovery errors $\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 / \|\mathbf{X}\|_F^2$ as a function of measurements taken. As before, the proposed method `MaxEnt` is compared with uniformly random masks. For *initial* recovery, `MaxEnt` yields markedly lower errors to random sampling for both images, which again shows the effectiveness of well-packed subspaces for maximizing initial information gain. As before, `ini.kk` provides a greater initial error reduction to `ini.flip`, which is expected. For *sequential* recovery, `MaxEnt` maintains a sizable improvement gap over random sampling for both images, particularly in the early stages of the sequential procedure. This illustrates the ability of `MaxEnt` to learn and target important image features via adaptive mask design. Such results are in line with the observations in Section VII-A.

For a visual comparison, Figure 8 shows the original images for 'peppers' and 'flare', and the recovered images using `MaxEnt` and random masks. For 'peppers', `MaxEnt` provides noticeably improved recovery of key image characteristics, such as the distinct shape and lighting of each individual pepper, whereas the same characteristics appear more blurred for random masks. Likewise, for 'flare', `MaxEnt` yields a clearer recovery of key solar flare features, such as the intensity and spray of each flare eruption, whereas the same features appear more blurred for random masks. This nicely visualizes the ability of `MaxEnt` to actively learn important image features, using the adaptive power allocation in the sequential construction (28).

2) Text document indexing: We now explore the usefulness of `MaxEnt` for compressing (or indexing) large text databases into smaller, representative datasets. The database to compress, $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$, is compiled using police reports provided by the Atlanta Police Department on crimes committed between the years 2014 – 2017. Each row of $\mathbf{X}$ corresponds to a separate police report (with $m_1 = 497$ report narratives in total), and each column of $\mathbf{X}$ records the frequency of a specific term in a bag-of-words representation [54] of these reports (with $m_2 = 100$ terms of interest); see [55] for details. For `MaxEnt`, we first construct an initial design of $n_{ini} = 1,000$ masks using `ini.flip` with $R_{ini} = 8$, then run the sequential part of the algorithm for $n_{seq} = 11,000$ samples, resulting in a compressed dataset with $n_{ini} + n_{seq} = 12,000$ samples. As before, the compression from `MaxEnt` is compared with the compression from uniformly-random mask projections, with both methods compared on how well the reduced dataset recovers the original police report database.

Table II summarizes the normalized recovery errors $\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 / \|\mathbf{X}\|_F^2$ for `MaxEnt` and random masks, where $\hat{\mathbf{X}}$ is the nuclear-norm-recovered database from the sketched dataset. Consider first the compression performance using $n_{ini} = 1,000$ initial samples. We see that the initial masks from `ini.flip` yield lower recovery errors compared to random

Table II
NORMALIZED RECOVERY ERRORS $\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 / \|\mathbf{X}\|_F^2$ FOR THE COMPRESSED TEXT DATABASE USING MAXENT AND RANDOM MASKS.

Compressed data size	$n_{ini} = 1,000$	$n_{ini} + n_{seq} = 12,000$
MaxEnt	88.4%	2.8%
Random	91.6%	4.2%

masks, which demonstrates the importance of well-packed row and column frames for initial data compression. Consider next the compression performance using $n_{ini}+n_{seq} = 12,000$ total measurements. We see that MaxEnt maintains a considerable reduction in compression error over random masks, which illustrates the proposed method's ability to learn and target important subspace properties of $\mathbf{X}$ via an adaptive mask design scheme. This is particularly insightful in light of the fact that, for latent semantic analysis (see, e.g., [54]), the SVD of the document-term matrix $\mathbf{X}$ encodes important information on (i) different document types found in the database, and (ii) typical terms found within each document type. Viewed this way, the sequential procedure in MaxEnt allows the algorithm to first *learn* this latent document-term structure, then *exploit* such structure to design effective masks which compress the large database at hand.

VIII. CONCLUSION

In this paper, we proposed a novel information-theoretic approach called MaxEnt, for designing measurement masks for low-rank matrix recovery. This method is motivated by the fact that, for many real-world applications, the desired matrix $\mathbf{X}$ to recover can be very high-dimensional, and the cost of obtaining measurements from $\mathbf{X}$ can be expensive. For such applications, the proposed method MaxEnt can offer excellent recovery performance of $\mathbf{X}$ using a limited number of observations from designed masks. MaxEnt relies on two important ingredients: (i) the singular matrix-variate Gaussian stochastic model on the low-rank matrix $\mathbf{X}$, and (ii) the maximum entropy principle, which states that a mask design scheme which *maximizes* the entropy of collected samples also *maximizes* information gain on $\mathbf{X}$. With these ingredients in hand, we showed several novel and interesting insights on the link between mask design, compressive sensing and coding theory. Using such insights, we then developed an algorithm (MaxEnt) for efficiently constructing initial and adaptive measurement masks which maximize information gain on $\mathbf{X}$. Lastly, we demonstrated the recovery effectiveness of MaxEnt over randomly-sampled masks in several numerical experiments, and for two real-world problems in image processing and text document indexing.

Looking forward, there are several interesting research directions to pursue next. First, while the numerical results in Section VII show a considerable advantage of MaxEnt compared to randomly-sampled masks, it would be nice to quantify this improvement via a theoretical convergence rate. Such a task does not appear to be straight-forward, and recent results on information-theoretic regret bounds in online optimization [56] can prove to be useful. Second, given the promising applications to image processing and text document indexing in this work, we are very much interested in exploring the usefulness of MaxEnt for tackling other important, low-rank modeling problems in engineering and statistics, perhaps involving additional structural constraints on measurement masks. Such applications include covariance sketching [42], system identification [10], physics-extraction from engineering flows [57], and gene association studies [2], [58]. Lastly, the conditional expressions in Lemma 3 can have further applications for the uncertainty quantification of $\mathbf{X}$, and we look forward to exploring this further in a future work.

ACKNOWLEDGMENTS

The authors would like to thank Shixiang Zhu for his help in cleaning and compiling the Atlanta Police Department dataset for the text document indexing application in Section VII-B2.

REFERENCES

[1] J. Bennett and S. Lanning, "The Netflix prize," in *Proceedings of KDD Cup and Workshop*, 2007, www.netflixprize.com.
[2] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene–disease associations," *Bioinformatics*, vol. 30, no. 12, pp. 60–68, 2014.
[3] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.
[4] M. F. Duarte, M. A. Davenport, D. Takbar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, 2008.
[5] A. Rajwade, D. Kittle, T.-H. Tsai, D. Brady, and L. Carin, "Coded hyperspectral imaging and blind compressive sensing," *SIAM Journal on Imaging Sciences*, vol. 6, no. 2, pp. 782–812, 2013.
[6] D. J. Brady, *Optical Imaging and Spectroscopy*. John Wiley & Sons, 2009.
[7] W. D. Pesnell, *Solar Dynamics Observatory (SDO)*. Springer, 2015.
[8] D. Achlioptas, "Database-friendly random projections," in *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM, 2001, pp. 274–281.
[9] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2001, pp. 245–250.
[10] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.
[11] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.
[12] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.
[13] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi, "A simplified approach to recovery conditions for low rank matrices," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*. IEEE, 2011, pp. 2318–2322.
[14] T. T. Cai and A. Zhang, "Compressed sensing and affine rank minimization under restricted isometry," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3279–3290, 2013.
[15] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
[16] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.
[17] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[18] B. Recht, "A simpler approach to matrix completion," *Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[19] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*. Springer Science & Business Media, 2013.

[20] S. Chakraborty, J. Zhou, V. Balasubramanian, S. Panchanathan, I. Davidson, and J. Ye, "Active matrix completion," in *IEEE 13th International Conference on Data Mining*, 2013, pp. 81–90.

[21] S. Mak and Y. Xie, "Uncertainty quantification and design for noisy matrix completion-a unified framework," *arXiv preprint arXiv:1706.08037*, 2017.

[22] D. J. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.

[23] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 141–154, 2006.

[24] W. R. Carson, M. Chen, M. R. Rodrigues, R. Calderbank, and L. Carin, "Communications-inspired projection design with application to compressive sensing," *SIAM Journal on Imaging Sciences*, vol. 5, no. 4, pp. 1185–1212, 2012.

[25] L. Wang, A. Razi, M. Rodrigues, R. Calderbank, and L. Carin, "Nonlinear information-theoretic compressive measurement design," in *International Conference on Machine Learning*, 2014, pp. 1161–1169.

[26] N. Shlezinger, R. Dabora, and Y. C. Eldar, "Measurement matrix design for phase retrieval based on mutual information," *arXiv preprint arXiv:1704.08021*, 2017.

[27] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 4240–4248.

[28] C. F. J. Wu and M. S. Hamada, *Experiments: Planning, Analysis, and Optimization*. John Wiley & Sons, 2009.

[29] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. CRC Press, 1999.

[30] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC Press, 2014, vol. 2.

[31] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[32] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2011, pp. 185–212.

[33] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[34] M. C. Shewry and H. P. Wynn, "Maximum entropy sampling," *Journal of Applied Statistics*, vol. 14, no. 2, pp. 165–170, 1987.

[35] P. Sebastiani and H. P. Wynn, "Maximum entropy sampling and optimal Bayesian experimental design," *Journal of the Royal Statistical Society, Series B*, vol. 62, no. 1, pp. 145–157, 2000.

[36] D. Blackwell, "Comparison of experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

[37] D. V. Lindley, "On a measure of the information provided by an experiment," *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 986–1005, 1956.

[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[39] S. Prasad, "Certain relations between mutual information and fidelity of statistical estimation," *arXiv preprint arXiv:1010.1508*, 2010.

[40] W. U. Bajwa and D. G. Mixon, "Group model selection using marginal correlations: The good, the bad and the ugly," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 494–501.

[41] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[42] R. Calderbank, A. Thompson, and Y. Xie, "On block coherence of frames," *Applied and Computational Harmonic Analysis*, vol. 38, no. 1, pp. 50–71, 2015.

[43] K. Hoffman and R. Kunze, *Linear Algebra*. Englewood Cliffs, New Jersey, 1971.

[44] L. L. Scharf, *Statistical Signal Processing*. Addison-Wesley, 1991.

[45] X. G. Doukopoulos and G. V. Moustakides, "Fast and stable subspace tracking," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1452–1465, 2008.

[46] W. U. Bajwa, R. Calderbank, and S. Jafarpour, "Revisiting model selection and recovery of sparse signals using one-step thresholding," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 977–984.

[47] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. Elsevier, 1977.

[48] P. W. Lemmens and J. J. Seidel, "Equi-isoclinic subspaces of Euclidean spaces," in *Indagationes Mathematicae (Proceedings)*, vol. 76, no. 2. Elsevier, 1973, pp. 98–107.

[49] A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. Sloane, and P. Solé, "The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes," *IEEE Transactions on Information Theory*, vol. 40, no. 2, pp. 301–319, 1994.

[50] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2008.

[51] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[52] Y. Xie, J. Huang, and R. Willett, "Change-point detection for high-dimensional time series with missing data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 12–27, 2013.

[53] Y. Cao and Y. Xie, "Poisson matrix recovery and completion," *IEEE Transactions on Signal Processing*, vol. 64, no. 6, pp. 1609–1620, 2016.

[54] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[55] S. Zhu and Y. Xie, "Crime incidents embedding using restricted boltzmann machines," *arXiv preprint arXiv:1710.10513*, 2017.

[56] D. Russo and B. Van Roy, "An information-theoretic analysis of Thompson sampling," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2442–2471, 2016.

[57] S. Mak, C. L. Sung, X. Wang, S. T. Yeh, Y. H. Chang, V. R. Joseph, V. Yang, and C. F. J. Wu, "An efficient surrogate model for emulation and physics extraction of large eddy simulations," *Journal of the American Statistical Association*, 2017, to appear.

[58] S. Mak and C. F. J. Wu, "cmenet: a new method for bi-level variable selection of conditional main effects," *arXiv preprint arXiv:1701.05547*, 2017.

[59] Y. L. Tong, *The Multivariate Normal Distribution*. Springer Science & Business Media, 2012.

[60] S. A. Gershgorin, "Uber die abgrenzung der eigenwerte einer matrix," *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk*, no. 6, pp. 749–754, 1931.

APPENDIX A
PROOFS

Proof of Lemma 1. The proof follows from a direct application of Lemma 2 in [21]. □

Proof of Lemma 2. Let $\mathbf{x}_1, \cdots, \mathbf{x}_{m_2} \in \mathbb{R}^{m_1}$ denote the m_2 columns in $\mathbf{X}$, and let $\mathbf{a}_{i,1}, \cdots, \mathbf{a}_{i,m_2} \in \mathbb{R}^{m_1}$ denote the m_2 columns in $\mathbf{A}_i$. Note that:

$$\begin{aligned}
\mathrm{Var}(y_i) &= \mathrm{Cov}\{\mathrm{tr}(\mathbf{A}_i^T\mathbf{X}) + \epsilon_i, \mathrm{tr}(\mathbf{A}_i^T\mathbf{X}) + \epsilon_i\} \\
&= \mathrm{Cov}\left\{\sum_{k_1=1}^{m_2} \mathbf{a}_{i,k_1}^T\mathbf{x}_{k_1}, \sum_{k_2=1}^{m_2} \mathbf{a}_{i,k_2}^T\mathbf{x}_{k_2}\right\} + \eta^2 \\
&= \sum_{k_1=1}^{m_2}\sum_{k_2=1}^{m_2} \mathbf{a}_{i,k_1}^T \mathrm{Cov}(\mathbf{x}_{k_1}, \mathbf{x}_{k_2})\mathbf{a}_{i,k_2} + \eta^2 \\
&= \sigma^2 \sum_{k_1=1}^{m_2}\sum_{k_2=1}^{m_2} [\mathcal{P}_\mathcal{V}]_{k_1,k_2}(\mathbf{a}_{i,k_1}^T\mathcal{P}_\mathcal{U}\mathbf{a}_{i,k_2}) + \eta^2 \\
&= \sigma^2\,\mathrm{tr}\{(\mathcal{P}_\mathcal{U}^{1/2}\mathbf{A}_i)\mathcal{P}_\mathcal{V}(\mathcal{P}_\mathcal{U}^{1/2}\mathbf{A}_i)^T\} + \eta^2 \\
&= \sigma^2\|\mathcal{P}_\mathcal{U}\mathbf{A}_i\mathcal{P}_\mathcal{V}\|_F^2 + \eta^2.
\end{aligned}$$

Using analogous steps, it follows that $\mathrm{Cov}(y_i, y_j) = \sigma^2\,\mathrm{tr}(\mathcal{P}_\mathcal{U}\mathbf{A}_i\mathcal{P}_\mathcal{V}\mathbf{A}_j^T) = \sigma^2\langle\mathcal{P}_\mathcal{U}\mathbf{A}_i\mathcal{P}_\mathcal{V}, \mathcal{P}_\mathcal{U}\mathbf{A}_j\mathcal{P}_\mathcal{V}\rangle_F$ for $i \neq j$, which completes the proof. □

Proof of Lemma 3. The proof follows from a straight-forward extension of Lemma 2, and the closed-form conditional distribution of a multivariate Gaussian random vector (see, e.g., Theorem 3.3.4 in [59]). □

Proof of Lemma 4. The proof of this lemma requires the following lemma:

Lemma 10 (Gershgorin's Circle Theorem; [60])**.** *Let $\mathbf{R} \in \mathbb{C}^{n\times n}$, and let $\mathcal{B}_i := \mathcal{B}(R_{i,i}, r_i)$ be the closed ball in the complex plane with center $R_{i,i}$ and radius $r_i = \sum_{j:j\neq i}|R_{i,j}|$. Then all the eigenvalues of $\mathbf{R}$ lie in the union $\cup_{i=1}^n \mathcal{B}_i$.*

Consider now the matrix $\sigma^2\mathbf{R}_n(\mathbf{A}_{1:n}) + \eta^2\mathbf{I}$, which is symmetric and positive-definite. By Lemma 10, the eigenvalues of this matrix (which are real-valued and positive) can be lower bounded by $\min_i\{\mathrm{Var}(y_i) - \sum_{j:j\neq i}\mathrm{Cov}(y_i, y_j)\}$ (note that covariances can always be made positive by an appropriate flipping of measurement masks). Viewing the determinant as a product of eigenvalues, it follows that:

$$\begin{aligned}
\mathrm{H}(\mathbf{A}_{1:n}) &= \det\{\sigma^2\mathbf{R}_n(\mathbf{A}_{1:n}) + \eta^2\mathbf{I}\} \\
&\geq \left[\min_i\left\{\mathrm{Var}(y_i) - \sum_{j:j\neq i}\mathrm{Cov}(y_i, y_j)\right\}\right]^n \\
&= \left[\min_i\left\{\mathrm{Var}(y_i) - \sigma^2\left\langle\mathcal{P}_\mathcal{U}\mathbf{A}_i\mathcal{P}_\mathcal{V}, \sum_{j:j\neq i}\mathcal{P}_\mathcal{U}\mathbf{A}_j\mathcal{P}_\mathcal{V}\right\rangle_F\right\}\right]^n \\
&\qquad\qquad \text{(Lemma 2)}
\end{aligned}$$

which completes the proof. □

Proof of Lemma 5. Under assumption **(A2)**, we have $\mathbf{A}_i = R^{-1/2}\mathbf{R}_i\mathbf{S}_i^T$ for $i = 1, \cdots, n$. The inner-product term then becomes:

$$\begin{aligned}
&\left\langle\mathcal{P}_\mathcal{U}\mathbf{A}_i\mathcal{P}_\mathcal{V}, \sum_{j:j\neq i}\mathcal{P}_\mathcal{U}\mathbf{A}_j\mathcal{P}_\mathcal{V}\right\rangle_F \\
&= \frac{1}{R}\sum_{j:j\neq i}\mathrm{tr}\left\{(\mathcal{P}_\mathcal{U}\mathbf{R}_i)^T(\mathcal{P}_\mathcal{U}\mathbf{R}_j)(\mathcal{P}_\mathcal{V}\mathbf{S}_j)^T(\mathcal{P}_\mathcal{V}\mathbf{S}_i)\right\} \\
&\leq \frac{1}{2R}\sum_{j:j\neq i}\left\{\|(\mathcal{P}_\mathcal{U}\mathbf{R}_i)^T(\mathcal{P}_\mathcal{U}\mathbf{R}_j)\|_F^2 + \|(\mathcal{P}_\mathcal{V}\mathbf{S}_j)^T(\mathcal{P}_\mathcal{V}\mathbf{S}_i)\|_F^2\right\} \\
&\leq C\left\{\max_{j:j\neq i}\|(\mathcal{P}_\mathcal{U}\mathbf{R}_i)^T(\mathcal{P}_\mathcal{U}\mathbf{R}_j)\|_F^2 + \max_{j:j\neq i}\|(\mathcal{P}_\mathcal{V}\mathbf{S}_j)^T(\mathcal{P}_\mathcal{V}\mathbf{S}_i)\|_F^2\right\} \\
&\leq C\left\{\max_{j:j\neq i}\|(\mathcal{P}_\mathcal{U}\mathbf{R}_i)^T(\mathcal{P}_\mathcal{U}\mathbf{R}_j)\|_2^2 + \max_{j:j\neq i}\|(\mathcal{P}_\mathcal{V}\mathbf{S}_j)^T(\mathcal{P}_\mathcal{V}\mathbf{S}_i)\|_2^2\right\},
\end{aligned}$$

where $C = (n-1)/(2R)$. □

Proof of Lemma 6. First, write the matrix $\mathbf{R}_{n+1}([\mathbf{A}_{1:n}\ \ \mathbf{A}]) + \gamma^2\mathbf{I}$ as:

$$\mathbf{R}_{n+1}([\mathbf{A}_{1:n}\ \ \mathbf{A}]) + \gamma^2\mathbf{I} = \begin{pmatrix}\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I} & \mathbf{r}_n(\mathbf{A}) \\ \mathbf{r}_n(\mathbf{A}) & \|\mathcal{P}_\mathcal{U}\mathbf{A}\mathcal{P}_\mathcal{V}\|_F^2 + \gamma^2\end{pmatrix}.$$

Using the determinant formula for the Schur complement (see, e.g., [43]), it follows that:

$$\begin{aligned}
&\mathrm{H}(\mathbf{A}_{1:n}, \mathbf{A}) \\
&= (\sigma^2)^{n+1}\det\{\mathbf{R}_{n+1}([\mathbf{A}_{1:n}\ \ \mathbf{A}]) + \gamma^2\mathbf{I}\} \\
&= (\sigma^2)^{n+1}\det\{\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}\}\cdot \\
&\quad\left[\|\mathcal{P}_\mathcal{U}\mathbf{A}\mathcal{P}_\mathcal{V}\|_F^2 + \gamma^2 - \mathbf{r}_n^T(\mathbf{A})(\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I})^{-1}\mathbf{r}_n(\mathbf{A})\right] \\
&= \sigma^2\mathrm{H}(\mathbf{A}_{1:n})\cdot \\
&\quad\left[\|\mathcal{P}_\mathcal{U}\mathbf{A}\mathcal{P}_\mathcal{V}\|_F^2 + \gamma^2 - \mathbf{r}_n^T(\mathbf{A})(\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I})^{-1}\mathbf{r}_n(\mathbf{A})\right],
\end{aligned}$$

which completes the proof. □

Proof of Theorem 9. The proof of this theorem requires the following lemmas:

Lemma 11. *For fixed projection matrices $\mathcal{P}_\mathcal{U}$ and $\mathcal{P}_\mathcal{V}$, define the operator $\mathcal{P}_{\mathcal{U},\mathcal{V}} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}^{m_1\times m_2}$ as $\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z}) = \mathcal{P}_\mathcal{U}\mathbf{Z}\mathcal{P}_\mathcal{V}$, and consider the linear space of matrices:*

$$\mathcal{T}_{\mathcal{U},\mathcal{V}} = \bigcup_{u_k\in\mathcal{U}, v_k\in\mathcal{V}}\mathrm{span}(\{\mathbf{u}_k\mathbf{v}_k^T\}_{k=1}^R). \tag{29}$$

It follows that $\mathcal{P}_{\mathcal{U},\mathcal{V}}$ is an orthogonal projection operator onto $\mathcal{T}_{\mathcal{U},\mathcal{V}}$ under the Frobenius inner-product $\langle\cdot,\cdot\rangle_F$.

Proof. This can be shown from first principles. Note that:

1) $\mathcal{P}_{\mathcal{U},\mathcal{V}}$ is idempotent, i.e., $\mathcal{P}_{\mathcal{U},\mathcal{V}}\{\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z})\} = \mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z})$ for all $\mathbf{Z} \in \mathbb{R}^{m_1\times m_2}$,
2) $\mathcal{P}_{\mathcal{U},\mathcal{V}}$ is the identity operator on $\mathcal{T}_{\mathcal{U},\mathcal{V}}$,
3) $\mathbf{Z}$ can be uniquely decomposed as $\mathbf{Z} = \mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z}) + [\mathcal{I} - \mathcal{P}_{\mathcal{U},\mathcal{V}}](\mathbf{Z})$, where $\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z}) \in \mathcal{T}_{\mathcal{U},\mathcal{V}}$, $[\mathcal{I} - \mathcal{P}_{\mathcal{U},\mathcal{V}}](\mathbf{Z}) \in \mathcal{T}_{\mathcal{U},\mathcal{V}}^\perp$, and $\perp$ is the orthogonal complement.

By definition, $\mathcal{P}_{\mathcal{U},\mathcal{V}}$ must be a projection operator. Moreover, for any $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{R}^{m_1\times m_2}$, $\langle\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z}_1), \mathbf{Z}_2 - \mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z}_2)\rangle_F = \langle\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z}_2), \mathbf{Z}_1 - \mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{Z}_1)\rangle_F = 0$, so $\mathcal{P}_{\mathcal{U},\mathcal{V}}$ must also be an orthogonal projection operator under $\langle\cdot,\cdot\rangle_F$. By Lemma 1(a) of [21], the range of $\mathcal{P}_{\mathcal{U},\mathcal{V}}$ is $\mathcal{T}_{\mathcal{U},\mathcal{V}}$, which completes the proof. □

Lemma 12. *Let* $f(\mathbf{A}) := \|\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{A})\|_F^2$, *where* $\|\mathbf{A}\|_F^2 = 1$. *If* $\mathbf{A} \in \mathcal{T}_{\mathcal{U},\mathcal{V}}$, *then* $f(\mathbf{A}) = 1$; *otherwise,* $f(\mathbf{A}) < 1$.

Proof. We know from Lemma 11 that $\mathcal{P}_{\mathcal{U},\mathcal{V}}$ is an orthogonal projection operator onto $\mathcal{T}_{\mathcal{U},\mathcal{V}}$ under $\langle\cdot,\cdot\rangle_F$. It follows that:

$$\begin{aligned} 1 = \|\mathbf{A}\|_F^2 &= \|\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{A}) + [\mathcal{I} - \mathcal{P}_{\mathcal{U},\mathcal{V}}](\mathbf{A})\|_F^2 \\ &= \|\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{A})\|_F^2 + \|[\mathcal{I} - \mathcal{P}_{\mathcal{U},\mathcal{V}}](\mathbf{A})\|_F^2 \geq \|\mathcal{P}_{\mathcal{U},\mathcal{V}}(\mathbf{A})\|_F^2, \end{aligned}$$

with equality holding iff $\|[\mathcal{I} - \mathcal{P}_{\mathcal{U},\mathcal{V}}](\mathbf{A})\|_F^2 = 0$, or equivalently, $\mathbf{A} \in \mathcal{T}_{\mathcal{U},\mathcal{V}}$. This completes the proof. □

Using these two lemmas, the proof for Theorem 9 is straight-forward. Consider first the maximization of the first term in (26), $f(\mathbf{A}) = \|\mathcal{P}_{\mathcal{U}}\mathbf{A}\mathcal{P}_{\mathcal{V}}\|_F^2$. By Lemma 12, $f(\mathbf{A})$ is maximized at 1 for any choice of $\mathbf{A} \in \mathcal{T}_{\mathcal{U},\mathcal{V}}$, or equivalently, any $\mathbf{A}$ of the form $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U} \in \mathcal{U}$ and $\mathbf{V} \in \mathcal{V}$. Consider next the minimization of the second term $g(\mathbf{A}) := \mathbf{r}_n^T(\mathbf{A})[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{r}_n(\mathbf{A})$. Note that, for any feasible solution $\mathbf{A}$ satisfying $\|\mathbf{A}\|_F^2 \leq 1$, the matrix $\widetilde{\mathbf{A}} = \mathcal{P}_{\mathcal{U}}\mathbf{A}\mathcal{P}_{\mathcal{V}}$ must also be a feasible solution with the same objective $g(\widetilde{\mathbf{A}}) = g(\mathbf{A})$, since $\|\widetilde{\mathbf{A}}\|_F^2 \leq 1$ by Lemma 12. Hence, the optimal solution for (26) must take the form $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ for some $\mathbf{\Sigma} \in \mathbb{R}^{R\times R}$. Moreover, because $f(\mathbf{A}) = 1$ for all $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, the problem reduces to finding an optimal choice of $\mathbf{\Sigma}$ which minimizes the second term $g(\mathbf{A})$.

With this in mind, $\mathbf{r}_n(\mathbf{A})$ can be rewritten as:

$$\begin{aligned} \mathbf{r}_n(\mathbf{A}) &= [\langle \mathcal{P}_{\mathcal{U}}\mathbf{A}_i\mathcal{P}_{\mathcal{V}}, \mathcal{P}_{\mathcal{U}}\mathbf{A}\mathcal{P}_{\mathcal{V}}\rangle_F]_{i=1}^n \\ &= [\text{tr}(\mathbf{\Sigma}_i^T\mathbf{\Sigma})]_{i=1}^n \qquad (\mathbf{\Sigma}_i := \mathbf{U}^T\mathbf{A}_i\mathbf{V}) \\ &= \left[\text{vec}(\mathbf{\Sigma}_i)^T\text{vec}(\mathbf{\Sigma})\right]_{i=1}^n \\ &= \mathbf{D}\boldsymbol{\nu}, \end{aligned}$$

where $\mathbf{D} = [\text{vec}(\mathbf{\Sigma}_1)^T; \cdots; \text{vec}(\mathbf{\Sigma}_n)^T] \in \mathbb{R}^{n\times R^2}$ and $\boldsymbol{\nu} = \text{vec}(\mathbf{\Sigma}) \in \mathbb{R}^{R^2}$. The desired objective $g(\mathbf{A})$ can then be rearranged as:

$$\begin{aligned} g(\mathbf{A}) &= \mathbf{r}_n^T(\mathbf{A})[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{r}_n(\mathbf{A}) \\ &= \boldsymbol{\nu}^T\mathbf{D}^T[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{D}\boldsymbol{\nu}, \end{aligned}$$

so the optimal $\mathbf{\Sigma}$ which minimizes $g(\mathbf{A})$ corresponds to the unit eigenvector for the smallest eigenvalue of $\mathbf{D}^T[\mathbf{R}_n(\mathbf{A}_{1:n}) + \gamma^2\mathbf{I}]^{-1}\mathbf{D}$. Letting $\mathbf{\Sigma}^*$ denote this optimal matrix, it follows that $\mathbf{A}_{n+1}^* = \mathbf{U}\mathbf{\Sigma}^*\mathbf{V}^T$, which completes the proof. □