10

Matrix estimation with rank constraints

In Chapter 8, we discussed the problem of principal component analysis, which can be understood as a particular type of low-rank estimation problem. In this chapter, we turn to other classes of matrix problems involving rank and other related constraints. We show how the general theory of Chapter 9 can be brought to bear in a direct way so as to obtain theoretical guarantees for estimators based on nuclear norm regularization, as well as various extensions thereof, including methods for additive matrix decomposition.

10.1 Matrix regression and applications

In previous chapters, we have studied various forms of vector-based regression, including standard linear regression (Chapter 7) and extensions based on generalized linear models (Chapter 9). As suggested by its name, matrix regression is the natural generalization of such vector-based problems to the matrix setting. The analog of the Euclidean inner product on the matrix space $\mathbb{R}^{d_1 \times d_2}$ is the *trace inner product*

$$\langle\langle \mathbf{A}, \mathbf{B} \rangle\rangle := \operatorname{trace}(\mathbf{A}^{\mathrm{T}}\mathbf{B}) = \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} A_{j_1 j_2} B_{j_1 j_2}. \tag{10.1}$$

This inner product induces the *Frobenius norm* $|||\mathbf{A}|||_{\mathrm{F}} = \sqrt{\sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} (A_{j_1 j_2})^2}$, which is simply the Euclidean norm on a vectorized version of the matrix.

In a matrix regression model, each observation takes the form $\mathbf{Z}_i = (\mathbf{X}_i, y_i)$, where $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ is a matrix of covariates, and $y_i \in \mathbb{R}$ is a response variable. As usual, the simplest case is the linear model, in which the response–covariate pair are linked via the equation

$$y_i = \langle\langle \mathbf{X}_i, \mathbf{\Theta}^* \rangle\rangle + w_i, \tag{10.2}$$

where w_i is some type of noise variable. We can also write this observation model in a more compact form by defining the *observation operator* $\mathfrak{X}_n \colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$ with elements $[\mathfrak{X}_n(\mathbf{\Theta})]_i = \langle\langle \mathbf{X}_i, \mathbf{\Theta} \rangle\rangle$, and then writing

$$y = \mathfrak{X}_n(\mathbf{\Theta}^*) + w, \tag{10.3}$$

where $y \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$ are the vectors of response and noise variables, respectively. The adjoint of the observation operator, denoted $\mathfrak{X}_n^*$, is the linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^{d_1 \times d_2}$ given by $u \mapsto \sum_{i=1}^n u_i \mathbf{X}_i$. Note that the operator $\mathfrak{X}_n$ is the natural generalization of the design matrix $\mathbf{X}$, viewed as a mapping from $\mathbb{R}^d$ to $\mathbb{R}^n$ in the usual setting of vector regression.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

As illustrated by the examples to follow, there are many applications in which the regression matrix $\mathbf{\Theta}^*$ is either low-rank, or well approximated by a low-rank matrix. Thus, if we were to disregard computational costs, an appropriate estimator would be a rank-penalized form of least squares. However, including a rank penalty makes this a non-convex form of least squares so that—apart from certain special cases—it is computationally difficult to solve. This obstacle motivates replacing the rank penalty with the nuclear norm, which leads to the convex program

$$\widehat{\mathbf{\Theta}} \in \arg\min_{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \|y - \mathfrak{X}_n(\mathbf{\Theta})\|_2^2 + \lambda_n |||\mathbf{\Theta}|||_{\text{nuc}} \right\}. \tag{10.4}$$

Recall that the nuclear norm of $\mathbf{\Theta}$ is given by the sum of its singular values—namely,

$$|||\mathbf{\Theta}|||_{\text{nuc}} = \sum_{j=1}^{d'} \sigma_j(\mathbf{\Theta}), \quad \text{where } d' = \min\{d_1, d_2\}. \tag{10.5}$$

See Example 9.8 for our earlier discussion of this matrix norm.

Let us illustrate these definitions with some examples, beginning with the problem of multivariate regression.

Example 10.1 (Multivariate regression as matrix regression) As previously introduced in Example 9.6, the multivariate regression observation model can be written as $\mathbf{Y} = \mathbf{Z}\mathbf{\Theta}^* + \mathbf{W}$, where $\mathbf{Z} \in \mathbb{R}^{n \times p}$ is the regression matrix, and $\mathbf{Y} \in \mathbb{R}^{n \times T}$ is the matrix of responses. The tth column $\Theta^*_{\bullet,t}$ of the $(p \times T)$-dimensional regression matrix $\mathbf{\Theta}^*$ can be thought of as an ordinary regression vector for the tth component of the response. In many applications, these vectors lie on or close to a low-dimensional subspace, which means that the matrix $\mathbf{\Theta}^*$ is low-rank, or well approximated by a low-rank matrix. A direct way of estimating $\mathbf{\Theta}^*$ would be via *reduced rank regression*, in which one minimizes the usual least-squares cost $|||\mathbf{Y} - \mathbf{Z}\mathbf{\Theta}|||_{\text{F}}^2$ while imposing a rank constraint directly on the regression matrix $\mathbf{\Theta}$. Even though this problem is non-convex due to the rank constraint, it is easily solvable in this special case; see the bibliographic section and Exercise 10.1 for further details. However, this ease of solution is very fragile and will no longer hold if other constraints, in addition to a bounded rank, are added. In such cases, it can be useful to apply nuclear norm regularization in order to impose a "soft" rank constraint.

Multivariate regression can be recast as a form of the matrix regression model (10.2) with $N = nT$ observations in total. For each $j = 1, \ldots, n$ and $\ell = 1, \ldots, T$, let $\mathbf{E}_{j\ell}$ be an $n \times T$ mask matrix, with zeros everywhere except for a one in position (j, ℓ). If we then define the matrix $\mathbf{X}_{j\ell} := \mathbf{Z}^{\mathrm{T}}\mathbf{E}_{j\ell} \in \mathbb{R}^{p \times T}$, the multivariate regression model is based on the $N = nT$ observations $(\mathbf{X}_{j\ell}, y_{j\ell})$, each of the form

$$y_{j\ell} = \langle\langle \mathbf{X}_{j\ell}, \mathbf{\Theta}^* \rangle\rangle + W_{j\ell}, \qquad \text{for } j = 1, \ldots, n \text{ and } \ell = 1, \ldots, T.$$

Consequently, multivariate regression can be analyzed via the general theory that we develop for matrix regression problems. ♣

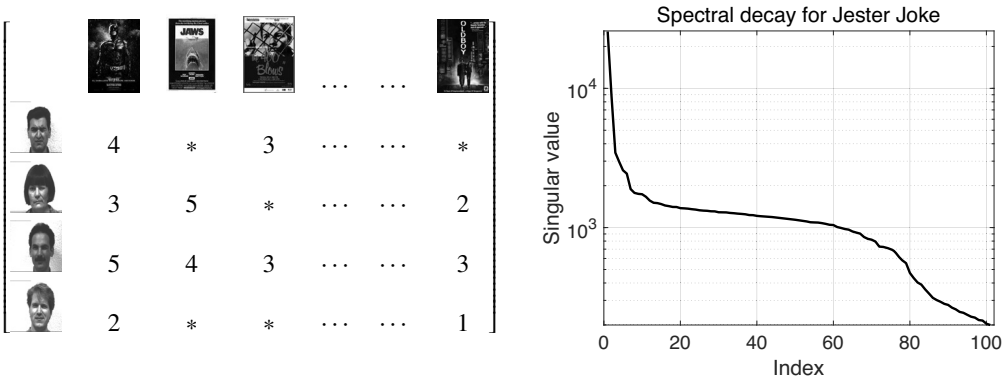
Another example of matrix regression is the problem of matrix completion.

Example 10.2 (Low-rank matrix completion) Matrix completion refers to the problem of

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

estimating an unknown matrix $\mathbf{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ based on (noisy) observations of a subset of its entries. Of course, this problem is ill-posed unless further structure is imposed, and so there are various types of matrix completion problems, depending on this underlying structure. One possibility is that the unknown matrix has a low-rank, or more generally can be well approximated by a low-rank matrix.

As one motivating application, let us consider the "Netflix problem", in which the rows of $\mathbf{\Theta}^*$ correspond to people, and columns correspond to movies. Matrix entry $\Theta^*_{a,b}$ represents the rating assigned by person a (say "Alice") to a given movie b that she has seen. In this setting, the goal of matrix completion is to make recommendations to Alice—that is, to suggest other movies that she has not yet seen but would be to likely to rate highly. Given the large corpus of movies stored by Netflix, most entries of the matrix $\mathbf{\Theta}^*$ are unobserved, since any given individual can only watch a limited number of movies over his/her lifetime. Consequently, this problem is ill-posed without further structure. See Figure 10.1(a) for an illustration of this observation model. Empirically, if one computes the singular values of recommender matrices, such as those that arise in the Netflix problem, the singular value spectrum tends to exhibit a fairly rapid decay—although the matrix itself is not exactly low-rank, it can be well-approximated by a matrix of low rank. This phenomenon is illustrated for a portion of the Jester joke data set (Goldberg et al., 2001), in Figure 10.1(b).



(a) (b)

Figure 10.1 (a) Illustration of the Netflix problem. Each user (rows of the matrix) rates a subset of movies (columns of the matrix) on a scale of 1 to 5. All remaining entries of the matrix are unobserved (marked with $*$), and the goal of matrix completion is to fill in these missing entries. (b) Plot of the singular values for a portion of the Jester joke data set (Goldberg et al., 2001), corresponding to ratings of jokes on a scale of $[-10, 10]$, and available at `http://eigentaste.berkeley.edu/`. Although the matrix is not exactly low-rank, it can be well approximated by a low-rank matrix.

In this setting, various observation models are possible, with the simplest being that we are given noiseless observations of a subset of the entries of $\mathbf{\Theta}^*$. A slightly more realistic

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

model allows for noisiness—for instance, in the linear case, we might assume that

$$\widetilde{y}_i = \Theta_{a(i),b(i)} + \frac{w_i}{\sqrt{d_1 d_2}}, \tag{10.6}$$

where w_i is some form[1] of observation noise, and $(a(i), b(i))$ are the row and column indices of the ith observation.

How to reformulate the observations as an instance of matrix regression? For sample index i, define the mask matrix $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$, which is zero everywhere *except* for position $(a(i), b(i))$, where it takes the value $\sqrt{d_1 d_2}$. Then by defining the rescaled observation $y_i := \sqrt{d_1 d_2}\ \widetilde{y}_i$, the observation model can be written in the trace regression form as

$$y_i = \langle\!\langle \mathbf{X}_i,\ \mathbf{\Theta}^* \rangle\!\rangle + w_i. \tag{10.7}$$

We analyze this form of matrix completion in the sequel.

Often, matrices might take on discrete values, such as for yes/no votes coded in the set $\{-1, 1\}$, or ratings belonging to some subset of the positive integers (e.g., $\{1, \ldots, 5\}$), in which case a generalized version of the basic linear model (10.6) would be appropriate. For instance, in order to model binary-valued responses $y \in \{-1, 1\}$, it could be appropriate to use the logistic model

$$\mathbb{P}(y_i \mid \mathbf{X}_i, \mathbf{\Theta}^*) = \frac{e^{y \langle\!\langle \mathbf{X}_i,\, \mathbf{\Theta}^* \rangle\!\rangle}}{1 + e^{y_i \langle\!\langle \mathbf{X}_i,\, \mathbf{\Theta}^* \rangle\!\rangle}}. \tag{10.8}$$

In this context, the parameter $\Theta^*_{a,b}$ is proportional to the log-odds ratio for whether user a likes (or dislikes) item b. ♣

We now turn to the matrix analog of the compressed sensing observation model, originally discussed in Chapter 7 for vectors. It is another special case of the matrix regression problem.

Example 10.3 (Compressed sensing for low-rank matrices) Working with the linear observation model (10.3), suppose that the design matrices $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ are drawn i.i.d from a random Gaussian ensemble. In the simplest of settings, the design matrix is chosen from the standard Gaussian ensemble, meaning that each of its $D = d_1 d_2$ entries is an i.i.d. draw from the $\mathcal{N}(0, 1)$ distribution. In this case, the random operator $\mathfrak{X}_n$ provides n random projections of the unknown matrix $\mathbf{\Theta}^*$—namely

$$y_i = \langle\!\langle \mathbf{X}_i,\ \mathbf{\Theta}^* \rangle\!\rangle \qquad \text{for } i = 1, \ldots, n. \tag{10.9}$$

In this noiseless setting, it is natural to ask how many such observations suffice to recover $\mathbf{\Theta}^*$ exactly. We address this question in Corollary 10.9 to follow in the sequel. ♣

The problem of signal phase retrieval leads to a variant of the low-rank compressed sensing problem:

Example 10.4 (Phase retrieval) Let $\theta^* \in \mathbb{R}^d$ be an unknown vector, and suppose that we make measurements of the form $\widetilde{y}_i = \left|\langle x_i, \theta^* \rangle\right|$ where $x_i \sim \mathcal{N}(0, \mathbf{I}_d)$ is a standard normal vector. This set-up is a real-valued idealization of the problem of phase retrieval in image

[1] Our choice of normalization by $1/\sqrt{d_1 d_2}$ is for later theoretical convenience, as clarified in the sequel—see equation (10.36).

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

processing, in which we observe the magnitude of complex inner products, and want the retrieve the phase of the associated complex vector. In this idealized setting, the "phase" can take only two possible values, namely the possible signs of $\langle x_i, \theta^* \rangle$.

A standard semidefinite relaxation is based on lifting the observation model to the space of matrices. Taking squares on both sides yields the equivalent observation model

$$\widetilde{y}_i^2 = \big(\langle x_i, \theta^* \rangle \big)^2 = \langle\!\langle x_i \otimes x_i, \ \theta^* \otimes \theta^* \rangle\!\rangle \qquad \text{for } i = 1, \ldots, n,$$

where $\theta^* \otimes \theta^* = \theta^* (\theta^*)^{\mathrm{T}}$ is the rank-one outer product. By defining the scalar observation $y_i := \widetilde{y}_i^2$, as well as the matrices $\mathbf{X}_i := x_i \otimes x_i$ and $\mathbf{\Theta}^* := \theta^* \otimes \theta^*$, we obtain an equivalent version of the noiseless phase retrieval problem—namely, to find a rank-one solution to the set of matrix-linear equations $y_i = \langle\!\langle \mathbf{X}_i, \ \mathbf{\Theta}^* \rangle\!\rangle$ for $i = 1, \ldots, n$. This problem is non-convex, but by relaxing the rank constraint to a nuclear norm constraint, we obtain a tractable semidefinite program (see equation (10.29) to follow).

Overall, the phase retrieval problem is a variant of the compressed sensing problem from Example 10.3, in which the random design matrices $\mathbf{X}_i$ are no longer Gaussian, but rather the outer product $x_i \otimes x_i$ of two Gaussian vectors. In Corollary 10.13 to follow, we show that the solution of the semidefinite relaxation coincides with the rank-constrained problem with high probability given $n \gtrsim d$ observations. ♣

Matrix estimation problems also arise in modeling of time series, where the goal is to describe the dynamics of an underlying process.

Example 10.5 (Time-series and vector autoregressive processes) A vector autoregressive (VAR) process in d dimensions consists of a sequence of d-dimensional random vectors $\{z^t\}_{t=1}^N$ that are generated by first choosing the random vector $z^1 \in \mathbb{R}^d$ according to some initial distribution, and then recursively setting

$$z^{t+1} = \mathbf{\Theta}^* z^t + w^t, \qquad \text{for } t = 1, 2, \ldots, N-1. \tag{10.10}$$

Here the sequence of d-dimensional random vectors $\{w^t\}_{t=1}^{N-1}$ forms the driving noise of the process; we model them as i.i.d. zero-mean random vectors with covariance $\mathbf{\Gamma} \succ 0$. Of interest to us is the matrix $\mathbf{\Theta}^* \in \mathbb{R}^{d \times d}$ that controls the dependence between successive samples of the process. Assuming that w^t is independent of z^t for each t, the covariance matrix $\mathbf{\Sigma}^t = \mathrm{cov}(z^t)$ of the process evolves according to the recursion $\mathbf{\Sigma}^{t+1} := \mathbf{\Theta}^* \mathbf{\Sigma}^t (\mathbf{\Theta}^*)^{\mathrm{T}} + \mathbf{\Gamma}$. Whenever $|||\mathbf{\Theta}^*|||_2 < 1$, it can be shown that the process is stable, meaning that the eigenvalues of $\mathbf{\Sigma}^t$ stay bounded independently of t, and the sequence $\{\mathbf{\Sigma}^t\}_{t=1}^{\infty}$ converges to a well-defined limiting object. (See Exercise 10.2.)

Our goal is to estimate the system parameters, namely the d-dimensional matrices $\mathbf{\Theta}^*$ and $\mathbf{\Gamma}$. When the noise covariance $\mathbf{\Gamma}$ is known and strictly positive definite, one possible estimator for $\mathbf{\Theta}^*$ is based on a sum of quadratic losses over successive samples—namely,

$$\mathcal{L}_n(\mathbf{\Theta}) = \frac{1}{2N} \sum_{t=1}^{N-1} \|z^{t+1} - \mathbf{\Theta} z^t\|_{\mathbf{\Gamma}^{-1}}^2, \tag{10.11}$$

where $\|a\|_{\mathbf{\Gamma}^{-1}} := \sqrt{\langle a, \mathbf{\Gamma}^{-1} a \rangle}$ is the quadratic norm defined by $\mathbf{\Gamma}$. When the driving noise w^t is zero-mean Gaussian with covariance $\mathbf{\Gamma}$, then this cost function is equivalent to the negative log-likelihood, disregarding terms not depending on $\mathbf{\Theta}^*$.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

In many applications, among them subspace tracking and biomedical signal processing, the system matrix $\mathbf{\Theta}^*$ can be modeled as being low-rank, or well approximated by a low-rank matrix. In this case, the nuclear norm is again an appropriate choice of regularizer, and when combined with the loss function (10.11), we obtain another form of semidefinite program to solve.

Although different on the surface, this VAR observation model can be reformulated as a particular instance of the matrix regression model (10.2), in particular one with $n = d\,(N-1)$ observations in total. At each time $t = 2, \ldots, N$, we receive a total of d observations. Letting $e_j \in \mathbb{R}^d$ denote the canonical basis vector with a single one in position j, the jth observation in the block has the form

$$z_j^t = \left\langle e_j,\, z^t \right\rangle = \left\langle e_j,\, \mathbf{\Theta}^* z^{t-1} \right\rangle + w_j^{t-1} = \langle\!\langle e_j \otimes z^{t-1},\; \mathbf{\Theta}^* \rangle\!\rangle + w_j^{t-1},$$

so that in the matrix regression observation model (10.2), we have $y_i = (z_t)_j$ and $\mathbf{X}_i = e_j \otimes z^{t-1}$ when i indexes the sample (t, j). ♣

10.2 Analysis of nuclear norm regularization

Having motivated problems of low-rank matrix regression, we now turn to the development and analysis of M-estimators based on nuclear norm regularization. Our goal is to bring to bear the general theory from Chapter 9. This general theory requires specification of certain subspaces over which the regularizer decomposes, as well as restricted strong convexity conditions related to these subspaces. This section is devoted to the development of these two ingredients in the special case of nuclear norm (10.5).

10.2.1 Decomposability and subspaces

We begin by developing appropriate choices of decomposable subspaces for the nuclear norm. For any given matrix $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, we let $\operatorname{rowspan}(\mathbf{\Theta}) \subseteq \mathbb{R}^{d_2}$ and $\operatorname{colspan}(\mathbf{\Theta}) \subseteq \mathbb{R}^{d_1}$ denote its row space and column space, respectively. For a given positive integer $r \leq d' := \min\{d_1, d_2\}$, let $\mathbb{U}$ and $\mathbb{V}$ denote r-dimensional subspaces of vectors. We can then define the two subspaces of matrices

$$\mathbb{M}(\mathbb{U}, \mathbb{V}) := \{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{rowspan}(\mathbf{\Theta}) \subseteq \mathbb{V},\;\; \operatorname{colspan}(\mathbf{\Theta}) \subseteq \mathbb{U}\} \tag{10.12a}$$

and

$$\overline{\mathbb{M}}^{\perp}(\mathbb{U}, \mathbb{V}) := \{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{rowspan}(\mathbf{\Theta}) \subseteq \mathbb{V}^{\perp},\;\; \operatorname{colspan}(\mathbf{\Theta}) \subseteq \mathbb{U}^{\perp}\}. \tag{10.12b}$$

Here $\mathbb{U}^{\perp}$ and $\mathbb{V}^{\perp}$ denote the subspaces orthogonal to $\mathbb{U}$ and $\mathbb{V}$, respectively. When the subspaces $(\mathbb{U}, \mathbb{V})$ are clear from the context, we omit them so as to simplify notation. From the definition (10.12a), any matrix in the model space $\mathbb{M}$ has rank at most r. On the other hand, equation (10.12b) defines the subspace $\overline{\mathbb{M}}(\mathbb{U}, \mathbb{V})$ implicitly, via taking the orthogonal complement. We show momentarily that unlike other regularizers considered in Chapter 9, this definition implies that $\overline{\mathbb{M}}(\mathbb{U}, \mathbb{V})$ is a *strict superset* of $\mathbb{M}(\mathbb{U}, \mathbb{V})$.

To provide some intuition for the definition (10.12), it is helpful to consider an explicit matrix-based representation of the subspaces. Recalling that $d' = \min\{d_1, d_2\}$, let $\mathbf{U} \in \mathbb{R}^{d_1 \times d'}$

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

and $\mathbf{V} \in \mathbb{R}^{d_2 \times d'}$ be a pair of orthonormal matrices. These matrices can be used to define r-dimensional spaces: namely, let $\mathbb{U}$ be the span of the first r columns of $\mathbf{U}$, and similarly, let $\mathbb{V}$ be the span of the first r columns of $\mathbf{V}$. In practice, these subspaces correspond (respectively) to the spaces spanned by the top r left and right singular vectors of the target matrix $\boldsymbol{\Theta}^*$.

With these choices, any pair of matrices $\mathbf{A} \in \mathbb{M}(\mathbb{U}, \mathbb{V})$ and $\mathbf{B} \in \bar{\mathbb{M}}^{\perp}(\mathbb{U}, \mathbb{V})$ can be represented in the form

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \mathbf{0}_{r \times (d'-r)} \\ \mathbf{0}_{(d'-r) \times r} & \mathbf{0}_{(d'-r) \times (d'-r)} \end{bmatrix} \mathbf{V}^{\mathrm{T}} \quad \text{and} \quad \mathbf{B} = \mathbf{U} \begin{bmatrix} \mathbf{0}_{r \times r} & \mathbf{0}_{r \times (d'-r)} \\ \mathbf{0}_{(d'-r) \times r} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \mathbf{V}^{\mathrm{T}}, \tag{10.13}$$

where $\boldsymbol{\Gamma}_{11} \in \mathbb{R}^{r \times r}$ and $\boldsymbol{\Gamma}_{22} \in \mathbb{R}^{(d'-r) \times (d'-r)}$ are arbitrary matrices. Thus, we see that $\mathbb{M}$ corresponds to the subspace of matrices with non-zero left and right singular vectors contained within the span of first r columns of $\mathbf{U}$ and $\mathbf{V}$, respectively.

On the other hand, the set $\bar{\mathbb{M}}^{\perp}$ corresponds to the subspace of matrices with non-zero left and right singular vectors associated with the remaining $d' - r$ columns of $\mathbf{U}$ and $\mathbf{V}$. Since the trace inner product defines orthogonality, any member $\bar{\mathbf{A}}$ of $\bar{\mathbb{M}}(\mathbb{U}, \mathbb{V})$ must take the form

$$\bar{\mathbf{A}} = \mathbf{U} \begin{bmatrix} \bar{\boldsymbol{\Gamma}}_{11} & \bar{\boldsymbol{\Gamma}}_{12} \\ \bar{\boldsymbol{\Gamma}}_{21} & \mathbf{0} \end{bmatrix} \mathbf{V}^{\mathrm{T}}, \tag{10.14}$$

where all three matrices $\bar{\boldsymbol{\Gamma}}_{11} \in \mathbb{R}^{r \times r}$, $\bar{\boldsymbol{\Gamma}}_{12} \in \mathbb{R}^{r \times (d'-r)}$ and $\bar{\boldsymbol{\Gamma}}_{21} \in \mathbb{R}^{(d'-r) \times r}$ are arbitrary. In this way, we see explicitly that $\bar{\mathbb{M}}$ is a strict superset of $\mathbb{M}$ whenever $r < d'$. An important fact, however, is that $\bar{\mathbb{M}}$ is not substantially larger than $\mathbb{M}$. Whereas any matrix in $\mathbb{M}$ has rank at most r, the representation (10.14) shows that any matrix in $\bar{\mathbb{M}}$ has rank at most $2r$.

The preceding discussion also demonstrates the decomposability of the nuclear norm. Using the representation (10.13), for an arbitrary pair of matrices $\mathbf{A} \in \mathbb{M}$ and $\mathbf{B} \in \bar{\mathbb{M}}^{\perp}$, we have

$$\begin{aligned} |||\mathbf{A} + \mathbf{B}|||_{\mathrm{nuc}} &\overset{\text{(i)}}{=} \left|\left|\left| \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \right|\right|\right|_{\mathrm{nuc}} = \left|\left|\left| \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right|\right|\right|_{\mathrm{nuc}} + \left|\left|\left| \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \right|\right|\right|_{\mathrm{nuc}} \\ &\overset{\text{(ii)}}{=} |||\mathbf{A}|||_{\mathrm{nuc}} + |||\mathbf{B}|||_{\mathrm{nuc}}, \end{aligned}$$

where steps (i) and (ii) use the invariance of the nuclear norm to orthogonal transformations corresponding to multiplication by the matrices $\mathbf{U}$ or $\mathbf{V}$, respectively.

When the target matrix $\boldsymbol{\Theta}^*$ is of rank r, then the "best" choice of the model subspace (10.12a) is clear. In particular, the low-rank condition on $\boldsymbol{\Theta}^*$ means that it can be factored in the form $\boldsymbol{\Theta}^* = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$, where the diagonal matrix $\mathbf{D} \in \mathbb{R}^{d' \times d'}$ has the r non-zero singular values of $\boldsymbol{\Theta}^*$ in its first r diagonal entries. The matrices $\mathbf{U} \in \mathbb{R}^{d_1 \times d'}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times d'}$ are orthonormal, with their first r columns corresponding to the left and right singular vectors, respectively, of $\boldsymbol{\Theta}^*$. More generally, even when $\boldsymbol{\Theta}^*$ is not exactly of rank r, matrix subspaces of this form are useful: we simply choose the first r columns of $\mathbf{U}$ and $\mathbf{V}$ to index the singular vectors associated with the largest singular values of $\boldsymbol{\Theta}^*$, a subspace that we denote by $\mathbb{M}(\mathbb{U}^r, \mathbb{V}^r)$.

With these details in place, let us state for future reference a consequence of Proposition 9.13 for M-estimators involving the nuclear norm. Consider an M-estimator of the form

$$\widehat{\boldsymbol{\Theta}} \arg \min_{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \{\mathcal{L}_n(\boldsymbol{\Theta}) + \lambda_n |||\boldsymbol{\Theta}|||_{\mathrm{nuc}}\},$$

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

where $\mathcal{L}_n$ is some convex and differentiable cost function. Then for any choice of regularization parameter $\lambda_n \geq 2|||\nabla \mathcal{L}_n(\mathbf{\Theta}^*)|||_2$, the error matrix $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ must satisfy the cone-like constraint

$$|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}^\perp}|||_{\text{nuc}} \leq 3|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}}|||_{\text{nuc}} + 4|||\mathbf{\Theta}^*_{\mathbb{M}^\perp}|||_{\text{nuc}}, \tag{10.15}$$

where $\mathbb{M} = \mathbb{M}(\mathbb{U}^r, \mathbb{V}^r)$ and $\bar{\mathbb{M}} = \bar{\mathbb{M}}(\mathbb{U}^r, \mathbb{V}^r)$. Here the reader should recall that $\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}}$ denotes the projection of the matrix $\widehat{\mathbf{\Delta}}$ onto the subspace $\bar{\mathbb{M}}$, with the other terms defined similarly.

10.2.2 Restricted strong convexity and error bounds

We begin our exploration of nuclear norm regularization in the simplest setting, namely when it is coupled with a least-squares objective function. More specifically, given observations $(y, \mathfrak{X}_n)$ from the matrix regression model (10.3), consider the estimator

$$\widehat{\mathbf{\Theta}} \in \arg\min_{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \|y - \mathfrak{X}_n(\mathbf{\Theta})\|_2^2 + \lambda_n |||\mathbf{\Theta}|||_{\text{nuc}} \right\}, \tag{10.16}$$

where $\lambda_n > 0$ is a user-defined regularization parameter. As discussed in the previous section, the nuclear norm is a decomposable regularizer and the least-squares cost is convex, and so given a suitable choice of λ_n, the error matrix $\widehat{\mathbf{\Delta}} := \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ must satisfy the cone-like constraint (10.15).

The second ingredient of the general theory from Chapter 9 is restricted strong convexity of the loss function. For this least-squares cost, restricted strong convexity amounts to lower bounding the quadratic form $\frac{\|\mathfrak{X}_n(\mathbf{\Delta})\|_2^2}{2n}$. In the sequel, we show the random operator $\mathfrak{X}_n$ satisfies a uniform lower bound of the form

$$\frac{\|\mathfrak{X}_n(\mathbf{\Delta})\|_2^2}{2n} \geq \frac{\kappa}{2} |||\mathbf{\Delta}|||_{\text{F}}^2 - c_0 \frac{(d_1 + d_2)}{n} |||\mathbf{\Delta}|||_{\text{nuc}}^2, \qquad \text{for all } \mathbf{\Delta} \in \mathbb{R}^{d_1 \times d_2}, \tag{10.17}$$

with high probability. Here the quantity $\kappa > 0$ is a *curvature constant*, and c_0 is another universal constant of secondary importance. In the notation of Chapter 9, this lower bound implies a form of restricted strong convexity—in particular, see Definition 9.15—with curvature κ and tolerance $\tau_n^2 = c_0 \frac{(d_1+d_2)}{n}$. We then have the following corollary of Theorem 9.19:

Proposition 10.6 *Suppose that the observation operator $\mathfrak{X}_n$ satisfies the restricted strong convexity condition* (10.17) *with parameter $\kappa > 0$. Then conditioned on the event $\mathbb{G}(\lambda_n) = \{|||\frac{1}{n}\sum_{i=1}^n w_i \mathbf{X}_i|||_2 \leq \frac{\lambda_n}{2}\}$, any optimal solution to nuclear norm regularized least squares* (10.16) *satisfies the bound*

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\text{F}}^2 \leq \frac{9}{2}\frac{\lambda_n^2}{\kappa^2}\, r \;+\; \frac{1}{\kappa}\left\{ 2\lambda_n \sum_{j=r+1}^{d'} \sigma_j(\mathbf{\Theta}^*) + \frac{32 c_0 (d_1 + d_2)}{n} \Big[\sum_{j=r+1}^{d'} \sigma_j(\mathbf{\Theta}^*) \Big]^2 \right\}, \tag{10.18}$$

valid for any $r \in \{1, \ldots, d'\}$ such that $r \leq \frac{\kappa n}{128\, c_0\, (d_1+d_2)}$.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Remark: As with Theorem 9.19, the result of Proposition 10.6 is a type of *oracle inequality*: it applies to any matrix $\mathbf{\Theta}^*$, and involves a natural splitting into estimation and approximation error, parameterized by the choice of r. Note that the choice of r can be optimized so as to obtain the tightest possible bound.

The bound (10.18) takes a simpler form in special cases. For instance, suppose that $\operatorname{rank}(\mathbf{\Theta}^*) < d'$ and moreover that $n > 128\,\frac{c_0}{\kappa}\,\operatorname{rank}(\mathbf{\Theta}^*)\,(d_1 + d_2)$. We then may apply the bound (10.18) with $r = \operatorname{rank}(\mathbf{\Theta}^*)$. Since $\sum_{j=r+1}^{d'} \sigma_j(\mathbf{\Theta}^*) = 0$, Proposition 10.6 implies the upper bound

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}}^2 \leq \frac{9}{2}\frac{\lambda_n^2}{\kappa^2}\ \operatorname{rank}(\mathbf{\Theta}^*). \tag{10.19}$$

We make frequent use of this simpler bound in the sequel.

Proof For each $r \in \{1, \ldots, d'\}$, let $(\mathbb{U}^r, \mathbb{V}^r)$ be the subspaces spanned by the top r left and right singular vectors of $\mathbf{\Theta}^*$, and recall the subspaces $\mathbb{M}(\mathbb{U}^r, \mathbb{V}^r)$ and $\bar{\mathbb{M}}^{\perp}(\mathbb{U}^r, \mathbb{V}^r)$ previously defined in (10.12). As shown previously, the nuclear norm is decomposable with respect to any such subspace pair. In general, the "good" event from Chapter 9 is given by $\mathbb{G}(\lambda_n) = \{\Phi^*(\nabla \mathcal{L}_n(\mathbf{\Theta}^*)) \leq \frac{\lambda_n}{2}\}$. From Table 9.1, the dual norm to the nuclear norm is the ℓ_2-operator norm. For the least-squares cost function, we have $\nabla \mathcal{L}_n(\mathbf{\Theta}^*) = \frac{1}{n}\sum_{i=1}^n w_i \mathbf{X}_i$, so that the statement of Proposition 10.6 involves the specialization of this event to the nuclear norm and least-squares cost.

The assumption (10.17) is a form of restricted strong convexity with tolerance parameter $\tau_n^2 = c_0\,\frac{d_1+d_2}{n}$. It only remains to verify the condition $\tau_n^2 \Psi^2(\bar{\mathbb{M}}) \leq \frac{\kappa}{64}$. The representation (10.14) reveals that any matrix $\mathbf{\Theta} \in \bar{\mathbb{M}}(\mathbb{U}^r, \mathbb{V}^r)$ has rank at most $2r$, and hence

$$\Psi(\bar{\mathbb{M}}(\mathbb{U}^r, \mathbb{V}^r)) := \sup_{\mathbf{\Theta} \in \bar{\mathbb{M}}(\mathbb{U}^r, \mathbb{V}^r)\setminus\{0\}} \frac{|||\mathbf{\Theta}|||_{\mathrm{nuc}}}{|||\mathbf{\Theta}|||_{\mathrm{F}}}\ \leq\ \sqrt{2r}.$$

Consequently, the final condition of Theorem 9.19 holds whenever the target rank r is bounded as in the statement of Proposition 10.6, which completes the proof. □

10.2.3 Bounds under operator norm curvature

In Chapter 9, we also proved a general result—namely, Theorem 9.24—that, for a given regularizer Φ, provides a bound on the estimation error in terms of the dual norm Φ^*. Recall from Table 9.1 that the dual to the nuclear norm is the ℓ_2-operator norm or spectral norm. For the least-squares cost function, the gradient is given by

$$\nabla \mathcal{L}_n(\mathbf{\Theta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^{\mathrm{T}}\Big(y_i - \langle\!\langle \mathbf{X}_i,\ \mathbf{\Theta} \rangle\!\rangle\Big) = \frac{1}{n}\mathfrak{X}_n^*\Big(y - \mathfrak{X}_n(\mathbf{\Theta})\Big),$$

where $\mathfrak{X}_n^*\colon \mathbb{R}^n \to \mathbb{R}^{d_1 \times d_2}$ is the adjoint operator. Consequently, in this particular case, the Φ^*-curvature condition from Definition 9.22 takes the form

$$|||\frac{1}{n}\mathfrak{X}_n^*\mathfrak{X}_n(\mathbf{\Delta})|||_2 \geq \kappa |||\mathbf{\Delta}|||_2 - \tau_n\, |||\mathbf{\Delta}|||_{\mathrm{nuc}} \qquad \text{for all } \mathbf{\Delta} \in \mathbb{R}^{d_1 \times d_2}, \tag{10.20}$$

where $\kappa > 0$ is the curvature parameter, and $\tau_n \geq 0$ is the tolerance parameter.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Proposition 10.7 *Suppose that the observation operator $\mathfrak{X}_n$ satisfies the curvature condition* (10.20) *with parameter $\kappa > 0$, and consider a matrix $\mathbf{\Theta}^*$ with* $\operatorname{rank}(\mathbf{\Theta}^*) < \frac{\kappa}{64\tau_n}$. *Then, conditioned on the event $\mathbb{G}(\lambda_n) = \{|||\frac{1}{n}\mathfrak{X}_n^*(w)|||_2 \leq \frac{\lambda_n}{2}\}$, any optimal solution to the M-estimator* (10.16) *satisfies the bound*

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2 \leq 3\sqrt{2}\,\frac{\lambda_n}{\kappa}. \tag{10.21}$$

Remark: Note that this bound is smaller by a factor of $\sqrt{r}$ than the Frobenius norm bound (10.19) that follows from Proposition 10.6. Such a scaling is to be expected, since the Frobenius norm of a rank-r matrix is at most $\sqrt{r}$ times larger than its operator norm. The operator norm bound (10.21) is, in some sense, stronger than the earlier Frobenius norm bound. More specifically, in conjunction with the cone-like inequality (10.15), inequality (10.21) implies a bound of the form (10.19). See Exercise 10.5 for verification of these properties.

Proof In order to apply Theorem 9.24, the only remaining condition to verify is the inequality $\tau_n \Psi^2(\overline{\mathbb{M}}) < \frac{\kappa}{32}$. We have previously calculated that $\Psi^2(\overline{\mathbb{M}}) \leq 2r$, so that the stated upper bound on r ensures that this inequality holds. □

10.3 Matrix compressed sensing

Thus far, we have derived some general results on least squares with nuclear norm regularization, which apply to any model that satisfies the restricted convexity or curvature conditions. We now turn to consequences of these general results for more specific observation models that arise in particular applications. Let us begin this exploration by studying compressed sensing for low-rank matrices, as introduced previously in Example 10.3. There we discussed the standard Gaussian observation model, in which the observation matrices $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ are drawn i.i.d., with all entries of each observation matrix drawn i.i.d. from the standard Gaussian $\mathcal{N}(0, 1)$ distribution. More generally, one might draw random observation matrices $\mathbf{X}_i$ with dependent entries, for instance with $\operatorname{vec}(\mathbf{X}_i) \sim \mathcal{N}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma} \in \mathbb{R}^{(d_1 d_2) \times (d_1 d_2)}$ is the covariance matrix. In this case, we say that $\mathbf{X}_i$ is drawn from the $\mathbf{\Sigma}$-*Gaussian ensemble*.

In order to apply Proposition 10.6 to this ensemble, our first step is to establish a form of restricted strong convexity. The following result provides a high-probability lower bound on the Hessian of the least-squares cost for this ensemble. It involves the quantity

$$\rho^2(\mathbf{\Sigma}) := \sup_{\|u\|_2 = \|v\|_2 = 1} \operatorname{var}(\langle\langle \mathbf{X}, uv^{\mathrm{T}} \rangle\rangle).$$

Note that $\rho^2(\mathbf{I}_d) = 1$ for the special case of the identity ensemble.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Theorem 10.8 *Given n i.i.d. draws $\{\mathbf{X}_i\}_{i=1}^n$ of random matrices from the $\boldsymbol{\Sigma}$-Gaussian ensemble, there are positive constants $c_1 < 1 < c_2$ such that*

$$\frac{\|\mathfrak{X}_n(\boldsymbol{\Delta})\|_2^2}{n} \geq c_1 \, \| \sqrt{\boldsymbol{\Sigma}} \, \mathrm{vec}(\boldsymbol{\Delta})\|_2^2 - c_2 \, \rho^2(\boldsymbol{\Sigma}) \left\{ \frac{d_1 + d_2}{n} \right\} |||\boldsymbol{\Delta}|||_{\mathrm{nuc}}^2 \quad \forall \, \boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2} \tag{10.22}$$

with probability at least $1 - \frac{e^{-\frac{n}{32}}}{1 - e^{-\frac{n}{32}}}$.

This result can be understood as a variant of Theorem 7.16, which established a similar result for the case of sparse vectors and the ℓ_1-norm. As with this earlier theorem, Theorem 10.8 can be proved using the Gordon–Slepian comparison lemma for Gaussian processes. In Exercise 10.6, we work through a proof of a slightly simpler form of the bound.

Theorem 10.8 has an immediate corollary for the *noiseless* observation model, in which we observe $(y_i, \mathbf{X}_i)$ pairs linked by the linear equation $y_i = \langle\langle \mathbf{X}_i, \, \boldsymbol{\Theta}^* \rangle\rangle$. In this setting, the natural analog of the basis pursuit program from Chapter 7 is the following convex program:

$$\min_{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}} |||\boldsymbol{\Theta}|||_{\mathrm{nuc}} \quad \text{such that } \langle\langle \mathbf{X}_i, \, \boldsymbol{\Theta} \rangle\rangle = y_i \text{ for all } i = 1, \ldots, n. \tag{10.23}$$

That is, we search over the space of matrices that match the observations perfectly to find the solution with minimal nuclear norm. As with the estimator (10.16), it can be reformulated as an instance of semidefinite program.

Corollary 10.9 *Given $n > 16 \frac{c_2}{c_1} \frac{\rho^2(\boldsymbol{\Sigma})}{\gamma_{\min}(\boldsymbol{\Sigma})} r\,(d_1 + d_2)$ i.i.d. samples from the $\boldsymbol{\Sigma}$-ensemble, the estimator* (10.23) *recovers the rank-r matrix $\boldsymbol{\Theta}^*$ exactly—i.e., it has a unique solution $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^*$—with probability at least $1 - \frac{e^{-\frac{n}{32}}}{1 - e^{-\frac{n}{32}}}$.*

The requirement that the sample size n is larger than $r\,(d_1 + d_2)$ is intuitively reasonable, as can be seen by counting the degrees of freedom required to specify a rank-r matrix of size $d_1 \times d_2$. Roughly speaking, we need r numbers to specify its singular values, and rd_1 and rd_2 numbers to specify its left and right singular vectors.[2] Putting together the pieces, we conclude that the matrix has of the order $r(d_1 + d_2)$ degrees of freedom, consistent with the corollary. Let us now turn to its proof.

Proof Since $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^*$ are optimal and feasible, respectively, for the program (10.23), we have $|||\widehat{\boldsymbol{\Theta}}|||_{\mathrm{nuc}} \leq |||\boldsymbol{\Theta}^*|||_{\mathrm{nuc}} = |||\boldsymbol{\Theta}^*_{\mathbb{M}}|||_{\mathrm{nuc}}$. Introducing the error matrix $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$, we have

$$|||\widehat{\boldsymbol{\Theta}}|||_{\mathrm{nuc}} \; = |||\boldsymbol{\Theta}^* + \widehat{\boldsymbol{\Delta}}|||_{\mathrm{nuc}} = |||\boldsymbol{\Theta}^*_{\mathbb{M}} + \widehat{\boldsymbol{\Delta}}_{\bar{\mathbb{M}}^\perp} + \widehat{\boldsymbol{\Delta}}_{\bar{\mathbb{M}}}|||_{\mathrm{nuc}} \overset{\text{(i)}}{\geq} |||\boldsymbol{\Theta}^*_{\mathbb{M}} + \widehat{\boldsymbol{\Delta}}_{\bar{\mathbb{M}}^\perp}|||_{\mathrm{nuc}} - |||\widehat{\boldsymbol{\Delta}}_{\bar{\mathbb{M}}}|||_{\mathrm{nuc}}$$

by the triangle inequality. Applying decomposability this yields $|||\boldsymbol{\Theta}^*_{\mathbb{M}} + \widehat{\boldsymbol{\Delta}}_{\bar{\mathbb{M}}^\perp}|||_{\mathrm{nuc}} =$

[2] The orthonormality constraints for the singular vectors reduce the degrees of freedom, so we have just given an upper bound here.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

$|||\mathbf{\Theta}^*_{\bar{\mathbb{M}}}|||_{\text{nuc}} + |||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}^\perp}|||_{\text{nuc}}$. Combining the pieces, we find that $|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}^\perp}|||_{\text{nuc}} \leq |||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}}|||_{\text{nuc}}$. From the representation (10.14), any matrix in $\bar{\mathbb{M}}$ has rank at most $2r$, whence

$$|||\widehat{\mathbf{\Delta}}|||_{\text{nuc}} \leq 2|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}}|||_{\text{nuc}} \leq 2\sqrt{2r}\,|||\widehat{\mathbf{\Delta}}|||_{\text{F}}. \tag{10.24}$$

Now let us condition on the event that the lower bound (10.22) holds. When applied to $\widehat{\mathbf{\Delta}}$, and coupled with the inequality (10.24), we find that

$$\frac{\|\mathfrak{X}_n(\widehat{\mathbf{\Delta}})\|_2^2}{n} \geq \left\{ c_1 \gamma_{\min}(\mathbf{\Sigma}) - 8\, c_2\, \rho^2(\mathbf{\Sigma}) \frac{r\,(d_1+d_2)}{n} \right\} |||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2 \geq \frac{c_1}{2} \gamma_{\min}(\mathbf{\Sigma})\, |||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2,$$

where the final inequality follows by applying the given lower bound on n, and performing some algebra. But since both $\widehat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$ are feasible for the convex program (10.23), we have shown that $0 = \frac{\|\mathfrak{X}_n(\widehat{\mathbf{\Delta}})\|_2^2}{n} \geq \frac{c_1}{2}\, \gamma_{\min}(\mathbf{\Sigma})\, |||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2$, which implies that $\widehat{\mathbf{\Delta}} = 0$ as claimed. □

Theorem 10.8 can also be used to establish bounds for the least-squares estimator (10.16), based on noisy observations of the form $y_i = \langle\langle \mathbf{X}_i,\ \mathbf{\Theta}^* \rangle\rangle + w_i$. Here we state and prove a result that is applicable to matrices of rank at most r.

Corollary 10.10 *Consider $n > 64 \frac{c_2}{c_1} \frac{\rho^2(\mathbf{\Sigma})}{\gamma_{\min}(\mathbf{\Sigma})}\, r\,(d_1+d_2)$ i.i.d. samples $(y_i, \mathbf{X}_i)$ from the linear matrix regression model, where each $\mathbf{X}_i$ is drawn from the $\mathbf{\Sigma}$-Gaussian ensemble. Then any optimal solution to the program* (10.16) *with $\lambda_n = 10\,\sigma\,\rho(\mathbf{\Sigma})(\sqrt{\frac{d_1+d_2}{n}} + \delta)$ satisfies the bound*

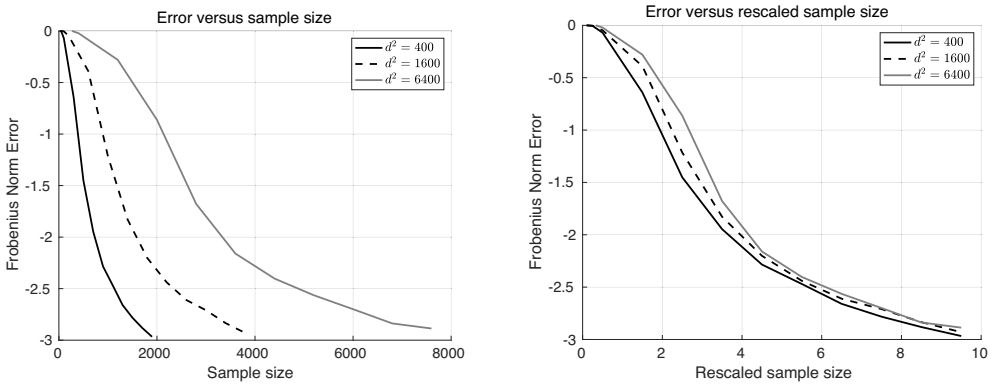
$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\text{F}}^2 \leq 125 \frac{\sigma^2 \rho^2(\mathbf{\Sigma})}{c_1^2\, \gamma_{\min}^2(\mathbf{\Sigma})} \left\{ \frac{r(d_1+d_2)}{n} + r\delta^2 \right\} \tag{10.25}$$

with probability at least $1 - 2e^{-2n\delta^2}$.

Figure 10.2 provides plots of the behavior predicted by Corollary 10.10. We generated these plots by simulating matrix regression problems with design matrices $\mathbf{X}_i$ chosen from the standard Gaussian ensemble, and then solved the convex program (10.16) with the choice of λ_n given in Corollary 10.10, and matrices of size $d \times d$, where $d^2 \in \{400, 1600, 6400\}$ and rank $r = \lceil \sqrt{d} \rceil$. In Figure 10.2(a), we plot the Frobenius norm error $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\text{F}}$, averaged over $T = 10$ trials, versus the raw sample size n. Each of these error plots tends to zero as the sample size increases, showing the classical consistency of the method. However, the curves shift to the right as the matrix dimension d (and hence the rank r) is increased, showing the effect of dimensionality. Assuming that the scaling of Corollary 10.10 is sharp, it predicts that, if we plot the same Frobenius errors versus the *rescaled sample size* $\frac{n}{rd}$, then all three curves should be relatively well aligned. These rescaled curves are shown in Figure 10.2(b): consistent with the prediction of Corollary 10.10, they are now all relatively well aligned, independently of the dimension and rank, consistent with the prediction.

Let us now turn to the proof of Corollary 10.10.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press



(a) (b)

Figure 10.2 Plots of the Frobenius norm error $|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_{\mathrm{F}}$ for the nuclear norm regularized least-squares estimator (10.16) with design matrices $\mathbf{X}_i$ drawn from the standard Gaussian ensemble. (a) Plots of Frobenius norm error versus sample size n for three different matrix sizes $d \in \{40, 80, 160\}$ and rank $r = \lceil \sqrt{d} \rceil$. (b) Same error measurements now plotted against the rescaled sample size $\frac{n}{rd}$. As predicted by the theory, all three curves are now relatively well-aligned.

Proof We prove the bound (10.25) via an application of Proposition 10.6, in particular in the form of the bound (10.19). Theorem 10.8 shows that the RSC condition holds with $\kappa = c_1$ and $c_0 = \frac{c_2 \rho^2(\boldsymbol{\Sigma})}{2}$, so that the stated lower bound on the sample size ensures that Proposition 10.6 can be applied with $r = \operatorname{rank}(\boldsymbol{\Theta}^*)$.

It remains to verify that the event $\mathbb{G}(\lambda_n) = \{|||\frac{1}{n}\sum_{i=1}^n w_i \mathbf{X}_i|||_2 \leq \frac{\lambda_n}{2}\}$ holds with high probability. Introduce the shorthand $\mathbf{Q} = \frac{1}{n}\sum_{i=1}^n w_i \mathbf{X}_i$, and define the event $\mathcal{E} = \left\{\frac{\|w\|_2^2}{n} \leq 2\sigma^2\right\}$. We then have

$$\mathbb{P}\Big[|||\mathbf{Q}|||_2 \geq \frac{\lambda_n}{2}\Big] \leq \mathbb{P}[\mathcal{E}^c] + \mathbb{P}\Big[|||\mathbf{Q}|||_2 \geq \frac{\lambda_n}{2} \mid \mathcal{E}\Big].$$

Since the noise variables $\{w_i\}_{i=1}^n$ are i.i.d., each zero-mean and sub-Gaussian with parameter σ, we have $\mathbb{P}[\mathcal{E}^c] \leq e^{-n/8}$. It remains to upper bound the second term, which uses conditioning on $\mathcal{E}$.

Let $\{u^1, \ldots, u^M\}$ and $\{v^1, \ldots, v^N\}$ be $1/4$-covers in Euclidean norm of the spheres $\mathbb{S}^{d_1-1}$ and $\mathbb{S}^{d_2-1}$, respectively. By Lemma 5.7, we can find such covers with $M \leq 9^{d_1}$ and $N \leq 9^{d_2}$ elements respectively. For any $v \in \mathbb{S}^{d_2-1}$, we can write $v = v^\ell + \Delta$ for some vector Δ with ℓ_2 at most $1/4$, and hence

$$|||\mathbf{Q}|||_2 = \sup_{v \in \mathbb{S}^{d_2-1}} \|\mathbf{Q}v\|_2 \leq \tfrac{1}{4}|||\mathbf{Q}|||_2 + \max_{\ell=1,\ldots,N} \|\mathbf{Q}v^\ell\|_2.$$

A similar argument involving the cover of $\mathbb{S}^{d_1-1}$ yields $\|\mathbf{Q}v^\ell\|_2 \leq \frac{1}{4}|||\mathbf{Q}|||_2 + \max_{j=1,\ldots,M} \left\langle u^j, \mathbf{Q}v^\ell \right\rangle$. Thus, we have established that

$$|||\mathbf{Q}|||_2 \leq 2 \max_{j=1,\ldots,M} \max_{\ell=1,\ldots,N} |Z^{j,\ell}| \qquad \text{where } Z^{j,\ell} = \left\langle u^j, \mathbf{Q}\, v^\ell \right\rangle.$$

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Fix some index pair (j, ℓ): we can then write $Z^{j,\ell} = \frac{1}{n}\sum_{i=1}^{n} w_i Y_i^{j,\ell}$ where $Y_i^{j,\ell} = \langle u^j, \mathbf{X}_i v^\ell \rangle$. Note that each variable $Y_i^{j,\ell}$ is zero-mean Gaussian with variance at most $\rho^2(\mathbf{\Sigma})$. Consequently, the variable $Z^{j,\ell}$ is zero-mean Gaussian with variance at most $\frac{2\sigma^2\rho^2(\mathbf{\Sigma})}{n}$, where we have used the conditioning on event $\mathcal{E}$. Putting together the pieces, we conclude that

$$\begin{aligned}\mathbb{P}\left[|||\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{X}_i|||_2 \geq \frac{\lambda_n}{2} \mid \mathcal{E}\right] &\leq \sum_{j=1}^{M}\sum_{\ell=1}^{N} \mathbb{P}\left[|Z^{j,\ell}| \geq \frac{\lambda_n}{4}\right] \\ &\leq 2\, e^{-\frac{n\lambda_n^2}{32\sigma^2\rho^2(\mathbf{\Sigma})} + \log M + \log N} \\ &\leq 2\, e^{-\frac{n\lambda_n^2}{32\sigma^2\rho^2(\mathbf{\Sigma})} + (d_1+d_2)\log 9}.\end{aligned}$$

Setting $\lambda_n = 10\sigma\rho(\mathbf{\Sigma})\left(\sqrt{\frac{(d_1+d_2)}{n}} + \delta\right)$, we find that $\mathbb{P}\left[|||\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{X}_i|||_2 \geq \frac{\lambda_n}{2}\right] \leq 2e^{-2n\delta^2}$ as claimed. □

Corollary 10.10 is stated for matrices that are exactly low-rank. However, Proposition 10.6 can also be used to derive error bounds for matrices that are not exactly low-rank, but rather *well approximated* by a low-rank matrix. For instance, suppose that $\mathbf{\Theta}^*$ belongs to the ℓ_q-"ball" of matrices given by

$$\mathbb{B}_q(R_q) := \left\{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j=1}^{d} (\sigma_j(\mathbf{\Theta}))^q \leq R_q\right\}, \tag{10.26}$$

where $q \in [0, 1]$ is a parameter, and R_q is the radius. Note that this is simply the set of matrices whose vector of singular values belongs to the usual ℓ_q-ball for vectors. See Figure 9.5 for an illustration.

When the unknown matrix $\mathbf{\Theta}^*$ belongs to $\mathbb{B}_q(R_q)$, Proposition 10.6 can be used to show that the estimator (10.16) satisfies an error bound of the form

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}}^2 \precsim R_q\left(\frac{\sigma^2(d_1+d_2)}{n}\right)^{1-\frac{q}{2}} \tag{10.27}$$

with high probability. Note that this bound generalizes Corollary 10.10, since in the special case $q = 0$, the set $\mathbb{B}_0(r)$ corresponds to the set of matrices with rank at most r. See Exercise 10.7 for more details.

As another extension, one can move beyond the setting of least squares, and consider more general non-quadratic cost functions. As an initial example, still in the context of matrix regression with samples $z = (\mathbf{X}, y)$, let us consider a cost function that satisfies a local L-Lipschitz condition of the form

$$\left|\mathcal{L}(\mathbf{\Theta}; z) - \mathcal{L}(\widetilde{\mathbf{\Theta}}; z)\right| \leq L\left|\langle\langle \mathbf{\Theta}, \mathbf{X}\rangle\rangle - \langle\langle \widetilde{\mathbf{\Theta}}, \mathbf{X}\rangle\rangle\right| \qquad \text{for all } \mathbf{\Theta}, \widetilde{\mathbf{\Theta}} \in \mathbb{B}_F(R).$$

For instance, if the response variables y were binary-valued, with the conditional distribution of the logistic form, as described in Example 9.2, then the log-likelihood would satisfy this condition with $L = 2$ (see Example 9.33). Similarly, in classification problems based on matrix-valued observations, the hinge loss that underlies the support vector machine would also satisfy this condition. In the following example, we show how Theorem 9.34 can be used

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

to establish restricted strong convexity with respect to the nuclear norm for such Lipschitz losses.

Example 10.11 (Lipschitz losses and nuclear norm) As a generalization of Corollary 10.10, suppose that the $d_1 \times d_2$ design matrices $\{\mathbf{X}_i\}_{i=1}^n$ are generated i.i.d. from a ν-sub-Gaussian ensemble, by which we mean that, for each pair of unit-norm vectors (u, v), the random variable $\langle u, \mathbf{X}_i v\rangle$ is zero-mean and ν-sub-Gaussian. Note that the $\mathbf{\Sigma}$-Gaussian ensemble is a special case with $\nu = \rho(\mathbf{\Sigma})$.

Now recall that

$$\mathcal{E}_n(\mathbf{\Delta}) := \mathcal{L}_n(\mathbf{\Theta}^* + \mathbf{\Delta}) - \mathcal{L}_n(\mathbf{\Theta}^*) - \langle\langle \nabla \mathcal{L}_n(\mathbf{\Theta}^*),\ \mathbf{\Delta}\rangle\rangle$$

denotes the error in the first-order Taylor-series expansion of the empirical cost function, whereas $\bar{\mathcal{E}}(\mathbf{\Delta})$ denotes the analogous quantity for the population cost function. We claim that for any $\delta > 0$, any cost function that is L-Lipschitz over the ball $\mathbb{B}_F(1)$ satisfies the bound

$$\left|\mathcal{E}_n(\mathbf{\Delta}) - \bar{\mathcal{E}}(\mathbf{\Delta})\right| \leq 16 L \nu |||\mathbf{\Delta}|||_{\text{nuc}} \left\{ 12 \sqrt{\frac{d_1 + d_2}{n}} + \epsilon \right\} \qquad \text{for all } \mathbf{\Delta} \in \mathbb{B}_F(1/d, 1) \tag{10.28}$$

with probability at least $1 - 4(\log d)^2\, e^{-\frac{n\epsilon^2}{12}}$.

In order to establish the bound (10.28), we need to verify the conditions of Theorem 9.34. For a matrix $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$, recall that we use $\{\sigma_j(\mathbf{\Theta})\}_{j=1}^d$ to denote its singular values. The dual to the nuclear norm is the ℓ_2-operator norm $|||\mathbf{\Theta}|||_2 = \max_{j=1,\ldots,d} \sigma_j(\mathbf{\Theta})$. Based on Theorem 9.34, we need to study the deviations of the random variable $|||\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{X}_i|||_2$, where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. sequence of Rademacher variables. Since the random matrices $\{\mathbf{X}_i\}_{i=1}^n$ are i.i.d., this random variable has the same distribution as $|||\mathbf{V}|||_2$, where $\mathbf{V}$ is a $\nu/\sqrt{n}$-sub-Gaussian random matrix. By the same discretization argument used in the proof of Corollary 10.10, for each $\lambda > 0$, we have $\mathbb{E}[e^{\lambda |||\mathbf{V}|||_2}] \leq \sum_{j=1}^M \sum_{\ell=1}^N \mathbb{E}[e^{2\lambda Z^{j,\ell}}]$, where $M \leq 9^{d_1}$ and $N \leq 9^{d_2}$, and each random variable $Z^{j,\ell}$ is sub-Gaussian with parameter at most $\sqrt{2}\nu/\sqrt{n}$. Consequently, for any $\delta > 0$,

$$\inf_{\lambda > 0} \mathbb{E}[e^{\lambda(|||\mathbf{V}|||_2 - \delta)}] \leq M N \inf_{\lambda > 0} e^{\frac{8\nu^2\lambda}{n} - \lambda\delta} = e^{-\frac{n\delta^2}{16\nu^2} + 9(d_1 + d_2)}.$$

Setting $\delta^2 = 144\nu^2 \frac{d_1+d_2}{n} + \nu^2\epsilon^2$ yields the claim (10.28). ♣

10.4 Bounds for phase retrieval

We now return to the problem of phase retrieval. In the idealized model previously introduced in Example 10.4, we make n observations of the form $\widetilde{y}_i = |\langle x_i, \theta^*\rangle|$, where the observation vector $x_i \sim \mathcal{N}(0, \mathbf{I}_d)$ are drawn independently. A standard lifting procedure leads to the semidefinite relaxation

$$\widehat{\mathbf{\Theta}} \in \arg\min_{\mathbf{\Theta} \in \mathcal{S}_+^{d \times d}} \text{trace}(\mathbf{\Theta}) \quad \text{such that} \quad \widetilde{y}_i^2 = \langle\langle \mathbf{\Theta},\ x_i \otimes x_i\rangle\rangle \quad \text{for all } i = 1, \ldots, n. \tag{10.29}$$

This optimization problem is known as a semidefinite program (SDP), since it involves optimizing over the cone $\mathcal{S}_+^{d \times d}$ of positive semidefinite matrices. By construction, the rank-one matrix $\mathbf{\Theta}^* = \theta^* \otimes \theta^*$ is feasible for the optimization problem (10.29), and our goal is to

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

understand when it is the unique optimal solution. Equivalently, our goal is to show that the error matrix $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ is equal to zero.

Defining the new response variables $y_i = \widetilde{y}_i^2$ and observation matrices $\mathbf{X}_i := x_i \otimes x_i$, the constraints in the SDP (10.29) can be written in the equivalent trace inner product form $y_i = \langle\langle \mathbf{X}_i, \Theta \rangle\rangle$. Since both $\widehat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$ are feasible and hence must satisfy these constraints, we see that the error matrix $\widehat{\mathbf{\Delta}}$ must belong to the nullspace of the linear operator $\mathfrak{X}_n : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ with components $[\mathfrak{X}_n(\mathbf{\Theta})]_i = \langle\langle \mathbf{X}_i, \mathbf{\Theta} \rangle\rangle$. The following theorem shows that this random operator satisfies a version of the restricted nullspace property (recall Chapter 7):

Theorem 10.12 (Restricted nullspace/eigenvalues for phase retrieval) *For each $i = 1, \ldots, n$, consider random matrices of the form $\mathbf{X}_i = x_i \otimes x_i$ for i.i.d. $\mathcal{N}(0, \mathbf{I}_d)$ vectors. Then there are universal constants (c_0, c_1, c_2) such that for any $\rho > 0$, a sample size $n > c_0 \rho d$ suffices to ensure that*

$$\frac{1}{n} \sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \mathbf{\Theta} \rangle\rangle^2 \geq \frac{1}{2} |||\mathbf{\Theta}|||_{\mathrm{F}}^2 \qquad \text{for all matrices such that } |||\mathbf{\Theta}|||_{\mathrm{F}}^2 \leq \rho |||\mathbf{\Theta}|||_{\mathrm{nuc}}^2 \tag{10.30}$$

with probability at least $1 - c_1 e^{-c_2 n}$.

Note that the lower bound (10.30) implies that there are no matrices in the intersection of nullspace of the operator $\mathfrak{X}_n$ with the matrix cone defined by the inequality $|||\mathbf{\Theta}|||_{\mathrm{F}}^2 \leq \rho |||\mathbf{\Theta}|||_{\mathrm{nuc}}^2$. Consequently, Theorem 10.12 has an immediate corollary for the exactness of the semidefinite programming relaxation (10.29):

Corollary 10.13 *Given $n > 2 c_0 d$ samples, the SDP* (10.29) *has the unique optimal solution $\widehat{\mathbf{\Theta}} = \mathbf{\Theta}^*$ with probability at least $1 - c_1 e^{-c_2 n}$.*

Proof Since $\widehat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$ are optimal and feasible (respectively) for the convex program (10.29), we are guaranteed that $\mathrm{trace}(\widehat{\mathbf{\Theta}}) \leq \mathrm{trace}(\mathbf{\Theta}^*)$. Since both matrices must be positive semidefinite, this trace constraint is equivalent to $|||\widehat{\mathbf{\Theta}}|||_{\mathrm{nuc}} \leq |||\mathbf{\Theta}^*|||_{\mathrm{nuc}}$. This inequality, in conjunction with the rank-one nature of $\mathbf{\Theta}^*$ and the decomposability of the nuclear norm, implies that the error matrix $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ satisfies the cone constraint $|||\widehat{\mathbf{\Delta}}|||_{\mathrm{nuc}} \leq \sqrt{2}\, |||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}}$. Consequently, we can apply Theorem 10.12 with $\rho = 2$ to conclude that

$$0 = \frac{1}{n} \sum_{i=1}^{n} \langle\langle \mathbf{X}_i, \widehat{\mathbf{\Delta}} \rangle\rangle \geq \frac{1}{2} |||\widehat{\mathbf{\Delta}}|||_2^2,$$

from which we conclude that $\widehat{\mathbf{\Delta}} = 0$ with the claimed probability. □

Let us now return to prove Theorem 10.12.

Proof For each matrix $\mathbf{\Delta} \in \mathcal{S}^{d \times d}$, consider the (random) function $f_{\Delta}(\mathbf{X}, v) = v \langle\langle \mathbf{X}, \mathbf{\Delta} \rangle\rangle$,

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

where $v \in \{-1, 1\}$ is a Rademacher variable independent of $\mathbf{X}$. By construction, we then have $\mathbb{E}[f_{\mathbf{\Delta}}(\mathbf{X}, v)] = 0$. Moreover, as shown in Exercise 10.9, we have

$$\|f_{\mathbf{\Delta}}\|_2^2 = \mathbb{E}[\langle\!\langle \mathbf{X}, \mathbf{\Delta} \rangle\!\rangle]^2 = |||\mathbf{\Delta}|||_{\mathrm{F}}^2 + 2(\operatorname{trace}(\mathbf{\Delta}))^2. \tag{10.31a}$$

As a consequence, if we define the set $\mathbb{A}_1(\sqrt{\rho}) = \{\mathbf{\Delta} \in \mathcal{S}^{d \times d} \mid |||\mathbf{\Delta}|||_{\mathrm{nuc}} \leq \sqrt{\rho}\, |||\mathbf{\Delta}|||_{\mathrm{F}}\}$, it suffices to show that

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \langle\!\langle \mathbf{X}, \mathbf{\Delta} \rangle\!\rangle^2}_{\|f_{\mathbf{\Delta}}\|_n^2} \geq \frac{1}{2} \underbrace{\mathbb{E}[\langle\!\langle \mathbf{X}, \mathbf{\Delta} \rangle\!\rangle^2]}_{\|f_{\mathbf{\Delta}}\|_2^2} \qquad \text{for all } \mathbf{\Delta} \in \mathbb{A}_1(\sqrt{\rho}) \tag{10.31b}$$

with probability at least $1 - c_1 e^{-c_2 n}$.

We prove claim (10.31b) as a corollary of a more general one-sided uniform law, stated as Theorem 14.12 in Chapter 14. First, observe that the function class $\mathscr{F} := \{f_{\mathbf{\Delta}} \mid \mathbf{\Delta} \in \mathbb{A}_1(\sqrt{\rho})\}$ is a cone, and so star-shaped around zero. Next we claim that the fourth-moment condition (14.22b) holds. From the result of Exercise 10.9, we can restrict attention to diagonal matrices without loss of generality. It suffices to show that $\mathbb{E}[f_{\mathbf{D}}^4(\mathbf{X}, v)] \leq C$ for all matrices such that $|||\mathbf{D}|||_{\mathrm{F}}^2 = \sum_{j=1}^d D_{jj}^2 \leq 1$. Since the Gaussian variables have moments of all orders, by Rosenthal's inequality (see Exercise 2.20), there is a universal constant c such that

$$\mathbb{E}[f_{\mathbf{D}}^4(\mathbf{X}, v)] = \mathbb{E}\Big[\Big(\sum_{j=1}^d D_{jj} x_j^2\Big)^4\Big] \leq c\left\{\sum_{j=1}^d D_{jj}^4 \mathbb{E}[x_j^8] + \Big(\sum_{j=1}^d D_{jj}^2 \mathbb{E}[x_j^4]\Big)^2\right\}.$$

For standard normal variates, we have $\mathbb{E}[x_j^4] = 4$ and $\mathbb{E}[x_j^8] = 105$, whence

$$\mathbb{E}[f_{\mathbf{D}}^4(\mathbf{X}, v)] \leq c\left\{105 \sum_{j=1}^d D_{jj}^4 + 16 |||\mathbf{D}|||_{\mathrm{F}}^4\right\}.$$

Under the condition $\sum_{j=1}^d D_{jj}^2 \leq 1$, this quantity is bounded by a universal constant C, thereby verifying the moment condition (14.22b).

Next, we need to compute the local Rademacher complexity, and hence the critical radius δ_n. As shown by our previous calculation (10.31a), the condition $\|f_{\mathbf{\Delta}}\|_2 \leq \delta$ implies that $|||\mathbf{\Delta}|||_{\mathrm{F}} \leq \delta$. Consequently, we have

$$\bar{\mathcal{R}}_n(\delta) \leq \mathbb{E}\Big[\sup_{\substack{\mathbf{\Delta} \in \mathbb{A}_1(\sqrt{\rho}) \\ |||\mathbf{\Delta}|||_{\mathrm{F}} \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_{\mathbf{\Delta}}(\mathbf{X}_i; v_i)\Big|\Big],$$

where $\{\varepsilon_i\}_{i=1}^n$ is another i.i.d. Rademacher sequence. Using the definition of $f_{\mathbf{\Delta}}$ and the duality between the operator and nuclear norms (see Exercise 9.4), we have

$$\bar{\mathcal{R}}_n(\delta) \leq \mathbb{E}\left[\sup_{\mathbf{\Delta} \in \mathbb{A}_1(\sqrt{\rho})} |||(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, (x_i \otimes x_i))|||_2\, |||\mathbf{\Delta}|||_{\mathrm{nuc}}\right] \leq \sqrt{\rho}\, \delta\, \mathbb{E}\left[|||\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, (x_i \otimes x_i)|||_2\right].$$

Finally, by our previous results on operator norms of random sub-Gaussian matrices (see Theorem 6.5), there is a constant c such that, in the regime $n > d$, we have

$$\mathbb{E}\Big[|||\frac{1}{n}\sum_{i=1}^n v_i\, (x_i \otimes x_i)|||_2\Big] \leq c\, \sqrt{\frac{d}{n}}.$$

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Putting together the pieces, we conclude that inequality (14.24) is satisfied for any $\delta_n \succsim \sqrt{\rho}\sqrt{\frac{d}{n}}$. Consequently, as long as $n > c_0 \rho d$ for a sufficiently large constant c_0, we can set $\delta_n = 1/2$ in Theorem 14.12, which establishes the claim (10.31b). □

10.5 Multivariate regression with low-rank constraints

The problem of multivariate regression, as previously introduced in Example 10.1, involves estimating a prediction function, mapping covariate vectors $z \in \mathbb{R}^p$ to output vectors $y \in \mathbb{R}^T$. In the case of linear prediction, any such mapping can be parameterized by a matrix $\mathbf{\Theta}^* \in \mathbb{R}^{p \times T}$. A collection of n observations can be specified by the model

$$\mathbf{Y} = \mathbf{Z}\mathbf{\Theta}^* + \mathbf{W}, \tag{10.32}$$

where $(\mathbf{Y}, \mathbf{Z}) \in \mathbb{R}^{n \times T} \times \mathbb{R}^{p \times T}$ are observed, and $\mathbf{W} \in \mathbb{R}^{n \times T}$ is a matrix of noise variables. For this observation model, the least-squares cost takes the form $\mathcal{L}_n(\mathbf{\Theta}) = \frac{1}{2n}|||\mathbf{Y} - \mathbf{Z}\mathbf{\Theta}|||_{\mathrm{F}}^2$.

The following result is a corollary of Proposition 10.7 in application to this model. It is applicable to the case of fixed design and so involves the minimum and maximum eigenvalues of the sample covariance matrix $\widehat{\mathbf{\Sigma}} := \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}{n}$.

Corollary 10.14 *Consider the observation model* (10.32) *in which* $\mathbf{\Theta}^* \in \mathbb{R}^{p \times T}$ *has rank at most r, and the noise matrix* $\mathbf{W}$ *has i.i.d. entries that are zero-mean and* σ*-sub-Gaussian. Then any solution to the program* (10.16) *with* $\lambda_n = 10\sigma\sqrt{\gamma_{\max}(\widehat{\mathbf{\Sigma}})}\,(\sqrt{\frac{p+T}{n}} + \delta)$ *satisfies the bound*

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2 \leq 30\sqrt{2}\,\frac{\sigma\sqrt{\gamma_{\max}(\widehat{\mathbf{\Sigma}})}}{\gamma_{\min}(\widehat{\mathbf{\Sigma}})}\left(\sqrt{\frac{p+T}{n}} + \delta\right) \tag{10.33}$$

with probability at least $1 - 2e^{-2n\delta^2}$. *Moreover, we have*

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}} \leq 4\sqrt{2r}|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2 \quad \text{and} \quad |||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{nuc}} \leq 32r|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2. \tag{10.34}$$

Note that the guarantee (10.33) is meaningful only when $n > p$, since the lower bound $\gamma_{\min}(\widehat{\mathbf{\Sigma}}) > 0$ cannot hold otherwise. However, even if the matrix $\mathbf{\Theta}^*$ were rank-one, it would have at least $p + T$ degrees of freedom, so this lower bound is unavoidable.

Proof We first claim that condition (10.20) holds with $\kappa = \gamma_{\min}(\widehat{\mathbf{\Sigma}})$ and $\tau_n = 0$. We have $\nabla\mathcal{L}_n(\mathbf{\Theta}) = \frac{1}{n}\mathbf{Z}^{\mathrm{T}}(y - \mathbf{Z}\mathbf{\Theta})$, and hence $\nabla\mathcal{L}_n(\mathbf{\Theta}^* + \mathbf{\Delta}) - \nabla\mathcal{L}_n(\mathbf{\Theta}^*) = \widehat{\mathbf{\Sigma}}\mathbf{\Delta}$ where $\widehat{\mathbf{\Sigma}} = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}{n}$ is the sample covariance. Thus, it suffices to show that

$$|||\widehat{\mathbf{\Sigma}}\mathbf{\Delta}|||_2 \geq \gamma_{\min}(\widehat{\mathbf{\Sigma}})|||\mathbf{\Delta}|||_2 \qquad \text{for all } \mathbf{\Delta} \in \mathbb{R}^{d \times T}.$$

For any vector $u \in \mathbb{R}^T$, we have $\|\widehat{\mathbf{\Sigma}}\mathbf{\Delta}u\|_2 \geq \gamma_{\min}(\widehat{\mathbf{\Sigma}})\|\mathbf{\Delta}u\|_2$, and thus

$$|||\widehat{\mathbf{\Sigma}}\mathbf{\Delta}|||_2 \sup_{\|u\|_2=1} \|\widehat{\mathbf{\Sigma}}\mathbf{\Delta}u\|_2 \geq \gamma_{\min}(\widehat{\mathbf{\Sigma}}) \sup_{\|u\|_2=1} \|\mathbf{\Delta}u\|_2 = \gamma_{\min}(\widehat{\mathbf{\Sigma}})\,|||\mathbf{\Delta}|||_2,$$

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

which establishes the claim.

It remains to verify that the inequality $|||\nabla \mathcal{L}_n(\mathbf{\Theta}^*)|||_2 \leq \frac{\lambda_n}{2}$ holds with high probability under the stated choice of λ_n. For this model, we have $\nabla \mathcal{L}_n(\mathbf{\Theta}^*) = \frac{1}{n}\mathbf{Z}^T\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{n \times T}$ is a zero-mean matrix of i.i.d. σ-sub-Gaussian variates. As shown in Exercise 10.8, we have

$$\mathbb{P}\left[|||\frac{1}{n}\mathbf{Z}^T\mathbf{W}|||_2 \geq 5\sigma\sqrt{\gamma_{\max}(\widehat{\mathbf{\Sigma}})}\,(\sqrt{\frac{d+T}{n}} + \delta) \geq\right] \leq 2\,e^{-2n\delta^2}, \tag{10.35}$$

from which the validity of λ_n follows. Thus, the bound (10.33) follows from Proposition 10.7.

Turning to the remaining bounds (10.34), with the given choice of λ_n, the cone inequality (10.15) guarantees that $|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}^\perp}|||_{\text{nuc}} \leq 3|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}}|||_{\text{nuc}}$. Since any matrix in $\bar{\mathbb{M}}$ has rank at most $2r$, we conclude that $|||\widehat{\mathbf{\Delta}}|||_{\text{nuc}} \leq 4\sqrt{2r}|||\widehat{\mathbf{\Delta}}|||_{\text{F}}$. Consequently, the nuclear norm bound in equation (10.34) follows from the Frobenius norm bound. We have

$$|||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2 = \langle\!\langle \widehat{\mathbf{\Delta}},\ \widehat{\mathbf{\Delta}} \rangle\!\rangle \overset{\text{(i)}}{\leq} |||\widehat{\mathbf{\Delta}}|||_{\text{nuc}}|||\widehat{\mathbf{\Delta}}|||_2 \overset{\text{(ii)}}{\leq} 4\sqrt{2r}|||\widehat{\mathbf{\Delta}}|||_{\text{F}}\,|||\widehat{\mathbf{\Delta}}|||_2,$$

where step (i) follows from Hölder's inequality, and step (ii) follows from our previous bound. Canceling out a factor of $|||\widehat{\mathbf{\Delta}}|||_{\text{F}}$ from both sides yields the Frobenius norm bound in equation (10.34), thereby completing the proof. □

10.6 Matrix completion

Let us now return to analyze the matrix completion problem previously introduced in Example 10.2. Recall that it corresponds to a particular case of matrix regression: observations are of the form $y_i = \langle\!\langle \mathbf{X}_i,\ \mathbf{\Theta}^* \rangle\!\rangle + w_i$, where $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ is a sparse mask matrix, zero everywhere except for a single randomly chosen entry $(a(i), b(i))$, where it is equal to $\sqrt{d_1 d_2}$. The sparsity of these regression matrices introduces some subtlety into the analysis of the matrix completion problem, as will become clear in the analysis to follow.

Let us now clarify why we chose to use rescaled mask matrices $\mathbf{X}_i$—that is, equal to $\sqrt{d_1 d_2}$ instead of 1 in their unique non-zero entry. With this choice, we have the convenient relation

$$\mathbb{E}\left[\frac{\|\mathfrak{X}_n(\mathbf{\Theta}^*)\|_2^2}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\langle\!\langle \mathbf{X}_i,\ \mathbf{\Theta}^* \rangle\!\rangle^2] = |||\mathbf{\Theta}^*|||_{\text{F}}^2, \tag{10.36}$$

using the fact that each entry of $\mathbf{\Theta}^*$ is picked out with probability $(d_1 d_2)^{-1}$.

The calculation (10.36) shows that, for any unit-norm matrix $\mathbf{\Theta}^*$, the squared Euclidean norm of $\|\mathfrak{X}_n(\mathbf{\Theta}^*)\|_2/\sqrt{n}$ has mean one. Nonetheless, in the high-dimensional setting of interest, namely, when $n \ll d_1 d_2$, there are many non-zero matrices $\mathbf{\Theta}^*$ of low rank such that $\mathfrak{X}_n(\mathbf{\Theta}^*) = 0$ with high probability. This phenomenon is illustrated by the following example.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Example 10.15 (Troublesome cases for matrix completion) Consider the matrix

$$\boldsymbol{\Theta}^{\mathrm{bad}} := e_1 \otimes e_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{10.37}$$

which is of rank one. Let $\mathfrak{X}_n : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ be the random observation operation based on n i.i.d. draws (with replacement) of rescaled mask matrices $\mathbf{X}_i$. As we show in Exercise 10.3, we have $\mathfrak{X}_n(\boldsymbol{\Theta}^{\mathrm{bad}}) = 0$ with probability converging to one whenever $n = o(d^2)$. ♣

Consequently, if we wish to prove non-trivial results about matrix completion in the regime $n \ll d_1 d_2$, we need to exclude matrices of the form (10.37). One avenue for doing so is by imposing so-called matrix incoherence conditions directly on the singular vectors of the unknown matrix $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$. These conditions were first introduced in the context of numerical linear algebra, in which context they are known as leverage scores (see the bibliographic section for further discussion). Roughly speaking, conditions on the leverage scores ensure that the singular vectors of $\boldsymbol{\Theta}^*$ are relatively "spread out".

More specifically, consider the singular value decomposition $\boldsymbol{\Theta}^* = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{D}$ is a diagonal matrix of singular values, and the columns of $\mathbf{U}$ and $\mathbf{V}$ contain the left and right singular vectors, respectively. What does it mean for the singular values to be spread out? Consider the matrix $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ of left singular vectors. By construction, each of its d_1-dimensional columns is normalized to Euclidean norm one; thus, if each singular vector were perfectly spread out, then each entry would have magnitude of the order $1/\sqrt{d_1}$. As a consequence, in this ideal case, each r-dimensional row of $\mathbf{U}$ would have Euclidean norm exactly $\sqrt{r/d_1}$. Similarly, the rows of $\mathbf{V}$ would have Euclidean norm $\sqrt{r/d_2}$ in the ideal case.

In general, the Euclidean norms of the rows of $\mathbf{U}$ and $\mathbf{V}$ are known as the left and right *leverage scores* of the matrix $\boldsymbol{\Theta}^*$, and matrix incoherence conditions enforce that they are relatively close to the ideal case. More specifically, note that the matrix $\mathbf{U}\mathbf{U}^{\mathrm{T}} \in \mathbb{R}^{d_1 \times d_1}$ has diagonal entries corresponding to the squared left leverage scores, with a similar observation for the matrix $\mathbf{V}\mathbf{V}^{\mathrm{T}} \in \mathbb{R}^{d_2 \times d_2}$. Thus, one way in which to control the leverage scores is via bounds of the form

$$\|\mathbf{U}\mathbf{U}^{\mathrm{T}} - \frac{r}{d_1}\mathbf{I}_{d_1 \times d_1}\|_{\max} \leq \mu \frac{\sqrt{r}}{d_1} \qquad \text{and} \qquad \|\mathbf{V}\mathbf{V}^{\mathrm{T}} - \frac{r}{d_2}\mathbf{I}_{d_2 d_2}\|_{\max} \leq \mu \frac{\sqrt{r}}{d_2}, \tag{10.38}$$

where $\mu > 0$ is the *incoherence parameter*. When the unknown matrix $\boldsymbol{\Theta}^*$ satisfies conditions of this type, it is possible to establish exact recovery results for the noiseless version of the matrix completion problem. See the bibliographic section for further discussion.

In the more realistic setting of noisy observations, the incoherence conditions (10.38) have an unusual property, in that they have no dependence on the singular values. In the presence of noise, one cannot expect to recover the matrix exactly, but rather only an estimate that captures all "significant" components. Here significance is defined relative to the noise level. Unfortunately, the incoherence conditions (10.38) are non-robust, and so less suitable in application to noisy problems. An example is helpful in understanding this issue.

Example 10.16 (Non-robustness of singular vector incoherence) Define the d-dimensional

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

vector $z = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \end{bmatrix}$, and the associated matrix $\mathbf{Z}^* := (z \otimes z)/d$. By construction, the matrix $\mathbf{Z}^*$ is rank-one, and satisfies the incoherence conditions (10.38) with constant μ. But now suppose that we "poison" this incoherent matrix with a small multiple of the "bad" matrix from Example 10.15, in particular forming the matrix

$$\mathbf{\Gamma}^* = (1 - \delta)\mathbf{Z}^* + \delta \mathbf{\Theta}^{\text{bad}} \qquad \text{for some } \delta \in (0, 1]. \tag{10.39}$$

As long as $\delta > 0$, then the matrix $\mathbf{\Gamma}^*$ has $e_1 \in \mathbb{R}^d$ as one of its eigenvectors, and so violates the incoherence conditions (10.38). But for the non-exact recovery results of interest in a statistical setting, very small values of δ need not be a concern, since the component $\delta \mathbf{\Theta}^{\text{bad}}$ has Frobenius norm δ, and so can be ignored. ♣

There are various ways of addressing this deficiency of the incoherence conditions (10.38). Possibly the simplest is by bounding the maximum absolute value of the matrix, or rather in order to preserve the scale of the problem, by bounding the ratio of the maximum value to its Frobenius norm. More precisely, for any non-zero matrix $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, we define the *spikiness ratio*

$$\alpha_{\text{sp}}(\mathbf{\Theta}) = \frac{\sqrt{d_1 d_2}\, \|\mathbf{\Theta}\|_{\max}}{|||\mathbf{\Theta}|||_{\text{F}}}, \tag{10.40}$$

where $\|\cdot\|_{\max}$ denotes the elementwise maximum absolute value. By definition of the Frobenius norm, we have

$$|||\mathbf{\Theta}|||_{\text{F}}^2 = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \Theta_{jk}^2 \leq d_1 d_2 \, \|\mathbf{\Theta}\|_{\max}^2,$$

so that the spikiness ratio is lower bounded by 1. On the other hand, it can also be seen that $\alpha_{\text{sp}}(\mathbf{\Theta}) \leq \sqrt{d_1 d_2}$, where this upper bound is achieved (for instance) by the previously constructed matrix (10.37). Recalling the "poisoned" matrix (10.39), note that unlike the incoherence condition, its spikiness ratio degrades as δ increases, but not in an abrupt manner. In particular, for any $\delta \in [0, 1]$, we have $\alpha_{\text{sp}}(\mathbf{\Gamma}^*) \leq \frac{(1-\delta)+\delta d}{1-2\delta}$.

The following theorem establishes a form of restricted strong convexity for the random operator that underlies matrix completion. To simplify the theorem statement, we adopt the shorthand $d = d_1 + d_2$.

Theorem 10.17 *Let $\mathfrak{X}_n : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$ be the random matrix completion operator formed by n i.i.d. samples of rescaled mask matrices $\mathbf{X}_i$. Then there are universal positive constants (c_1, c_2) such that*

$$\left| \frac{1}{n} \frac{\|\mathfrak{X}_n(\mathbf{\Theta})\|_2^2}{|||\mathbf{\Theta}|||_{\text{F}}^2} - 1 \right| \leq c_1 \, \alpha_{\text{sp}}(\mathbf{\Theta}) \frac{|||\mathbf{\Theta}|||_{\text{nuc}}}{|||\mathbf{\Theta}|||_{\text{F}}} \sqrt{\frac{d \log d}{n}} + c_2 \, \alpha_{\text{sp}}^2(\mathbf{\Theta}) \left(\sqrt{\frac{d \log d}{n}} + \delta \right)^2 \tag{10.41}$$

for all non-zero $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$, uniformly with probability at least $1 - 2e^{-\frac{1}{2} d \log d - n\delta}$.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

In order to interpret this claim, note that the ratio $\beta(\mathbf{\Theta}) := \frac{|||\mathbf{\Theta}|||_{\text{nuc}}}{|||\mathbf{\Theta}|||_{\text{F}}}$ serves as a "weak" measure of the rank. For any rank-r matrix, we have $\beta(\mathbf{\Theta}) \leq \sqrt{r}$, but in addition, there are many other higher-rank matrices that also satisfy this type of bound. On the other hand, recall the "bad" matrix $\mathbf{\Theta}^{\text{bad}}$ from Example 10.15. Although it has rank one, its spikiness ratio is maximal—that is, $\alpha_{\text{sp}}(\mathbf{\Theta}^{\text{bad}}) = d$. Consequently, the bound (10.41) does not provide any interesting guarantee until $n \gg d^2$. This prediction is consistent with the result of Exercise 10.3.

Before proving Theorem 10.17, let us state and prove one of its consequences for noisy matrix completion. Given n i.i.d. samples $\widetilde{y}_i$ from the noisy linear model (10.6), consider the nuclear norm regularized estimator

$$\widehat{\mathbf{\Theta}} \in \arg \min_{|||\mathbf{\Theta}|||_{\max} \leq \frac{\alpha}{\sqrt{d_1 d_2}}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} d_1 d_2 \{\widetilde{y}_i - \Theta_{a(i),b(i)}\}^2 + \lambda_n |||\mathbf{\Theta}|||_{\text{nuc}} \right\}, \tag{10.42}$$

where Theorem 10.17 motivates the addition of the extra side constraint on the infinity norm of $\mathbf{\Theta}$. As before, we use the shorthand notation $d = d_1 + d_2$.

Corollary 10.18 *Consider the observation model* (10.6) *for a matrix* $\mathbf{\Theta}^*$ *with rank at most* r*, elementwise bounded as* $|||\mathbf{\Theta}^*|||_{\max} \leq \alpha/\sqrt{d_1 d_2}$*, and i.i.d. additive noise variables* $\{w_i\}_{i=1}^n$ *that satisfy the Bernstein condition with parameters* (σ, b)*. Given a sample size* $n > \frac{100 b^2}{\sigma^2} d \log d$*, if we solve the program* (10.42) *with* $\lambda_n^2 = 25 \frac{\sigma^2 d \log d}{n} + \delta^2$ *for some* $\delta \in (0, \frac{\sigma^2}{2b})$*, then any optimal solution* $\widehat{\mathbf{\Theta}}$ *satisfies the bound*

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\text{F}}^2 \leq c_1 \max\{\sigma^2, \alpha^2\} r \left\{ \frac{d \log d}{n} + \delta^2 \right\} \tag{10.43}$$

with probability at least $1 - e^{-\frac{n\delta^2}{16 d}} - 2e^{-\frac{1}{2} d \log d - n\delta}$.

Remark: Note that the bound (10.43) implies that the squared Frobenius norm is small as long as (apart from a logarithmic factor) the sample size n is larger than the degrees of freedom in a rank-r matrix—namely, $r(d_1 + d_2)$.

Proof We first verify that the good event $\mathbb{G}(\lambda_n) = \{|||\nabla \mathcal{L}_n(\mathbf{\Theta}^*)|||_2 \leq \frac{\lambda_n}{2}\}$ holds with high probability. Under the observation model (10.6), the gradient of the least-squares objective (10.42) is given by

$$\nabla \mathcal{L}_n(\mathbf{\Theta}^*) = \frac{1}{n} \sum_{i=1}^{n} (d_1 d_2) \frac{w_i}{\sqrt{d_1 d_2}} \mathbf{E}_i = \frac{1}{n} \sum_{i=1}^{n} w_i \mathbf{X}_i,$$

where we recall the rescaled mask matrices $\mathbf{X}_i := \sqrt{d_1 d_2}\, \mathbf{E}_i$. From our calculations in Ex-

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

ample 6.18, we have[3]

$$\mathbb{P}\Big[|||\frac{1}{n}\sum_{i=1}^{n} w_i \mathbf{X}_i|||_2 \geq \epsilon\Big] \leq 4d\, e^{-\frac{n\epsilon^2}{8d(\sigma^2+b\epsilon)}} \leq 4d\, e^{-\frac{n\epsilon^2}{16d\sigma^2}},$$

where the second inequality holds for any $\epsilon > 0$ such that $b\epsilon \leq \sigma^2$. Under the stated lower bound on the sample size, we are guaranteed that $b\lambda_n \leq \sigma^2$, from which it follows that the event $\mathbb{G}(\lambda_n)$ holds with the claimed probability.

Next we use Theorem 10.17 to verify a variant of the restricted strong convexity condition. Under the event $\mathbb{G}(\lambda_n)$, Proposition 9.13 implies that the error matrix $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ satisfies the constraint $|||\widehat{\mathbf{\Delta}}|||_{\text{nuc}} \leq 4|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}}|||_{\text{nuc}}$. As noted earlier, any matrix in $\bar{\mathbb{M}}$ has rank at most $2r$, whence $|||\widehat{\mathbf{\Delta}}|||_{\text{nuc}} \leq 4\sqrt{2r}\,|||\widehat{\mathbf{\Delta}}|||_{\text{F}}$. By construction, we also have $|||\widehat{\mathbf{\Delta}}|||_{\max} \leq \frac{2\alpha}{\sqrt{d_1 d_2}}$. Putting together the pieces, Theorem 10.17 implies that, with probability at least $1 - 2e^{-\frac{1}{2}d\log d - n\delta}$, the observation operator $\mathfrak{X}_n$ satisfies the lower bound

$$\begin{aligned}\frac{|||\mathfrak{X}_n(\widehat{\mathbf{\Delta}})|||_2^2}{n} &\geq |||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2 - 8\sqrt{2}\,c_1\alpha\sqrt{\frac{rd\log d}{n}}|||\widehat{\mathbf{\Delta}}|||_{\text{F}} - 4c_2\alpha^2\left(\sqrt{\frac{d\log d}{n}} + \delta\right)^2 \\ &\geq |||\widehat{\mathbf{\Delta}}|||_{\text{F}}\left\{|||\widehat{\mathbf{\Delta}}|||_{\text{F}} - 8\sqrt{2}\,c_1\alpha\sqrt{\frac{rd\log d}{n}}\right\} - 8c_2\alpha^2\left(\frac{d\log d}{n} + \delta^2\right). \qquad (10.44)\end{aligned}$$

In order to complete the proof using this bound, we only need to consider two possible cases.

Case 1: On one hand, if either

$$|||\widehat{\mathbf{\Delta}}|||_{\text{F}} \leq 16\sqrt{2}c_1\alpha\sqrt{\frac{rd\log d}{n}} \quad \text{or} \quad |||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2 \leq 64c_2\alpha^2\left(\frac{d\log d}{n} + \delta^2\right),$$

then the claim (10.43) follows.

Case 2: Otherwise, we must have

$$|||\widehat{\mathbf{\Delta}}|||_{\text{F}} - 8\sqrt{2}\,c_1\alpha\sqrt{\frac{rd\log d}{n}} > \frac{|||\widehat{\mathbf{\Delta}}|||_{\text{F}}}{2} \quad \text{and} \quad 8c_2\alpha^2\left(\frac{d\log d}{n} + \delta^2\right) < \frac{|||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2}{4},$$

and hence the lower bound (10.44) implies that

$$\frac{|||\mathfrak{X}_n(\widehat{\mathbf{\Delta}})|||_2^2}{n} \geq \frac{1}{2}|||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2 - \frac{1}{4}|||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2 = \frac{1}{4}|||\widehat{\mathbf{\Delta}}|||_{\text{F}}^2.$$

This is the required restricted strong convexity condition, and so the proof is then complete. □

Finally, let us return to prove Theorem 10.17.

[3] Here we have included a factor of 8 (as opposed to 2) in the denominator of the exponent, to account for the possible need of symmetrizing the random variables w_i.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Proof Given the invariance of the inequality to rescaling, we may assume without loss of generality that $|||\mathbf{\Theta}|||_{\mathrm{F}} = 1$. For given positive constants (α, ρ), define the set

$$\mathbb{S}(\alpha, \rho) = \left\{ \mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2} \mid |||\mathbf{\Theta}|||_{\mathrm{F}} = 1, \quad \|\mathbf{\Theta}\|_{\max} \leq \frac{\alpha}{\sqrt{d_1 d_2}} \quad \text{and} \quad |||\mathbf{\Theta}|||_{\mathrm{nuc}} \leq \rho \right\}, \tag{10.45}$$

as well as the associated random variable $Z(\alpha, \rho) := \sup_{\mathbf{\Theta} \in \mathbb{S}(\alpha,\rho)} \left| \frac{1}{n} \|\mathfrak{X}_n(\mathbf{\Theta})\|_2^2 - 1 \right|$. We begin by showing that there are universal constants (c_1, c_2) such that

$$\mathbb{P}\left[Z(\alpha, \rho) \geq \frac{c_1}{4} \alpha\rho \sqrt{\frac{d \log d}{n}} + \frac{c_2}{4} (\alpha \sqrt{\frac{d \log d}{n}})^2 \right] \leq e^{-d \log d}. \tag{10.46}$$

Here our choice of the rescaling by $1/4$ is for later theoretical convenience. Our proof of this bound is divided into two steps.

Concentration around mean: Introducing the convenient shorthand notation $F_{\mathbf{\Theta}}(\mathbf{X}) := \langle\!\langle \mathbf{\Theta}, \mathbf{X} \rangle\!\rangle^2$, we can write

$$Z(\alpha, r) = \sup_{\Theta \in \mathbb{S}(\alpha,\rho)} \left| \frac{1}{n} \sum_{i=1}^{n} F_{\mathbf{\Theta}}(\mathbf{X}_i) - \mathbb{E}[F_{\mathbf{\Theta}}(\mathbf{X}_i)] \right|,$$

so that concentration results for empirical processes from Chapter 3 can be applied. In particular, we will apply the Bernstein-type bound (3.86): in order to do, we need to bound $\|F_{\mathbf{\Theta}}\|_{\max}$ and $\mathrm{var}(F_{\mathbf{\Theta}}(\mathbf{X}))$ uniformly over the class. On one hand, for any rescaled mask matrix $\mathbf{X}$ and parameter matrix $\mathbf{\Theta} \in \mathbb{S}(\alpha, r)$, we have

$$|F_{\mathbf{\Theta}}(\mathbf{X})| \leq \|\mathbf{\Theta}\|_{\max}^2 \|\mathbf{X}\|_1^2 \leq \frac{\alpha^2}{d_1 d_2} d_1 d_2 = \alpha^2,$$

where we have used the fact that $\|\mathbf{X}\|_1^2 = d_1 d_2$ for any rescaled mask matrix. Turning to the variance, we have

$$\mathrm{var}(F_{\mathbf{\Theta}}(\mathbf{X})) \leq \mathbb{E}[F_{\mathbf{\Theta}}^2(\mathbf{X})] \leq \alpha^2 \mathbb{E}[F_{\mathbf{\Theta}}(\mathbf{X})] = \alpha^2,$$

a bound which holds for any $\mathbf{\Theta} \in \mathbb{S}(\alpha, \rho)$. Consequently, applying the bound (3.86) with $\epsilon = 1$ and $t = d \log d$, we conclude that there are universal constants (c_1, c_2) such that

$$\mathbb{P}\left[Z(\alpha, \rho) \geq 2\mathbb{E}[Z(\alpha, r)] + \frac{c_1}{8} \alpha \sqrt{\frac{d \log d}{n}} + \frac{c_2}{4} \alpha^2 \frac{d \log d}{n} \right] \leq e^{-d \log d}. \tag{10.47}$$

Bounding the expectation: It remains to bound the expectation. By Rademacher symmetrization (see Proposition 4.11), we have

$$\mathbb{E}[Z(\alpha, \rho)] \leq 2\mathbb{E}\left[\sup_{\mathbf{\Theta} \in \mathbb{S}(\alpha,\rho)} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle\!\langle \mathbf{X}_i, \mathbf{\Theta} \rangle\!\rangle^2 \right| \right] \overset{\text{(ii)}}{\leq} 4\alpha \, \mathbb{E}\left[\sup_{\mathbf{\Theta} \in \mathbb{S}(\alpha,\rho)} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle\!\langle \mathbf{X}_i, \mathbf{\Theta} \rangle\!\rangle \right| \right],$$

where inequality (ii) follows from the Ledoux–Talagrand contraction inequality (5.61) for Rademacher processes, using the fact that $|\langle\!\langle \mathbf{\Theta}, \mathbf{X}_i \rangle\!\rangle| \leq \alpha$ for all pairs $(\mathbf{\Theta}, \mathbf{X}_i)$. Next we apply

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

Hölder's inequality to bound the remaining term: more precisely, since $|||\mathbf{\Theta}|||_{\text{nuc}} \leq \rho$ for any $\mathbf{\Theta} \in \mathbb{S}(\alpha, \rho)$, we have

$$\mathbb{E}\left[\sup_{\mathbf{\Theta} \in \mathbb{S}(\alpha,\rho)} \left|\langle\!\langle \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i, \mathbf{\Theta} \rangle\!\rangle\right|\right] \leq \rho \, \mathbb{E}\Big[|||\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i|||_2\Big].$$

Finally, note that each matrix $\varepsilon_i \mathbf{X}_i$ is zero-mean, has its operator norm upper bounded as $|||\varepsilon_i \mathbf{X}_i|||_2 \leq \sqrt{d_1 d_2} \leq d$, and its variance bounded as

$$|||\operatorname{var}(\varepsilon_i \mathbf{X}_i)|||_2 = \frac{1}{d_1 d_2} |||d_1 d_2 \, (1 \otimes 1)|||_2 = \sqrt{d_1 d_2}.$$

Consequently, the result of Exercise 6.10 implies that

$$\mathbb{P}\Big[|||\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i|||_2 \geq \delta\Big] \leq 2d \exp\left\{\frac{n\delta^2}{2d(1+\delta)}\right\}.$$

Next, applying the result of Exercise 2.8(a) with $C = 2d$, $v^2 = \frac{d}{n}$ and $B = \frac{d}{n}$, we find that

$$\mathbb{E}\Big[|||\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i|||_2\Big] \leq 2\sqrt{\frac{d}{n}}\left(\sqrt{\log(2d)} + \sqrt{\pi}\right) + \frac{4d\log(2d)}{n} \overset{\text{(i)}}{\leq} 16\sqrt{\frac{d\log d}{n}}.$$

Here the inequality (i) uses the fact that $n > d \log d$. Putting together the pieces, we conclude that

$$\mathbb{E}[Z(\alpha, \rho)] \leq \frac{c_1}{16} \alpha\rho \sqrt{\frac{d\log d}{n}},$$

for an appropriate definition of the universal constant c_1. Since $\rho \geq 1$, the claimed bound (10.46) follows.

Note that the bound (10.46) involves the fixed quantities (α, ρ), as opposed to the arbitrary quantities $(\sqrt{d_1 d_2}|\mathbf{\Theta}|_{\max}, |||\mathbf{\Theta}|||_{\text{nuc}})$ that would arise in applying the result to an arbitrary matrix. Extending the bound (10.46) to the more general bound (10.41) requires a technique known as peeling.

Extension via peeling: Let $\mathbb{B}_F(1)$ denote the Frobenius ball of norm one in $\mathbb{R}^{d_1 \times d_2}$, and let $\mathcal{E}$ be the event that the bound (10.41) is violated for some $\mathbf{\Theta} \in \mathbb{B}_F(1)$. For $k, \ell = 1, 2, \ldots$, let us define the sets

$$\mathbb{S}_{k,\ell} := \Big\{\mathbf{\Theta} \in \mathbb{B}_F(1) \mid 2^{k-1} \leq d|\mathbf{\Theta}|_{\max} \leq 2^k \text{ and } 2^{\ell-1} \leq |||\mathbf{\Theta}|||_{\text{nuc}} \leq 2^{\ell}\Big\},$$

and let $\mathcal{E}_{k,\ell}$ be the event that the bound (10.41) is violated for some $\mathbf{\Theta} \in \mathbb{S}_{k,\ell}$. We first claim that

$$\mathcal{E} \subseteq \bigcup_{k,\ell=1}^{M} \mathcal{E}_{k,\ell}, \qquad \text{where } M = \lceil \log d \rceil. \tag{10.48}$$

Indeed, for any matrix $\Theta \in \mathbb{S}(\alpha, \rho)$, we have

$$|||\mathbf{\Theta}|||_{\text{nuc}} \geq |||\mathbf{\Theta}|||_{\text{F}} = 1 \quad \text{and} \quad |||\mathbf{\Theta}|||_{\text{nuc}} \leq \sqrt{d_1 d_2}|||\mathbf{\Theta}|||_{\text{F}} \leq d.$$

Thus, we may assume that $|||\mathbf{\Theta}|||_{\text{nuc}} \in [1, d]$ without loss of generality. Similarly, for any

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

matrix of Frobenius norm one, we must have $d|||\mathbf{\Theta}|||_{\max} \geq \sqrt{d_1 d_2}|||\mathbf{\Theta}|||_{\max} \geq 1$ and $d|||\mathbf{\Theta}|||_{\max} \leq d$, showing that we may also assume that $d|||\mathbf{\Theta}|||_{\max} \in [1, d]$. Thus, if there exists a matrix $\mathbf{\Theta}$ of Frobenius norm one that violates the bound (10.41), then it must belong to some set $\mathbb{S}_{k,\ell}$ for $k, \ell = 1, 2 \ldots, M$, with $M = \lceil \log d \rceil$.

Next, for $\alpha = 2^k$ and $\rho = 2^\ell$, define the event

$$\widetilde{\mathcal{E}}_{k,\ell} := \left\{ Z(\alpha, \rho) \geq \frac{c_1}{4} \alpha\rho \sqrt{\frac{d \log d}{n}} + \frac{c_2}{4} \left(\alpha \sqrt{\frac{d \log d}{n}} \right)^2 \right\}.$$

We claim that $\mathcal{E}_{k,\ell} \subseteq \widetilde{\mathcal{E}}_{k,\ell}$. Indeed, if event $\mathcal{E}_{k,\ell}$ occurs, then there must exist some $\mathbf{\Theta} \in \mathbb{S}_{k,\ell}$ such that

$$\begin{aligned}
\left| \frac{1}{n} |||\mathfrak{X}_n(\mathbf{\Theta})|||_2^2 - 1 \right| &\geq c_1 d |||\mathbf{\Theta}|||_{\max} \, |||\mathbf{\Theta}|||_{\text{nuc}} \sqrt{\frac{d \log d}{n}} + c_2 \left(d \, |||\mathbf{\Theta}|||_{\max} \sqrt{\frac{d \log d}{n}} \right)^2 \\
&\geq c_1 2^{k-1} 2^{\ell-1} \sqrt{\frac{d \log d}{n}} + c_2 \left(2^{k-1} \sqrt{\frac{d \log d}{n}} \right)^2 \\
&\geq \frac{c_1}{4} 2^k 2^\ell \sqrt{\frac{d \log d}{n}} + \frac{c_2}{4} \left(2^k \sqrt{\frac{d \log d}{n}} \right)^2,
\end{aligned}$$

showing that $\widetilde{\mathcal{E}}_{k,\ell}$ occurs.

Putting together the pieces, we have

$$\mathbb{P}[\mathcal{E}] \overset{\text{(i)}}{\leq} \sum_{k,\ell=1}^{M} \mathbb{P}[\widetilde{\mathcal{E}}_{k,\ell}] \overset{\text{(ii)}}{\leq} M^2 e^{-d \log d} \leq e^{-\frac{1}{2} d \log d},$$

where inequality (i) follows from the union bound applied to the inclusion $\mathcal{E} \subseteq \bigcup_{k,\ell=1}^{M} \widetilde{\mathcal{E}}_{k,\ell}$; inequality (ii) is a consequence of the earlier tail bound (10.46); and inequality (iii) follows since $\log M^2 = 2 \log \log d \leq \frac{1}{2} d \log d$. □

10.7 Additive matrix decompositions

In this section, we turn to the problem of additive matrix decomposition. Consider a pair of matrices $\mathbf{\Lambda}^*$ and $\mathbf{\Gamma}^*$, and suppose that we observe a vector $y \in \mathbb{R}^n$ of the form

$$y = \mathfrak{X}_n(\mathbf{\Lambda}^* + \mathbf{\Gamma}^*) + w, \tag{10.49}$$

where $\mathfrak{X}_n$ is a known linear observation operator, mapping matrices in $\mathbb{R}^{d_1 \times d_2}$ to a vector in $\mathbb{R}^n$. In the simplest case, the observation operator performs a simple vectorization—that is, it maps a matrix $\mathbf{M}$ to the vectorized version vec($\mathbf{M}$). In this case, the sample size n is equal to the product $d_1 d_2$ of the dimensions, and we observe noisy versions of the sum $\mathbf{\Lambda}^* + \mathbf{\Gamma}^*$.

How to recover the two components based on observations of this form? Of course, this problem is ill-posed without imposing any structure on the components. One type of structure that arises in various applications is the combination of a low-rank matrix $\mathbf{\Lambda}^*$ with a sparse matrix $\mathbf{\Gamma}^*$. We have already encountered one instance of this type of decomposition in our discussion of multivariate regression in Example 9.6. The problem of Gaussian graphical selection with hidden variables, to be discussed at more length in Section 11.4.2,

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

provides another example of a low-rank and sparse decomposition. Here we consider some additional examples of such matrix decompositions.

Example 10.19 (Factor analysis with sparse noise) Factor analysis is a natural generalization of principal component analysis (see Chapter 8 for details on the latter). In factor analysis, we have i.i.d. random vectors $z \in \mathbb{R}^d$ assumed to be generated from the model

$$z_i = \mathbf{L}u_i + \varepsilon_i, \qquad \text{for } i = 1, 2, \ldots, N, \tag{10.50}$$

where $\mathbf{L} \in \mathbb{R}^{d \times r}$ is a loading matrix, and the vectors $u_i \sim \mathcal{N}(0, \mathbf{I}_r)$ and $\varepsilon_i \sim \mathcal{N}(0, \mathbf{\Gamma}^*)$ are independent. Given n i.i.d. samples from the model (10.50), the goal is to estimate the loading matrix $\mathbf{L}$, or the matrix $\mathbf{L}\mathbf{L}^T$ that projects onto the column span of $\mathbf{L}$. A simple calculation shows that the covariance matrix of Z_i has the form $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Gamma}^*$. Consequently, in the special case when $\mathbf{\Gamma}^* = \sigma^2 \mathbf{I}_d$, then the range of $\mathbf{L}$ is spanned by the top r eigenvectors of Σ, and so we can recover it via standard principal components analysis.

In other applications, we might no longer be guaranteed that $\mathbf{\Gamma}^*$ is the identity, in which case the top r eigenvectors of $\mathbf{\Sigma}$ need not be close to the column span of $\mathbf{L}$. Nonetheless, when $\mathbf{\Gamma}^*$ is a sparse matrix, the problem of estimating $\mathbf{L}\mathbf{L}^T$ can be understood as an instance of our general observation model (10.3) with $n = d^2$. In particular, letting the observation vector $y \in \mathbb{R}^n$ be the vectorized version of the sample covariance matrix $\frac{1}{N}\sum_{i=1}^N z_i z_i^T$, then some algebra shows that $y = \text{vec}(\mathbf{\Lambda}^* + \mathbf{\Gamma}^*) + \text{vec}(\mathbf{W})$, where $\mathbf{\Lambda}^* = \mathbf{L}\mathbf{L}^T$ is of rank r, and the random matrix $\mathbf{W}$ is a Wishart-type noise—viz.

$$\mathbf{W} := \frac{1}{N}\sum_{i=1}^N (z_i \otimes z_i) - \{\mathbf{L}\mathbf{L}^T + \mathbf{\Gamma}^*\}. \tag{10.51}$$

When $\mathbf{\Gamma}^*$ is assumed to be sparse, then this constraint can be enforced via the elementwise ℓ_1-norm. ♣

Other examples of matrix decomposition involve the combination of a low-rank matrix with a column or row-sparse matrix.

Example 10.20 (Matrix completion with corruptions) Recommender systems, as previously discussed in Example 10.2, are subject to various forms of corruption. For instance, in 2002, the Amazon recommendation system for books was compromised by a simple attack. Adversaries created a large number of false user accounts, amounting to additional rows in the matrix of user–book recommendations. These false user accounts were populated with strong positive ratings for a spiritual guide and a sex manual. Naturally enough, the end effect was that those users who liked the spiritual guide would also be recommended to read the sex manual.

If we again model the unknown true matrix of ratings as being low-rank, then such adversarial corruptions can be modeled in terms of the addition of a relatively sparse component. In the case of the false user attack described above, the adversarial component $\mathbf{\Gamma}^*$ would be relatively row-sparse, with the active rows corresponding to the false users. We are then led to the problem of recovering a low-rank matrix $\mathbf{\Lambda}^*$ based on partial observations of the sum $\mathbf{\Lambda}^* + \mathbf{\Gamma}^*$. ♣

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

As discussed in Chapter 6, the problem of covariance estimation is fundamental. A robust variant of the problem leads to another form of matrix decomposition, as discussed in the following example:

Example 10.21 (Robust covariance estimation) For $i = 1, 2, \ldots, N$, let $u_i \in \mathbb{R}^d$ be samples from a zero-mean distribution with unknown covariance matrix $\mathbf{\Lambda}^*$. When the vectors u_i are observed without any form of corruption, then it is straightforward to estimate $\mathbf{\Lambda}^*$ by performing PCA on the sample covariance matrix. Imagining that $j \in \{1, 2, \ldots, d\}$ indexes different individuals in the population, now suppose that the data associated with some subset S of individuals is arbitrarily corrupted. This adversarial corruption can be modeled by assuming that we observe the vectors $z_i = u_i + \gamma_i$ for $i = 1, \ldots, N$, where each $\gamma_i \in \mathbb{R}^d$ is a vector supported on the subset S. Letting $\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^N (z_i \otimes z_i)$ be the sample covariance matrix of the corrupted samples, some algebra shows that it can be decomposed as $\widehat{\mathbf{\Sigma}} = \mathbf{\Lambda}^* + \mathbf{\Delta} + \mathbf{W}$, where $\mathbf{W} := \frac{1}{N}\sum_{i=1}^N (u_i \otimes u_i) - \mathbf{\Lambda}^*$ is again a type of recentered Wishart noise, and the remaining term can be written as

$$\mathbf{\Delta} := \frac{1}{N}\sum_{i=1}^N (\gamma_i \otimes \gamma_i) + \frac{1}{N}\sum_{i=1}^N (u_i \otimes \gamma_i + \gamma_i \otimes u_i). \tag{10.52}$$

Thus, defining $y = \mathrm{vec}(\widehat{\mathbf{\Sigma}})$, we have another instance of the general observation model with $n = d^2$—namely, $y = \mathrm{vec}(\mathbf{\Lambda}^* + \mathbf{\Delta}) + \mathrm{vec}(\mathbf{W})$.

Note that $\mathbf{\Delta}$ itself is not a column-sparse or row-sparse matrix; however, since each vector $v_i \in \mathbb{R}^d$ is supported only on some subset $S \subset \{1, 2, \ldots, d\}$, we can write $\mathbf{\Delta} = \mathbf{\Gamma}^* + (\mathbf{\Gamma}^*)^{\mathrm{T}}$, where $\mathbf{\Gamma}^*$ is a column-sparse matrix with entries only in columns indexed by S. This structure can be enforced by the use of the column-sparse regularizer, as discussed in the sequel. ♣

Finally, as we discuss in Chapter 11 to follow, the problem of Gaussian graphical model selection with hidden variables also leads to a problem of additive matrix decomposition (see Section 11.4.2).

Having motivated additive matrix decompositions, let us now consider efficient methods for recovering them. For concreteness, we focus throughout on the case of low-rank plus elementwise-sparse matrices. First, it is important to note that—like the problem of matrix completion—we need somehow to exclude matrices that are simultaneously low-rank and sparse. Recall the matrix $\mathbf{\Theta}^{\mathrm{bad}}$ from Example 10.16: since it is both low-rank and sparse, it could be decomposed either as a low-rank matrix plus the all-zeros matrix as the sparse component, or as a sparse matrix plus the all-zeros matrix as the low-rank component.

Thus, it is necessary to impose further assumptions on the form of the decomposition. One possibility is to impose incoherence conditions (10.38) directly on the singular vectors of the low-rank matrix. As noted in Example 10.16, these bounds are not robust to small perturbations of this problem. Thus, in the presence of noise, it is more natural to consider a bound on the "spikiness" of the low-rank component, which can be enforced by bounding the maximum absolute value over its elements. Accordingly, we consider the following estimator:

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

$$(\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Lambda}}) = \arg \min_{\substack{\mathbf{\Gamma} \in \mathbb{R}^{d_1 \times d_2} \\ \|\mathbf{\Lambda}\|_{\max} \leq \frac{\alpha}{\sqrt{d_1 d_2}}}} \left\{ \frac{1}{2} |||\mathbf{Y} - (\mathbf{\Gamma} + \mathbf{\Lambda})|||_{\mathrm{F}}^2 + \lambda_n \Big(\|\mathbf{\Gamma}\|_1 + \omega_n |||\mathbf{\Lambda}|||_2 \Big) \right\}. \tag{10.53}$$

It is parameterized by two regularization parameters, namely λ_n and ω_n. The following corollary provides suitable choices of these parameters that ensure the estimator is well behaved; the guarantee is stated in terms of the squared Frobenius norm error

$$e^2(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*, \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*) := |||\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*|||_{\mathrm{F}}^2 + |||\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*|||_{\mathrm{F}}^2. \tag{10.54}$$

Corollary 10.22 *Suppose that we solve the convex program* (10.53) *with parameters*

$$\lambda_n \geq 2 \, \|\mathbf{W}\|_{\max} + 4 \frac{\alpha}{\sqrt{d_1 d_2}} \quad \text{and} \quad \omega_n \geq \frac{2 |||\mathbf{W}|||_2}{\lambda_n}. \tag{10.55}$$

Then there are universal constants c_j such that for any matrix pair $(\mathbf{\Lambda}^*, \mathbf{\Gamma}^*)$ with $\|\mathbf{\Lambda}^*\|_{\max} \leq \frac{\alpha}{\sqrt{d_1 d_2}}$ and for all integers $r = 1, 2, \ldots, \min\{d_1, d_2\}$ and $s = 1, 2, \ldots, (d_1 d_2)$, the squared Frobenius error (10.54) *is upper bounded as*

$$c_1 \, \omega_n^2 \, \lambda_n^2 \left\{ r + \frac{1}{\omega_n \lambda_n} \sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\mathbf{\Lambda}^*) \right\} + c_2 \, \lambda_n^2 \left\{ s + \frac{1}{\lambda_n} \sum_{(j,k) \notin S} |\mathbf{\Gamma}^*_{jk}| \right\}, \tag{10.56}$$

where S is an arbitrary subset of matrix indices of cardinality at most s.

As with many of our previous results, the bound (10.56) is a form of oracle inequality, meaning that the choices of target rank r and subset S can be optimized so as to achieve the tightest possible bound. For instance, when the matrix $\mathbf{\Lambda}^*$ is exactly low-rank and $\mathbf{\Gamma}^*$ is sparse, then setting $r = \operatorname{rank}(\mathbf{\Lambda}^*)$ and $S = \operatorname{supp}(\mathbf{\Gamma}^*)$ yields

$$e^2(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*, \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*) \leq \lambda_n^2 \Big\{ c_1 \, \omega_n^2 \, \operatorname{rank}(\mathbf{\Lambda}^*) + c_2 \, |\operatorname{supp}(\mathbf{\Gamma}^*)| \Big\}.$$

In many cases, this inequality yields optimal results for the Frobenius error of the low-rank plus sparse problem. We consider a number of examples in the exercises.

Proof We prove this claim as a corollary of Theorem 9.19. Doing so requires three steps: (i) verifying a form of restricted strong convexity; (ii) verifying the validity of the regularization parameters; and (iii) computing the subspace Lipschitz constant from Definition 9.18.

We begin with restricted strong convexity. Define the two matrices $\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}} = \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*$ and $\mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}} := \widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*$, corresponding to the estimation error in the sparse and low-rank components, respectively. By expanding out the quadratic form, we find that the first-order error in the Taylor series is given by

$$\mathcal{E}_n(\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}}, \mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}}) = \tfrac{1}{2} |||\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}} + \mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}}|||_{\mathrm{F}}^2 = \tfrac{1}{2} \underbrace{\{ |||\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}}|||_{\mathrm{F}}^2 + |||\mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}}|||_{\mathrm{F}}^2 \}}_{e^2(\mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}}, \mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}})} + \langle\!\langle \mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}}, \, \mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}} \rangle\!\rangle.$$

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

By the triangle inequality and the construction of our estimator, we have

$$\|\mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}}\|_{\max} \leq \|\widehat{\mathbf{\Lambda}}\|_{\max} + \|\mathbf{\Lambda}^*\|_{\max} \leq \frac{2\alpha}{\sqrt{d_1 d_2}}.$$

Combined with Hölder's inequality, we see that

$$\mathcal{E}_n(\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}}, \mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}}) \geq \frac{1}{2} e^2(\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}}, \mathbf{\Delta}_{\widehat{\mathbf{\Lambda}}}) - \frac{2\alpha}{\sqrt{d_1 d_2}} \|\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}}\|_1,$$

so that restricted strong convexity holds with $\kappa = 1$, but along with an extra error term. Since it is proportional to $\|\mathbf{\Delta}_{\widehat{\mathbf{\Gamma}}}\|_1$, the proof of Theorem 9.19 shows that it can be absorbed without any consequence as long as $\lambda_n \geq \frac{4\alpha}{\sqrt{d_1 d_2}}$.

Verifying event $\mathbb{G}(\lambda_n)$: A straightforward calculation gives $\nabla \mathcal{L}_n(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*) = (\mathbf{W}, \mathbf{W})$. From the dual norm pairs given in Table 9.1, we have

$$\Phi^*_{\omega_n}(\nabla \mathcal{L}_n(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*)) = \max\left\{ \|\mathbf{W}\|_{\max}, \frac{|||\mathbf{W}|||_2}{\omega_n} \right\}, \tag{10.57}$$

so that the choices (10.55) guarantee that $\lambda_n \geq 2\Phi^*_{\omega_n}(\nabla \mathcal{L}_n(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*))$.

Choice of model subspaces: For any subset S of matrix indices of cardinality at most s, define the subset $\mathbb{M}(S) := \{\mathbf{\Gamma} \in \mathbb{R}^{d_1 \times d_2} \mid \Gamma_{ij} = 0 \text{ for all } (i,j) \notin S\}$. Similarly, for any $r = 1, \ldots, \min\{d_1, d_2\}$, let $\mathbb{U}_r$ and $\mathbb{V}_r$ be (respectively) the subspaces spanned by the top r left and right singular vectors of $\mathbf{\Lambda}^*$, and recall the subspaces $\bar{\mathbb{M}}(\mathbb{U}_r, \mathbb{V}_r)$ and $\mathbb{M}^\perp(\mathbb{U}_r, \mathbb{V}_r)$ previously defined in equation (10.12). We are then guaranteed that the regularizer $\Phi_{\omega_n}(\mathbf{\Gamma}, \mathbf{\Lambda}) = \|\mathbf{\Gamma}\|_1 + \omega_n |||\mathbf{\Lambda}|||_{\text{nuc}}$ is decomposable with respect to the model subspace $\mathbb{M} := \mathbb{M}(S) \times \bar{\mathbb{M}}(\mathbb{U}_r, \mathbb{V}_r)$ and deviation space $\mathbb{M}^\perp(S) \times \mathbb{M}^\perp(\mathbb{U}_r, \mathbb{V}_r)$. It then remains to bound the subspace Lipschitz constant. We have

$$\begin{aligned} \Psi(\mathbb{M}) = \sup_{(\mathbf{\Gamma}, \mathbf{\Lambda}) \in \mathbb{M}(S) \times \bar{\mathbb{M}}(\mathbb{U}_r, \mathbb{V}_r)} \frac{\|\mathbf{\Gamma}\|_1 + \omega_n |||\mathbf{\Lambda}|||_{\text{nuc}}}{\sqrt{|||\mathbf{\Gamma}|||_{\text{F}}^2 + |||\mathbf{\Lambda}|||_{\text{F}}^2}} &\leq \sup_{(\mathbf{\Gamma}, \mathbf{\Lambda})} \frac{\sqrt{s} |||\mathbf{\Gamma}|||_{\text{F}} + \omega_n \sqrt{2r} |||\mathbf{\Lambda}|||_{\text{F}}}{\sqrt{|||\mathbf{\Gamma}|||_{\text{F}}^2 + |||\mathbf{\Lambda}|||_{\text{F}}^2}} \\ &\leq \sqrt{s} + \omega_n \sqrt{2r}. \end{aligned}$$

Putting together the pieces, the overall claim (10.56) now follows as a corollary of Theorem 9.19. □

10.8 Bibliographic details and background

In her Ph.D. thesis, Fazel (2002) studied various applications of the nuclear norm as a surrogate for a rank constraint. Recht et al. (2010) studied the use of nuclear norm regularization for the compressed sensing variant of matrix regression, with noiseless observations and matrices $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ drawn independently, each with i.i.d. $\mathcal{N}(0,1)$ entries. They established sufficient conditions for exact recovery in the noiseless setting (observation model (10.2) with $w_i = 0$) when the covariates $\mathbf{X}_i$ are drawn from the standard Gaussian ensemble (each entry of $\mathbf{X}_i$ distributed as $\mathcal{N}(0,1)$, drawn independently). In the noisy setting, this particular ensemble was also studied by Candès and Plan (2010) and Negahban and Wainwright (2011a),

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

who both gave sharp conditions on the required sample size. The former paper applies to sub-Gaussian but isotropic ensembles (identity covariance), whereas the latter paper established Theorem 10.8 that applies to Gaussian ensembles with arbitrary covariance matrices. Recht et al. (2009) provide precise results on the threshold behavior for the identity version of this ensemble.

Nuclear norm regularization has also been studied for more general problem classes. Rohde and Tsybakov (2011) impose a form of the restricted isometry condition (see Chapter 7), adapted to the matrix setting, whereas Negabahn and Wainwright (2011a) work with a milder lower curvature condition, corresponding to the matrix analog of a restricted eigenvalue condition in the special case of quadratic losses. Rohde and Tsybakov (2011) also provide bounds on the nuclear norm estimate in various other Schatten matrix norms. Bounds for multivariate (or multitask) regression, as in Corollary 10.14, have been proved by various authors (Lounici et al., 2011; Negahban and Wainwright, 2011a; Rohde and Tsybakov, 2011). The use of reduced rank estimators for multivariate regression has a lengthy history; see Exercise 10.1 for its explicit form as well as the references (Izenman, 1975, 2008; Reinsel and Velu, 1998) for some history and more details. See also Bunea et al. (2011) for non-asymptotic analysis of a class of reduced rank estimators in multivariate regression.

There are wide number of variants of the matrix completion problem; see the survey chapter by Laurent (2001) and references therein for more details. Srebro and his co-authors (2004; 2005a; 2005b) proposed low-rank matrix completion as a model for recommender systems, among them the Netflix problem described here. Srebro et al. (2005b) provide error bounds on the prediction error using nuclear norm regularization. Candès and Recht (2009) proved exact recovery guarantees for the nuclear norm estimator, assuming noiseless observations and certain incoherence conditions on the matrix involving the leverage scores. Leverage scores also play an important role in approximating low-rank matrices based on random subsamples of its rows or columns; see the survey by Mahoney (2011) and references therein. Gross (2011) provided a general scheme for exact recovery based on a dual witness construction, and making use of Ahlswede–Winter matrix bound from Section 6.4.4; see also Recht (2011) for a relatively simple argument for exact recovery. Keshavan et al. (2010a; 2010b) studied both methods based on the nuclear norm (SVD thresholding) as well as heuristic iterative methods for the matrix completion problem, providing guarantees in both the noiseless and noisy settings. Negahban and Wainwright (2012) study the more general setting of weighted sampling for both exactly low-rank and near-low-rank matrices, and provided minimax-optimal bounds for the ℓ_q-"balls" of matrices with control on the "spikiness" ratio (10.40). They proved a weighted form of Theorem 10.17; the proof given here for the uniformly sampled setting is more direct. Koltchinski et al. (2011) assume that the sampling design is known, and propose a variant of the matrix Lasso. In the case of uniform sampling, it corresponds to a form of SVD thresholding, an estimator that was also analyzed by Keshavan et al. (2010a; 2010b). See Exercise 10.11 for some analysis of this type of estimator.

The problem of phase retrieval from Section 10.4 has a lengthy history and various applications (e.g., Grechberg and Saxton, 1972; Fienup, 1982; Griffin and Lim, 1984; Fienup and Wackerman, 1986; Harrison, 1993). The idea of relaxing a non-convex quadratic program to a semidefinite program is a classical one (Shor, 1987; Lovász and Schrijver, 1991; Nesterov, 1998; Laurent, 2003). The semidefinite relaxation (10.29) for phase retrieval was proposed

by Chai et al. (2011). Candès et al. (2013) provided the first theoretical guarantees on exact recovery, in particular for Gaussian measurement vectors. See also Waldspurger et al. (2015) for discussion and analysis of a closely related but different SDP relaxation.

The problem of additive matrix decompositions with sparse and low-rank matrices was first formalized by Chandrasekaran et al. (2011), who analyzed conditions for exact recovery based on deterministic incoherence conditions between the sparse and low-rank components. Candès et al. (2011) provided related guarantees for random ensembles with milder incoherence conditions. Chandrasekaran et al. (2012b) showed that the problem of Gaussian graphical model selection with hidden variables can be tackled within this framework; see Section 11.4.2 of Chapter 11 for more details on this problem. Agarwal et al. (2012) provide a general analysis of regularization-based methods for estimating matrix decompositions for noisy observations; their work uses the milder bounds on the maximum entry of the low-rank matrix, as opposed to incoherence conditions, but guarantees only approximate recovery. See Ren and Zhou (2012) for some two-stage approaches for estimating matrix decompositions. Fan et al. (2013) study a related class of models for covariance matrices involving both sparse and low-rank components.

10.9 Exercises

Exercise 10.1 (Reduced rank regression) Recall the model of multivariate regression from Example 10.1, and, for a target rank $r \leq T \leq p$, consider the reduced rank regression estimate

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{RR}} := \arg \min_{\substack{\boldsymbol{\Theta} \in \mathbb{R}^{p \times T} \\ \operatorname{rank}(\boldsymbol{\Theta}) \leq r}} \left\{ \frac{1}{2n} |||\mathbf{Y} - \mathbf{Z}\boldsymbol{\Theta}|||_{\mathrm{F}}^2 \right\}.$$

Define the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{ZZ} = \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$, and the sample cross-covariance matrix $\widehat{\boldsymbol{\Sigma}}_{ZY} = \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Y}$. Assuming that $\widehat{\boldsymbol{\Sigma}}_{ZZ}$ is invertible, show that the reduced rank estimate has the explicit form

$$\widehat{\boldsymbol{\Theta}}_{\mathrm{RR}} = \widehat{\boldsymbol{\Sigma}}_{ZZ}^{-1}\widehat{\boldsymbol{\Sigma}}_{XY}\mathbf{V}\mathbf{V}^{\mathrm{T}},$$

where the matrix $\mathbf{V} \in \mathbb{R}^{T \times r}$ has columns consisting of the top r eigenvectors of the matrix $\widehat{\boldsymbol{\Sigma}}_{YZ}\widehat{\boldsymbol{\Sigma}}_{ZZ}^{-1}\widehat{\boldsymbol{\Sigma}}_{ZY}$.

Exercise 10.2 (Vector autogressive processes) Recall the vector autoregressive (VAR) model described in Example 10.5.

(a) Suppose that we initialize by choosing $z^1 \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where the symmetric matrix $\boldsymbol{\Sigma}$ satisfies the equation

$$\boldsymbol{\Sigma} - \boldsymbol{\Theta}^* \boldsymbol{\Sigma} (\boldsymbol{\Theta}^*)^{\mathrm{T}} - \boldsymbol{\Gamma} = 0. \tag{10.58}$$

Here $\boldsymbol{\Gamma} \succ 0$ is the covariance matrix of the driving noise. Show that the resulting stochastic process $\{z^t\}_{t=1}^{\infty}$ is stationary.

(b) Suppose that there exists a strictly positive definite solution $\boldsymbol{\Sigma}$ to equation (10.58). Show that $|||\boldsymbol{\Theta}^*|||_2 < 1$.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

(c) Conversely, supposing that $|||\mathbf{\Theta}^*|||_2 < 1$, show that there exists a strictly positive definite solution $\mathbf{\Sigma}$ to equation (10.58).

Exercise 10.3 (Nullspace in matrix completion) Consider the random observation operator $\mathfrak{X}_n : \mathbb{R}^{d\times d} \to \mathbb{R}$ formed by n i.i.d. draws of rescaled mask matrices (zero everywhere except for d in an entry chosen uniformly at random). For the "bad" matrix $\mathbf{\Theta}^{\text{bad}}$ from equation (10.37), show that $\mathbb{P}[\mathfrak{X}_n(\mathbf{\Theta}^{\text{bad}}) = 0] = 1 - o(1)$ whenever $n = o(d^2)$.

Exercise 10.4 (Cone inequalities for nuclear norm) Suppose that $|||\widehat{\mathbf{\Theta}}|||_{\text{nuc}} \leq |||\mathbf{\Theta}^*|||_{\text{nuc}}$, where $\mathbf{\Theta}^*$ is a rank-r matrix. Show that $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ satisfies the cone constraint $|||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}^\perp}|||_{\text{nuc}} \leq |||\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}}|||_{\text{nuc}}$, where the subspace $\bar{\mathbb{M}}^\perp$ was defined in equation (10.14).

Exercise 10.5 (Operator norm bounds)

(a) Verify the specific form (10.20) of the Φ^*-curvature condition.

(b) Assume that $\mathbf{\Theta}^*$ has rank r, and that $\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ satisfies the cone constraint (10.15), where $\mathbb{M}(\mathbb{U}, \mathbb{V})$ is specified by subspace $\mathbb{U}$ and $\mathbb{V}$ of dimension r. Show that

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\text{F}} \leq 4\sqrt{2r}\,|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2.$$

Exercise 10.6 (Analysis of matrix compressed sensing) In this exercise, we work through part of the proof of Theorem 10.8 for the special case $\mathbf{\Sigma} = \mathbf{I}_D$, where $D = d_1 d_2$. In particular, defining the set

$$\mathbb{B}(t) := \Big\{\mathbf{\Delta} \in \mathbb{R}^{d_1\times d_1} \mid |||\mathbf{\Delta}|||_{\text{F}} = 1,\ |||\mathbf{\Delta}|||_{\text{nuc}} \leq t\Big\},$$

for some $t > 0$, we show that

$$\inf_{\mathbf{\Delta}\in\mathbb{B}(t)} \sqrt{\frac{1}{n}\sum_{i=1}^n \langle\!\langle \mathbf{X}_i,\ \mathbf{\Delta}\rangle\!\rangle^2} \geq \frac{1}{2} - \delta - 2\left(\sqrt{\frac{d_1}{n}} + \sqrt{\frac{d_2}{n}}\right) t$$

with probability greater than $1 - e^{-n\delta^2/2}$. (This is a weaker result than Theorem 10.8, but the argument sketched here illustrates the essential ideas.)

(a) Reduce the problem to lower bounding the random variable

$$Z_n(t) := \inf_{\mathbf{\Delta}\in\mathbb{B}(t)} \sup_{\|u\|_2=1} \frac{1}{\sqrt{n}} \sum_{i=1}^n u_i \langle\!\langle \mathbf{X}_i,\ \mathbf{\Delta}\rangle\!\rangle.$$

(b) Show that the expectation can be lower bounded as

$$\mathbb{E}[Z_n(t)] \geq \frac{1}{\sqrt{n}}\Big\{\mathbb{E}[\|w\|_2] - \mathbb{E}[|||\mathbf{W}|||_2]\, t\Big\},$$

where $w \in \mathbb{R}^n$ and $\mathbf{W} \in \mathbb{R}^{d_1\times d_2}$ are populated with i.i.d. $\mathcal{N}(0, 1)$ variables. (*Hint:* The Gordon–Slepian comparison principle from Chapter 5 could be useful here.)

(c) Complete the proof using concentration of measure and part (b).

Exercise 10.7 (Bounds for approximately low-rank matrices) Consider the observation

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

model $y = \mathfrak{X}_n(\mathbf{\Theta}^*) + w$ with $w \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, and consider the nuclear norm constrained estimator

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta} \in \mathbb{R}^{d \times d}} \left\{ \frac{1}{2n} |\!|y - \mathfrak{X}_n(\mathbf{\Theta})|\!|_2^2 \right\} \qquad \text{subject to } |\!|\!|\mathbf{\Theta}|\!|\!|_{\text{nuc}} \leq |\!|\!|\mathbf{\Theta}^*|\!|\!|_{\text{nuc}}.$$

Suppose that $\mathbf{\Theta}^*$ belongs to the ℓ_q-"ball" of near-low-rank matrices (10.26).

In this exercise, we show that the estimate $\widehat{\mathbf{\Theta}}$ satisfies an error bound of the form (10.27) when the random operator $\mathfrak{X}_n$ satisfies the lower bound of Theorem 10.8.

(a) For an arbitrary $r \in \{1, 2, \ldots, d\}$, let $\mathbb{U}$ and $\mathbb{V}$ be subspaces defined by the top r left and right singular vectors of $\mathbf{\Theta}^*$, and consider the subspace $\bar{\mathbb{M}}(\mathbb{U}, \mathbb{V})$. Prove that the error matrix $\widehat{\mathbf{\Delta}}$ satisfies the inequality

$$|\!|\!|\widehat{\mathbf{\Delta}}_{\bar{\mathbb{M}}^\perp}|\!|\!|_{\text{nuc}} \leq 2\sqrt{2r} |\!|\!|\widehat{\mathbf{\Delta}}|\!|\!|_{\text{F}} + 2 \sum_{j=r+1}^{d} \sigma_j(\mathbf{\Theta}^*).$$

(b) Consider an integer $r \in \{1, \ldots, d\}$ such that $n > C\, rd$ for some sufficiently large but universal constant C. Using Theorem 10.8 and part (a), show that

$$|\!|\!|\widehat{\mathbf{\Delta}}|\!|\!|_{\text{F}}^2 \precsim \underbrace{\max\{T_1(r), T_1^2(r)\}}_{\text{approximation error}} + \underbrace{\sigma \sqrt{\frac{r\,d}{n}}\, |\!|\!|\widehat{\mathbf{\Delta}}|\!|\!|_{\text{F}}}_{\text{estimationerror}},$$

where $T_1(r) := \sigma \sqrt{\frac{d}{n}} \sum_{j=r+1}^{d} \sigma_j(\mathbf{\Theta}^*)$. (*Hint:* You may assume that an inequality of the form $|\!|\!|\frac{1}{n} \sum_{i=1}^{n} w_i \mathbf{X}_i|\!|\!|_2 \precsim \sigma \sqrt{\frac{d}{n}}$ holds.)

(c) Specify a choice of r that trades off the estimation and approximation error optimally.

Exercise 10.8 Under the assumptions of Corollary 10.14, prove that the bound (10.35) holds.

Exercise 10.9 (Phase retrieval with Gaussian masks) Recall the real-valued phase retrieval problem, based on the functions $f_{\mathbf{\Theta}}(\mathbf{X}) = \langle\!\langle \mathbf{X}, \mathbf{\Theta} \rangle\!\rangle$, for a random matrix $\mathbf{X} = x \otimes x$ with $x \sim \mathcal{N}(0, \mathbf{I}_n)$.

(a) Letting $\mathbf{\Theta} = \mathbf{U}^{\mathrm{T}} \mathbf{D} \mathbf{U}$ denote the singular value decomposition of $\mathbf{\Theta}$, explain why the random variables $f_{\mathbf{\Theta}}(\mathbf{X})$ and $f_{\mathbf{D}}(\mathbf{X})$ have the same distributions.

(b) Prove that

$$\mathbb{E}[f_{\mathbf{\Theta}}^2(\mathbf{X})] \;=\; |\!|\!|\mathbf{\Theta}|\!|\!|_{\text{F}}^2 + 2(\,\text{trace}(\mathbf{\Theta}))^2.$$

Exercise 10.10 (Analysis of noisy matrix completion) In this exercise, we work through the proof of Corollary 10.18.

(a) Argue that with the setting $\lambda_n \geq |\!|\!|\frac{1}{n} \sum_{i=1}^{n} w_i \mathbf{E}_i|\!|\!|_2$, we are guaranteed that the error matrix $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ satisfies the bounds

$$\frac{|\!|\!|\widehat{\mathbf{\Delta}}|\!|\!|_{\text{nuc}}}{|\!|\!|\widehat{\mathbf{\Delta}}|\!|\!|_{\text{F}}} \leq 2\sqrt{2r} \quad \text{and} \quad |\!|\widehat{\mathbf{\Delta}}|\!|_{\max} \leq 2\alpha.$$

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press

(b) Use part (a) and results from the chapter to show that, with high probability, at least one of the following inequalities must hold:

$$|||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}}^2 \leq \frac{c_2}{2}\alpha^2\frac{d\log d}{n} + 128\, c_1^2\,\alpha^2\frac{rd\log d}{n} \quad \text{or} \quad \frac{\|\mathfrak{X}_n(\widehat{\mathbf{\Delta}})\|_2^2}{n} \geq \frac{|||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}}^2}{4}.$$

(c) Use part (c) to establish the bound.

Exercise 10.11 (Alternative estimator for matrix completion) Consider the problem of noisy matrix completion, based on observations $y_i = \langle\langle \mathbf{X}_i,\ \mathbf{\Theta}^* \rangle\rangle + w_i$, where $\mathbf{X}_i \in \mathbb{R}^{d\times d}$ is a d-rescaled mask matrix (i.e., with a single entry of d in one location chosen uniformly at random, and zeros elsewhere). Consider the estimator

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}\in\mathbb{R}^{d\times d}} \left\{ \frac{1}{2}|||\mathbf{\Theta}|||_{\mathrm{F}}^2 - \langle\langle \mathbf{\Theta},\ \frac{1}{n}\sum_{i=1}^n y_i\mathbf{X}_i \rangle\rangle + \lambda_n |||\mathbf{\Theta}|||_{\mathrm{nuc}} \right\}.$$

(a) Show that the optimal solution $\widehat{\mathbf{\Theta}}$ is unique, and can be obtained by soft thresholding the singular values of the matrix $\mathbf{M} := \frac{1}{n}\sum_{i=1}^n y_i\mathbf{X}_i$. In particular, if $\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ denotes the SVD of $\mathbf{M}$, then $\widehat{\mathbf{\Theta}} = \mathbf{U}\,[T_{\lambda_n}(\mathbf{D})]\,\mathbf{V}^{\mathrm{T}}$, where $T_{\lambda_n}(\mathbf{D})$ is the matrix formed by soft thresholding the diagonal matrix of singular values $\mathbf{D}$.

(b) Suppose that the unknown matrix $\mathbf{\Theta}^*$ has rank r. Show that, with the choice

$$\lambda_n \geq 2\max_{|||\mathbf{U}|||_{\mathrm{nuc}}\leq 1} \Big|\frac{1}{n}\sum_{i=1}^n \langle\langle \mathbf{U},\ \mathbf{X}_i\rangle\rangle\langle\langle \mathbf{X}_i,\ \mathbf{\Theta}^*\rangle\rangle - \langle\langle \mathbf{U},\ \mathbf{\Theta}^*\rangle\rangle\Big| + 2|||\frac{1}{n}\sum_{i=1}^n w_i\mathbf{X}_i|||_2,$$

the optimal solution $\widehat{\mathbf{\Theta}}$ satisfies the bound

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}} \leq \frac{3}{\sqrt{2}}\,\sqrt{r}\,\lambda_n.$$

(c) Suppose that the noise vector $w \in \mathbb{R}^n$ has i.i.d. σ-sub-Gaussian entries. Specify an appropriate choice of λ_n that yields a useful bound on $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}}$.

https://doi.org/10.1017/9781108627771.010 Published online by Cambridge University Press