Chapter 4

Convex Methods for Low-Rank Matrix Recovery

In this chapter, we will branch out from sparse signals to a broader class of models: the low-rank matrices. Similar to the problem of recovering sparse signals, we consider how to recover a matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ from linear measurements $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}] \in \mathbb{R}^m$. This problem can be phrased as searching for a solution $\boldsymbol{X}$ to a linear system of equations

$$\mathcal{A}\Big[\underbrace{\boldsymbol{X}}_{\text{unknown}}\Big] = \underbrace{\boldsymbol{y}}_{\text{observation}}. \tag{4.0.1}$$

Here, $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear map.

We will see that much of the mathematical structure in the sparse vector recovery problem carries over in a very natural way to this more general setting. In particular, in many interesting instances, we need to recover $\boldsymbol{X}$ from far fewer measurements than the number of entries in the matrix, i.e., $m \ll n_1 \times n_2$. Unless we can leverage some additional prior information about $\boldsymbol{X}$, the problem recovering $\boldsymbol{X}$ from the linear measurements $\boldsymbol{y}$ is ill-posed.

We will consider applications in which we can leverage the following powerful piece of structural information: the target matrix $\boldsymbol{X}$ is *low-rank*. Recall that the rank of a matrix $\boldsymbol{X}$ is the dimension of the linear subspace $\mathrm{col}(\boldsymbol{X})$ spanned by the columns of $\boldsymbol{X}$. If $\boldsymbol{X} = [\boldsymbol{x}_1 \mid \cdots \mid \boldsymbol{x}_{n_2}] \in \mathbb{R}^{n_1 \times n_2}$ is a data matrix whose columns are n_1-dimensional vectors, then $\mathrm{rank}(\boldsymbol{X}) = r \ll n_1$ if and only if the columns of $\boldsymbol{X}$ lie on an r-dimensional linear subspace of the data space $\mathbb{R}^{n_1}$ – see Figure 4.1. Low-rank matrix recovery

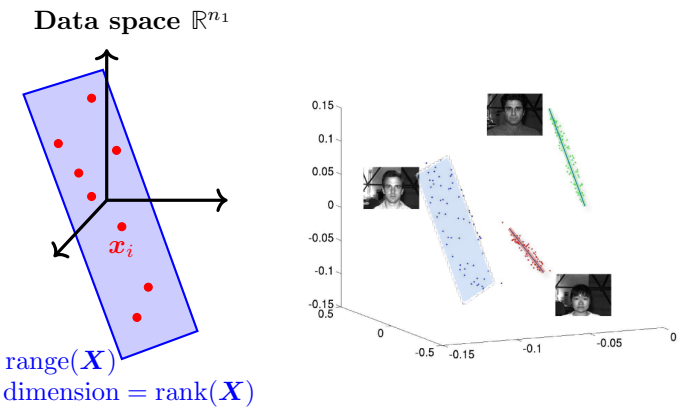

Figure 4.1. **Low-rank data matrices.** If matrix $\boldsymbol{X}$ with columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_2}$ has rank r, its columns lie on an r-dimensional subspace range($\boldsymbol{X}$). Many naturally ocurring data matrices approximately satisfy this property. Right: low-dimensional approximations to images of faces under different lighting conditions.

problems arise in a broad range of application areas. We sketch a few of these below.

4.1 Motivating Examples of Low-Rank Modeling

4.1.1 Latent Semantic Analysis

Low-dimensional models are very popular in document analysis. Consider an idealized problem in search or document retrieval. The system has access to n_2 documents (say, news articles), each of which is viewed as a collection of words in a dictionary of size n_1. For the j-th document, we compute a histogram of word occurrences, giving an n_1-dimensional vector $\boldsymbol{y}_j$ whose i-th entry is the fraction of occurrences of word i in document j. Set

$$\underset{\text{Word occurances}}{\boldsymbol{Y}} = \overset{\text{Words}}{\underset{\text{Documents}}{\left[\boldsymbol{y}_1 \mid \cdots \mid \boldsymbol{y}_{n_2}\right]}}.$$

We model these observations as follows. We imagine that there exist a set of "topics" $\boldsymbol{t}_1 \ldots \boldsymbol{t}_r$. Each topic is a probability distribution on $\{1, 2, \ldots, n_1\}$. We may imagine that the $\boldsymbol{t}_l$ correspond loosely with our informal notion of what a topic is – say, architecture or New York city. An article on architecture in New York would involve multiple topics. We

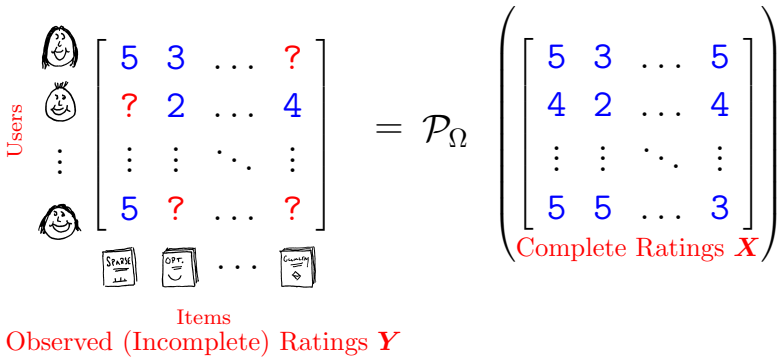

Figure 4.2. **Collaborative filtering as a matrix completion problem.** Consider a universe of n_1 users and n_2 items. Users experience items, and then rate their experience. Our observation $\boldsymbol{Y}$ consists of those ratings that user have provided: Y_{ij} is user i's rating of item j. We wish to predict users ratings of items that they have not yet rated. This can be viewed as attempting to recover a large matrix $\boldsymbol{X}$ from a subset $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X})$ of its entries.

model this as a mixture distribution, writing

$$\underset{\text{Word distribution for document } j}{\boldsymbol{p}_j} = \sum_{l=1}^{r} \underset{\text{topic}}{\boldsymbol{t}_l} \underset{\text{abundance}}{\alpha_{l,j}},$$

where $\alpha_{1,j} + \alpha_{2,j} + \cdots + \alpha_{r,j} = 1$. We imagine that $\boldsymbol{y}_j$ is generated by sampling words independently at random from the mixture distribution $\boldsymbol{p}_j$ and computing a histogram.[1] If the number of words sampled is large, we can imagine $\boldsymbol{y}_j \approx \boldsymbol{p}_j$. So, if we write $\boldsymbol{T} = [\boldsymbol{t}_1 \dots \boldsymbol{t}_r]$ and $\boldsymbol{A} = [\boldsymbol{\alpha}_1 \dots \boldsymbol{\alpha}_n]$, then we have

$$\underset{\text{Word occurances}}{\boldsymbol{Y}} \approx \underset{\text{Topics}}{\boldsymbol{T}} \underset{\text{Abundances}}{\boldsymbol{A}}. \tag{4.1.1}$$

Notice that $\operatorname{rank}(\boldsymbol{TA}) \leq r$: *the rank is bounded by the number of topics. Latent semantic analysis* computes a best low-rank approximation to $\boldsymbol{Y}$ and then uses it for search and indexing [Deerwester et al., 1990].

4.1.2 Recommendation Systems

Here, we imagine that we have n_2 products of interest, and n_1 users. Users consume products and rate them based on the quality of their experience. Our goal is to use the information of all the users' ratings to predict which

[1] In practice, researchers have observed that more complicated methods of constructing $\boldsymbol{Y}$ (say, using the *TF-IDF* weighting) improves performance compared to just using the histogram.

products will appeal to a given user. Formally, our object of interest is a large, unknown matrix

$$\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2},$$

whose (i, j) entry contains user i's degree of preference for item j. If we let

$$\Omega = \Big\{(i, j) \mid \text{user } i \text{ has rated product } j\Big\},$$

then we observe

$$\underset{\text{Observed ratings}}{\boldsymbol{Y}} = \mathcal{P}_\Omega \left[\underset{\text{Complete ratings}}{\boldsymbol{X}}\right].$$

Here, $\mathcal{P}_\Omega$ is the projection operator onto the subset Ω:

$$\mathcal{P}_\Omega[\boldsymbol{X}](i, j) = \begin{cases} X_{ij} & (i, j) \in \Omega, \\ 0 & \text{else.} \end{cases}$$

See Figure 4.2 for a schematic representation of this scenario. Our goal is to fill in the missing entries of $\boldsymbol{X}$. This problem is encountered in online recommendation systems – the most famous recent instance being the "Netflix Prize" competition conducted between 2006 and 2009. Obviously, with no additional assumptions, the problem of filling in the missing entries of $\boldsymbol{X}$ is ill-posed. One popular assumption is that the ratings of distinct users (or distinct products) are correlated, and hence the target matrix $\boldsymbol{X}$ is low-rank, or approximately so. The relevant mathematical problem then becomes filling in the missing entries of a low-rank matrix, or, somewhat equivalently, looking for the matrix $\boldsymbol{X}$ of minimum rank that is consistent with our given observations:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}. \end{array} \tag{4.1.2}$$

This problem is often referred to as *matrix completion* [Candes and Recht, 2009].

4.1.3 3D Shape from Photometric Measurments.

There are also situations in which low-rank models arise due to the physical processes that generate the data. For example, in computer vision, rank constraints arise in a number of problems in reconstructing the three-dimensional shape of a scene from two-dimensional measurements (images). In *photometric stereo* [Woodham, 1980], we obtain images $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_2} \in \mathbb{R}^{n_1}$ of an object illuminated by different distant point light sources. Write $\boldsymbol{Y} = [\boldsymbol{y}_1 \mid \cdots \mid \boldsymbol{y}_{n_2}] \in \mathbb{R}^{n_1 \times n_2}$. Let $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{n_2} \in \mathbb{S}^2$ denote the directions of these light sources. The *Lambertian model* for reflectance models the reflected light intensity as

$$Y_{ij} = \alpha_i [\langle \boldsymbol{\nu}_i, \boldsymbol{l}_j \rangle]_+,$$

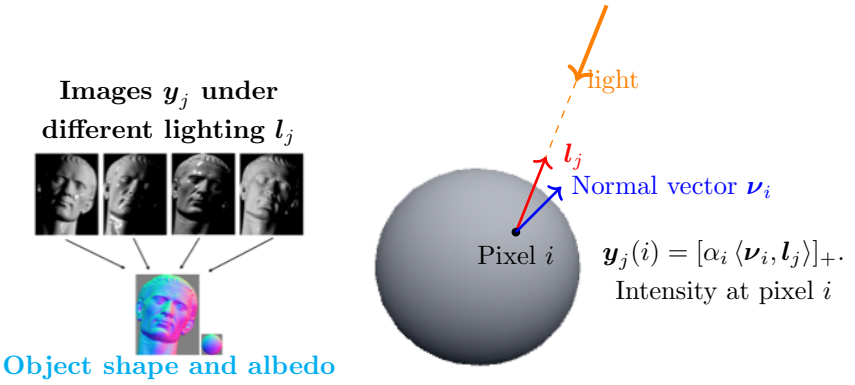

Figure 4.3. **Photometric stereo as low-rank recovery.** Photometric stereo (left) seeks to recover object shape from images taken under different illuminations. Under a diffuse reflective (Lambertian) model (right), this leads directly to a low-rank recovery problem.

where $\boldsymbol{\nu}_i \in \mathbb{S}^2$ is the surface normal at the i-th pixel, α_i is a nonnegative scalar known as the *albedo*, and $[\cdot]_+$ takes the positive part of its argument. This model is appropriate for matte objects. See Figure 4.3 for a visualization of this model.

Under this model, if we let

$$\boldsymbol{N} = \begin{bmatrix} \alpha_1 \boldsymbol{\nu}_1^* \\ \vdots \\ \alpha_m \boldsymbol{\nu}_m^* \end{bmatrix} \in \mathbb{R}^{n_1 \times 3}, \quad \text{and} \quad \boldsymbol{L} = \begin{bmatrix} \boldsymbol{l}_1 \mid \cdots \mid \boldsymbol{l}_n \end{bmatrix} \in \mathbb{R}^{3 \times n_2},$$

then we have

$$\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{N}\boldsymbol{L}],$$

where $\Omega = \{(i,j) \mid \langle \boldsymbol{\nu}_i, \boldsymbol{l}_j \rangle \geq 0\}$. If we can recover the low-rank matrix $\boldsymbol{X} = \boldsymbol{N}\boldsymbol{L}$, we can then recover information about the shape and reflectance of the object. Again, a useful heuristic is to look for a solution of minimum rank consistent with the observations [Wu et al., 2010]:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}. \end{array} \tag{4.1.3}$$

The reader can obtain an open-source implementation of this example from: https://github.com/yasumat/RobustPhotometricStereo. More detailed discussion will be covered in Part III, Chapter 13.

4.1.4 Euclidean Distance Matrix Embedding

The abstract problem can be stated as follows: first imagine that we have n points $\boldsymbol{X} = [\boldsymbol{x}_1 \mid \cdots \mid \boldsymbol{x}_n]$ living in $\mathbb{R}^d$. We can define a matrix $\boldsymbol{D}$ via

$$D_{ij} = d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2.$$

Such a $\boldsymbol{D}$ is known as a *Euclidean distance matrix*. Now imagine the following scenario: rather than observing the $\boldsymbol{x}_i$ themselves, we instead see their pairwise distances $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$. How can we tell if these distances were generated by some configuration of points living in $\mathbb{R}^d$? A necessary and sufficient condition is given by the following classical result:

Theorem 4.1.1 (Schoenberg)**.** *$\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is a Euclidean distance matrix for some set of n points in $\mathbb{R}^d$ if and only if the following conditions hold:*

- *$\boldsymbol{D}$ is symmetric.*
- *$D_{ii} = 0$ for all i.*
- *$\boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Phi}^* \preceq \mathbf{0}$, where $\boldsymbol{\Phi} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^*$ is the* centering matrix *(here $\mathbf{1}$ is the vector whose entries are all one).*
- *$\operatorname{rank}(\boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Phi}^*) \leq d$.*

Now imagine we only know D_{ij} for some subset $\Omega \subset \{1\dots n\} \times \{1\dots n\}$, i.e., we observe $\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{D}]$. We can cast the problem of looking for a Euclidean distance matrix that agrees with our observations as a rank minimization problem:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Phi}^*), \\ \text{subject to} & \boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Phi}^* \preceq \mathbf{0},\ \boldsymbol{D} = \boldsymbol{D}^*,\ \mathcal{P}_\Omega[\boldsymbol{D}] = \boldsymbol{Y},\ \forall i\, D_{ii} = 0. \end{array} \tag{4.1.4}$$

Many additional examples arise, for example in solving positioning problems, problems in system identification, quantum state tomography, image and video alignment, ect. We will survey more of these in the coming application chapters.

4.2 Representing Low-rank Matrix via SVD

In all of the applications described above, our goal is to recover an unknown $\boldsymbol{X}$ whose columns live on an r-dimensional linear subspace of the data space $\mathbb{R}^{n_1}$. This subspace can be characterized via the *singular value decomposition* (SVD) of $\boldsymbol{X}$ (see Appendix A.8 for a more detailed review):

Theorem 4.2.1. *Let $\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}$ be a matrix, and $r = \operatorname{rank}(\boldsymbol{X})$. Then there exist numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and matrices $\boldsymbol{U} \in \mathbb{R}^{n_1\times r}$, $\boldsymbol{V} \in \mathbb{R}^{n_2\times r}$, such that $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}$ and if we set $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$,*

and

$$\boldsymbol{X} \;=\; \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* \;=\; \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*. \tag{4.2.1}$$

Exercise 4.2 gives a guided proof of this result. This construction turns out to be a very versatile tool both for theory and for numerical computation. The *full* singular value decomposition extends the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ to complete orthonormal bases for $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively, by adding bases for the left and right null spaces of $\boldsymbol{X}$:

Theorem 4.2.2. *Let $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ be a matrix. Then there exist orthogonal matrices $\boldsymbol{U} \in O(n_1)$ and $V \in O(n_2)$, and numbers*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{n_1,n_2\}}$$

such that if we let $\boldsymbol{\Sigma} \in \mathbb{R}^{n_1 \times n_2}$ with $\boldsymbol{\Sigma}_{ii} = \sigma_i$ and $\boldsymbol{\Sigma}_{ij} = 0$ for $i \neq j$,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*. \tag{4.2.2}$$

Fact 4.2.3 (Properties of the SVD)**.** *We note the following properties of the construction in Theorem A.8.1:*

- *The right singular vectors $\boldsymbol{v}_i$ are the eigenvectors of $\boldsymbol{X}^*\boldsymbol{X}$.*
- *The nonzero singular values σ_i are the positive square roots of the positive eigenvalues λ_i of $\boldsymbol{X}^*\boldsymbol{X}$.*
- *The left singular vectors $\boldsymbol{u}_i$ are the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^*$ (check this!).*
- *The nonzero singular values σ_i are also the positive square roots of the positive eigenvalues λ_i of $\boldsymbol{X}\boldsymbol{X}^*$.*

Notice that since $\boldsymbol{U}$ and $\boldsymbol{V}$ are nonsingular, the $\operatorname{rank}(\boldsymbol{X}) = \operatorname{rank}(\boldsymbol{\Sigma})$. Since $\boldsymbol{\Sigma}$ is diagonal, this quantity is especially simple – it is simply the number of nonzero entries σ_i! Here, and below, we will let $\boldsymbol{\sigma}(\boldsymbol{X}) = (\sigma_1, \ldots, \sigma_{\min\{n_1,n_2\}}) \in \mathbb{R}^{\min\{n_1,n_2\}}$ denote the vector of singular values of $\boldsymbol{X}$. Then, in the language that we've been developing thus far,

$$\operatorname{rank}(\boldsymbol{X}) = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0 . \tag{4.2.3}$$

Hence any problem that minimizes the rank of an unknown matrix $\boldsymbol{X}$ is essentially minimizing the number of nonzero singular values of $\boldsymbol{X}$ – the "sparsity" of singular values, subject to data constraints.

4.2.1 Singular Vectors via Nonconvex Optimization

The SVD can be computed in time $O(\max\{n_1, n_2\} \min\{n_1, n_2\}^2)$. The first r singular value/vector triples can be computed in time $O(n_1 n_2 r)$. Hence, the problem of finding a linear subspace that best fits a given set of data can be solved in polynomial time. On the surface this is quite remarkable – the

problem of computing singular vectors is nonconvex. We briefly describe why this nonconvex problem can be solved globally in an efficient manner.

We give a brief indication of *why* it is possible to efficiently compute singular vectors of a matrix $\boldsymbol{Y}$. Consider the matrix $\boldsymbol{\Gamma} \doteq \boldsymbol{Y}\boldsymbol{Y}^*$. Let $\boldsymbol{\Gamma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$ be the eigenvalue decomposition of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{n_1})$ be the eigenvalues. It is obvious that the left singular vectors $\boldsymbol{u}_i$ of $\boldsymbol{Y}$ are the eigenvectors of $\boldsymbol{\Gamma}$. Because our goal in this paragraph is merely to convey intuition, we make the simplifying assumption that $\boldsymbol{\Gamma}$ has no repeated eigenvalues and λ_1 is the largest. We show how to use nonconvex optimization to compute the leading eigenvector $\boldsymbol{u}_1$ – see Exercise 4.5 for extensions.

Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & \varphi(\boldsymbol{q}) \equiv -\frac{1}{2}\boldsymbol{q}^*\boldsymbol{\Gamma}\boldsymbol{q}, \\ \text{subject to} & \|\boldsymbol{q}\|_2^2 = 1. \end{array} \tag{4.2.4}$$

The gradient and Hessian of the function $\varphi(\boldsymbol{q})$ are

$$\nabla\varphi(\boldsymbol{q}) = -\boldsymbol{\Gamma}\boldsymbol{q} \quad \text{and} \quad \nabla^2\varphi(\boldsymbol{q}) = -\boldsymbol{\Gamma}, \tag{4.2.5}$$

respectively. A point $\boldsymbol{q}$ is a *critical point* of the function φ over $\mathbb{S}^{n-1} = \left\{\boldsymbol{q} \mid \|\boldsymbol{q}\|_2^2 = 1\right\}$ if there is no direction $\boldsymbol{v} \perp \boldsymbol{q}$ (i.e., no direction that is tangent to the sphere at $\boldsymbol{q}$) along which the function decreases. Equivalently, $\boldsymbol{q}$ is a critical point of φ over the sphere if and only if the gradient is proportional to $\boldsymbol{q}$:

$$\nabla\varphi(\boldsymbol{q}) \propto \boldsymbol{q}. \tag{4.2.6}$$

Using our expression for $\nabla\varphi$, this is true if and only if $\boldsymbol{\Gamma}\boldsymbol{q} = \lambda\boldsymbol{q}$ for some λ: *The critical points of φ over $\mathbb{S}^{n-1}$ are precisely the eigenvectors $\pm\boldsymbol{u}_i$ of $\boldsymbol{\Gamma}$.*

Which critical points $\pm\boldsymbol{u}_i$ are actual local or global minimizers (instead of saddle points)? To answer this question, we need to study the curvature of the function $\varphi(\boldsymbol{q})$ around a critical point $\bar{\boldsymbol{q}}$. In Euclidean space, the correct tool for studying curvature is the Hessian, as is justified by the second order Taylor expansion of the function along the curve $\boldsymbol{x}(t) = \boldsymbol{x} + t\boldsymbol{v}$:

$$f(\boldsymbol{x} + t\boldsymbol{v}) = f(\boldsymbol{x}) + \underbrace{t\,\langle \nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle}_{=0 \text{ at any critical point}} + \frac{1}{2}t^2\boldsymbol{v}^*\nabla^2 f(\boldsymbol{x})\boldsymbol{v} + o(t^2).$$

In Euclidean space, a critical point $\bar{\boldsymbol{x}}$ is a local minimizer if $\nabla^2 f(\bar{\boldsymbol{x}}) \succ \boldsymbol{0}$. Conversely, if $\nabla^2 f(\bar{\boldsymbol{x}})$ has a negative eigenvalue, the point is not a local minimizer.

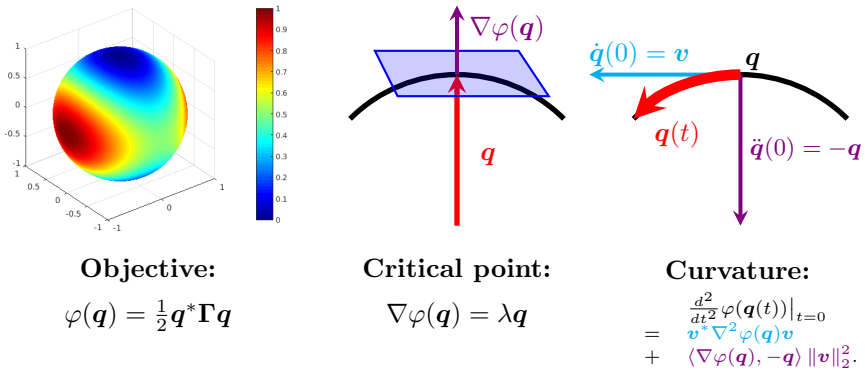

Figure 4.4. **Eigenvector Computation as Nonconvex Optimization over the Sphere.** We plot $\varphi(\boldsymbol{q}) = -\frac{1}{2}\boldsymbol{q}^*\boldsymbol{\Gamma}\boldsymbol{q}$ over the sphere, for one particular $\boldsymbol{\Gamma}$. Critical points (middle) are points $\boldsymbol{q}$ for which $\nabla\varphi(\boldsymbol{q})$ is proportional to $\boldsymbol{q}$. Every critical point is an eigenvector of $\boldsymbol{\Gamma}$; the only local minimizers are eigenvectors that correspond to the largest eigenvalue $\lambda_1(\boldsymbol{\Gamma})$. Right: curvature of the φ over the sphere comes from both curvature $\nabla^2\varphi$ of φ and the curvature of the sphere.

Over the sphere, we can perform a similar Taylor expansion, but we need to replace the straight line $\boldsymbol{x}(t) = \boldsymbol{x} + t\boldsymbol{v}$ with a great circle[2]

$$\boldsymbol{q}(t) = \boldsymbol{q}\cos(t) + \boldsymbol{v}\sin(t), \tag{4.2.7}$$

where $\boldsymbol{v} \perp \boldsymbol{q}$ and $\|\boldsymbol{v}\|_2 = 1$. Calculus shows that the second derivative of $\varphi(\boldsymbol{q}(t))$ is given by

$$\frac{d^2}{dt^2}\varphi(\boldsymbol{q}(t))\Big|_{t=0} = \underbrace{\boldsymbol{v}^*\nabla^2\varphi(\boldsymbol{q})\boldsymbol{v}}_{\text{Curvature of } \varphi} - \underbrace{\langle\nabla\varphi(\boldsymbol{q}), \boldsymbol{q}\rangle\, \boldsymbol{v}^*\boldsymbol{v}}_{\text{Curvature of the sphere}}. \tag{4.2.8}$$

This formula contains two terms, which combine the usual hessian of φ (accounting for the curvature of φ) and a second correction term involving $-\langle\nabla\varphi(\boldsymbol{q}), \boldsymbol{q}\rangle$ which accounts for the fact that the curve $\boldsymbol{q}(t)$ curves in the $-\boldsymbol{q}$ direction in order to stay on the sphere.

Noting that $\nabla^2\varphi(\boldsymbol{q}) = -\boldsymbol{\Gamma}$. So we have $\langle\nabla\varphi(\boldsymbol{u}_i), \boldsymbol{u}_i\rangle = -\boldsymbol{u}_i^*\boldsymbol{\Gamma}\boldsymbol{u}_i = -\lambda_i$, we observe that at a critical point $\bar{\boldsymbol{q}} = \pm\boldsymbol{u}_i$, the second derivative in the $\boldsymbol{v}$ direction is

$$\frac{d^2}{dt^2}\varphi(\boldsymbol{q}(t))\Big|_{t=0} = \boldsymbol{v}^*\Big(-\boldsymbol{\Gamma} + \lambda_i\boldsymbol{I}\Big)\boldsymbol{v}. \tag{4.2.9}$$

The eigenvalues of the operator $-\boldsymbol{\Gamma} + \lambda_i$ take the form $-\lambda_j + \lambda_i$; there is a strictly negative eigenvalue if $\boldsymbol{u}_i$ is an eigenvector that does not correspond to λ_1: $\pm\boldsymbol{u}_1$ are the only local minimizers of φ. All other critical points have

[2] Curves of this form are *geodesics* on $\mathbb{S}^{n-1}$.

a direction of strict negative curvature. This benign geometry implies that a simple projected gradient method converges to a global optimizer from almost any initialization. Exercise 4.6 gives a more algorithmic analysis, using the power method. We will return more formally to questions about global optimization of nonconvex functions in Chapter 9 in the context of learning sparse models for data. For now, we take these observations as an intuitive indication of why the SVD is amenable to efficient computation.

Implications and history.

Whichever rationale we adopt, the fact that the SVD is both optimal (in a precisely defined and often quite relevant sense) and efficient (at least for moderate problems) makes it a very useful element in the numerical computing toolbox. The canonical example application of the SVD is *Principal Component Analysis* (PCA). Outlined in 1901 and 1933 papers by Pearson and Hotelling [Pearson, 1901, Hotelling, 1933], respectively, PCA finds a best-fitting low-dimensional subspace, which can be computed via the SVD, as suggested by Theorem 4.2.4 below. Remarkably, Pearson's 1901 paper asserts that PCA is "well-suited to numerical computation" – meaning hand calculations!

4.2.2 Best Low-Rank Matrix Approximation

We are interested in recovering a low-rank matrix that is consistent with certain linear observations. Because the rank has a similar characteristic to the ℓ^0 norm, one should expect that these problems would be computationally difficult in general, as in the case with recovering a sparse solution (see Theorem 2.2.8).

Remarkably, there *are* however a few special instances of rank minimization that we *can* solve efficiently, with virtually no assumptions on the input. The most important is the *best rank-r approximation* problem, in which we try to approximate an arbitrary input matrix $\boldsymbol{Y}$ with a matrix $\boldsymbol{X}$ of rank at most r such that the approximation error $\|\boldsymbol{X} - \boldsymbol{Y}\|_F$ is as small as possible. The optimal solution to this problem can be obtained by simply retaining the first r leading singular values/vectors of $\boldsymbol{Y}$:

Theorem 4.2.4. *Let $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$, and consider the following optimization problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X} - \boldsymbol{Y}\|_F, \\ \text{subject to} & \operatorname{rank}(\boldsymbol{X}) \leq r. \end{array} \tag{4.2.10}$$

Every optimal solution $\hat{\boldsymbol{X}}$ to the above problem has the form $\hat{\boldsymbol{X}} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$, where $\boldsymbol{Y} = \sum_{i=1}^{\min(n_1,n_2)} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$ is a (full) singular value decomposition of $\boldsymbol{Y}$.

In fact, the same solution (truncating the SVD) also solves the low-rank approximation problem when the error is measured in the operator norm,

or any other orthogonal-invariant matrix norm. Please see Exercise 4.3 for guidance on how to prove Theorem 4.2.4.

The problem (4.2.10) can be turned around and cast as one of minimizing the rank of the unknown matrix, subject to a data fidelity constraint:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \|\boldsymbol{X} - \boldsymbol{Y}\|_F \leq \varepsilon. \end{array} \tag{4.2.11}$$

This is an example of a *matrix rank minimization* problem – we seek a matrix of minimum rank that is consistent with some given observations. Because of its very special nature, this particular rank minimization can be solved optimally via the SVD. We leave the solution to this problem as an exercise to the reader (see Exercise 4.4).[3]

4.3 Recovering a Low-Rank Matrix

4.3.1 General Rank Minimization Problems

In the previous section, we saw that for certain very specific rank minimization problems, globally optimal solutions could be obtained using efficient algorithms based on the singular value decomposition. However, all of the applications discussed above (and many others!) force us to attempt to minimize the rank of $\boldsymbol{X}$ over much more complicated sets. One model example problem is the *affine rank minimization* [Fazel et al., 2004] problem

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array} \tag{4.3.1}$$

Here $\boldsymbol{y} \in \mathbb{R}^m$ is an observation, and $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear map. When $m \ll n_1 n_2$, the linear system of equations $\mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}$ is underdetermined. The notion of a linear map $\mathcal{A}$ from $n_1 \times n_2$ matrices to m-dimensional vectors may seem somewhat abstract. Any linear map of this form can be represented using the matrix inner product[4]:

$$\mathcal{A}[\boldsymbol{X}] = (\langle \boldsymbol{A}_1, \boldsymbol{X} \rangle, \ldots, \langle \boldsymbol{A}_m, \boldsymbol{X} \rangle). \tag{4.3.2}$$

Here, the set of matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{R}^{n_1 \times n_2}$ define our "measurements" $\boldsymbol{y}$, through their inner products with the unknown matrix $\boldsymbol{X}$.[5]

A mathematically simple and natural assumption on these "measurement" matrices is that they are i.i.d. Gaussian matrices. Such an

[3] Hint: you may first try to guess what the optimal solution is and then show its optimality.

[4] Recall that the standard inner product between matrices $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{R}^{n_1 \times n_2}$ is defined by $\langle \boldsymbol{P}, \boldsymbol{Q} \rangle = \sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{Q}_{ij} = \operatorname{trace}[\boldsymbol{Q}^* \boldsymbol{P}]$.

[5] You can think of the measurements $\boldsymbol{A}_i$ as analogous to the *rows* of the matrix $\boldsymbol{A}$ in the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ studied in Chapters 2-3.

assumption will allow us to understand the conditions under which one could expect to recover a low-rank matrix with generic measurements. Hence, our first attempt to understand the low-rank recovery problem will rely on such a simplifying assumption. However, in many practical problems of interest, the operator $\mathcal{A}$ has particular structures that make it behave differently. As a concrete example, in the matrix completion problems discussed above, we would have $p = |\Omega|$, and $\boldsymbol{A}_l = \boldsymbol{e}_{i_l}\boldsymbol{e}_{j_l}^*$, with $\Omega = \{(i_1, j_1), \ldots, (i_m, j_m)\}$. We will also thoroughly analyze this important special case and provide conditions under which the recovery can be successful.

Connection to ℓ^0, NP-hardness.

To make the connection to sparse recovery explicit, using the observation that $\operatorname{rank}(\boldsymbol{X}) = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0$, we can rewrite the affine rank minimization problem as

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0 \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array} \tag{4.3.3}$$

Moreover, *if* $\boldsymbol{X}$ is a diagonal matrix, then $\operatorname{rank}(\boldsymbol{X}) = \|\boldsymbol{X}\|_0$. So, every ℓ^0 minimization problem can be converted into a rank minimization problem with a diagonal constraint. This means that in the worst case, the rank minimization problem is at least as hard as the ℓ^0 minimization problem: it is NP-hard (as shown in Theorem 2.2.8).

As was the case for ℓ^0 minimization, we could simply give up here in searching for tractable algorithms. However, given the close analogy between rank minimization and ℓ^0 minimization, we might hope that there could be some fairly broad class of "nice enough" problems that we *can* solve efficiently.

4.3.2 Convex Relaxation of Rank Minimization

The close analogy to ℓ^0 minimization suggests a natural strategy: replace the rank, which is the ℓ^0 norm $\boldsymbol{\sigma}(\boldsymbol{X})$ with the ℓ^1 norm of $\boldsymbol{\sigma}(\boldsymbol{X})$:

$$\|\boldsymbol{\sigma}(\boldsymbol{X})\|_1 = \sum_i \sigma_i(\boldsymbol{X}). \tag{4.3.4}$$

We call this function the *nuclear norm* of $\boldsymbol{X}$, and reserve the special notation

$$\|\boldsymbol{X}\|_* = \sum_i \sigma_i(\boldsymbol{X}). \tag{4.3.5}$$

When $\boldsymbol{X}$ is a symmetric positive semidefinite matrix, $\boldsymbol{X}$ has real nonnegative eigenvalues, and $\sigma_i(\boldsymbol{X}) = \lambda_i(\boldsymbol{X})$. Since $\sum_i \lambda_i(\boldsymbol{X}) = \operatorname{trace}(\boldsymbol{X})$, in the special case when $\boldsymbol{X}$ is semidefinite, $\|\boldsymbol{X}\|_* = \operatorname{trace}[\boldsymbol{X}]$. For this reason, the nuclear norm is sometimes also referred to as the *trace norm*. Other names

in various literatures include the *Schatten* 1*-norm* and *Ky-Fan* m*-norm*.[6] When $\boldsymbol{X}$ is not a semidefinite matrix, the function $\|\boldsymbol{X}\|_*$ depends on the entries in a very complicated way. It may not be at all obvious that this function is a norm, or even a convex function.

To allay any suspicion, we give a quick proof that $\|\cdot\|_*$ is indeed a norm:

Theorem 4.3.1. *For $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$, let $\|\boldsymbol{M}\|_* = \sum_{i=1}^{\min\{n_1,n_2\}} \sigma_i(\boldsymbol{M})$. Then $\|\cdot\|_*$ is a norm. Moreover, the nuclear norm and the ℓ^2 operator norm are* dual norms*:*

$$\|\boldsymbol{M}\|_* = \sup_{\|\boldsymbol{N}\| \le 1} \langle \boldsymbol{M}, \boldsymbol{N} \rangle, \quad \textit{and} \quad \|\boldsymbol{M}\| = \sup_{\|\boldsymbol{N}\|_* \le 1} \langle \boldsymbol{M}, \boldsymbol{N} \rangle. \tag{4.3.6}$$

Proof. We begin by proving the first equality in (4.3.6). Let

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* \tag{4.3.7}$$

be a full singular value decomposition of $\boldsymbol{M}$, with $\boldsymbol{U} \in O(n_1)$, $\boldsymbol{V} \in O(n_2)$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{n_1 \times n_2}$, and note that

$$\begin{aligned} \sup_{\|\boldsymbol{N}\| \le 1} \langle \boldsymbol{N}, \boldsymbol{M} \rangle &= \sup_{\|\boldsymbol{N}\| \le 1} \langle \boldsymbol{N}, \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* \rangle \\ &= \sup_{\|\boldsymbol{N}\| \le 1} \left\langle \boldsymbol{U}^*\boldsymbol{N}\boldsymbol{V}, \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{n_2} \\ 0 & 0 & 0 \\ & \vdots & \end{bmatrix} \right\rangle \\ &\ge \sum_{i=1}^{n_2} \sigma_i, \end{aligned} \tag{4.3.8}$$

where the last line follows by making a particular choice

$$\boldsymbol{N} = \boldsymbol{U} \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ 0 & 0 & 0 \\ & \vdots & \end{bmatrix} \boldsymbol{V}^*. \tag{4.3.9}$$

So, $\sup_{\boldsymbol{N}} \langle \boldsymbol{N}, \boldsymbol{M} \rangle \ge \|\boldsymbol{M}\|_*$.

For the opposite direction, notice if matrix $\boldsymbol{N} \in \mathbb{R}^{n_1 \times n_2}$ satisfies $\|\boldsymbol{N}\| \le 1$, then $\bar{\boldsymbol{N}} \doteq \boldsymbol{U}^*\boldsymbol{N}\boldsymbol{V}$ has columns of ℓ^2 norm at most one. Thus, for each

[6] For $p \in [1, \infty]$, the Schatten p-norm of a matrix is $\|\boldsymbol{X}\|_{S_p} = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_p$. The Ky-Fan k-norm is $\|\boldsymbol{X}\|_{KF_k} = \sum_{i=1}^{k} \sigma_i(\boldsymbol{X})$. Both of these functions are examples of *orthogonal invariant matrix norms*.

i, $\bar{\boldsymbol{N}}_{ii} \leq 1$, and

$$\langle \boldsymbol{N}, \boldsymbol{M} \rangle = \left\langle \bar{\boldsymbol{N}}, \boldsymbol{\Sigma} \right\rangle = \sum_{i=1}^{n_2} \bar{\boldsymbol{N}}_{ii} \sigma_i \;\leq\; \sum_i \sigma_i = \|\boldsymbol{M}\|_* . \tag{4.3.10}$$

This establishes the result.

For the second equality in (4.3.6), notice that for any nonzero $\boldsymbol{M}$,

$$\langle \boldsymbol{M}, \boldsymbol{N} \rangle \;=\; \|\boldsymbol{M}\| \left\langle \frac{\boldsymbol{M}}{\|\boldsymbol{M}\|}, \boldsymbol{N} \right\rangle \;\leq\; \|\boldsymbol{M}\| \, \|\boldsymbol{N}\|_* . \tag{4.3.11}$$

Hence, $\sup_{\|\boldsymbol{N}\|_* \leq 1} \langle \boldsymbol{M}, \boldsymbol{N} \rangle \leq \|\boldsymbol{M}\|$. To show that this inequality is actually an equality, let take $\boldsymbol{N} = \boldsymbol{u}_1 \boldsymbol{v}_1^*$, and notice that $\|\boldsymbol{N}\|_* = 1$ and $\langle \boldsymbol{M}, \boldsymbol{N} \rangle = \boldsymbol{u}_1^* \boldsymbol{M} \boldsymbol{v}_1 = \sigma_1(\boldsymbol{M}) = \|\boldsymbol{M}\|$. This completes the proof of (4.3.6).

To see that $\|\cdot\|_*$ is indeed a norm, we just use (4.3.6) to verify that the three axioms of a norm are satisfied. Since the singular values are nonnegative, and $\sigma_1(\boldsymbol{M}) = 0$ if and only if $\boldsymbol{M} = \boldsymbol{0}$, it is immediate that $\|\boldsymbol{M}\|_* \geq 0$ with equality iff $\boldsymbol{M} = \boldsymbol{0}$. For nonnegative homogeneity, notice that for $t \in \mathbb{R}_+$,

$$\|t\boldsymbol{M}\|_* = \sup_{\|\boldsymbol{N}\| \leq 1} \langle t\boldsymbol{M}, \boldsymbol{N} \rangle = t \sup_{\|\boldsymbol{N}\| \leq 1} \langle \boldsymbol{M}, \boldsymbol{N} \rangle = t \, \|\boldsymbol{M}\|_* . \tag{4.3.12}$$

Finally, for the triangle inequality, consider two matrices $\boldsymbol{M}$ and $\boldsymbol{M}'$, and notice that

$$\begin{aligned} \left\|\boldsymbol{M} + \boldsymbol{M}'\right\|_* \;&=\; \sup_{\|\tilde{\boldsymbol{N}}\| \leq 1} \left\langle \boldsymbol{M} + \boldsymbol{M}', \tilde{\boldsymbol{N}} \right\rangle \\ &\leq\; \sup_{\|\boldsymbol{N}\| \leq 1} \langle \boldsymbol{M}, \boldsymbol{N} \rangle \;+\; \sup_{\|\boldsymbol{N}'\| \leq 1} \left\langle \boldsymbol{M}', \boldsymbol{N}' \right\rangle \\ &=\; \|\boldsymbol{M}\|_* + \left\|\boldsymbol{M}'\right\|_* , \end{aligned} \tag{4.3.13}$$

verifying the triangle inequality. This shows that $\|\cdot\|_*$ is indeed a norm. □

The above proof highlights a useful fact about $\|\cdot\|_*$: it is the dual norm of the operator norm $\|\boldsymbol{X}\| = \sigma_1(\boldsymbol{X})$. The fact that $\|\cdot\|_*$ is the dual of $\|\cdot\|$ explains the $*$ notation – this symbol is often used for duality.

Because $\|\cdot\|_*$ is a norm, it is convex. Hence, a natural convex replacement for the rank minimization problem is the *nuclear norm minimization* problem

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* , \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array} \tag{4.3.14}$$

This problem is convex, and moreover is efficiently solvable. In Chapter 8, we will see how to use the special structure of this problem to give practical, efficient algorithms which work well at moderate scales.

4.3.3 Nuclear Norm as A Convex Envelope of Rank

From the analogy to ℓ^0/ℓ^1 minimization, we might guess that the nuclear norm is a good convex surrogate for the rank, over some appropriate set. Recall that we have proved in Theorem 2.3.3 that the ℓ^1 norm was the convex envelope of the ℓ^0 norm over the ℓ^∞ ball. Since for a matrix $\boldsymbol{X}$, $\|\boldsymbol{\sigma}(\boldsymbol{X})\|_\infty = \sigma_1(\boldsymbol{X}) = \|\boldsymbol{X}\|$, you might guess the following relationship:

Theorem 4.3.2. *$\|\boldsymbol{M}\|_*$ is the convex envelope of $f(\boldsymbol{M}) = \operatorname{rank}(\boldsymbol{M})$ over*

$$\mathsf{B}_{op} = \{\boldsymbol{M} \mid \|\boldsymbol{M}\| \leq 1\}. \tag{4.3.15}$$

Proof. We prove that any convex function f which satisfies

$$f(\boldsymbol{M}) \leq \operatorname{rank}(\boldsymbol{M}) \tag{4.3.16}$$

for all $\boldsymbol{M} \in \mathsf{B}_{op}$, is majorized by the nuclear norm: $f(\boldsymbol{M}) \leq \|\boldsymbol{M}\|_*$.

Write the SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. Notice that

$$\boldsymbol{\Sigma} \in \operatorname{conv}\left\{\operatorname{diag}(\boldsymbol{w}) \mid \boldsymbol{w} \in \{0,1\}^{\min\{n_1,n_2\}}\right\}, \tag{4.3.17}$$

and for any $\boldsymbol{w} \in \{0,1\}^{\min\{n_1,n_2\}}$,

$$\|\boldsymbol{U}\operatorname{diag}(\boldsymbol{w})\boldsymbol{V}^*\|_* = \sum_i w_i = \operatorname{rank}(\boldsymbol{U}\operatorname{diag}(\boldsymbol{w})\boldsymbol{V}^*). \tag{4.3.18}$$

Writing

$$\boldsymbol{\Sigma} = \sum_i \lambda_i \operatorname{diag}(\boldsymbol{w}_i) \tag{4.3.19}$$

with $\boldsymbol{w}_i \in \{0,1\}^{\min\{n_1,n_2\}}$ with $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$, and applying Jensen's inequality, we obtain

$$\begin{aligned}
f(\boldsymbol{M}) &= f\left(\boldsymbol{U}\sum_i \lambda_i \operatorname{diag}(\boldsymbol{w}_i)\boldsymbol{V}^*\right) && (4.3.20)\\
&\leq \sum_i \lambda_i f\left(\boldsymbol{U}\operatorname{diag}(\boldsymbol{w}_i)\boldsymbol{V}^*\right) && (4.3.21)\\
&\leq \sum_i \lambda_i \operatorname{rank}\left(\boldsymbol{U}\operatorname{diag}(\boldsymbol{w}_i)\boldsymbol{V}^*\right) && (4.3.22)\\
&= \sum_i \lambda_i \|\boldsymbol{w}_i\|_1 && (4.3.23)\\
&= \left\|\boldsymbol{U}\sum_i \lambda_i \operatorname{diag}(\boldsymbol{w}_i)\boldsymbol{V}^*\right\|_* && (4.3.24)\\
&= \|\boldsymbol{M}\|_* && (4.3.25)
\end{aligned}$$

as desired. □

Note that this proof essentially mirrored our argument for ℓ^1 and ℓ^∞. This is not a coincidence!

4.3.4 Success of Nuclear Norm under Rank-RIP

For now, content that we can solve nuclear norm minimization problems efficiently, we turn our attention to whether nuclear norm minimization actually gives the correct answers. Namely, if we know that $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}]$, with $r = \text{rank}(\boldsymbol{X}_{\text{true}}) \ll n$, is it true that $\boldsymbol{X}_{\text{true}}$ is the unique optimal solution to the nuclear norm minimization problem (4.3.14)? What we can say depends strongly on what we know about the operator $\mathcal{A}$.

By analogy to the *sparse* recovery problem, we can ask if it is enough for $\mathcal{A}$ to preserve the geometry of a small set of structured objects – here, the low-rank matrices. Formally, we can define a *rank-restricted isometry property*, under which for every rank-r $\boldsymbol{X}$, $\|\mathcal{A}[\boldsymbol{X}]\|_2 \approx \|\boldsymbol{X}\|_F$.

Definition 4.3.3 (Rank-Restricted Isometry Property [Recht et al., 2010]). *The operator $\mathcal{A}$ has the rank-restricted isometry property of rank r with constant δ, if $\forall\, \boldsymbol{X}$ s.t. $\text{rank}(\boldsymbol{X}) \leq r$, we have*

$$(1-\delta)\|\boldsymbol{X}\|_F^2 \;\leq\; \|\mathcal{A}[\boldsymbol{X}]\|_2^2 \;\leq\; (1+\delta)\|\boldsymbol{X}\|_F^2. \tag{4.3.26}$$

The rank-r restricted isometry constant $\delta_r(\mathcal{A})$ is the smallest δ such that the above property holds.

As with the RIP for sparse vectors, the rank-RIP implies uniqueness of structured (low-rank) solutions:

Theorem 4.3.4. *If $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}]$, with $r = \text{rank}(\boldsymbol{X}_{\text{true}})$ and $\delta_{2r}(\mathcal{A}) < 1$, then $\boldsymbol{X}_{\text{true}}$ is the unique optimal solution to the rank minimization problem*

$$\begin{array}{ll} \text{minimize} & \text{rank}(\boldsymbol{X}) \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array} \tag{4.3.27}$$

We leave the proof of this claim as an exercise to the reader (see Exercise 4.14). The key property is the subadditivity of the matrix rank, namely,

$$\text{rank}(\boldsymbol{X} + \boldsymbol{X}') \;\leq\; \text{rank}(\boldsymbol{X}) + \text{rank}(\boldsymbol{X}'). \tag{4.3.28}$$

Moreover, like the RIP for sparse vectors, when the rank-RIP holds with sufficiently small constant δ, we can conclude that nuclear norm minimization will recover the desired low-rank solution:

Theorem 4.3.5 (Recht, Fazel, Parillo [Recht et al., 2010]). *Suppose that $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}]$ with $\text{rank}(\boldsymbol{X}_{\text{true}}) \leq r$, and that $\delta_{4r}(\mathcal{A}) \leq \sqrt{2}-1$. Then $\boldsymbol{X}_{\text{true}}$ is the unique optimal solution to the nuclear norm minimization problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \end{array} \tag{4.3.29}$$

There is nothing so special here about the numbers $4r$ and $\sqrt{2}-1$. The interesting part is the qualitative statement: if $\mathcal{A}$ respects the geometry of low-rank matrices in a sufficiently strong sense, then nuclear norm minimization succeeds. The proof is analogous to the proof we have given in the

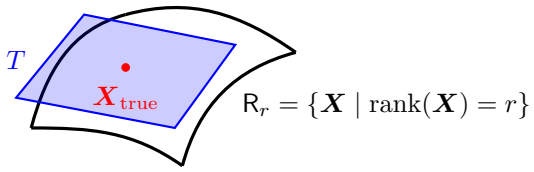

Figure 4.5. **"Support" of a Low-rank Matrix $\boldsymbol{X}_{\text{true}}$.** Consider a rank-$r$ matrix $\boldsymbol{X}_0$ with compact singular value decomposition $\boldsymbol{X}_{\text{true}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. The subspace $T = \{\boldsymbol{U}\boldsymbol{W}^* + \boldsymbol{Q}\boldsymbol{V}^*\}$ can be interpreted as the *tangent space* to the collection R_r of rank-r matrices at $\boldsymbol{X}_0$.

previous chapter for the success of ℓ^1 minimization for recovering sparse signals. However, to extend the proof techniques from ℓ^1 to nuclear norm, we need to generalize a few concepts from vectors to matrices.

"Support" and "Signs" of a Low-rank Matrix $\boldsymbol{X}_0$.

Let $\boldsymbol{X}_{\text{true}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ denote the compact SVD of the true solution $\boldsymbol{X}_{\text{true}}$. Let

$$T \doteq \left\{\boldsymbol{U}\boldsymbol{W}^* + \boldsymbol{Q}\boldsymbol{V}^* \mid \boldsymbol{W} \in \mathbb{R}^{n_2 \times r},\ \boldsymbol{Q} \in \mathbb{R}^{n_1 \times r}\right\} \subseteq \mathbb{R}^{n_1 \times n_2}. \tag{4.3.30}$$

Notice that T is a linear subspace. In the analogy between ℓ^1 minimization and nuclear norm minimization, the subspace T plays the role of the "support" of $\boldsymbol{X}_{\text{true}}$. Geometrically, T represents the *tangent space* to the set of rank-r matrices at $\boldsymbol{X}_{\text{true}}$ – see Figure 4.5 and Exercise 4.9. The subspace T is generated by matrices $\boldsymbol{U}\boldsymbol{W}^*$ whose column space is contained in $\operatorname{col}(\boldsymbol{X}_{\text{true}})$ and matrices $\boldsymbol{Q}\boldsymbol{V}^*$ whose row space is contained in $\operatorname{row}(\boldsymbol{X}_{\text{true}})$. Notice that elements in T have rank no more than $2r$. Meanwhile the matrix $\boldsymbol{U}\boldsymbol{V}^*$ plays the role of the "signs" of $\boldsymbol{X}_{\text{true}}$ since $\boldsymbol{U}\boldsymbol{V}^* \in T$ and

$$\langle \boldsymbol{X}_{\text{true}}, \boldsymbol{U}\boldsymbol{V}^* \rangle = \|\boldsymbol{X}_{\text{true}}\|_* . \tag{4.3.31}$$

The orthogonal complement of T is

$$T^\perp \doteq \{\boldsymbol{M} \mid \operatorname{col}(\boldsymbol{M}) \perp \operatorname{col}(\boldsymbol{X}), \operatorname{row}(\boldsymbol{M}) \perp \operatorname{row}(\boldsymbol{X})\} . \tag{4.3.32}$$

Let $\boldsymbol{P}_U = \boldsymbol{U}\boldsymbol{U}^*$ and $\boldsymbol{P}_V = \boldsymbol{V}\boldsymbol{V}^*$ be the orthogonal projections onto the column space and row space of $\boldsymbol{X}_{\text{true}}$, respectively. Then the orthogonal projections onto these subspaces are given by[7]

$$\mathcal{P}_T[\boldsymbol{M}] = \boldsymbol{P}_U\boldsymbol{M} + \boldsymbol{M}\boldsymbol{P}_V - \boldsymbol{P}_U\boldsymbol{M}\boldsymbol{P}_V, \tag{4.3.33}$$

and

$$\mathcal{P}_{T^\perp}[\boldsymbol{M}] = (\boldsymbol{I} - \boldsymbol{P}_U)\boldsymbol{M}(\boldsymbol{I} - \boldsymbol{P}_V). \tag{4.3.34}$$

[7]Equations (4.3.33) and (4.3.34) can be derived from the condition that at $\mathcal{P}_T[\boldsymbol{M}]$, the error $\boldsymbol{M} - \mathcal{P}_T[\boldsymbol{M}]$ is orthogonal to T.

Notice that because the orthogonal projections $\boldsymbol{P}_{U^\perp} = \boldsymbol{I} - \boldsymbol{P}_U$ and $\boldsymbol{P}_{V^\perp} = \boldsymbol{I} - \boldsymbol{P}_V$ have norm at most one, $\mathcal{P}_{T^\perp}$ does not increase the operator norm:

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{M}]\| \leq \|\boldsymbol{M}\| . \tag{4.3.35}$$

Feasible Cone Restriction.

Note that any matrix $\boldsymbol{M} \in T^\perp$ has columns that are orthogonal to the columns of $\boldsymbol{U}$ and rows that are orthogonal to the rows of $\boldsymbol{V}^*$. This implies that

$$\|\boldsymbol{M} + \boldsymbol{U}\boldsymbol{V}^*\| = \max\{\|\boldsymbol{M}\|, \|\boldsymbol{U}\boldsymbol{V}^*\|\} = \max\{\|\boldsymbol{M}\|, 1\} . \tag{4.3.36}$$

So, for any matrix $\boldsymbol{X}$,

$$\begin{aligned}
\|\boldsymbol{X}\|_* &= \sup_{\|\boldsymbol{Q}\| \leq 1} \langle \boldsymbol{X}, \boldsymbol{Q} \rangle && (4.3.37)\\
&\geq \sup_{\|\boldsymbol{M}\| \leq 1} \langle \boldsymbol{X}, \boldsymbol{U}\boldsymbol{V}^* + \mathcal{P}_{T^\perp}[\boldsymbol{M}] \rangle && (4.3.38)\\
&= \langle \boldsymbol{X}, \boldsymbol{U}\boldsymbol{V}^* \rangle + \sup_{\|\boldsymbol{M}\| \leq 1} \langle \mathcal{P}_{T^\perp}[\boldsymbol{X}], \boldsymbol{M} \rangle && (4.3.39)\\
&= \langle \boldsymbol{X}, \boldsymbol{U}\boldsymbol{V}^* \rangle + \|\mathcal{P}_{T^\perp}[\boldsymbol{X}]\|_* . && (4.3.40)
\end{aligned}$$

Let $\hat{\boldsymbol{X}}$ be any optimal solution to our problem (4.3.29). It can be written as $\hat{\boldsymbol{X}} = \boldsymbol{X}_{\text{true}} + \boldsymbol{H}$, with $\boldsymbol{H} = \hat{\boldsymbol{X}} - \boldsymbol{X}_{\text{true}} \in \text{null}(\mathcal{A})$. From the above calculation, we have

$$\begin{aligned}
\|\boldsymbol{X}_{\text{true}} + \boldsymbol{H}\|_* &\geq \langle \boldsymbol{X}_{\text{true}} + \boldsymbol{H}, \boldsymbol{U}\boldsymbol{V}^* \rangle + \left\| \mathcal{P}_{T^\perp}[\hat{\boldsymbol{X}}] \right\|_* && (4.3.41)\\
&= \|\boldsymbol{X}_{\text{true}}\|_* + \langle \boldsymbol{H}, \boldsymbol{U}\boldsymbol{V}^* \rangle + \|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* && (4.3.42)\\
&\geq \|\boldsymbol{X}_{\text{true}}\|_* - \|\mathcal{P}_T[\boldsymbol{H}]\|_* + \|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* . && (4.3.43)
\end{aligned}$$

So, if a better solution than $\boldsymbol{X}_{\text{true}}$ exists, the feasible perturbation $\boldsymbol{H}$ must satisfy the following cone restriction:

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* \leq \|\mathcal{P}_T[\boldsymbol{H}]\|_* . \tag{4.3.44}$$

Matrix Restricted Strong Convexity Property.

As in the proof of ℓ^1 success, we want to show that feasible perturbations $\boldsymbol{H} \in \text{null}(\mathcal{A})$ must have $\|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* \gg \|\mathcal{P}_T[\boldsymbol{H}]\|_*$. This is true if the operator $\mathcal{A}$ satisfies the following (uniform) *matrix restricted strong convexity* property (RSC) property:

Definition 4.3.6 (Matrix Restricted Strong Convexity)**.** *The linear operator $\mathcal{A}$ satisfies the matrix* restricted strong convexity *(RSC) condition of rank r with constant C if for every support T of a matrix of rank k and for all nonzero $\boldsymbol{H}$ satisfying*

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* \leq C \cdot \|\mathcal{P}_T[\boldsymbol{H}]\|_* . \tag{4.3.45}$$

with some constant $C \geq 1$, we have

$$\|\mathcal{A}[\boldsymbol{H}]\|_2^2 > \mu \cdot \|\boldsymbol{H}\|_F^2 \tag{4.3.46}$$

for some constant $\mu > 0$.

The following theorem says that if $\mathcal{A}$ satisfies the rank-RIP, then it satisfies the matrix RSC:

Theorem 4.3.7 (Rank-RIP Implies Matrix RSC)**.** *If a linear operator $\mathcal{A}$ satisfies rank-RIP with $\delta_{4r} < \frac{1}{1+C\sqrt{2}}$, then $\mathcal{A}$ satisfies the matrix-RSC condition of rank r with constant C.*

Both the statement and proof of Theorem 4.3.7 parallel Theorem 3.3.10 for the ℓ^1 norm. Theorem 4.3.7 involves δ_{4r}, as opposed to δ_{2k} for k-sparse vectors. The bigger constant in $4r = r + 3r$ reflects the need to account for all three components of the singular value decomposition in the proof:

Proof. Using the parallelogram identity, similar to Lemma 3.3.9, it is not difficult to show that for any $\boldsymbol{Z}$, $\boldsymbol{Z}'$ such that $\boldsymbol{Z} \perp \boldsymbol{Z}'$, and $\operatorname{rank}(\boldsymbol{Z}) + \operatorname{rank}(\boldsymbol{Z}') \leq 4r$,

$$\left|\left\langle \mathcal{A}[\boldsymbol{Z}], \mathcal{A}[\boldsymbol{Z}'] \right\rangle\right| \leq \delta_{4r}(\mathcal{A}) \, \|\boldsymbol{Z}\|_F \, \left\|\boldsymbol{Z}'\right\|_F . \tag{4.3.47}$$

Let T denote the support subspace for some matrix of rank r. Take any $\boldsymbol{H}$ that satisfies the cone restriction $\|\mathcal{P}_{T^\perp}[\boldsymbol{Z}]\|_* \leq C \cdot \|\mathcal{P}_T \boldsymbol{Z}\|_*$, and write

$$\boldsymbol{H} = \mathcal{P}_T[\boldsymbol{H}] + \mathcal{P}_{T^\perp}[\boldsymbol{H}]. \tag{4.3.48}$$

Let $\boldsymbol{H}_T$ denote $\mathcal{P}_T[\boldsymbol{H}]$. For the second term, $\mathcal{P}_{T^\perp}[\boldsymbol{H}]$, write its compact singular value decomposition

$$\mathcal{P}_{T^\perp}[\boldsymbol{H}] = \sum_i \eta_i \boldsymbol{\phi}_i \boldsymbol{\zeta}_i^*, \tag{4.3.49}$$

where $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots$ are the left singular vectors, $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots$ the right singular vectors, and $\eta_1 \geq \eta_2 \geq \cdots > 0$ the singular values. From the variational characterization of the singular vectors, each $\boldsymbol{\phi}_i$ is orthogonal to the columns of $\boldsymbol{U}$, and each $\boldsymbol{\zeta}_i$ is orthogonal to the columns of $\boldsymbol{V}$. So, if we partition $\mathcal{P}_{T^\perp}[\boldsymbol{H}]$ as

$$\mathcal{P}_{T^\perp}[\boldsymbol{H}] \;=\; \underbrace{\sum_{i=1}^{r} \eta_i \boldsymbol{\phi}_i \boldsymbol{\zeta}_i^*}_{\doteq \boldsymbol{\Phi}_1} + \underbrace{\sum_{i=r+1}^{2r} \eta_i \boldsymbol{\phi}_i \boldsymbol{\zeta}_i^*}_{\doteq \boldsymbol{\Phi}_2} + \ldots, \tag{4.3.50}$$

we have $\boldsymbol{\Phi}_i \perp \boldsymbol{\Phi}_j$ for $i \neq j$, $\boldsymbol{\Phi}_i \perp \boldsymbol{H}_T$ for every T.

Since the singular values η_i are non-increasing, the largest singular value of the $(i+1)$-th block is bounded by the average of the singular values in the i-th block.

$$\forall i \geq 1, \quad \|\boldsymbol{\Phi}_{i+1}\| \leq \frac{\|\boldsymbol{\Phi}_i\|_*}{r}. \tag{4.3.51}$$

So, noting that as an element in T, we have $\operatorname{rank}(\boldsymbol{H}_T) \leq 2r$ and so $\operatorname{rank}(\boldsymbol{H}_T + \boldsymbol{\Phi}_1) \leq 3r$. Notice that

$$\mathcal{A}[\boldsymbol{H}_T] + \mathcal{A}[\boldsymbol{\Phi}_1] = \mathcal{A}[\boldsymbol{H}] - \mathcal{A}[\boldsymbol{\Phi}_2] - \mathcal{A}[\boldsymbol{\Phi}_3] - \cdots . \tag{4.3.52}$$

Then, very similar to the derivation of inequalities (3.3.32) in Theorem 3.3.10, and by applying the rank-RIP to matrices of rank bounded by at most $4r$, we have

$$\begin{aligned}
& (1-\delta_{4r}) \left\|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\right\|_F^2 \\
\leq\ & \langle \mathcal{A}[\boldsymbol{H}_T + \boldsymbol{\Phi}_1], \mathcal{A}[\boldsymbol{H}_T + \boldsymbol{\Phi}_1] \rangle \\
=\ & \langle \mathcal{A}[\boldsymbol{H}_T + \boldsymbol{\Phi}_1], \mathcal{A}[\boldsymbol{H}] - \mathcal{A}[\boldsymbol{\Phi}_2] - \mathcal{A}[\boldsymbol{\Phi}_3] - \ldots \rangle \\
\leq\ & \sum_{j\geq 2} |\langle \mathcal{A}[\boldsymbol{H}_T], \mathcal{A}[\boldsymbol{\Phi}_j] \rangle| + |\langle \mathcal{A}[\boldsymbol{\Phi}_1], \mathcal{A}[\boldsymbol{\Phi}_j] \rangle| + \langle \mathcal{A}[\boldsymbol{H}_T + \boldsymbol{\Phi}_1], \mathcal{A}[\boldsymbol{H}] \rangle \\
\leq\ & \delta_{4r} \left(\|\boldsymbol{H}_T\|_F + \|\boldsymbol{\Phi}_1\|_F\right) \sum_{j\geq 2} \|\boldsymbol{\Phi}_j\|_F + \|\mathcal{A}[\boldsymbol{H}_T + \boldsymbol{\Phi}_1]\|_2 \, \|\mathcal{A}[\boldsymbol{H}]\|_2 \\
\leq\ & \delta_{4r}\sqrt{2} \left\|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\right\|_F \sum_{j\geq 2} \|\boldsymbol{\Phi}_j\|_F + (1+\delta_{4r}) \left\|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\right\|_F \|\mathcal{A}[\boldsymbol{H}]\|_2 \\
\leq\ & \delta_{4r}\sqrt{2} \left\|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\right\|_F \frac{\|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_*}{\sqrt{r}} + (1+\delta_{4r}) \left\|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\right\|_F \|\mathcal{A}[\boldsymbol{H}]\|_2 .
\end{aligned}$$

Note that $\boldsymbol{H}$ is restricted by the cone condition (4.3.45), which leads to:

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* \leq C\sqrt{r} \left\|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\right\|_F . \tag{4.3.53}$$

Combining this with the previous inequality, we obtain:

$$\|\mathcal{A}[\boldsymbol{H}]\|_2 \geq \frac{1-\delta_{4r}(1+C\sqrt{2})}{1+\delta_{4r}} \left\|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\right\|_F . \tag{4.3.54}$$

Since the singular values η_i are non-increasing, the ith singular value in $\boldsymbol{\Phi}_2 + \boldsymbol{\Phi}_3 + \cdots$ is no larger than the mean of the first i singular values in $P_{T^\perp}[\boldsymbol{H}]$. So we have

$$\forall i \geq r+1,\ \eta_i \leq \|P_{T^\perp}[\boldsymbol{H}]\|_* / i. \tag{4.3.55}$$

This leads to

$$\begin{aligned}
\|\boldsymbol{\Phi}_2 + \boldsymbol{\Phi}_3 + \cdots\|_F^2 &= \sum_{i=r+1}^{\infty} \eta_i^2 && (4.3.56)\\
&\leq \|P_{T^\perp}[\boldsymbol{H}]\|_*^2 \sum_{i=r+1}^{\infty} \frac{1}{i^2} && (4.3.57)\\
&\leq \frac{\|P_{T^\perp}[\boldsymbol{H}]\|_*^2}{r} \leq \frac{C^2 \|\boldsymbol{H}_T\|_*^2}{r} && (4.3.58)\\
&\leq C^2 \|\boldsymbol{H}_T\|_F^2 \leq C^2 \|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\|_F^2 . && (4.3.59)
\end{aligned}$$

Since $\boldsymbol{\Phi}_i$ with $i \geq 2$ are orthogonal to $\boldsymbol{H}_T + \boldsymbol{\Phi}_1$, this gives us

$$\|\boldsymbol{H}\|_F^2 \leq (1 + C^2) \|\boldsymbol{H}_T + \boldsymbol{\Phi}_1\|_F^2 . \tag{4.3.60}$$

Combining this with the previous bound (4.3.54) on $\|\mathcal{A}[\boldsymbol{H}]\|_2$, we obtain

$$\|\mathcal{A}[\boldsymbol{H}]\|_2 \geq \frac{1 - \delta_{4r}(1 + C\sqrt{2})}{(1 + \delta_{4r})\sqrt{1 + C^2}} \|\boldsymbol{H}\|_F . \tag{4.3.61}$$

This concludes the proof. □

Note that for the nuclear norm minimization problem, the feasible perturbation $\boldsymbol{H}$ satisfies the cone restriction (4.3.44). Thus Theorem 4.3.5 is essentially a corollary to Theorem 4.3.7 with constant $C = 1$ for the cone restriction.

4.3.5 Rank-RIP of Random Measurements

Theorem 4.3.5 indicates that the rank-RIP implies a very strong conclusion: nuclear norm minimization exactly recovers low-rank matrices. Moreover the recovery is *uniform* in the sense that a single set of measurements $\mathcal{A}$ suffices to recover any sufficiently low-rank matrix $\boldsymbol{X}_{\text{true}}$. The remaining question is what measurement operators satisfy the rank-RIP?

Random Gaussian Measurements

A simple and natural choice is to consider the random Gaussian measurements:

$$\mathcal{A}[\boldsymbol{X}] = (\langle \boldsymbol{A}_1, \boldsymbol{X} \rangle, \ldots, \langle \boldsymbol{A}_m, \boldsymbol{X} \rangle), \tag{4.3.62}$$

where the entries of the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{R}^{n_1 \times n_2}$ are all i.i.d. Gaussian $\mathcal{N}(0, \frac{1}{m})$. This is equivalent to viewing $\mathcal{A}$ as an $m \times n_1 n_2$ matrix with entries $\mathcal{A}_{ij}$ sampled i.i.d. $\mathcal{N}(0, \frac{1}{m})$. We demonstrate that such random maps satisfy the rank-RIP with high probability, using ideas and techniques similar to the proof of the (regular) RIP of random Gaussian matrices in Section 3.4.2:

Theorem 4.3.8 (Rank-RIP of Gaussian Measurements)**.** *If the measurement operator $\mathcal{A}$ is a random Gaussian map with entries i.i.d. $\mathcal{N}(0, \frac{1}{m})$, then $\mathcal{A}$ satisfies the rank-RIP with constant $\delta_r(\mathcal{A}) \leq \delta$ with high probability, provided $m \geq Cr(n_1 + n_2) \times \delta^{-2} \log \delta^{-1}$, where $C > 0$ is a numerical constant.*

Proof. Let

$$S_r = \{\boldsymbol{X} \mid \operatorname{rank}(\boldsymbol{X}) \leq r, \|\boldsymbol{X}\|_F = 1\}.$$

Notice that $\delta_r(\mathcal{A}) \leq \delta$ if and only if

$$\sup_{\boldsymbol{X} \in S_r} | \langle \mathcal{A}(\boldsymbol{X}), \mathcal{A}(\boldsymbol{X}) \rangle - 1 | \leq \delta. \tag{4.3.63}$$

We complete the rest of the proof in three steps.

1. Constructing a Covering ϵ-Net for S_r.

Notice that for any rank-r matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, it can be represented by its SVD; $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. So to construct a covering of all rank-r matrices, we can try to construct a covering for each of the terms $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{\Sigma}$, respectively.

Lemma 4.3.9. *There is a covering ε-net N_U for the $H = \{\boldsymbol{U} \in \mathbb{R}^{n_1 \times r} | \boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}\}$, of size $|N_U| \leq (6/\varepsilon)^{n_1 r}$.*

Proof. Let N' be an $\varepsilon/2$-net for $\{\boldsymbol{U} \in \mathbb{R}^{n_1 \times r} | \|\boldsymbol{U}\|_F^2 = 1\}$ of size $|N'| \leq (6/\varepsilon)^{n_1 r}$ (from Lemma 3.4.5). Let

$$Q = \{\boldsymbol{U}' \in N' | \exists \boldsymbol{U} \in H \text{ with } \|\boldsymbol{U} - \boldsymbol{U}'\|_F \leq \varepsilon/2\}.$$

For each $\boldsymbol{U}' \in Q$, let $\hat{\boldsymbol{U}}(\boldsymbol{U}')$ be the nearest element of H. Set $N_U = \{\hat{\boldsymbol{U}}(\boldsymbol{U}') | \boldsymbol{U}' \in Q\} \subseteq H$. By the triangle inequality, N_U is an ε-net for H. □

Similarly, one can construct an ε-net N_V for $H' = \{\boldsymbol{V} \in \mathbb{R}^{n_2 \times r} | \boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}\}$ of size $|N_V| \leq (6/\varepsilon)^{n_2 r}$. With this lemma, we have the following result.

Lemma 4.3.10. *There is a covering ε-net N_r for the set S_r, of size $|N_r| \leq \exp\big((n_1 + n_2)r \log(18/\varepsilon) + r \log(9/\varepsilon)\big)$.*

Proof. Choose covering $\varepsilon/3$-nets N_U and N_V for H and H', respectively. According to the above lemma, the sizes of the nets can be less than $(18/\varepsilon)^{n_1 r}$ and $(18/\varepsilon)^{n_2 r}$, respectively. Form a covering $\varepsilon/3$-net N_Σ for

$$D = \{\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r} | \boldsymbol{\Sigma} \text{ diagonal}, \|\boldsymbol{\Sigma}\|_F = 1\}.$$

According to Lemma 3.4.5, the size of the net can be less than $|N_\Sigma| \leq (9/\varepsilon)^r$.

Now consider the following net for the whole set S_r:

$$N_r = \{\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* | \boldsymbol{U} \in N_U, \boldsymbol{\Sigma} \in N_S, \boldsymbol{V} \in N_V\}.$$

Its size is bounded by the product of all three nets, hence the expression in the Lemma. Now we only have to show that this is indeed a covering ε-net for S_r. For any given $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, we can find $\hat{\boldsymbol{X}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^* \in N_r$ with $\|\boldsymbol{U} - \hat{\boldsymbol{U}}\|_F \leq \varepsilon/3$, $\|\boldsymbol{V} - \hat{\boldsymbol{V}}\|_F \leq \varepsilon/3$, and $\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon/3$.

The triangle inequality gives

$$\begin{aligned} & \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F \\ \leq\ & \|\boldsymbol{U} - \hat{\boldsymbol{U}}\|_F \|\boldsymbol{\Sigma}\boldsymbol{V}^*\| + \|\hat{\boldsymbol{U}}\| \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F \|\boldsymbol{V}^*\| + \|\hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\| \|\boldsymbol{V}^* - \hat{\boldsymbol{V}}^*\|_F \\ \leq\ & \varepsilon, \end{aligned}$$

where we have used that each of the approximation errors is bounded by $\varepsilon/3$, and each of the operator norms in the above expression are bounded by 1. □

2. Discretization.

As in the ℓ^1 case for sparse signals in Section 3.4.2, the goal of discretization is trying to show that if $\mathcal{A}$ is restricted isometric on the finite set of (discrete) points in the covering net N_r with a constant δ_{N_r}, so is $\mathcal{A}$ on the whole set S_r, with a constant δ_r possibly slightly larger than δ_{N_r}.

Now consider a point $\boldsymbol{X}$ in S_r and its closest point $\hat{\boldsymbol{X}}$ in N_r. Thus, we have $\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F \leq \varepsilon$. Also we have[8]

$$\begin{aligned} & |\langle \mathcal{A}[\boldsymbol{X}], \mathcal{A}[\boldsymbol{X}]\rangle - \langle \mathcal{A}[\hat{\boldsymbol{X}}], \mathcal{A}[\hat{\boldsymbol{X}}]\rangle| \\ = \; & |\langle \mathcal{A}[\boldsymbol{X}], \mathcal{A}[\boldsymbol{X} - \hat{\boldsymbol{X}}\boldsymbol{P}_V]\rangle + \langle \mathcal{A}[\boldsymbol{X} - \boldsymbol{P}_{\hat{U}}\boldsymbol{X}], \mathcal{A}[\hat{\boldsymbol{X}}\boldsymbol{P}_V]\rangle \\ & + \langle \mathcal{A}[\boldsymbol{P}_{\hat{U}}\boldsymbol{X} - \hat{\boldsymbol{X}}], \mathcal{A}[\hat{\boldsymbol{X}}\boldsymbol{P}_V]\rangle + \langle \mathcal{A}[\hat{\boldsymbol{X}}], \mathcal{A}[\hat{\boldsymbol{X}}\boldsymbol{P}_V - \hat{\boldsymbol{X}}]\rangle|. \end{aligned}$$

To bound the first term in the above expression, notice that

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\boldsymbol{P}_V\|_F = \|(\boldsymbol{X} - \hat{\boldsymbol{X}})\boldsymbol{P}_V\|_F \leq \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F \leq \varepsilon.$$

Also, $\boldsymbol{X} - \hat{\boldsymbol{X}}\boldsymbol{P}_V$ is of rank r. So we have

$$|\langle \mathcal{A}[\boldsymbol{X}], \mathcal{A}[\boldsymbol{X} - \hat{\boldsymbol{X}}\boldsymbol{P}_V]\rangle| \leq (1 + \delta_r(\mathcal{A}))\varepsilon.$$

For the second term, since $\boldsymbol{P}_{\hat{U}}$ is an orthogonal projection onto the space of matrices whose columns are the same as $\hat{\boldsymbol{X}}$, we have

$$\|\boldsymbol{X} - \boldsymbol{P}_{\hat{U}}\boldsymbol{X}\|_F \leq \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F \leq \varepsilon.$$

Also, since $\boldsymbol{X}$ and $\boldsymbol{P}_{\hat{U}}\boldsymbol{X}$ have the same row space, so $\boldsymbol{X} - \boldsymbol{P}_{\hat{U}}\boldsymbol{X}$ is of rank r or less. Therefore, we also have

$$|\langle \mathcal{A}[\boldsymbol{X} - \boldsymbol{P}_{\hat{U}}\boldsymbol{X}], \mathcal{A}[\hat{\boldsymbol{X}}\boldsymbol{P}_V]\rangle| \leq (1 + \delta_r(\mathcal{A}))\varepsilon.$$

Similarly for the third and fourth terms, each is bounded by the same bound. Therefore, we get

$$|\langle \mathcal{A}[\boldsymbol{X}], \mathcal{A}[\boldsymbol{X}]\rangle - \langle \mathcal{A}[\hat{\boldsymbol{X}}], \mathcal{A}[\hat{\boldsymbol{X}}]\rangle| \leq 4(1 + \delta_r(\mathcal{A}))\varepsilon.$$

From this we have

$$\delta_r(\mathcal{A}) - \delta_{N_r} \leq 4(1 + \delta_r(\mathcal{A}))\varepsilon. \tag{4.3.64}$$

This gives

$$\delta_r(\mathcal{A}) \leq \frac{4\varepsilon + \delta_{N_r}}{1 - 4\varepsilon}. \tag{4.3.65}$$

3. Union Bound.

For each $\boldsymbol{X} \in N_r$, $\mathcal{A}[\boldsymbol{X}] \in \mathbb{R}^m$ is a random vector with entries independent $\mathcal{N}(0, 1/m)$. We have

$$\mathbb{P}\left[\left|\|\mathcal{A}[\boldsymbol{X}]\|_2^2 - 1\right| > t\right] \;\leq\; 2\exp(-mt^2/8). \tag{4.3.66}$$

[8]Notice that here the derivation is more subtle than the ℓ^1 case because $\boldsymbol{X} - \hat{\boldsymbol{X}}$ is not necessarily of rank r.

Hence, summing the probabilities over all elements of N_r, we have

$$\begin{aligned}\mathbb{P}\left[\delta_{N_r} > t\right] &\leq 2\left|N_r\right| \exp\left(-mt^2/8\right) \\ &= 2\exp\left(-\frac{mt^2}{8} + (n_1+n_2)r\log(18/\varepsilon) + r\log(9/\varepsilon)\right).\end{aligned}$$

If we choose $\varepsilon = c\cdot\delta$ and $t = c\cdot\delta$ for some small constant c and ensure $m \geq Cr(n_1+n_2)\delta^{-2}\log\delta^{-1}$ for some large enough C, the above failure probability is bounded by $2\exp(-c'm\delta^2)$. On the complement of this "failure" event, $\delta_{N_r} \leq c\cdot\delta$, and due to (4.3.65) we have $\delta_r(\mathcal{A}) \leq \delta$. This concludes the proof of the Theorem 4.3.8. □

The number of measurements $m = O(r(n_1+n_2))$ required is nearly optimal, since an $n_1 \times n_2$ rank-r matrix has $r(n_1+n_2-r)$ degrees of freedom. Of course, the big O notation hides a numerical constant. Like the ℓ^1 minimization for sparse recovery, when the dimension is high, nuclear norm minimization exhibits a phase transition between success and failure. Identifying this transition yields more precise estimates of the number m of measurements required to reconstruct a low rank matrix. We discuss this issue in more detail below.

Random Submatrix of a Unitary Basis

Although random Gaussian measurements have very nice properties such as (rank) RIP, the lack of structure in such measurements makes it rather expensive to generate, store and apply such operators in practice. Hence it is natural to ask if there exist other more structured measurements that have similarly good RIP properties. In Section 3.4.3, we have seen that given any unitary matrix that is incoherent from sparse signals, then a randomly selected subset of its rows will satisfy the RIP with high probability. An important special case that has been widely used in practice for compressive sensing is a randomly chosen submatrix of the discrete Fourier transform basis. It is then natural to ask what are the Fourier-type bases for matrices.

In the case of sparse recovery, we start with a unitary basis $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ and show that if the rows $\{\boldsymbol{u}_i\}_{i=1}^n$ of the basis is incoherent with sparse signals:

$$\forall i \quad \|\boldsymbol{u}_i\|_\infty = \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_0=1} \langle \boldsymbol{u}_i, \boldsymbol{x}\rangle \leq \zeta/\sqrt{n}$$

for some constant ζ, then a randomly selected (sufficient) number of rows of $\boldsymbol{U}$ will satisfy RIP.

To simplify the discussion of matrices, we will assume $n_1 = n_2 = n$ for the rest of this subsection; a similar approach applies when $n_1 \neq n_2$. Let us assume $\{\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_{n^2}\} \subset \mathbb{C}^{n\times n}$ form a unitary basis for the matrix space $\mathbb{C}^{n\times n}$. Similarly we want each of the matrix $\boldsymbol{U}_i$ to be incoherent with

low-rank matrices. Note that for any $\boldsymbol{X} \in \mathbb{C}^{n\times n}$,

$$\|\boldsymbol{U}_i\| = \sup_{\boldsymbol{X}:\|\boldsymbol{X}\|_2=1,\mathrm{rank}(\boldsymbol{X})=1} \langle \boldsymbol{U}_i, \boldsymbol{X}\rangle. \tag{4.3.67}$$

Hence in order for each $\boldsymbol{U}_i$ to be incoherent with low-rank matrices, we could require:

$$\forall i \quad \|\boldsymbol{U}_i\| \le \zeta/\sqrt{n}. \tag{4.3.68}$$

Then to construct the measurement operator $\mathcal{A}$, we randomly select a subset of m bases from $\{\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_{n^2}\}$ and properly scale them as:[9]

$$\mathcal{A}: \quad \boldsymbol{A}_i = \frac{n}{\sqrt{m}}\boldsymbol{U}_i, \quad i = 1, \ldots, m. \tag{4.3.69}$$

Then one should expect that when m is large enough, with high probability, the so-defined $\mathcal{A}$ satisfies the rank-RIP. The following theorem makes this precise:

Theorem 4.3.11. *Let us assume $\{\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_{n^2}\} \subset \mathbb{C}^{n\times n}$ be a unitary basis for the matrix space $\mathbb{C}^{n\times n}$ and with $\|\boldsymbol{U}_i\| \le \zeta/\sqrt{n}$ for some constant ζ. Let $\mathcal{A}$ to be defined as per (4.3.69). Then if*

$$m \ge C\zeta^2 \cdot rn \log^6 n, \tag{4.3.70}$$

then with high probability, $\mathcal{A}$ satisfies the rank-RIP over the set of all rank-r matrices.

The proof of this theorem is out of the scope of this book and interested readers may refer to the work of [Liu, 2011].

According to this statement, from an incoherent unitary basis, with high probability we could find a (compressive) sensing operator $\mathcal{A}$ such that it is rank-RIP. Hence with this operator, one can recover all rank-r matrices via the nuclear norm minimization. The remaining question is what type of structured bases (of the matrix space) are rank-incoherent as per (4.3.68)? To this end, one should seek a matrix analogue to the Fourier basis.

In the case of MRI imaging, we have seen that measurements that one can physically take are essentially the Fourier coefficients of the brain image. As it turns out, the matrix analogue to Fourier basis also has a natural origin from physics. In quantum-state tomography, a system of k qubits is of dimension $n = 2^k$. The quantum state of such a system is described by a density matrix $\boldsymbol{X}_0 \in \mathbb{C}^{n\times n}$ which is positive semidefinite with trace 1. When the state is early pure, $\boldsymbol{X}_0$ is a very low-rank matrix with $\mathrm{rank}\,(\boldsymbol{X}_0) = r \ll n$.

One problem in quantum physics is how to recover the quantum state $\boldsymbol{X}_0$ of a system from linear measurements. As it turns out, a set of experimentally feasible measurements are given by the so-called *Pauli observables*.

[9] The scaling is to ensure that the "column" of $\mathcal{A}$ to be of unit norm.

Each Pauli measurement is given by the inner product of $\boldsymbol{X}_0$ with matrices of the form $\boldsymbol{P}_1 \otimes \cdots \otimes \boldsymbol{P}_k$ where $\otimes$ is the tensor (Kronecker) product and each $\boldsymbol{P}_i = \frac{1}{\sqrt{2}}\boldsymbol{\sigma}$ where $\boldsymbol{\sigma}$ is a 2×2 matrix chosen from the following four possibilities:

$$\boldsymbol{\sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\sigma}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{\sigma}_3 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \boldsymbol{\sigma}_4 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

It is easy to see that there are a total of 4^k possible choices for the tensor product, denoted as $\{\boldsymbol{U}_i\}_{i=1}^{4^k}$ and they together form an orthonormal basis for the matrix space $\mathbb{C}^{n\times n}$ where $n = 2^k$.

One can show that for each basis $\boldsymbol{U}_i = \boldsymbol{P}_1 \otimes \cdots \otimes \boldsymbol{P}_k$, its operator norm is bounded as $\|\boldsymbol{U}_i\| \leq 1/\sqrt{n}$ hence incoherent with low-rank matrices. Then according to Theorem 4.3.11, a randomly selected $m \geq Crn \log^6 n$ rows of the Pauli bases will satisfy the rank-RIP property with high probability. Hence, such a sensing operator will be able to uniformly recover all pure quantum states less than rank r.

4.3.6 Noise, Inexact Low Rank, and Phase Transition

Above, we established that, under fairly broad conditions, nuclear norm minimization correctly recovers a low-rank matrix $\boldsymbol{X}_{\text{true}}$ from ideal measurements $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}]$. In practice, the measurements can be corrupted by noise or measurement errors. In some cases, $\boldsymbol{X}_{\text{true}}$ may not be exactly low rank. It is desirable to understand whether nuclear norm minimization still gives reasonably good estimates of $\boldsymbol{X}_{\text{true}}$ in these situations.

In Section 3.5, we established that ℓ^1 minimization accurately estimates sparse signals under deterministic noise, random noise, and even inexact sparsity. As we will see in this section, essentially the same analysis and results generalize to the case of nuclear norm minimization for recovering low-rank matrices.

Deterministic Noise.

Here we still assume the matrix $\boldsymbol{X}_{\text{true}}$ is perfectly low-rank, but the measurements $\boldsymbol{y}$ is corrupted by small additive noise:

$$\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}] + \boldsymbol{z}, \qquad \|\boldsymbol{z}\|_2 \leq \varepsilon. \tag{4.3.71}$$

Similar to Theorem 3.5.1, for recovering low-rank matrices with (deterministic) noise, we have the following result.

Theorem 4.3.12 (Stable Low-rank Recovery via BPDN)**.** *Suppose that* $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}] + \boldsymbol{z}$*, with* $\|\boldsymbol{z}\|_2 \leq \varepsilon$*, and let* $\operatorname{rank}(\boldsymbol{X}_{\text{true}}) = r$*. If* $\delta_{4r}(\mathcal{A}) < \sqrt{2} - 1$*, then any solution* $\hat{\boldsymbol{X}}$ *to the optimization problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \|\mathcal{A}[\boldsymbol{X}] - \boldsymbol{y}\|_2 \leq \varepsilon. \end{array} \tag{4.3.72}$$

satisfies

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}_{\text{true}}\right\|_F \leq C\varepsilon. \tag{4.3.73}$$

Here, C is a numerical constant.

Proof. The proof of this theorem parallels that for Theorem 3.5.1 and we leave the details for the reader as an exercise (see Exercise 4.15). □

Random Noise.

Now let us consider the case when the noise in the above measurement model (4.3.71) is random (Gaussian):

$$\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}] + \boldsymbol{z}, \tag{4.3.74}$$

where entries of $\boldsymbol{z}$ are random i.i.d. Gaussian $\mathcal{N}(0, \frac{\sigma^2}{m})$. Then we have the following theorem that parallels Theorem 3.5.3 for the ℓ^1 case.

Theorem 4.3.13 (Stable Sparse Recovery via Lasso)**.** *Suppose that $\mathcal{A} \sim_{\text{iid}} \mathcal{N}(0, \frac{1}{m})$, and $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}] + \boldsymbol{z}$, with $\boldsymbol{X}_{\text{true}}$ of rank r and $\boldsymbol{z} \sim_{\text{iid}} \mathcal{N}(0, \frac{\sigma^2}{m})$. Solve the matrix Lasso*

$$\text{minimize} \;\; \tfrac{1}{2}\left\|\boldsymbol{y} - \mathcal{A}[\boldsymbol{X}]\right\|_2^2 + \lambda_m \left\|\boldsymbol{X}\right\|_*, \tag{4.3.75}$$

with regularization parameter $\lambda = c \cdot 2\sigma\sqrt{\frac{(n_1+n_2)}{m}}$ for a large enough c. Then with high probability,

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}_{\text{true}}\right\|_F \;\leq\; C'\sigma\sqrt{\frac{r(n_1+n_2)}{m}}. \tag{4.3.76}$$

Notice in contrast to deterministic noise, random noise leads to a much favorable scaling $\sqrt{\frac{r(n_1+n_2)}{m}}$ in the estimation error. In a typical compressive sensing setting (as suggested by Theorem 4.3.8), the sampling dimension m needs to be at least $C \cdot r(n_1 + n_2)$ for some large constant C. Hence the scaling factor is proportional to $1/\sqrt{C}$ and it becomes small when C is large.

Proof. The overall proof strategy is quite similar to that of Theorem 3.5.3 for the stability of Lasso estimate. We will lay out the key places that are different from the ℓ^1 case and leave the details to the reader an exercise.

In the proof of Lemma 3.5.2, we see that in order to establish the cone condition for the Lasso type minimization, one of the key steps is to bound $|\langle \boldsymbol{A}^T\boldsymbol{z}, \boldsymbol{h}\rangle|$ via

$$|\langle \boldsymbol{A}^T\boldsymbol{z}, \boldsymbol{h}\rangle| \leq \left\|\boldsymbol{A}^T\boldsymbol{z}\right\|_\infty \|\boldsymbol{h}\|_1.$$

Following similar arguments, in the matrix Lasso case here, we need to bound $|\langle \mathcal{A}^*\boldsymbol{z}, \boldsymbol{H}\rangle|$ instead as

$$|\langle \mathcal{A}^*\boldsymbol{z}, \boldsymbol{H}\rangle| \leq \|\mathcal{A}^*\boldsymbol{z}\| \, \|\boldsymbol{H}\|_*,$$

where $\|\mathcal{A}^*\boldsymbol{z}\|$ is the operator norm (largest singular value) of the matrix $\mathcal{A}^*\boldsymbol{z} = \sum_{i=1}^m z_i \boldsymbol{A}_i$. To this end, we need to provide a tight bound for the operator norm of $\mathcal{A}^*\boldsymbol{z}$.

Notice that

$$M \doteq \left\| \sum_{i=1}^m z_i \boldsymbol{A}_i \right\| = \sup_{\boldsymbol{u}\in\mathbb{S}^{n_1-1}, v\in\mathbb{S}^{n_2-1}} \boldsymbol{u}^T \sum_{i=1}^m z_i \boldsymbol{A}_i \boldsymbol{v} \tag{4.3.77}$$

$$= \sup_{\boldsymbol{u}\in\mathbb{S}^{n_1-1}, \boldsymbol{v}\in\mathbb{S}^{n_2-1}} \langle \boldsymbol{z}, \mathcal{A}[\boldsymbol{u}\boldsymbol{v}^T]\rangle. \tag{4.3.78}$$

The $\boldsymbol{u}_\star$ and $\boldsymbol{v}_\star$ that achieve the maximum value in (4.3.78) depend on $\boldsymbol{z}$ and $\mathcal{A}$. So in order to eliminate this dependency and provide a bound for $\|\sum_{i=1}^m z_i \boldsymbol{A}_i\|$, we cover the two spheres $\mathbb{S}^{n_1-1}$ and $\mathbb{S}^{n_2-1}$ with two ε-nets N_1 and N_2 respectively. According to Lemma 3.4.5, the sizes of the nets can be less than $(3/\varepsilon)^{n_1}$ and $(3/\varepsilon)^{n_2}$ respectively.

Let us denote

$$M_N \doteq \sup_{\boldsymbol{u}\in N_1, \boldsymbol{v}\in N_2} \boldsymbol{u}^T \sum_{i=1}^m z_i \boldsymbol{A}_i \boldsymbol{v},$$

and then it is easy to show that[10]

$$M \le \frac{M_N}{1-2\varepsilon}. \tag{4.3.79}$$

Notice that given any $\boldsymbol{u} \in N_1, \boldsymbol{v} \in N_2$, $\langle \boldsymbol{z}, \mathcal{A}[\boldsymbol{u}\boldsymbol{v}^T]\rangle$ is a Gaussian variable of distribution $\mathcal{N}(0, \|\mathcal{A}[\boldsymbol{u}\boldsymbol{v}^T]\|_2^2(\sigma^2/m))$. Since $\mathcal{A}$ is rank-RIP and uv^T is a rank-1 matrix of unit Frobenius norm, we have

$$\left\|\mathcal{A}[\boldsymbol{u}\boldsymbol{v}^T]\right\|_2^2 \le (1+\delta) \le 2. \tag{4.3.80}$$

Thus, we have

$$\mathbb{P}\left[\left|\boldsymbol{u}^T \sum_{i=1}^m z_i \boldsymbol{A}_i \boldsymbol{v}\right| > t\right] \le 2\exp\left(-\frac{mt^2}{4\sigma^2}\right). \tag{4.3.81}$$

Apply the union bound on all possible pairs of $(\boldsymbol{u}, \boldsymbol{v})$ from the two nets and choose $t = \alpha\sigma\sqrt{\frac{n_1+n_2}{m}}$ for some large enough α, then we have $M_N > t$ with diminishing probability as n_1 or n_2 becomes large. Therefore, we have

$$M = \left\| \sum_{i=1}^m z_i \boldsymbol{A}_i \right\| \le \beta\sigma\sqrt{\frac{n_1+n_2}{m}} \tag{4.3.82}$$

for some constant β with high probability.

Now, similar to the proof of Lemma 3.5.2, if we choose λ_m to be in the order of $O\big(\sigma\sqrt{\frac{n_1+n_2}{m}}\big)$, then the feasible perturbation $\boldsymbol{H}$ satisfies the

[10] We leave the details of proving this inequality to the reader as an exercise.

cone restriction. Since $\mathcal{A}$ is rank-RIP, it implies that $\mathcal{A}$ satisfies the matrix restricted strong convexity (RSC) property. That leads to the bound on the estimation error:

$$\|\boldsymbol{H}\|_F = \left\|\hat{\boldsymbol{X}} - \boldsymbol{X}_{\text{true}}\right\|_F \;\leq\; C'\sigma\sqrt{\frac{r(n_1+n_2)}{m}}. \tag{4.3.83}$$

The details of the proof for this follows essentially the same steps as those in the proof of Theorem 3.5.3 for the ℓ^1 case. We leave those to the reader as an exercise (see Exercise 4.17.) □

The error bound given in the above theorem can actually be shown to be nearly optimal as it is close to the best error that can be achieved by any estimator over all rank-r matrices. The following theorem, due to [Candes and Plan, 2011] makes this precise:

Theorem 4.3.14. *Suppose that* $\mathcal{A} \sim_{\text{iid}} \mathcal{N}(0, \frac{1}{m})$ *and we observe* $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}] + \boldsymbol{z}$ *where entries of* $\boldsymbol{z}$ *are i.i.d.* $\mathcal{N}(0, \frac{\sigma^2}{m})$ *random variables. Set*

$$M^\star(\mathcal{A}) = \inf_{\hat{\boldsymbol{X}}(\boldsymbol{y})} \sup_{\text{rank}(\boldsymbol{X})\leq r} \mathbb{E}\left\|\hat{\boldsymbol{X}}(\boldsymbol{y}) - \boldsymbol{X}\right\|_F^2. \tag{4.3.84}$$

Then we have

$$M^\star(\mathcal{A}) \;\geq\; c\sigma^2\frac{rn}{m}, \tag{4.3.85}$$

for $n = \max\{n_1, n_2\}$*, where* $c > 0$ *is a numerical constant.*

The proof of this theorem is beyond the scope of this book; we refer interested readers to [Candes and Plan, 2011] for a proof. According to Theorem 4.3.13, the worst error of the matrix Lasso matches the best achievable by any estimator, up to constants.

Inexact Low-rank Matrices.

In the case when $\boldsymbol{X}_{\text{true}}$ is not exactly low-rank, let $[\boldsymbol{X}_{\text{true}}]_r$ be the best rank-r approximation of $\boldsymbol{X}_{\text{true}}$. We can rewrite the observation model

$$\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}] + \boldsymbol{z}, \quad \|\boldsymbol{z}\|_2 \leq \varepsilon. \tag{4.3.86}$$

as:

$$\boldsymbol{y} = \mathcal{A}\left[[\boldsymbol{X}_{\text{true}}]_r\right] + \mathcal{A}\left[\boldsymbol{X}_{\text{true}} - [\boldsymbol{X}_{\text{true}}]_r\right] + \boldsymbol{z}, \quad \|\boldsymbol{z}\|_2 \leq \varepsilon.$$

Theorem 4.3.15. *Let* $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}_{\text{true}}] + \boldsymbol{z}$*, with* $\|\boldsymbol{z}\|_2 \leq \varepsilon$*. Let* $\hat{\boldsymbol{X}}$ *solve the denoising problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \|\boldsymbol{y} - \mathcal{A}[\boldsymbol{X}]\|_2 \leq \varepsilon. \end{array} \tag{4.3.87}$$

Then for any r *such that* $\delta_{4r}(\boldsymbol{A}) \;<\; \sqrt{2} - 1$*,*

$$\left\|\hat{\boldsymbol{X}} - \boldsymbol{X}_{\text{true}}\right\|_2 \;\leq\; C\frac{\|\boldsymbol{X}_{\text{true}} - [\boldsymbol{X}_{\text{true}}]_r\|_*}{\sqrt{r}} + C'\varepsilon \tag{4.3.88}$$

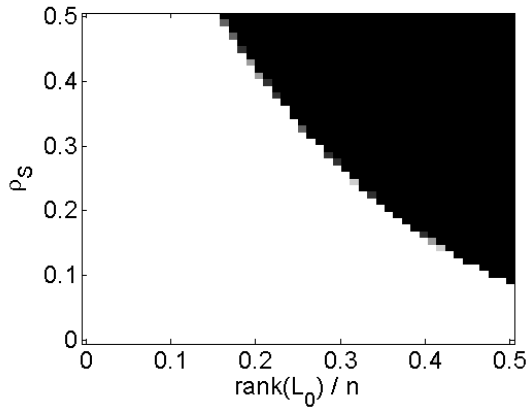

Figure 4.6. **Phase transitions in Low Rank Matrix Recovery.** We plot the probability of successfully recovering an $n \times n$ matrix from Gaussian measurements. Vertical axis: undersampling factor $1 - m/n^2$. Horizontal axis rank r/n. The success of nuclear norm minimization exhibits a very sharp transition from success to failure.

for some constants C and C'.

Proof. The proof of this theorem parallels that for Theorem 3.5.5 for the inexact sparse recovery problem. We here only setup some analogous concepts and key ideas that allow us to extend that proof to the matrix case here. But we leave details of the proof as an exercise to the reader.

Let $\boldsymbol{X}_{\text{true}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ denote the compact SVD of the true solution $\boldsymbol{X}_{\text{true}}$. Then its best rank-$r$ approximation is $[\boldsymbol{X}_{\text{true}}]_r = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^*$. Now let

$$T \doteq \left\{\boldsymbol{U}_r\boldsymbol{W}^* + \boldsymbol{Q}\boldsymbol{V}_r^* \mid \boldsymbol{W} \in \mathbb{R}^{n_2\times r},\ \boldsymbol{Q} \in \mathbb{R}^{n_1\times r}\right\} \subseteq \mathbb{R}^{n_1\times n_2}. \tag{4.3.89}$$

Show that in the inexact low-rank case, instead of the cone restriction (4.3.44), we have the following restriction for the feasible perturbation $\boldsymbol{H} = \hat{\boldsymbol{X}} - \boldsymbol{X}_{\text{true}}$:

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* \le \|\mathcal{P}_T[\boldsymbol{H}]\|_* + 2\,\|\mathcal{P}_{T^\perp}[\boldsymbol{X}_{\text{true}}]\|_*\,. \tag{4.3.90}$$

Notice that $\mathcal{P}_{T^\perp}[\boldsymbol{X}_{\text{true}}] = \boldsymbol{X}_{\text{true}} - [\boldsymbol{X}_{\text{true}}]_r$. Then, similar to the proof of Theorem 3.5.5, simply carry the extra term $2\,\|\mathcal{P}_{T^\perp}[\boldsymbol{X}_{\text{true}}]\|_*$ at the places in the proof of Theorem 4.3.7 where the cone restriction is applied. One can reach the conclusion of the theorem. We leave details of the proof as an exercise to the reader (see Exercise 4.16). □

Phase Transition for Low Rank Matrix Recovery.

Thus far, we have seen strong parallels between sparse vector recovery using ℓ^1 norm minimization and low-rank matrix recovery using nuclear norm minimization. In both cases, we saw how an appropriate notion of

restricted isometry property could be used to guarantee exact recovery from a near-minimal number of random measurements – about $k \log(n/k)$ for k-sparse vectors, and about nr for rank-r matrices. However, just like in the sparse vector case, this tool does not yield sharp constants.

In fact, there is a phase transition phenomenon for low-rank recovery, which mirrors that for sparse recovery: as the dimension grows, the transition between success and failure in low-rank recovery becomes increasingly sharp. Figure 4.6 illustrates this.

Just as we did for sparse recovery, we can use the "coefficient space" geometry of the low-rank recovery problem to derive very sharp estimates of this transition. This geometry is phrased in terms of the descent cone D of the nuclear norm at the target solution $\boldsymbol{X}_{\text{true}}$:

$$\mathsf{D} = \{\boldsymbol{H} \mid \|\boldsymbol{X}_{\text{true}} + \boldsymbol{H}\|_* \leq \|\boldsymbol{X}_{\text{true}}\|_*\}\,. \tag{4.3.91}$$

As for sparse recovery, $\boldsymbol{X}_{\text{true}}$ is the unique optimal solution to the nuclear norm minimization problem if and only if $\mathsf{D} \cap \mathsf{null}(\mathcal{A}) = \{\boldsymbol{0}\}$. Hence, quantifying the probability of success under a random linear projection becomes equivalent to quantifying the probability that the two convex cones D and $\mathsf{null}(\mathcal{A})$ have only trivial intersection. Deploying Theorem 3.6.4, we find that there is a sharp transition between success and failure around

$$m^\star \sim \delta(\mathsf{D}), \tag{4.3.92}$$

the statistical dimension of the descent cone. Moreover, the theorem tells us that the width of the transition region is roughly $O(\sqrt{n_1 n_2})$. The *location* of the transition region can be characterized using the same machinery that we deployed in Section 3.6 to estimate the statistical dimension of the descent cone of the ℓ^1 norm. This machinery involves estimating the expected squared distance of a random vector (here, random matrix) to the polar cone, which is spanned by the subdifferential of the nuclear norm. For convenience, for a matrix $\boldsymbol{M}$ with singular value decomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, let us define the *singular value thresholding operator* as

$$\mathcal{D}_\tau[\boldsymbol{M}] \doteq \boldsymbol{U}\mathcal{S}_\tau[\boldsymbol{\Sigma}]\boldsymbol{V}^*, \tag{4.3.93}$$

where $\mathcal{S}_\tau[\cdot]$ is the entry-wise *soft thresholding* operator:

$$\forall \boldsymbol{X}, \quad \mathcal{S}_\tau[\boldsymbol{X}] = \operatorname{sign}(\boldsymbol{X}) \circ (|\boldsymbol{X}| - \tau)_+,$$

where $\circ$ is the entry-wise (Hadamard) product of two matrices. An intermediate result produced by these calculations is as follows:

Theorem 4.3.16 (Phase Transition in Low-rank Recovery)**.** *Let* D *denote the descent cone of the nuclear norm at any matrix* $\boldsymbol{X}_{\text{true}} \in \mathbb{R}^{n_1 \times n_2}$ *of rank* r*. Let* $\boldsymbol{G}$ *be an* $(n_1 - r) \times (n_2 - r)$ *matrix with entries i.i.d.* $\mathcal{N}(0, 1)$*. Set*

$$\psi(n_1, n_2, r) = \inf_{\tau \geq 0} \left\{ r(n_1 + n_2 - r + \tau^2) + \mathbb{E}_{\boldsymbol{G}} \left[\|\mathcal{D}_\tau[\boldsymbol{G}]\|_F^2 \right] \right\}. \tag{4.3.94}$$

Then

$$\psi(n_1, n_2, r) - 2\sqrt{n_2/r} \;\leq\; \delta(\mathsf{D}) \;\leq\; \psi(n_1, n_2, r). \tag{4.3.95}$$

This theorem identifies a sharp transition in low-rank recovery. It is possible to use asymptotic results on the limiting distribution of the singular values of a random matrix to give a formula for $\psi(n_1, n_2, r)/(n_1 n_2)$, which is valid when $n_1 \to \infty$, $n_1/n_2 \to \alpha \in (0, \infty)$ and $r/n_1 \to \rho \in (0, 1)$. In the exercises, we guide the interested reader through this derivation. Here, we merely display the result of this calculation in Figure 4.6, and note the excellent agreement between this theoretical prediction and numerical experiment: *for the idealized setting of "generic" measurements, we have a very precise prediction of the phase transition!*

4.4 Low Rank Matrix Completion

We have seen how concepts from sparse recovery transpose directly to the low-rank recovery problem. The concept of sparsity had a natural analogue in the concept of rank-deficiency. The ℓ^1 minimization problem for sparse recovery had a natural analogue in the nuclear norm minimization problem for low-rank recovery. Moreover, these convex relaxations succeed under analogous conditions involving restricted isometry properties of the observation operator.

However, in many of the most interesting applications of nuclear norm minimization, the RIP does not hold! In the introduction to this chapter, we sketched applications to recommender systems, in which we had access to *a subset* of the entries of a low-rank user-item matrix. We also sketched problems in reconstructing 3D shape, in which we observed *a subset* of the pixels of the rank-3 matrix $\boldsymbol{NL}$. Finally, we sketched a problem in Euclidean embedding, in which we observe *a subset* of the distances between some objects of interest. In all of these problems, the object of interest is a low-rank matrix $\boldsymbol{X}_{\text{true}} \in \mathbb{R}^{n \times n}$; the observation selects a subset $\Omega \subset [n] \times [n]$ of the entries of $\boldsymbol{X}_{\text{true}}$. The *matrix completion* asks us to fill in the missing entries:

Problem 4.4.1 (Matrix Completion)**.** *Let $\boldsymbol{X}_{\text{true}} \in \mathbb{R}^{n \times n}$ be a low-rank matrix. Suppose we are given $\boldsymbol{y} = \mathcal{P}_\Omega[\boldsymbol{X}_{\text{true}}]$, where $\Omega \subseteq [n] \times [n]$. Fill in the mssing entries of $\boldsymbol{X}_{\text{true}}$.*

In matrix completion $\mathcal{A} = \mathcal{P}_\Omega$ is the restriction onto some small subset $\Omega \subseteq [n] \times [n]$ of the entries. In this situation, if $(i, j) \notin \Omega$, $\mathcal{P}_\Omega[\boldsymbol{E}_{ij}] = \boldsymbol{0}$. That is to say, if Ω is a strict subset of $[n] \times [n]$, then $\mathcal{P}_\Omega$ has matrices of rank one in its null space! So, the rank-RIP cannot hold for any positive rank r with any nontrivial $\delta < 1$.

At a more basic level, the example of $\boldsymbol{X}_{\text{true}} = \boldsymbol{E}_{ij}$ suggests that there are some (very sparse) matrices that are impossible to complete from only a few

entries. This is in contrast to our discussion of low-rank matrix recovery thus far, in which the only factor that dictates the ease or difficulty of recovering a target $\boldsymbol{X}_{\text{true}}$ is the complexity $\text{rank}(\boldsymbol{X}_{\text{true}})$. Nevertheless, our development thus far suggests that even for the more challenging problem of matrix completion, there may be some class of *well-structured* matrices $\boldsymbol{X}_{\text{true}}$ of interest for applications, which *can* be efficiently completed from just a few entries. In this section, we will see that this is indeed the case.

4.4.1 Nuclear Norm Minimization for Matrix Completion

In light of our previous study of matrix recovery, a natural approach to completing a low-rank matrix from a small subset $\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{X}_{\text{true}}]$ of its entries is to look for the matrix $\boldsymbol{X}$ of minimum nuclear norm that agrees with the observation:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}. \end{array} \tag{4.4.1}$$

This is a special instance of the general nuclear norm minimization problem (4.3.14), with observation operator $\mathcal{A} = \mathcal{P}_\Omega$. As such, it is a semidefinite program, and can be solved to high accuracy in polynomial time. In practice, though, it is more important to have methods that scale to large problem instances. In the next section, we sketch one approach to achieving this, using Lagrange multiplier techniques. This approach has pedagogical value: it introduces several objects that will be used for analyzing when we can solve matrix completion problems efficiently. It also yields reasonably scalable algorithms. For practical matrix completion at the scale of $n \sim 10^6$ and beyond, even more scalable methods are needed; we discuss these issues in Chapters 8-9.

4.4.2 Algorithm via Augmented Lagrangian Method

There are two basic challenges in solving problem (4.4.1) at large scale. The first arises from the nonsmoothness of the nuclear norm $\|\cdot\|_*$; the second is due to the need to satisfy the constraint $\mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}$ exactly.[11] The fundamental technology for handling constraints in optimization is Lagrange duality.

The basic object is the *Lagrangian*, which introduces a matrix $\boldsymbol{\Lambda}$ of Lagrange multipliers for the constraint $\mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}$. The Lagrangian for (4.4.1) is

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\Lambda}) = \|\boldsymbol{X}\|_* + \langle \boldsymbol{\Lambda}, \mathcal{P}_\Omega[\boldsymbol{Y}] - \mathcal{P}_\Omega[\boldsymbol{X}] \rangle \,. \tag{4.4.2}$$

[11] In practice, when observations are noisy, exactly satisfying $\mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}$ is neither necessary nor desirable. We study the noisy matrix completion in Section 4.4.5, and develop dedicated algorithms for it in Chapter 8.

Optimal $\boldsymbol{X}$ are characterized by *saddle points* at which the Lagrangian is *minimized* with respect to $\boldsymbol{X}$, and maximized with respect to $\boldsymbol{\Lambda}$. A basic approach to solving a constrained problem such as (4.4.1) is to seek such a saddle point. In practice, more robustly convergent algorithms can be derived by instead working with the *augmented* Lagrangian

$$\mathcal{L}_\mu(\boldsymbol{X}, \boldsymbol{\Lambda}) = \|\boldsymbol{X}\|_* + \langle \boldsymbol{\Lambda}, \mathcal{P}_\Omega[\boldsymbol{Y}] - \mathcal{P}_\Omega[\boldsymbol{X}] \rangle + \tfrac{\mu}{2} \|\mathcal{P}_\Omega[\boldsymbol{Y}] - \mathcal{P}_\Omega[\boldsymbol{X}]\|_F^2 , \quad (4.4.3)$$

which encourages satisfaction of the constraint by adding an additional quadratic penalty term $\frac{\mu}{2} \|\mathcal{P}_\Omega[\boldsymbol{Y}] - \mathcal{P}_\Omega[\boldsymbol{X}]\|_F^2$. A more general introduction to augmented Lagrangian method (ALM) is given in section sec:convex:ALM of Chapter 8.

The augmented Lagrangian method seeks a saddle point of $\mathcal{L}_\mu$ by alternating between minimizing with respect to the "primal variables" $\boldsymbol{X}$ and taking one step of gradient ascent to increase $\mathcal{L}_\mu$ using the "dual variables" $\boldsymbol{\Lambda}$:

$$\boldsymbol{X}^{(k+1)} \in \arg\min_{\boldsymbol{X}} \mathcal{L}_\mu(\boldsymbol{X}, \boldsymbol{\Lambda}^{(k)}), \quad (4.4.4)$$

$$\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \mu \mathcal{P}_\Omega\big[\boldsymbol{Y} - \boldsymbol{X}^{(k+1)}\big]. \quad (4.4.5)$$

Here, $\mathcal{P}_\Omega\big[\boldsymbol{Y} - \boldsymbol{X}^{(k+1)}\big] = \nabla_{\boldsymbol{\Lambda}} \mathcal{L}_\mu(\boldsymbol{X}^{(k+1)}, \boldsymbol{\Lambda})$. The ALM algorithm makes a very special choice of the step size (μ) for updating $\boldsymbol{\Lambda}$. This choice is important in general: it ensures that $\boldsymbol{\Lambda}$ stays dual feasible, an issue that we will explain in more depth in section 8.4 of Chapter 8.

Under very general conditions, the iteration (4.4.4)-(4.4.5) converges to a primal dual optimal pair $(\boldsymbol{X}_\star, \boldsymbol{\Lambda}_\star)$, and hence yields a solution to (4.4.1). While this algorithm appears simple, some caution is necessary: the first step is itself a nontrivial optimization problem! This subproblem has a characteristic form, which we encountered in our study of sparse recovery in noise: the objective function is a sum of a smooth convex term $f(\boldsymbol{X})$, and a nonsmooth convex function $g(\boldsymbol{X}) = \|\boldsymbol{X}\|_*$:

$$\min_{\boldsymbol{X}} \underbrace{\|\boldsymbol{X}\|_*}_{g(\boldsymbol{X}) \text{ convex}} + \underbrace{\langle \boldsymbol{\Lambda}, \mathcal{P}_\Omega[\boldsymbol{Y}] - \mathcal{P}_\Omega[\boldsymbol{X}] \rangle + \tfrac{\mu}{2} \|\mathcal{P}_\Omega[\boldsymbol{Y}] - \mathcal{P}_\Omega[\boldsymbol{X}]\|_F^2}_{f(\boldsymbol{X}) \text{ smooth, convex}} . \quad (4.4.6)$$

Here,

$$\nabla f(\boldsymbol{X}) = -\mathcal{P}_\Omega[\boldsymbol{\Lambda}] + \mu \mathcal{P}_\Omega[\boldsymbol{X} - \boldsymbol{Y}]. \quad (4.4.7)$$

This is μ-Lipschitz, in the sense that for any pair of matrices $\boldsymbol{X}$ and $\boldsymbol{X}'$,

$$\big\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}')\big\|_F \le \mu \big\|\boldsymbol{X} - \boldsymbol{X}'\big\|_F . \quad (4.4.8)$$

This class of problem is amenable to the *proximal gradient method*.

The general proximal gradient iteration applies to objectives of the form $F(\boldsymbol{X}) = g(\boldsymbol{X}) + f(\boldsymbol{X})$, where g is convex, and f is convex, smooth, and has L-Lipschitz gradient. See section 8.2 of Chapter 8. Here we have the

Algorithm 4.1 (Matrix Completion by Augmented Lagrange Multiplier (ALM))

1: **initialize:** $\boldsymbol{X}^{(0)} = \boldsymbol{\Lambda}^{(0)} = 0, \mu > 0$.
2: **while** not converged **do**
3: compute $\boldsymbol{X}^{(k+1)} \in \arg\min_{\boldsymbol{X}} \mathcal{L}_\mu\big(\boldsymbol{X}, \boldsymbol{\Lambda}^{(k)}\big)$;
4: compute $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \mu\big(\mathcal{P}_\Omega[\boldsymbol{Y}] - \mathcal{P}_\Omega\big[\boldsymbol{X}^{(k+1)}\big]\big)$;
5: **end while**

Lipschitz constant $L = \mu$. So the iteration takes the form

$$\boldsymbol{X}^{(k+1)} = \arg\min_{\boldsymbol{X}} \left\{ g(\boldsymbol{X}) + \tfrac{\mu}{2} \left\| \boldsymbol{X} - \big(\boldsymbol{X}^{(k)} - \tfrac{1}{\mu}\nabla f(\boldsymbol{X}^{(k)})\big) \right\|_F^2 \right\}. \quad (4.4.9)$$

In particular, it requires us to solve a sequence of "proximal problems"

$$\min_{\boldsymbol{X}} \left\{ g(\boldsymbol{X}) + \tfrac{\mu}{2} \|\boldsymbol{X} - \boldsymbol{M}\|_F^2 \right\}, \quad (4.4.10)$$

for particular choices of the matrix $\boldsymbol{M}$. When g is the nuclear norm, this problem can be solved in closed form from the SVD of $\boldsymbol{M}$. Recall from (4.3.93), for a matrix $\boldsymbol{M}$ with the singular value decomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, its singular value thresholding operator is defined to be

$$\mathcal{D}_\tau[\boldsymbol{M}] = \boldsymbol{U}\mathcal{S}_\tau[\boldsymbol{\Sigma}]\boldsymbol{V}^*,$$

where $\mathcal{S}_\tau[\boldsymbol{X}] = \operatorname{sign}(\boldsymbol{X}) \circ (|\boldsymbol{X}| - \tau)_+$ is the soft thresholding operator.

Theorem 4.4.2. *The unique solution $\boldsymbol{X}_\star$ to the program:*

$$\min_{\boldsymbol{X}} \left\{ \|\boldsymbol{X}\|_* + \tfrac{\mu}{2} \|\boldsymbol{X} - \boldsymbol{M}\|_F^2 \right\}, \quad (4.4.11)$$

is given by

$$\boldsymbol{X}_\star = \mathcal{D}_{\mu^{-1}}[\boldsymbol{M}]. \quad (4.4.12)$$

The proof of this result follows from Exercise 4.11. The resulting procedures are stated as Algorithms 4.1-4.2. We have neglected important issues such as the choice of stopping conditions, and the effect of inexact solution to subproblem (4.4.4) on the convergence of the basic ALM iteration in Algorithm 4.1.

To understand when the convex program (4.4.1) and the above algorithm correctly recover a matrix $\boldsymbol{X}$ from a part of its entries, we vary the rank of matrix $\boldsymbol{X}$ as a fraction of the dimension n and a fraction $\rho_s \in (0, 1)$ of (randomly chosen) unobserved entries. In other words, ρ_s is the probability that an entry is omitted from the observation. Figure 4.7 shows the simulation results of using the above algorithm to recover a matrix $\boldsymbol{X} = \boldsymbol{L}_0$ under different settings.

Algorithm 4.2 (Minimizing the Augmented Lagrangian via Proximal Gradient)

1: **initialize:** $\boldsymbol{X}^{(0)}$ starts with the $\boldsymbol{X}^{(k)}$ from the outer loop.
2: **while** not converged **do**
3: compute

$$\begin{aligned} \boldsymbol{X}^{(\ell+1)} &= \operatorname{prox}_{g/\mu}\big(\boldsymbol{X}^{(\ell)} - \mu^{-1}\nabla f(\boldsymbol{X}^{(\ell)})\big) \\ &= \mathcal{D}_{\mu^{-1}}\Big[\mathcal{P}_{\Omega^c}[\boldsymbol{X}^{(\ell)}] + \mathcal{P}_{\Omega}[\boldsymbol{Y}] + \mu^{-1}\mathcal{P}_{\Omega}\big[\boldsymbol{\Lambda}^{(k)}\big]\Big]. \end{aligned}$$

4: **end while**

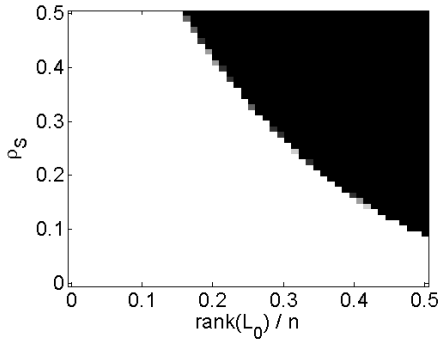

Figure 4.7. **Correct matrix completion for varying rank and sparsity.** Fraction of correct recoveries across 10 trials, as a function of rank($\boldsymbol{L}_0$) (x-axis) and fraction ρ_s of unobserved entries (y-axis). Here, $n = 400$. In all cases, $\boldsymbol{L}_0 = \boldsymbol{A}\boldsymbol{B}^*$ is a product of independent $n \times r$ i.i.d. $\mathcal{N}(0, 1/n)$ matrices. Trials are considered successful if $\|\hat{\boldsymbol{X}} - \boldsymbol{L}_0\|_F / \|\boldsymbol{L}_0\|_F < 10^{-3}$.

We may draw a few observations from the above simulations: 1. the convex program (4.4.1) and the above algorithm indeed succeed under a surprisingly wide range of conditions, as long as the rank of the matrix is relatively low and a fraction of the entries are observed. 2. the success and failure of the convex program (4.4.1) exhibit a sharp phase-transition phenomenon.

4.4.3 When Nuclear Norm Minimization Succeed?

The above simulations encourage us to understand the conditions under which nuclear norm minimization (4.4.1) is guaranteed to succeed for matrix completion?[12] It may be easier to first think about when it fails. It may

[12] Or ultimately, if possible, to precisely characterize the phase transition behavior we have observed through experiments.

fail if (i) $\boldsymbol{X}_0$ is *sparse* (as in the example of $\boldsymbol{E}_{ij}$), or (ii) if the sampling pattern Ω is chosen adversarially (e.g., if we miss an entire row or column of $\boldsymbol{X}_0$). Below, we will state a theorem that makes this intuition precise – namely, if $\boldsymbol{X}_0$ is low-rank, and not too "spiky", and Ω is chosen at random, then nuclear norm minimization succeeds with high probability. Below we make these assumptions precise.

Incoherent Low-rank Matrices.

Although our intuition is that $\boldsymbol{X}_0$ itself should not be too "sparse", for technical reasons it will be necessary to enforce this condition on the singular vectors of $\boldsymbol{X}_0$, rather than on $\boldsymbol{X}_0$ itself. Let $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the (reduced) singular value decomposition of $\boldsymbol{X}_0$. We say that $\boldsymbol{X}_0$ is ν-*incoherent* if the following hold:

$$\forall\, i \in [n], \quad \|\boldsymbol{e}_i^* \boldsymbol{U}\|_2^2 \leq \nu r/n, \tag{4.4.13}$$

$$\forall\, j \in [n], \quad \|\boldsymbol{e}_j^* \boldsymbol{V}\|_2^2 \leq \nu r/n. \tag{4.4.14}$$

These two conditions control the "spikiness" of the singular vectors of $\boldsymbol{X}_0$. To understand them better, note that $\boldsymbol{U}$ is an $n \times r$ matrix whose columns have unit ℓ^2 norm. Hence, $\sum_i \|\boldsymbol{e}_i^* \boldsymbol{U}\|_2^2 = \|\boldsymbol{U}\|_F^2 = r$. There are n rows, and so at least one of them must have ℓ^2 norm at least as large as the average, r/n. Hence, for any matrix $\boldsymbol{U}$ with unit norm columns, $\max_i \|\boldsymbol{e}_i^* \boldsymbol{U}\|_2^2 \geq r/n$. The *incoherence parameter* ν quantifies how much we lose with respect to this optimal bound. So, if ν is small, the singular vectors are, in a sense, spread around. To give a sense of scale, notice that it is always true that

$$1 \leq \nu \leq n/r. \tag{4.4.15}$$

If $\boldsymbol{U}$ and $\boldsymbol{V}$ are chosen uniformly at random (say by orthogonalizing the columns of a Gaussian matrix), then with high probability ν is bounded by $C \log(n)$. However, the definition does not require $\boldsymbol{U}$ and $\boldsymbol{V}$ to be random.

One important implication of this definition for matrix completion is that *when ν is small, there are no sparse matrices close to the tangent space T.* Indeed, let $\boldsymbol{E}_{ij} = \boldsymbol{e}_i \boldsymbol{e}_j^*$ denote the one-sparse matrix whose nonzero element occurs in entry (i, j). Then, using the expression (4.3.33) for the projection operator $\mathcal{P}_T$ onto the tangent space T, we have that

$$\begin{aligned} \|\mathcal{P}_T[\boldsymbol{E}_{ij}]\|_F^2 &= \|\boldsymbol{U}\boldsymbol{U}^* \boldsymbol{E}_{ij}\|_F^2 + \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^*)\boldsymbol{E}_{ij}\boldsymbol{V}\boldsymbol{V}^*\|_F^2 \\ &\leq \|\boldsymbol{U}^* \boldsymbol{e}_i\|_2^2 + \left\|\boldsymbol{e}_j^* \boldsymbol{V}\right\|_2^2 \\ &\leq \frac{2\nu r}{n}. \end{aligned} \tag{4.4.16}$$

This indicates that no standard basis matrix $\boldsymbol{E}_{ij}$ is too close to the subspace T. Strangely enough, this implies the standard basis $\{\boldsymbol{E}_{ij}\}$ is a good choice for reconstructing elements from T. This is similar in spirit to our observations on incoherent operator bases: if no $\boldsymbol{E}_{ij}$ is too close to T, information about any particular element $\boldsymbol{X}_0 \in T$ must be spread across many different

$\boldsymbol{E}_{ij}$. It will only take a few of these projections to be able to reconstruct $\boldsymbol{X}_0$. Note, however, a crucial difference between this notion of incoherence and our previous notions for matrix and vector recovery: here, the subspace T depends on $\boldsymbol{X}_0$ itself. The discussion in this section suggests that random sampling will be effective for reconstructing the particular matrix $\boldsymbol{X}_0$. We make this intuition formal below.

Exact Matrix Completion from Random Samples.

We assume that each entry (i, j) belongs to the set Ω independently with probability p. We call this a *Bernoulli* sampling model, since the indicators $\mathbb{1}_{(i,j)\in\Omega}$ are independent $\mathrm{Ber}(p)$ random variables. Under this model, the expected number of observed entries is

$$m = \mathbb{E}\big[\,|\Omega|\,\big] = pn^2. \tag{4.4.17}$$

Under this model, nuclear norm minimization succeeds even when the number m of observations is close to the number of intrinsic degrees of freedom in the rank-r matrix $\boldsymbol{X}_0$. The following theorem makes this precise:

Theorem 4.4.3 (Matrix Completion via Nuclear Norm Minimization)**.** *Let $\boldsymbol{X}_0 \in \mathbb{R}^{n\times n}$ be a rank-r matrix with coherence ν. Suppose that we observe $\boldsymbol{y} = \mathcal{P}_\Omega[\boldsymbol{X}_0]$, with Ω sampled according to the Bernoulli model with probability*

$$p \geq C_1 \frac{\nu r \log^2(n)}{n}. \tag{4.4.18}$$

Then with probability at least $1 - C_2 n^{-c_3}$, $\boldsymbol{X}_0$ is the unique optimal solution to

$$\textit{minimize} \quad \|\boldsymbol{X}\|_* \quad \textit{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{y}. \tag{4.4.19}$$

There are several things to notice about the above theorem. First, the expected number of measurements is

$$m = pn^2 = C_1 \nu n r \log^2(n). \tag{4.4.20}$$

Since a rank-r matrix has $O(nr)$ degrees of freedom, the oversampling factor is only about $C\nu \log^2(n)$ – the number of samples we must see is nearly minimal.[13] Second, the number of samples required scales with the coherence of the matrix $\boldsymbol{X}_0$. So, if we want to recover a very coherent (think, "nearly sparse") $\boldsymbol{X}_0$, we will simply need more observations. Finally, the probability of success is in all the possible choices of the observed subset but

[13] According to Theorem 1.7 of [Candes and Tao, 2009], if the sampling probability $p < \frac{\nu r \log(2n)}{2n}$, there will be infinitely many matrices of rank at most r that satisfy the incoherence condition and all have the same entries on Ω.

is only for a given low-rank matrix $\boldsymbol{X}_0$.[14] Of course, the precise conditions of the above theorem can only be interpreted as an idealized mathematical abstraction of real matrix completion or collaborative filtering problems. In particular, in real problems there may be noise in the observation, and, more importantly the observations may not be uniformly distributed. This has been a very intense area of recent investigation.

4.4.4 Proving Correctness of Nuclear Norm Minimization

In this section, we prove Theorem 4.4.3. Our approach is analogous to our proof that ℓ^1 recovers sparse vectors under incoherence (in Section 3.2.2) – we simply write down the optimality conditions and try to show that they are satisfied! Carrying this program through will be trickier, though.

Subdifferential of the Nuclear Norm.

To get started, we need an optimality condition for the nuclear norm minimization problem (4.4.19). As for the ℓ^1 norm, this means we need an expression for the subdifferential. The following lemma provides one:

Lemma 4.4.4. *Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ have compact singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. The subdifferential of the nuclear norm at $\boldsymbol{X}$ is given by*

$$\partial \left\|\cdot\right\|_* (\boldsymbol{X}) = \left\{\boldsymbol{Z} \mid \mathcal{P}_T[\boldsymbol{Z}] = \boldsymbol{U}\boldsymbol{V}^*,\ \left\|\mathcal{P}_{T^\perp}[\boldsymbol{Z}]\right\| \leq 1\right\}. \tag{4.4.21}$$

Proof. Consider any $\boldsymbol{Z}$ satisfying $\mathcal{P}_T[\boldsymbol{Z}] = \boldsymbol{U}\boldsymbol{V}^*$, and $\left\|\mathcal{P}_{T^\perp}[\boldsymbol{Z}]\right\| \leq 1$. Notice that $\|\boldsymbol{Z}\| = 1$. Since $\boldsymbol{X} \in T$,

$$\langle \boldsymbol{X}, \boldsymbol{Z} \rangle = \langle \boldsymbol{X}, \boldsymbol{U}\boldsymbol{V}^* \rangle = \langle \boldsymbol{U}^*\boldsymbol{X}\boldsymbol{V}, \boldsymbol{I} \rangle = \langle \boldsymbol{\Sigma}, \boldsymbol{I} \rangle = \left\|\boldsymbol{X}\right\|_*. \tag{4.4.22}$$

For every $\boldsymbol{X}'$,

$$\left\|\boldsymbol{X}\right\|_* + \left\langle \boldsymbol{Z}, \boldsymbol{X}' - \boldsymbol{X} \right\rangle = \left\langle \boldsymbol{Z}, \boldsymbol{X}' \right\rangle \leq \|\boldsymbol{Z}\| \left\|\boldsymbol{X}'\right\|_* = \left\|\boldsymbol{X}'\right\|_*. \tag{4.4.23}$$

Thus $\boldsymbol{Z}$ is a subgradient of the nuclear norm at $\boldsymbol{X}$: $\boldsymbol{Z} \in \partial \|\cdot\|_*(\boldsymbol{X})$. To complete the proof, we need to show that every element $\boldsymbol{Z} \in \partial \|\cdot\|_*(\boldsymbol{X})$ satisfies $\mathcal{P}_T[\boldsymbol{Z}] = \boldsymbol{U}\boldsymbol{V}^*$ and $\left\|\mathcal{P}_{T^\perp}[\boldsymbol{Z}]\right\| \leq 1$. We leave the converse as an exercise (see Exercise 4.18). □

If we compare to the expression for the subdifferential of the ℓ^1 norm, here, the subspace T plays the role of the *support* of the matrix, while the matrix $\boldsymbol{U}\boldsymbol{V}^*$ is playing the role of the *signs*. Indeed, in this language, $\partial \left\|\cdot\right\|_*$ consists of those $\boldsymbol{Z}$ that are equal to the "sign" $\boldsymbol{U}\boldsymbol{V}^*$ on the support T, and whose dual norm $\|\cdot\|$ is bounded by one on the orthogonal complement $T^\perp$ of the support.

[14] This is to contrast with the probability of success in the generic case, where an incoherent sampling operator is good for recovering the set of all matrices of rank less than r.

Optimality Conditions.

Once we have the subdifferential in hand, we can fairly immediately write down an optimality condition for the convex program of interest. Indeed, consider the optimization problem

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \mathcal{P}_\Omega[\boldsymbol{X}] = \mathcal{P}_\Omega[\boldsymbol{X}_0]. \end{array} \tag{4.4.24}$$

Any feasible $\boldsymbol{X}$ can be written as $\boldsymbol{X}_0 + \boldsymbol{H}$, where $\boldsymbol{H} \in \operatorname{null}(\mathcal{P}_\Omega)$, i.e., $\boldsymbol{H}$ is supported on the set Ω^c of entries that we do not observe. Similar to the ℓ^1 case in Section 3.2.2, if we can find some $\boldsymbol{\Lambda}$, called a dual certificate, such that it satisfies (the KKT condition):

- (i) $\boldsymbol{\Lambda}$ is supported on Ω and
- (ii) $\boldsymbol{\Lambda} \in \partial \|\cdot\|_* (\boldsymbol{X}_0)$ – i.e., $\mathcal{P}_T[\boldsymbol{\Lambda}] = \boldsymbol{U}\boldsymbol{V}^*$ and $\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}]\| \le 1$,

then we have

$$\|\boldsymbol{X}_0 + \boldsymbol{H}\|_* \ge \|\boldsymbol{X}_0\|_* + \langle \boldsymbol{\Lambda}, \boldsymbol{H} \rangle = \|\boldsymbol{X}_0\|_* , \tag{4.4.25}$$

where the final equality holds because $\boldsymbol{\Lambda}$ is supported on Ω and $\boldsymbol{H}$ is supported on Ω^c. In addition, if we further have $\|\mathcal{P}_{\Omega^c}\mathcal{P}_T\| < 1$ and $\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}]\| < 1$, then one can show that $\boldsymbol{X}_0$ is the *unique* optimal solution. The proof is similar to that in the ℓ^1 case (see the proof of Theorem 3.2.3) and we leave to the reader as an exercise (see Exercise 4.13).

A natural idea for constructing $\boldsymbol{\Lambda}$ might be to simply follow the program that has worked before (in the ℓ^1 minimization case) and look for a matrix $\boldsymbol{\Lambda}$ of smallest 2-norm that satisfies the equality constraints

$$\mathcal{P}_{\Omega^c}[\boldsymbol{\Lambda}] = \boldsymbol{0}, \quad \mathcal{P}_T[\boldsymbol{\Lambda}] = \boldsymbol{U}\boldsymbol{V}^*, \tag{4.4.26}$$

and then hope to check that it satisfies the inequality constraints $\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}]\| \le 1$. For example, we could take $\boldsymbol{\Lambda} = \mathcal{P}_\Omega[\boldsymbol{G}]$, with $\boldsymbol{G} = (\mathcal{P}_T\mathcal{P}_\Omega)^\dagger[\boldsymbol{U}\boldsymbol{V}^*]$, where $(\cdot)^\dagger$ denotes the pseudo inverse. We are then left to check that

$$\left\|\mathcal{P}_{T^\perp}\mathcal{P}_\Omega(\mathcal{P}_T\mathcal{P}_\Omega)^\dagger[\boldsymbol{U}\boldsymbol{V}^*]\right\| \tag{4.4.27}$$

is small. This is a random matrix, but it is an exceedingly complicated one. It actually *is* possible to analyze its norm, but the analysis is quite intricate. The challenge arises because the thing that is random here is the support Ω. It is repeated in several places, creating probabilistic dependencies, which complicates the analysis.

Relaxed Optimality Conditions.

As it is difficult to directly find a dual certificate satisfying the KKT conditions exactly, we might want to relax these conditions and see if we could still find another certificate for the optimality. The following proposition

suggests that we can ensure the optimality of $\boldsymbol{X}_0$ with an alternative set of (relaxed) conditions:

Proposition 4.4.5 (KKT Conditions – Approximate Version)**.** *The matrix $\boldsymbol{X}_0$ is the unique optimal solution to the nuclear minimization problem (4.4.19) if the following set of conditions hold*

1. *The operator norm of the operator $p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \mathcal{P}_T$ is small: $\left\|p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\| \leq \frac{1}{2}$.*
2. *There exists a dual certificate $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ that satisfies $\mathcal{P}_\Omega(\boldsymbol{\Lambda}) = \boldsymbol{\Lambda}$ and*
   (a) $\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}]\| \leq \frac{1}{2}$.
   (b) $\|\mathcal{P}_T[\boldsymbol{\Lambda}] - \boldsymbol{U}\boldsymbol{V}^*\|_F \leq \frac{1}{4n}$.

Conditions 2(a) and 2(b) above trade off between the degree of satisfaction of the equality constraint $\mathcal{P}_T[\boldsymbol{\Lambda}] = \boldsymbol{U}\boldsymbol{V}^*$ and the inequality constraint for the dual norm $\|\mathcal{P}_{T^\perp}\boldsymbol{\Lambda}\| \leq 1$ in the original KKT conditions. This is possible under the additional assumption that $\left\|p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\|$ is not too large. This assumption is satisfied whenever the sampling map $p^{-1}\mathcal{P}_\Omega$ nearly preserves the lengths of all elements $\boldsymbol{X} \in T$. It can be considered a strengthening of the condition that $T \cap \Omega^\perp = \{\boldsymbol{0}\}$, which was needed for unique optimality.

To prove Proposition 4.4.5, we will need another lemma. This says that provided $\mathcal{P}_\Omega$ acts nicely on matrices from T, every feasible perturbation $\boldsymbol{H}$ (i.e., $\boldsymbol{H}$ such that $\mathcal{P}_\Omega[\boldsymbol{H}] = \boldsymbol{0}$) must have a nonnegligible component along $T^\perp$:

Lemma 4.4.6. *Suppose that the operator $\mathcal{P}_\Omega$ satisfies*

$$\left\|\mathcal{P}_T - p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T\right\| \leq \frac{1}{2}. \tag{4.4.28}$$

Then for any $\boldsymbol{H}$ satisfying $\mathcal{P}_\Omega[\boldsymbol{H}] = \boldsymbol{0}$, we have

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_F \geq \sqrt{\frac{p}{2}}\,\|\mathcal{P}_T[\boldsymbol{H}]\|_F\,. \tag{4.4.29}$$

Proof. We have

$$\begin{aligned}
\langle \mathcal{P}_\Omega\mathcal{P}_T[\boldsymbol{H}], \mathcal{P}_\Omega\mathcal{P}_T[\boldsymbol{H}]\rangle &= \langle \mathcal{P}_T[\boldsymbol{H}], \mathcal{P}_\Omega\mathcal{P}_T[\boldsymbol{H}]\rangle \\
&= p\langle \mathcal{P}_T[\boldsymbol{H}], p^{-1}\mathcal{P}_\Omega\mathcal{P}_T[\boldsymbol{H}]\rangle \\
&= p\langle \mathcal{P}_T[\boldsymbol{H}], \mathcal{P}_T p^{-1}\mathcal{P}_\Omega\mathcal{P}_T\mathcal{P}_T[\boldsymbol{H}]\rangle \\
&\geq p\Big(1 - \left\|\mathcal{P}_T - \mathcal{P}_T p^{-1}\mathcal{P}_\Omega\mathcal{P}_T\right\|\Big)\,\|\mathcal{P}_T[\boldsymbol{H}]\|_F^2 \\
&\geq \frac{p}{2}\,\|\mathcal{P}_T[\boldsymbol{H}]\|_F^2\,,
\end{aligned} \tag{4.4.30}$$

Then from $\mathcal{P}_\Omega \mathcal{P}_T[\boldsymbol{H}] + \mathcal{P}_\Omega \mathcal{P}_{T^\perp}[\boldsymbol{H}] = \mathcal{P}_\Omega[\boldsymbol{H}] = \mathbf{0}$, we have

$$\begin{aligned} 0 &= \|\mathcal{P}_\Omega \mathcal{P}_T[\boldsymbol{H}] + \mathcal{P}_\Omega \mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_F && (4.4.31)\\ &\geq \|\mathcal{P}_\Omega \mathcal{P}_T[\boldsymbol{H}]\|_F - \|\mathcal{P}_\Omega \mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_F && (4.4.32)\\ &\geq \sqrt{\frac{p}{2}} \|\mathcal{P}_T[\boldsymbol{H}]\|_F - \|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_F , && (4.4.33) \end{aligned}$$

giving the conclusion. □

We are now ready to prove the optimality of $\boldsymbol{X}_0$ under the conditions given by Proposition 4.4.5.

Proof. We want to show that under the above conditions, for any feasible perturbation $\boldsymbol{H} \neq \mathbf{0}$ and $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{H}$, we have $\|\boldsymbol{X}\|_* > \|\boldsymbol{X}_0\|_*$. Let $\mathcal{P}_{T^\perp}[\boldsymbol{H}] = \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{V}}^*$. Then we have $\bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^* \in T^\perp$ and $\left\|\bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^*\right\| \leq 1$. Therefore, we have $\boldsymbol{U}\boldsymbol{V}^* + \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^* \in \partial \|\cdot\|_* (\boldsymbol{X}_0)$ is a subgradient of the nuclear norm at $\boldsymbol{X}_0$.

Also, we have $\langle \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^*, \mathcal{P}_{T^\perp}[\boldsymbol{H}] \rangle = \|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_*$ and $\langle \boldsymbol{\Lambda}, \boldsymbol{H} \rangle = 0$ and apply them to the following inequalities:

$$\begin{aligned} \|\boldsymbol{X}_0 + \boldsymbol{H}\|_* &\geq \|\boldsymbol{X}_0\|_* + \langle \boldsymbol{U}\boldsymbol{V}^* + \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^*, \boldsymbol{H} \rangle,\\ &= \|\boldsymbol{X}_0\|_* + \langle \boldsymbol{U}\boldsymbol{V}^* + \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^* - \boldsymbol{\Lambda}, \boldsymbol{H} \rangle,\\ &= \|\boldsymbol{X}_0\|_* + \langle \boldsymbol{U}\boldsymbol{V}^* - \mathcal{P}_T[\boldsymbol{\Lambda}], \boldsymbol{H} \rangle + \langle \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^* - \mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}], \boldsymbol{H} \rangle,\\ &\geq \|\boldsymbol{X}_0\|_* - \frac{1}{4n} \|\mathcal{P}_T[\boldsymbol{H}]\|_F + \frac{1}{2} \|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_* ,\\ &\geq \|\boldsymbol{X}_0\|_* + \underbrace{\left(\tfrac{1}{2} - \tfrac{1}{4n}\sqrt{\tfrac{2}{p}}\right)}_{>0, \text{ since } p > n^{-2}.} \|\mathcal{P}_{T^\perp}[\boldsymbol{H}]\|_F . \qquad (4.4.34) \end{aligned}$$

In the final inequality, we have invoked Lemma 4.4.6.

Hence, for feasible perturbations $\boldsymbol{H}$, $\|\boldsymbol{X}_0 + \boldsymbol{H}\|_* \geq \|\boldsymbol{X}_0\|_*$, with equality if and only if $\mathcal{P}_{T^\perp}[\boldsymbol{H}] = \mathbf{0}$. But via Lemma 4.4.6, $\mathcal{P}_{T^\perp}[\boldsymbol{H}] = \mathbf{0} \implies \boldsymbol{H} = \mathbf{0}$. Thus, for any nonzero feasible perturbation $\boldsymbol{H}$, $\|\boldsymbol{X}_0 + \boldsymbol{H}\|_* > \|\boldsymbol{X}_0\|_*$, establishing the desired condition. □

The Optimality Condition is Satisfied with High Probability.

To complete the proof, we simply need to show that the optimality condition can be satisfied with high probability. To do this, we need to verify two claims: first, that with high probability the sampling operator Ω acts nicely on T, in the sense that $\left\|p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\|$ is small. We then need to show that with high probability we can construct the desired dual certificate $\boldsymbol{\Lambda}$.

The sampling operator acts nicely on T

We next prove that the sampling operator $\mathcal{P}_\Omega$ preserves some part of every element of T, in the sense that $\left\|p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \mathcal{P}_T\right\|$ is small. This phenomenon is a consequence of the incoherence of the matrix $\boldsymbol{X}_0$ and the

uniform random model on Ω. The proof of the following lemma uses the matrix (operator) Bernstein inequality to show this rigorously.

Lemma 4.4.7. *Let $\mathcal{P}_\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ denote the operator*

$$\mathcal{P}_\Omega[\boldsymbol{X}] = \sum_{ij} \boldsymbol{X}_{ij} \mathbb{1}_{(i,j)\in\Omega} \boldsymbol{E}_{ij} \tag{4.4.35}$$

with $\mathbb{1}_{(i,j)\in\Omega}$ independent Bernoulli random variables with probability p. Fix any ε with $c\frac{\sqrt{\log n}}{n} \le \varepsilon \le 1$. There is a numerical constant C such that if $p > C\frac{\nu r \log n}{\varepsilon^2 n}$, then with high probability,

$$\left\|\mathcal{P}_T - p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T\right\| \le \varepsilon. \tag{4.4.36}$$

Proof. We apply the matrix Bernstein inequality in Theorem E.4.4 to bound the norm of

$$\mathcal{P}_T - p^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T \quad = \quad \sum_{ij} \underbrace{\mathcal{P}_T\left(\frac{\mathcal{I}}{n^2} - p^{-1}\mathbb{1}_{(i,j)\in\Omega}\boldsymbol{E}_{ij}\langle \boldsymbol{E}_{ij}, \cdot\rangle\right)\mathcal{P}_T}_{\doteq \mathcal{W}_{ij}}.$$

Here, $\mathcal{W}_{ij} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ are independent random linear maps, and $\mathbb{E}\left[\sum_{ij}\mathcal{W}_{ij}\right] = 0$. The matrix Bernstein inequality requires (i) an almost sure bound R on $\max_{ij}\|\mathcal{W}_{ij}\|$, and (ii) control of the "variance"

$$\sum_{ij}\mathbb{E}\left[\mathcal{W}_{ij}^*\mathcal{W}_{ij}\right]. \tag{4.4.37}$$

We provide these as follows.
(i) Almost sure control of the summands:

$$\begin{aligned}
\|\mathcal{W}_{ij}\| &\le \max\left\{\left\|n^{-2}\mathcal{P}_T\right\|, \left\|p^{-1}\mathcal{P}_T[\boldsymbol{E}_{ij}]\langle\mathcal{P}_T[\boldsymbol{E}_{ij}],\cdot\rangle\right\|\right\}, \quad \text{almost surely}\\
&= \max\left\{n^{-2}, p^{-1}\|\mathcal{P}_T[\boldsymbol{E}_{ij}]\|_F^2\right\},\\
&\le \max\left\{n^{-2}, \frac{2\nu r}{np}\right\},\\
&\le \max\left\{\frac{1}{n^2}, \frac{2\varepsilon^2}{C\log n}\right\},\\
&= \frac{2\varepsilon^2}{C\log n}.
\end{aligned} \tag{4.4.38}$$

We may take $R = \frac{2\varepsilon^2}{C\log n}$.

(ii) Control of the "operator variance". Note that

$$\begin{aligned}
&\sum_{ij} \mathbb{E}\left[\mathcal{W}_{ij}^* \mathcal{W}_{ij}\right] \\
&= \sum_{ij} \mathbb{E}\Big[\frac{1}{n^4}\mathcal{P}_T - \frac{2p^{-1}}{n^2} \mathbb{1}_{(i,j)\in\Omega} \mathcal{P}_T \boldsymbol{E}_{ij} \langle \boldsymbol{E}_{ij}, \cdot \rangle \mathcal{P}_T \\
&\qquad\qquad + \mathbb{1}_{(i,j)\in\Omega} p^{-2} \mathcal{P}_T \boldsymbol{E}_{ij} \|\mathcal{P}_T \boldsymbol{E}_{ij}\|_F^2 \langle \boldsymbol{E}_{ij}, \cdot \rangle \mathcal{P}_T \Big] \\
&\preceq p^{-1} \sum_{ij} \mathcal{P}_T \boldsymbol{E}_{ij} \|\mathcal{P}_T \boldsymbol{E}_{ij}\|_F^2 \langle \boldsymbol{E}_{ij}, \cdot \rangle \mathcal{P}_T \\
&\preceq p^{-1} \frac{2\nu r}{n} \sum_{ij} \mathcal{P}_T \boldsymbol{E}_{ij} \langle \boldsymbol{E}_{ij}, \cdot \rangle \mathcal{P}_T \\
&\preceq \frac{2\varepsilon^2}{C \log n} \mathcal{P}_T. \qquad (4.4.39)
\end{aligned}$$

The operator $\sum_{ij} \mathbb{E}\left[\mathcal{W}_{ij}^* \mathcal{W}_{ij}\right]$ is self-adjoint and positive semidefinite. The above calculation therefore implies that

$$\begin{aligned}
\sigma^2 &= \max\left\{ \left\| \sum_{ij} \mathbb{E}\left[\mathcal{W}_{ij}^* \mathcal{W}_{ij}\right] \right\|, \left\| \sum_{ij} \mathbb{E}\left[\mathcal{W}_{ij} \mathcal{W}_{ij}^*\right] \right\| \right\} \\
&\leq \frac{2\varepsilon^2}{C \log n}. \qquad (4.4.40)
\end{aligned}$$

Using these calculations, we obtain a bound

$$\mathbb{P}\left[\left\| \sum_{ij} \mathcal{W}_{ij} \right\| > t \right] \leq 2n \exp\left(\frac{-t^2/2}{\frac{2\varepsilon^2}{C\log n} + t\frac{2\varepsilon^2}{3C\log n}} \right). \qquad (4.4.41)$$

The probability of failure for $t = \varepsilon$ is bounded by $n^{-\rho}$; the exponent ρ can be made as large as desired by choosing C appropriately. □

Choosing $\varepsilon = 1/2$ in the statement of the above lemma, we obtain the desired condition needed for Lemma 4.4.6.

Construction of a dual certificate by the golfing scheme.

From the above discussion, in order to prove Theorem 4.4.3, we only have to show that under the conditions of the theorem, we can find a dual certificate that satisfies two conditions 2 (a) and 2 (b) of Proposition 4.4.5. In this section, we show how to construct such a dual certificate, $\boldsymbol{\Lambda}$. In the next chapter, we will reuse this construction to analyze the related problem of *robust matrix recovery*, in which a fraction of the entries of a low-rank matrix have been corrupted. For this purpose, we give a complete summary of the properties of our construction in the following proposition.

Here, properties (i) and (ii) are essential for matrix completion; property (iii) will be used in the following chapters for matrix recovery.

Proposition 4.4.8 (Dual certificate for low-rank recovery)**.** *Let $\boldsymbol{L}_0 \in \mathbb{R}^{n\times n}$ be a rank-r matrix, with coherence ν. Let $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$ be matrices whose columns are leading left- and right singular vectors of $\boldsymbol{L}_0$. Let*

$$T = \left\{ \boldsymbol{U}\boldsymbol{X}^* + \boldsymbol{Y}\boldsymbol{V}^* \mid \boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r} \right\}. \tag{4.4.42}$$

Then if $\Omega \sim \mathrm{Ber}(p)$, with

$$p > C_0 \frac{\nu r \log^2 n}{n}, \tag{4.4.43}$$

there exists a matrix $\boldsymbol{\Lambda}$ supported on Ω, satisfying

(i) $\|\mathcal{P}_T \boldsymbol{\Lambda} - \boldsymbol{U}\boldsymbol{V}^*\|_F \leq \frac{1}{4n}$,

(ii) $\|\mathcal{P}_{T^\perp} \boldsymbol{\Lambda}\| \leq \frac{1}{4}$,

(iii) $\|\boldsymbol{\Lambda}\|_\infty < \frac{C_1 \log n}{p} \times \|\boldsymbol{U}\boldsymbol{V}^*\|_\infty$,

with high probability. Here, C_1 is a positive numerical constant.

We prove this proposition using an iterative construction. Let

$$\Omega_1, \ldots, \Omega_k \tag{4.4.44}$$

be *independent* random subsets, chosen according to the Bernoulli model with parameter q. Set

$$\Omega = \bigcup_{i=1}^{k} \Omega_i. \tag{4.4.45}$$

Then Ω is *also* a Bernoulli subset, with parameter

$$p = 1 - (1-q)^k. \tag{4.4.46}$$

The parameter p is the probability that a given entry is in *at least one* of the subsets Ω_i. Hence, $p \leq kq$. The argument that we develop below will lead us to choose $k = C_g \log(n)$, with C_g a constant. Because k is not too large, this implies that the parameter q is also not too small:

$$q \geq \frac{p}{k} = \frac{C_0}{C_g} \frac{\nu r \log n}{n}. \tag{4.4.47}$$

Provided C_0 is large enough compared to C_g, the subsets Ω_i *all* satisfy the conditions of Lemma 4.4.7, and so with high probability

$$\left\| \mathcal{P}_T - q^{-1} \mathcal{P}_T \mathcal{P}_{\Omega_j} \mathcal{P}_T \right\| \leq \tfrac{1}{2}, \quad j = 1, \ldots, k. \tag{4.4.48}$$

We will construct a sequence of matrices $\boldsymbol{\Lambda}_0, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_k$, in which each $\boldsymbol{\Lambda}_j$ depends only on $\Omega_1, \ldots \Omega_j$. We let $\boldsymbol{\Lambda}_0 = \boldsymbol{0}$. And let

$$\boldsymbol{E}_j = \mathcal{P}_T[\boldsymbol{\Lambda}_j] - \boldsymbol{U}\boldsymbol{V}^*. \tag{4.4.49}$$

Since our goal is to obtain $\mathbf{\Lambda}$ such that $\mathcal{P}_T[\mathbf{\Lambda}] \approx \boldsymbol{U}\boldsymbol{V}^*$, $\boldsymbol{E}_j$ should be considered the *error* at iteration j. To get our next $\mathbf{\Lambda}$, we simply try to correct the error:

$$\mathbf{\Lambda}_j = \mathbf{\Lambda}_{j-1} - \left(q^{-1}\mathcal{P}_{\Omega_j}\right)[\boldsymbol{E}_{j-1}]. \tag{4.4.50}$$

This construction is known as the *golfing scheme*, as it tries to reach the goal by reducing error step by step.

There are several things worth noting about this construction. First, it produces $\mathbf{\Lambda}_j$ supported only on $\Omega_1 \cup \cdots \cup \Omega_j$. Thus, as desired, $\mathbf{\Lambda}_k$ is supported on Ω. Second, because $\boldsymbol{U}\boldsymbol{V}^* \in T$, $\boldsymbol{E}_j \in T$ for each j. This means that

$$\begin{aligned}
\boldsymbol{E}_j &= \mathcal{P}_T[\mathbf{\Lambda}_j] - \boldsymbol{U}\boldsymbol{V}^* \\
&= \mathcal{P}_T[\mathbf{\Lambda}_{j-1}] - \boldsymbol{U}\boldsymbol{V}^* - q^{-1}\mathcal{P}_T\mathcal{P}_{\Omega_j}[\boldsymbol{E}_{j-1}] \\
&= \boldsymbol{E}_j - q^{-1}\mathcal{P}_T\mathcal{P}_{\Omega_j}[\boldsymbol{E}_{j-1}] \\
&= (\mathcal{P}_T - q^{-1}\mathcal{P}_T\mathcal{P}_{\Omega_j}\mathcal{P}_T)[\boldsymbol{E}_{j-1}].
\end{aligned}$$

Since $\mathbb{E}\left[q^{-1}\mathcal{P}_{\Omega_j}\right] = \mathcal{I}$, in expectation, this iterative process drives the error to zero: $\mathbb{E}[\boldsymbol{E}_j] = \boldsymbol{0}$.

As it turns out, due to the fact that $\left\|\mathcal{P}_T - q^{-1}\mathcal{P}_T\mathcal{P}_{\Omega_j}\mathcal{P}_T\right\| \leq \frac{1}{2}$, after k steps, the error reduces to

$$\|\mathcal{P}_T[\mathbf{\Lambda}_k] - \boldsymbol{U}\boldsymbol{V}^*\|_F = \|\boldsymbol{E}_k\|_F \leq 2^{-k}\|\boldsymbol{E}_0\|_F \tag{4.4.51}$$

with high probability.

So, based on the golfing scheme, to achieve the desired accuracy as suggested by the above lemma, we want $2^{-k}\|\boldsymbol{E}_0\|_F = 2^{-k}\sqrt{r} \leq \frac{1}{4n}$. Since $r < n$, we only need to have $2^{-k} \sim O(1/n^2)$, that is to choose $k = C_g \log(n)$ for some large enough constant C_g, say $C_g = 20$. Therefore, under these conditions, the dual certificate constructed after k iterations $\mathbf{\Lambda}_k$ satisfies condition 2 (b) of Proposition 4.4.5:

$$\|\mathcal{P}_T[\mathbf{\Lambda}_k] - \boldsymbol{U}\boldsymbol{V}^*\|_F \leq \frac{1}{4n}. \tag{4.4.52}$$

Finally, to satisfy Condition 2(a) of Proposition 4.4.5, we need to show that the operator norm of the random matrix $\mathcal{P}_{T^\perp}[\mathbf{\Lambda}_k]$ is bounded as

$$\|\mathcal{P}_{T^\perp}[\mathbf{\Lambda}_k]\| \leq 1/4.$$

Notice that from the construction of $\mathbf{\Lambda}_k$, we have

$$\begin{aligned}
\mathbf{\Lambda}_k &= \sum_{j=1}^{k} -q^{-1}\mathcal{P}_{\Omega_j}[\boldsymbol{E}_{j-1}], \\
\boldsymbol{E}_j &= (\mathcal{P}_T - \mathcal{P}_T q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_T)[\boldsymbol{E}_{j-1}], \quad \text{with } \boldsymbol{E}_0 = -\boldsymbol{U}\boldsymbol{V}^*.
\end{aligned}$$

The matrix of interest can be expressed as

$$\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k] = \sum_{j=1}^{k} -q^{-1}\mathcal{P}_{T^\perp}\mathcal{P}_{\Omega_j}[\boldsymbol{E}_{j-1}] = \sum_{j=1}^{k} \mathcal{P}_{T^\perp}(\mathcal{P}_T - q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_T)[\boldsymbol{E}_{j-1}], \tag{4.4.53}$$

where the second identity is due to $\mathcal{P}_{T^\perp}\mathcal{P}_T = 0$ and $\mathcal{P}_T[\boldsymbol{E}_j] = \boldsymbol{E}_j$.

Since we are interested in bounding the norm of $\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]$, it would help if we know good bounds on various norms of $\mathcal{P}_{\Omega_j}$ and its interaction with the operator $\mathcal{P}_T$ or $\mathcal{P}_{T^\perp}$. Notice each $\mathcal{P}_{\Omega_j}$ is a summation of independent random operators. A very powerful tool we can use to bound the norm of summation of random matrices (or operators) is the so-called matrix Bernstein inequality introduced in the Appendix E.4, which we have used once before in Lemma 4.4.7.

To bound the norm of $\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]$, we need good bounds on three additional operators similar to that in Lemma 4.4.7. The proofs of these bounds[15] are all similar to that of Lemma 4.4.7 by utilizing the matrix Bernstein inequality. We hence leave their derivations as exercises to the reader to get familiar with the matrix Bernstein inequality.

We phrase these bounds in terms of

$$\|\boldsymbol{Z}\|_\infty = \max_{ij} |\boldsymbol{Z}_{ij}|, \tag{4.4.54}$$

and the maximum of the largest ℓ^2 norm of a row and the largest ℓ^2 norm of a column, which we denote by $\|\cdot\|_{rc}$:

$$\|\boldsymbol{Z}\|_{rc} = \max\left\{\max_i \|\boldsymbol{e}_i^*\boldsymbol{Z}\|_2\,,\ \max_j \|\boldsymbol{Z}\boldsymbol{e}_j\|_2\right\}. \tag{4.4.55}$$

Lemma 4.4.9. *Let $\boldsymbol{Z}$ be any fixed $n \times n$ matrix, and Ω a $\mathrm{Ber}(q)$ subset, with*

$$q > C_0 \frac{\nu r \log n}{n}. \tag{4.4.56}$$

Then with high probability

$$\left\|\left(q^{-1}\mathcal{P}_\Omega - \mathcal{I}\right)\boldsymbol{Z}\right\| \le C\Big(\frac{n}{C_0\nu r}\|\boldsymbol{Z}\|_\infty + \sqrt{\frac{n}{C_0\nu r}}\|\boldsymbol{Z}\|_{rc}\Big), \tag{4.4.57}$$

where C is a numerical constant.

Proof. Exercise 4.21. □

Lemma 4.4.10. *Let $\boldsymbol{Z}$ be any fixed $n \times n$ matrix. There exists a numerical constant C_0 such that if Ω is a $\mathrm{Ber}(q)$ subset with*

$$q > C_0 \frac{\nu r \log n}{n}, \tag{4.4.58}$$

[15] following the work of [Chen et al., 2013].

then with high probability

$$\left\|(q^{-1}\mathcal{P}_T\mathcal{P}_\Omega - \mathcal{P}_T)\boldsymbol{Z}\right\|_{rc} \le \frac{1}{2}\Big(\sqrt{\frac{n}{\nu r}}\,\|\boldsymbol{Z}\|_\infty + \|\boldsymbol{Z}\|_{rc}\Big). \tag{4.4.59}$$

Proof. Exercise 4.22. □

Lemma 4.4.11. *Suppose $\boldsymbol{Z}$ is a fixed $n \times n$ matrix in T. There exists a constant C_0 such that if Ω is a* Bernoulli(q) *subset with*

$$q > C_0 \frac{\nu r \log n}{n}. \tag{4.4.60}$$

Then with high probability we have

$$\left\|(\mathcal{P}_T - q^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T)\boldsymbol{Z}\right\|_\infty \le \tfrac{1}{2}\,\|\boldsymbol{Z}\|_\infty\,. \tag{4.4.61}$$

Proof. Exercise 4.23. □

With these three lemmas in hand, we are now ready to show that the spectral norm of $\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]$ is very small, in particular can be bounded as $\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]\| \le 1/4$:

Proof. From the golfing construction, $\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]$ can be expressed as the series given in (4.4.53). Hence we have

$$\begin{aligned}
\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]\| &\le \sum_{j=1}^{k} \left\|\mathcal{P}_{T^\perp}(\mathcal{P}_T - q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_T)[\boldsymbol{E}_{j-1}]\right\| \\
&\le \sum_{j=1}^{k} \left\|(\mathcal{P}_T - q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_T)[\boldsymbol{E}_{j-1}]\right\| \\
&= \sum_{j=1}^{k} \left\|(\mathcal{I} - q^{-1}\mathcal{P}_{\Omega_j})[\boldsymbol{E}_{j-1}]\right\|.
\end{aligned} \tag{4.4.62}$$

Notice that in the construction of the golfing scheme, we have ensured that each subset Ω_j is sampled according to the Bernoulli model, with parameter $q > C_0\frac{\nu r \log n}{n}$ for some large enough C_0. This means each of the k subsets Ω_j satisfies the conditions of the above lemmas. We first apply Lemma 4.4.9 to the right hand side of the last inequality and obtain (assuming $C_0 > 1$):

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]\| \le \frac{C}{\sqrt{C_0}} \sum_{j=1}^{k} \Big(\frac{n}{\nu r}\,\|\boldsymbol{E}_{j-1}\|_\infty + \sqrt{\frac{n}{\nu r}}\,\|\boldsymbol{E}_{j-1}\|_{rc}\Big). \tag{4.4.63}$$

To bound $\|\boldsymbol{E}_{j-1}\|_\infty$ we apply Lemma 4.4.11 and obtain

$$\begin{aligned}\|\boldsymbol{E}_{j-1}\|_\infty &= \left\|(\mathcal{P}_T - \tfrac{1}{q}\mathcal{P}_T\mathcal{P}_{\Omega_{j-1}}\mathcal{P}_T)\cdots(\mathcal{P}_T - \tfrac{1}{q}\mathcal{P}_T\mathcal{P}_{\Omega_1}\mathcal{P}_T)\boldsymbol{E}_0\right\|_\infty \\ &\le \left(\frac{1}{2}\right)^{j-1}\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty . \end{aligned} \tag{4.4.64}$$

Using this together with the fact that $\boldsymbol{\Lambda}_k = -\sum_j q^{-1}\mathcal{P}_{\Omega_j}\boldsymbol{E}_{j-1}$, we obtain

$$\begin{aligned}\|\boldsymbol{\Lambda}_k\|_\infty &\le q^{-1}\sum_j \|\boldsymbol{E}_{j-1}\|_\infty &(4.4.65)\\ &\le 2q^{-1}\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty . &(4.4.66)\end{aligned}$$

Since $q > p/C_q \log n$, this establishes property (iii) of Proposition 4.4.8 for $\boldsymbol{\Lambda}_k$.

To bound $\|\boldsymbol{E}_{j-1}\|_{rc}$ we apply Lemma 4.4.10 and obtain

$$\begin{aligned}\|\boldsymbol{E}_{j-1}\|_{rc} &= \left\|(\mathcal{P}_T - \tfrac{1}{q}\mathcal{P}_T\mathcal{P}_{\Omega_{j-1}}\mathcal{P}_T)\boldsymbol{E}_{j-2}\right\|_{rc} \\ &\le \frac{1}{2}\sqrt{\frac{n}{\nu r}}\|\boldsymbol{E}_{j-2}\|_\infty + \frac{1}{2}\|\boldsymbol{E}_{j-1}\|_{rc} . \end{aligned} \tag{4.4.67}$$

Combine the above two inequalities and apply them recursively to $j-1, j-2, \ldots, 0$ and we obtain

$$\|\boldsymbol{E}_{j-1}\|_{rc} \le j\left(\frac{1}{2}\right)^{j-1}\sqrt{\frac{n}{\nu r}}\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty + \left(\frac{1}{2}\right)^{j-1}\|\boldsymbol{U}V^*\|_{rc} . \tag{4.4.68}$$

Substitute the bounds (4.4.64) and (4.4.68) to the right and side of (4.4.63) and we obtain

$$\begin{aligned}\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]\| &\le \frac{C}{\sqrt{C_0}}\frac{n}{\nu r}\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty\sum_{j=1}^{k}(j+1)\left(\frac{1}{2}\right)^{j-1} \\ &\quad + \frac{C}{\sqrt{C_0}}\sqrt{\frac{n}{\nu r}}\|\boldsymbol{U}\boldsymbol{V}^*\|_{rc}\sum_{j=1}^{k}\left(\frac{1}{2}\right)^{j-1} \\ &\le \frac{6C}{\sqrt{C_0}}\frac{n}{\nu r}\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty + \frac{2C}{\sqrt{C_0}}\sqrt{\frac{n}{\nu r}}\|\boldsymbol{U}\boldsymbol{V}^*\|_{rc} . \end{aligned} \tag{4.4.69}$$

As the matrix $\boldsymbol{X}_0$ satisfies the incoherence conditions (4.4.13) and (4.4.14), we have

$$\begin{aligned}\|\boldsymbol{U}\boldsymbol{V}^*\|_\infty &\le \max_{i,j}\left\{\|\boldsymbol{U}^*\boldsymbol{e}_i\|_2 \times \|\boldsymbol{V}^*\boldsymbol{e}_j\|_2\right\} \le \frac{\nu r}{n}, \\ \|\boldsymbol{U}\boldsymbol{V}^*\|_{rc} &\le \max\left\{\max_i\|\boldsymbol{e}_i^*\boldsymbol{U}\boldsymbol{V}^*\|_2 , \max_j\|\boldsymbol{U}\boldsymbol{V}^*\boldsymbol{e}_j\|_2\right\} \le \sqrt{\frac{\nu r}{n}}. \end{aligned}$$

Therefore,

$$\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]\| \le \frac{6C}{\sqrt{C_0}} + \frac{2C}{\sqrt{C_0}} \le \frac{1}{4} \tag{4.4.70}$$

for large enough C_0. This establishes property (ii) of Proposition 4.4.8 for $\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}_k]$. □

The above derivations and results show that the relaxed KKT conditions in Proposition 4.4.5 can be satisfied with high probability, proving Theorem 4.4.3.

4.4.5 Stable Matrix Completion with Noise

So far in the matrix completion problem, we have assumed the observed entries are precise. In real world matrix completion problems, the observed entries are often corrupted with some noise:

$$\boldsymbol{Y}_{ij} = [\boldsymbol{X}_0]_{ij} + \boldsymbol{Z}_{ij}, \quad (i,j) \in \Omega, \tag{4.4.71}$$

where $\boldsymbol{Z}_{ij}$ can be some small noise. Or equivalently, we can write

$$\mathcal{P}_\Omega[\boldsymbol{Y}] = \mathcal{P}_\Omega[\boldsymbol{X}_0] + \mathcal{P}_\Omega[\boldsymbol{Z}], \tag{4.4.72}$$

where $\boldsymbol{Z}$ is an $n \times n$ matrix of noises. We may assume the overall noise level is small $\|\mathcal{P}_\Omega[\boldsymbol{Z}]\|_F < \delta$. As in the stable matrix recovery case, we could expect to recover a low rank matrix matrix close to $\boldsymbol{X}_0$ via solving the following convex program:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \|\mathcal{P}_\Omega[\boldsymbol{X}] - \mathcal{P}_\Omega[\boldsymbol{Y}]\|_F < \delta. \end{array} \tag{4.4.73}$$

The following theorem states that under the same conditions of Theorem 4.4.3 when the nuclear norm minimization recovers the correct low rank matrix from noiseless measurements, the above program gives a stable estimate $\hat{\boldsymbol{X}}$ of the true low-rank matrix $\boldsymbol{X}_0$:

Theorem 4.4.12 (Stable Matrix Completion)**.** *Let $\boldsymbol{X}_0 \in \mathbb{R}^{n\times n}$ be a rank-r, ν-incoherent matrix. Suppose that we observe $\mathcal{P}_\Omega[\boldsymbol{Y}] = \mathcal{P}_\Omega[\boldsymbol{X}_0] + \mathcal{P}_\Omega[\boldsymbol{Z}]$, where Ω is a subset of $[n] \times [n]$. If Ω is uniformly sampled from subsets of size*

$$m \geq C_1 \nu n r \log^2(n), \tag{4.4.74}$$

then with high probability, the optimal solution $\hat{\boldsymbol{X}}$ to the convex program (4.4.73) satisfies

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_F \leq c \frac{n\sqrt{n}\log(n)}{\sqrt{m}}\delta \leq c' \frac{n}{\sqrt{r}}\delta \tag{4.4.75}$$

for some constant $c > 0$.

Proof. Similar to the proof of Theorem 4.4.3 in the noiseless case which has the same incoherence condition on $\boldsymbol{X}_0$ and the sampling condition, we know the sampling operator $\mathcal{P}_\Omega$ and the dual certificate $\boldsymbol{\Lambda}_k$ constructed via the golfing scheme satisfies the properties in Proposition 4.4.5. All we

need to show here is that these properties also imply the conclusion of this theorem for the case with noisy measurements.

Let $\boldsymbol{H} = \hat{\boldsymbol{X}} - \boldsymbol{X}_0$. Notice that we can split $\boldsymbol{H}$ into two parts $\boldsymbol{H} = \mathcal{P}_\Omega[\boldsymbol{H}] + \mathcal{P}_{\Omega^c}[\boldsymbol{H}]$. For the first part, we have

$$\begin{aligned} \|\mathcal{P}_\Omega[\boldsymbol{H}]\|_F &= \|\mathcal{P}_\Omega[\hat{\boldsymbol{X}} - \boldsymbol{X}_0]\|_F \\ &\leq \|\mathcal{P}_\Omega[\hat{\boldsymbol{X}} - \boldsymbol{Y}]\|_F + \|\mathcal{P}_\Omega[\boldsymbol{Y} - \boldsymbol{X}_0]\|_F \\ &\leq 2\delta. \end{aligned} \tag{4.4.76}$$

Notice that the second part $\mathcal{P}_{\Omega^c}[\boldsymbol{H}]$ is a feasible perturbation to the noiseless matrix completion problem. From the proof of Proposition 4.4.5 and in particular (4.4.34), we have

$$\|\boldsymbol{X}_0 + \mathcal{P}_{\Omega^c}[\boldsymbol{H}]\|_* \geq \|\boldsymbol{X}_0\|_* + \left(\frac{1}{2} - \frac{1}{4C_2\sqrt{nr}}\right) \|\mathcal{P}_{T^\perp}[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]\|_F \tag{4.4.77}$$

and based on triangle inequality, we also have

$$\|\hat{\boldsymbol{X}}\|_* = \|\boldsymbol{X}_0 + \boldsymbol{H}\|_* \geq \|\boldsymbol{X}_0 + \mathcal{P}_{\Omega^c}[\boldsymbol{H}]\|_* - \|\mathcal{P}_\Omega[\boldsymbol{H}]\|_* . \tag{4.4.78}$$

Since $\|\hat{\boldsymbol{X}}\|_* \leq \|\boldsymbol{X}_0\|_*$, we have

$$\|\mathcal{P}_\Omega[\boldsymbol{H}]\|_* \geq \left(\frac{1}{2} - \frac{1}{4C_2\sqrt{nr}}\right) \|\mathcal{P}_{T^\perp}[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]\|_F . \tag{4.4.79}$$

This leads to

$$\|\mathcal{P}_{T^\perp}[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]\|_F \leq 4\,\|\mathcal{P}_\Omega[\boldsymbol{H}]\|_* \leq 4\sqrt{n}\,\|\mathcal{P}_\Omega[\boldsymbol{H}]\|_F \leq 4\sqrt{n}\delta. \tag{4.4.80}$$

Since $\mathcal{P}_{\Omega^c}[\boldsymbol{H}] = \mathcal{P}_{T^\perp}[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]] + \mathcal{P}_T[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]$, we remain to bound the term $\mathcal{P}_T[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]$. Applying the proof of Lemma 4.4.6 to $\mathcal{P}_{\Omega^c}[\boldsymbol{H}]$, we have $\|\mathcal{P}_{T^\perp}[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]\|_F \geq C_1 \frac{\sqrt{m}}{n\log(n)} \|\mathcal{P}_T[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]\|_F$ for some large enough C_1. Therefore, we have

$$\|\mathcal{P}_T[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]\|_F \leq \frac{n\log(n)}{C_1\sqrt{m}} \|\mathcal{P}_{T^\perp}[\mathcal{P}_{\Omega^c}[\boldsymbol{H}]]\|_F \leq c\frac{n\sqrt{n}\log(n)}{\sqrt{m}}\delta. \tag{4.4.81}$$

This bound dominates the bounds of all the other terms, leading to the conclusion of the theorem. □

4.5 Summary

In this chapter, we have studied the problem of recovering a low rank matrix from incomplete observations. This problem arises in a range of applications. It generalizes in a natural way the problem of recovering a sparse vector. We described a convex relaxation of the low rank recovery problem, in which we minimize the nuclear norm, which is the sum (ℓ^1 norm) of the singular values of a matrix. We proved that roughly $(n_1+n_2)r$ generic measurements suffice to recover all rank-r matrices. We also studied a specific,

more structured measurement model, in which we observe a subset of the entries of a matrix. This matrix completion problem captures the structure of many of the most important practical low rank recovery problems. It is mathematically more challenging, because certain very sparse low rank matrices cannot be completed without seeing almost all of their entries. Nevertheless, we saw that for low rank matrices whose singular vectors are not too concentrated on any coordinate, nuclear norm minimization indeed succeeds. Paralleling our development for sparse vectors, we showed that these algorithmic and theoretical results can be extended to cope with non-idealities, such as measurement noise. Moreover, in the next chapter we will see how these ideas combine naturally with ideas from sparse recovery to generate even richer classes of models and more robust algorithms.

4.6 Exercises

4.1 (Proof of Schoenberg's Theorem)**.** *In this exercise, we invite the interested reader to prove Schoenberg's Euclidean embedding theorem (Theorem 4.1.1). Let $\boldsymbol{D}$ be a Euclidean distance matrix for some point set $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$, i.e., $\boldsymbol{D}_{ij} = \|\boldsymbol{x}_i\|_2^2 + \|\boldsymbol{x}_j\|_2^2 - 2\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle$. Let $\mathbf{1} \in \mathbb{R}^n$ denote the vector of all ones, and $\boldsymbol{\Phi} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^*$. Using that $\boldsymbol{\Phi}\mathbf{1} = \mathbf{0}$, argue that $\boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Phi}^*$ satisfies the conditions of Schoenberg's theorem, i.e., it is negative semidefinite and has rank at most d.*

For the converse, let $\boldsymbol{D}$ be a symmetric matrix with zero diagonal, and suppose that $\boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Phi}^*$ is negative semidefinite and has rank at most d. Argue that there exists some matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ for which $\boldsymbol{D}_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$.

4.2 (Derivation of the SVD)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}$ be a matrix of rank r. Argue that there exists matrices $\boldsymbol{U} \in \mathbb{R}^{n_1\times r}$, $\boldsymbol{V} \in \mathbb{R}^{n_2\times r}$, with orthonormal columns and a diagonal matrix $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$, with $\sigma_1 \geq \cdots \geq \sigma_r > 0$, such that*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*. \tag{4.6.1}$$

Hint: what is the relationship between the singular values σ_i and singular vectors $\boldsymbol{v}_i$ and the eigenvalues / eigenvectors of the matrix $\boldsymbol{X}^*\boldsymbol{X}$?

4.3 (Best Rank-r Approximation)**.** *We prove Theorem 4.2.4. First, consider the special case in which $\boldsymbol{Y} = \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 > \sigma_2 > \cdots > \sigma_n$. An arbitrary rank-$r$ matrix $\boldsymbol{X}$ can be expressed as $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{G}^*$ with $\boldsymbol{F} \in \mathbb{R}^{n_1\times r}$, $\boldsymbol{F}^*\boldsymbol{F} = \boldsymbol{I}$ and $\boldsymbol{G} \in \mathbb{R}^{n_2\times r}$.*

1. *Argue that for any fixed $\boldsymbol{F}$, the solution to the optimization problem*

$$\min_{\boldsymbol{G}\in\mathbb{R}^{n_2\times r}} \|\boldsymbol{F}\boldsymbol{G}^* - \boldsymbol{\Sigma}\|_F^2 \tag{4.6.2}$$

is given by $\hat{\boldsymbol{G}} = \boldsymbol{\Sigma}^* \boldsymbol{F}$*, and the optimal cost is*

$$\|(\boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^*)\boldsymbol{\Sigma}\|_F^2 . \tag{4.6.3}$$

2. *Let* $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^*$*, and write* $\nu_i = \|\boldsymbol{P}\boldsymbol{e}_i\|_2^2$*. Argue that* $\sum_{i=1}^n \nu_i = n_1 - r$ *and* $\nu_i \in [0,1]$*. Conclude that*

$$\|\boldsymbol{P}\boldsymbol{\Sigma}\|_F^2 = \sum_{i=1}^{n_1} \sigma^2 \nu_i \geq \sum_{i=r+1}^{n_1} \sigma_i^2, \tag{4.6.4}$$

with equality if and only if $\nu_1 = \nu_2 = \cdots = \nu_r = 0$ *and* $\nu_{r+1} = \cdots = \nu_n$*. Conclude that Theorem 4.2.4 holds in the special case* $\boldsymbol{Y} = \boldsymbol{\Sigma}$*.*

3. *Extend your argument to the situation in which the* σ_i *are not distinct (i.e.,* $\sigma_i = \sigma_{i+1}$ *for some* i*).*

4. *Extend your argument to any* $\boldsymbol{Y} \in \mathbb{R}^{n \times n}$*. Hint: use the fact that the Frobenius norm* $\|\boldsymbol{M}\|_F$ *is unchanged by orthogonal transformations of the rows and columns:* $\|\boldsymbol{M}\|_F = \|\boldsymbol{R}\boldsymbol{M}\boldsymbol{S}\|_F$ *for any orthogonal matrices* $\boldsymbol{R}, \boldsymbol{S}$*.*

4.4 (Minimal Rank Approximation)**.** *We consider a variant of Theorem 4.2.4 in which we are given a data matrix* $\boldsymbol{Y}$ *and we want to find a matrix* $\boldsymbol{X}$ *of minimum rank that approximates* $\boldsymbol{Y}$ *up to some given fidelity:*

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{X}), \\ \text{subject to} & \|\boldsymbol{X} - \boldsymbol{Y}\|_F \leq \varepsilon. \end{array} \tag{4.6.5}$$

Give an expression for the optimal solution(s) to this problem, in terms of the SVD of $\boldsymbol{Y}$*. Prove that your expression is correct.*

4.5 (Multiple and repeated eigenvalues)**.** *Consider the eigenvector problem*

$$\textit{minimize} \quad -\tfrac{1}{2}\boldsymbol{q}^*\boldsymbol{\Gamma}\boldsymbol{q} \quad \textit{subject to} \quad \|\boldsymbol{q}\|_2^2 = 1, \tag{4.6.6}$$

where $\boldsymbol{\Gamma}$ *is a symmetric matrix. In the text, we argued that when the eigenvalues of* $\boldsymbol{\Gamma}$ *are distinct, every local minimizer of this problem is global. (i) Argue that even when* $\boldsymbol{\Gamma}$ *has repeated eigenvalues, every local minimum of this problem is global. (ii) Now suppose we wish to find multiple eigenvector/eigenvalue pairs. Consider the optimization problem*

$$\begin{array}{ll} \text{minimize} & -\frac{1}{2}\boldsymbol{Q}^*\boldsymbol{\Gamma}\boldsymbol{Q} \\ \text{subject to} & \boldsymbol{Q} \in \mathrm{St}(n,p) = \{\boldsymbol{Q} \in \mathbb{R}^{n \times p} \mid \boldsymbol{Q}^*\boldsymbol{Q} = \boldsymbol{I}\}. \end{array} \tag{4.6.7}$$

Argue that every local minimizer of this problem has the form

$$\boldsymbol{Q} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p]\,\boldsymbol{\Pi}, \tag{4.6.8}$$

where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ *are eigenvectors of* $\boldsymbol{\Gamma}$ *associated with the* p *largest eigenvalues, and* $\boldsymbol{\Pi}$ *is a permutation matrix.*

4.6 (The Power Method)**.** *We show how to compute eigenvectors (and hence singular vectors) using the power method. Let* $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times n}$ *be a symmetric positive semidefinite matrix. Let* $\boldsymbol{q}^{(0)}$ *be a random vector that is*

uniformly distributed on the sphere $\mathbb{S}^{n-1}$ *(we can generate such a random vector by taking an* n*-dimensional iid* $\mathcal{N}(0,1)$ *vector and then normalizing it to have unit* ℓ^2 *norm). Generate a sequence of vectors* $\boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)}, \ldots$ *via the iteration*

$$\boldsymbol{q}^{(k+1)} = \frac{\boldsymbol{\Gamma}\boldsymbol{q}^{(k)}}{\left\|\boldsymbol{\Gamma}\boldsymbol{q}^{(k)}\right\|_2}. \tag{4.6.9}$$

This iteration is called the power method.

Suppose that there is a gap between the first and second eigenvalues of $\boldsymbol{\Gamma}$*:* $\lambda_1(\boldsymbol{\Gamma}) > \lambda_2(\boldsymbol{\Gamma})$.

1. *What does* $\boldsymbol{q}^{(k)}$ *converge to? Hint: write* $\boldsymbol{\Gamma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*$ *in terms of its eigenvectors/values. How does* $\boldsymbol{V}^*\boldsymbol{q}^{(k)}$ *evolve?*
2. *Obtain a bound on the error* $\|\boldsymbol{q}^{(k)} - \boldsymbol{q}^{(\infty)}\|_2$ *in terms of the spectral gap* $\frac{\lambda_1 - \lambda_2}{\lambda_1}$.
3. *Your bound in (ii) should suggest that as long as there is a gap between* λ_1 *and* λ_2*, the power method converges rapidly. How does the method behave if* $\lambda_1 = \lambda_2$*?*
4. *How can we use the power method to compute the singular values of a matrix* $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$*?*

4.7 (Convex Envelope Property via the Bidual)**.** *In Theorem 4.3.2, we proved that the nuclear norm* $\|\boldsymbol{X}\|_*$ *is the convex envelope of* $\operatorname{rank}(\boldsymbol{X})$ *over the operator norm ball* $\mathsf{B}_{\text{op}} = \{\boldsymbol{X} \mid \|\boldsymbol{X}\| \leq 1\}$*. Here, we give an alternative derivation of this result, using the fact that the* biconjugate *of a function over a set* B *is the convex envelope. Let* $f(\boldsymbol{X}) = \operatorname{rank}(\boldsymbol{X})$ *denote the rank function.*

1. *Prove that the Fenchel* dual

$$f^*(\boldsymbol{Y}) = \sup_{\boldsymbol{X} \in \mathsf{B}} \{\langle \boldsymbol{X}, \boldsymbol{Y} \rangle - f(\boldsymbol{X})\}$$

can be expressed as

$$f^*(\boldsymbol{Y}) = \|\mathcal{D}_1[\boldsymbol{Y}]\|_*,$$

where $\mathcal{D}_\tau[\boldsymbol{M}]$ *is the singular value thresholding operator, given by* $\mathcal{D}_\tau[\boldsymbol{M}] = \boldsymbol{U}\mathcal{S}_\tau[\boldsymbol{S}]\boldsymbol{V}^*$ *for any singular value decomposition* $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^*$ *of* $\boldsymbol{M}$.

2. *Prove that the dual of* f^**,*

$$f^{**}(\boldsymbol{X}) = \sup_{\boldsymbol{Y}} \langle \boldsymbol{X}, \boldsymbol{Y} \rangle - f^*(\boldsymbol{Y})$$

can satisfies

$$f^{**}(\boldsymbol{X}) = \|\boldsymbol{X}\|_*.$$

3. *Use Proposition B.2.10 of Appendix B to conclude that* $\|\cdot\|_*$ *is the convex envelope of* $\operatorname{rank}(\cdot)$ *over* B.

4.8 (Convexifying Low-rank Approximation)**.** *Consider the following optimization problem:*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{\Pi}\boldsymbol{Y}\|_F^2 \\ \text{subject to} & \boldsymbol{0} \preceq \boldsymbol{\Pi} \preceq \boldsymbol{I},\ \operatorname{trace}[\boldsymbol{\Pi}] = m - r. \end{array} \tag{4.6.10}$$

Prove that if $\sigma_r(\boldsymbol{Y}) > \sigma_{r+1}(\boldsymbol{Y})$*, this problem has a unique optimal solution* $\boldsymbol{\Pi}_\star$*, which is the orthoprojector onto the linear span of the* $n_1 - r$ *trailing singular vectors* $\boldsymbol{u}_{r+1}, \boldsymbol{u}_{r+2}, \ldots, \boldsymbol{u}_{n_1}$*. The matrix* $(\boldsymbol{I} - \boldsymbol{\Pi}_\star)\boldsymbol{Y}$ *is the best rank-*r *approximation to* $\boldsymbol{Y}$*.*

4.9 (Tangent Space to the Rank-r Matrices)**.** *Consider a matrix* $\boldsymbol{X}_0$ *of rank* r *with compact singular value decomposition* $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$*. Argue that the* tangent space *to the collection* $\mathsf{R}_r = \{\boldsymbol{X} \mid \operatorname{rank}(\boldsymbol{X}) = r\}$ *at* $\boldsymbol{X}_0$ *is given by* $T = \{\boldsymbol{U}\boldsymbol{W}^* + \boldsymbol{H}\boldsymbol{V}^*\}$*. Hint: consider generating a nearby low-rank matrix by writing* $\boldsymbol{X}' = (\boldsymbol{U} + \boldsymbol{\Delta}_{\boldsymbol{U}})(\boldsymbol{\Sigma} + \boldsymbol{\Delta}_{\boldsymbol{\Sigma}})(\boldsymbol{V} + \boldsymbol{\Delta}_{\boldsymbol{V}})^*$*.*

4.10 (Quadratic Measurements)**.** *Consider a target vector* $\boldsymbol{x}_0 \in \mathbb{R}^{n \times n}$*. In many applications, the observation can be modeled as a* quadratic *function of the vector* $\boldsymbol{x}_0$*. In notation, we see the squares* $y_1 = \langle \boldsymbol{a}_1, \boldsymbol{x}_0 \rangle^2$*,* $y_2 = \langle \boldsymbol{a}_2, \boldsymbol{x}_0 \rangle^2$*,* $\ldots$*,* $y_m = \langle \boldsymbol{a}_m, \boldsymbol{x}_0 \rangle^2$ *of the projections of* $\boldsymbol{x}_0$ *onto vectors* $\boldsymbol{a}_1 \ldots \boldsymbol{a}_m$*. Notice that from this observation, it is only possible to reconstruct* $\boldsymbol{x}_0$ *up to a sign ambiguity:* $-\boldsymbol{x}_0$ *produces exactly the same observation.*

1. *Consider the quadratic problem*

$$\min_{\boldsymbol{x}} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle^2 \right)^2. \tag{4.6.11}$$

   Is this problem convex in $\boldsymbol{x}$*?*

2. *Convert this to a convex problem, by replacing the vector valued variable* $\boldsymbol{x}$ *with a matrix valued variable* $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*$*: convert the problem to*

$$\min_{\boldsymbol{X}} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle \right)^2. \tag{4.6.12}$$

   How should we choose the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$*? Show that if* $m < n^2$*,* $\boldsymbol{X}_0 = \boldsymbol{x}_0\boldsymbol{x}_0^*$ *is* not *the unique optimal solution to this problem. How can we use the fact that* $\operatorname{rank}(\boldsymbol{X}_0) = 1$ *to improve this?*

3. *In the absence of noise, we can attempt to solve for* $\boldsymbol{X}_0$ *by solving the convex program*

$$\min \|\boldsymbol{X}\|_* \quad \textit{such that} \quad \mathcal{A}[\boldsymbol{X}] = \boldsymbol{y}. \tag{4.6.13}$$

   Implement this optimization using a custom algorithm or CVX. Does it typically recover $\boldsymbol{X}_0$*?*

4. *Does the operator $\mathcal{A}$ satisfy the rank RIP?*

4.11 (Proof of Theorem 4.4.2)**.** *We prove Theorem 4.4.2. The goal here is to show that the solution to*

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_* + \tfrac{1}{2}\|\boldsymbol{X} - \boldsymbol{M}\|_F^2 \tag{4.6.14}$$

is given by $\mathcal{D}_1[\boldsymbol{M}]$

1. *Argue that Problem (4.6.14) is strongly convex, and hence has a unique optimal solution.*
2. *Show that a solution $\boldsymbol{X}_\star$ is optimal if and only if $\boldsymbol{X}_\star \in \boldsymbol{M} - \partial\|\cdot\|_*(\boldsymbol{X}_\star)$.*
3. *Using the condition from part (ii), show that if $\boldsymbol{M}$ is diagonal, i.e., $\boldsymbol{M}_{ij} = 0$ for $i \neq j$, then $\mathcal{S}_1[\boldsymbol{M}]$ is the unique optimal solution to (4.6.14).*
4. *Use the SVD to argue that in general, $\mathcal{D}_1[\boldsymbol{M}]$ is the unique optimal solution to (4.6.14).*

4.12 (Uniform Matrix Completion?)**.** *Let Ω be a strict subset of $[n] \times [n]$. Show that there exist two matrix $\boldsymbol{X}_0$ and $\boldsymbol{X}_0'$ of rank one such that $\mathcal{P}_\Omega[\boldsymbol{X}_0] = \mathcal{P}_\Omega[\boldsymbol{X}_0']$.* The implication of this is that it is not possible to reconstruct *all* low-rank matrices from the same observation Ω.

4.13 (Unique Optimality for Matrix Completion)**.** *Consider the optimization problem*

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{P}_\Omega[\boldsymbol{X}_0]. \end{array} \tag{4.6.15}$$

Suppose that $\|\mathcal{P}_{\Omega^c}\mathcal{P}_T\| < 1$. Assume we can find some $\boldsymbol{\Lambda}$ such that

1. *$\boldsymbol{\Lambda}$ is supported on Ω and*
2. *$\boldsymbol{\Lambda} \in \partial\|\cdot\|_*(\boldsymbol{X}_0)$ – i.e., $\mathcal{P}_T[\boldsymbol{\Lambda}] = \boldsymbol{U}\boldsymbol{V}^*$ and $\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}]\| < 1$.*

Show that $\boldsymbol{X}_0$ is the unique optimal solution to the optimization problem.

4.14. *Prove Theorem 4.3.5.*

4.15. *Prove Theorem 4.3.12*

4.16. *Fill in the detailed steps of proof for Theorem 4.3.15.*

4.17. *Derive detailed steps that prove the error bound (4.3.83) in the proof of Theorem 4.3.13.*

4.18. *Show that in Lemma 4.4.4, any subdifferential of nuclear norm must be of the form given in (4.4.21).*

4.19. *Let $\mathcal{R}_\Omega[\boldsymbol{X}_0] = \sum_{\ell=1}^{q} [\boldsymbol{X}_0]_{i_\ell,j_\ell} \boldsymbol{e}_{i_\ell}\boldsymbol{e}_{j_\ell}^*$ with each (i_ℓ, j_ℓ) chosen iid at random from the uniform distribution on $[n]\times[n]$. Use the matrix Bernstein inequality to show that if $q > C\nu nr \log n$ for sufficiently large C, we have*

$$\left\| \mathcal{P}_{T^\perp} \tfrac{n^2}{q} \mathcal{R}_\Omega \mathcal{P}_T \right\| \le t. \tag{4.6.16}$$

for any arbitrarily small constant t with high probability. [Hint: similar to the proof of Lemma 4.4.7.]

4.20. *For the dual certificate $\boldsymbol{\Lambda}$ constructed from the golfing scheme, use the fact in Exercise 4.19 and the fact that $\left\| \frac{n^2}{q} \mathcal{P}_{T^\perp} \mathcal{R}_{\Omega_j}[\boldsymbol{E}_j] \right\|_F \le \left\| \frac{n^2}{q} \mathcal{P}_{T^\perp} \mathcal{R}_{\Omega_j} \mathcal{P}_T \right\| \|\boldsymbol{E}_j\|_F$, show that if*

$$m > C\nu nr^2 \log^2 n$$

for a large enough constant C, we have $\|\mathcal{P}_{T^\perp}[\boldsymbol{\Lambda}]\| \le 1/2$ with high probability.

4.21. *Prove Lemma 4.4.9. Hint: write:*

$$(q^{-1}\mathcal{P}_\Omega - \mathcal{I})\boldsymbol{Z} = \sum_{ij} \underbrace{Z_{ij}\left(q^{-1}\mathbb{1}_{ij\in\Omega} - 1\right)\boldsymbol{E}_{ij}}_{\doteq \boldsymbol{W}_{ij}},$$

and apply the operator Bernstein inequality, controlling the operator norm of $\boldsymbol{W}_{ij}$ in terms of $\|\boldsymbol{Z}\|_\infty$ and controlling the matrix variance in terms of $\|\boldsymbol{Z}\|_{rc}$.

4.22. *Prove Lemma 4.4.10. Use the matrix Bernstein inequality to obtain a bound on the probability that the ℓ-th row $\|\boldsymbol{e}_\ell^*\left(q^{-1}\mathcal{P}_T\mathcal{P}_\Omega - \mathcal{P}_T\right)\boldsymbol{Z}\|$ is large, repeat for each column, and then sum the failure probabilities over all rows and columns to obtain a bound on the probability that the $\|\cdot\|_{rc}$ is large. Hint: apply the matrix Bernstein inequality to the random vector:*

$$\boldsymbol{e}_\ell^*\left(q^{-1}\mathcal{P}_T\mathcal{P}_\Omega - \mathcal{P}_T\right)\boldsymbol{Z} = \sum_{ij} \underbrace{Z_{ij}(q^{-1}\mathbb{1}_{ij\in\Omega} - 1)\boldsymbol{e}_\ell^*\mathcal{P}_T[\boldsymbol{E}_{ij}]}_{\doteq \boldsymbol{W}_{ij}}.$$

4.23. *Prove Lemma 4.4.11. Apply the standard Bernstein inequality to bound the probability that the k,l entry of $(\mathcal{P}_T - q^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T)\boldsymbol{Z}$ is large, and then sum this probability over all entries k,l to bound the probability that the ℓ^∞ norm is large. For the k,l entry work with the sum of independent random variables*

$$\begin{aligned}\left[\left(\mathcal{P}_T - q^{-1}\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T\right)\boldsymbol{Z}\right]_{kl} &= Z_{kl} - [q^{-1}\mathcal{P}_T\mathcal{P}_\Omega\boldsymbol{Z}]_{kl} \\ &= \sum_{ij} \underbrace{n^{-2}Z_{kl} - q^{-1}\mathbb{1}_{ij\in\Omega}\left\langle \mathcal{P}_T\boldsymbol{E}_{kl}, \mathcal{P}_T\boldsymbol{E}_{ij}\right\rangle Z_{ij}}_{\doteq \boldsymbol{W}_{ij}}.\end{aligned}$$