Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees

Yudong Chen* Martin J. Wainwright†,*

Department of Statistics† Department of Electrical Engineering and Computer Sciences*
University of California, Berkeley
Berkeley, CA 94720

yudong.chen@cornell.edu wainwrig@berkeley.edu

September 11, 2015

Abstract

Optimization problems with rank constraints arise in many applications, including matrix regression, structured PCA, matrix completion and matrix decomposition problems. An attractive heuristic for solving such problems is to factorize the low-rank matrix, and to run projected gradient descent on the nonconvex factorized optimization problem. The goal of this problem is to provide a general theoretical framework for understanding when such methods work well, and to characterize the nature of the resulting fixed point. We provide a simple set of conditions under which projected gradient descent, when given a suitable initialization, converges geometrically to a statistically useful solution. Our results are applicable even when the initial solution is outside any region of local convexity, and even when the problem is globally concave. Working in a non-asymptotic framework, we show that our conditions are satisfied for a wide range of concrete models, including matrix regression, structured PCA, matrix completion with real and quantized observations, matrix decomposition, and graph clustering problems. Simulation results show excellent agreement with the theoretical predictions.

1 Introduction

There are a variety of problems in statistics and machine learning that require estimating a matrix that is assumed—or desired—to be low-rank. For high-dimensional problems, the low-rank property is is useful as a form of regularization, and also can lead to more interpretable results in scientific settings. Low-rank matrix estimation can be formulated as a nonconvex optimization problem involving a cost function, measuring the fit to the data, along with a rank constraint. Even when the cost function is convex—such as in the ubiquitous case of least-squares fitting—solving a rank-constrained problem can be computationally difficult, with many interesting special cases known to have NP-hard complexity in the worst-case setting. However, statistical settings lead naturally to random ensembles, in which context such complexity concerns have been assuaged to some extent by the use of semidefinite programming (SDP) relaxations. These SDP relaxations are based on replacing the nonconvex rank constraint with a convex constraint based on the trace/nuclear norm. For many statistical ensembles of problems, among them multivariate regression, matrix completion and matrix decomposition, such types of SDP relaxations have been shown to have near-optimal performace (e.g., see the papers [18, 54, 51, 50, 22, 42] and references therein). Although in theory, any SDP can be solved to ϵ accuracy in polynomial-time [52], the associated computational cost is

often too high in practice. Letting d denote the dimension of the matrix, it can be as high as d^6 using standard interior point methods [11, 52]; such a scaling is not practical for many real-world applications involving high-dimensional matrices. More recent work has developed algorithms that are specifically tailored to certain classes of SDPs; however, even such specialized algorithms require at least d^2 time, since solving the SDP involves optimizing over the space of $d \times d$ matrices.

In practice, researchers often resort instead to heuristic methods that directly optimize over the space of low-rank matrices, using iterative algorithms such as alternating minimization, power iteration, expectation maximization (EM) and projected gradient descent. Letting r denote the rank, these factorized optimization problems live in an $\mathcal{O}(rd)$ dimensional space, as opposed to the $\mathcal{O}(d^2)$ space of the original problem. Such heuristic methods are quite effective in practice for some problems, but sometimes can also suffer from local optima. These intriguing phenomena motivate a recent and evolving line of work on understanding such iterative methods in the low-rank space. As we discuss in detail below, recent work has studied some of these algorithms in a number of specific settings. A natural question then arises: is there a general theory for understanding when low-rank iterative methods will succeed?

In this paper, we make progress on this general question by focusing on projected gradient descent in the low-rank space. We characterize a general set of conditions that govern the computational and statistical properties of the solutions, and then specialize this general theory to obtain corollaries for a broad range of problems. In more detail, suppose that we write a rank-r matrix $M \in \mathbb{R}^{d \times d}$ in its factorized form $F \otimes F = FF^\top$, where $F \in \mathbb{R}^{d \times r}$, and consider projected gradient descent methods in the variable F. The matrix quadratic form $F \otimes F$ makes the problem inherently nonconvex, and in many cases, the problem is not even locally convex. Nevertheless, our theory shows that given a suitable initialization, projected gradient descent converges geometrically to a statistically useful solution, under conditions that are much more general than convexity. Our results are applicable even when the initial solution is outside any region of local convexity, or when the problem is globally concave. Each iteration of projected gradient descent typically takes time that is linear in dr, the degrees of freedom of a low-rank matrix, as well as in the input size. Therefore, by directly enforcing low-rankness, our method simultaneously achieves two goals: we not only attain *statistical* consistency in the high-dimensional regime, but also gain *computational* advantages over convex relaxation methods that lift the problem to the space of $d \times d$ matrices.

For this approach to be relevant, an equally important question is when the above conditions for convergence are satisfied. We verify these conditions for a broad range of statistical and machine learning problems, including matrix sensing, matrix completion in both its standard and one-bit forms, sparse principal component analysis (SPCA), graph clustering, and matrix decomposition or robust PCA. For each of these problems, we show that a suitable initialization can be obtained efficiently using simple methods, and the projected gradient descent approach has sample complexity and statistical error bounds that are comparable (and sometimes better) to the best existing results (which are often achieved by convex relaxation methods). Notably, our approach does not require using fresh samples in each iteration—a heuristic known as sample splitting that is often used to simplify analysis—nor does it involve the computation of multiple singular value decompositions (SVDs).

Let us now put our contributions in a broader historical context. The seminal work in [12] studies the problem of obtaining low-rank solutions to SDPs using gradient descent on the factor space. Several subsequent papers aim to obtain rigorous guarantees for nonconvex gradient descent focused on specific classes of matrix estimation problems. For instance, the recent papers [66, 58] study

exact recovery in the setting of noiseless matrix sensing (i.e., solving linear matrix equalities with random designs). Focusing on the rank-one setting, De et al. [25, 26] study the noiseless matrix completion problem, and a stochastic version of nonconvex gradient descent; they prove global convergence with a constant success probability, assuming independence between the samples used by each iteration. The recent manuscript [57] studies several variants of nonconvex gradient descent algorithms, again for noiseless matrix completion. Another line of work [17, 23] considers the phase retrieval problem, which can be reformulated as recovering a rank one ($r = 1$) matrix from random quadratic measurements. The regularity conditions imposed in this work bear some similarity with our conditions, but their validation requires a very different analysis. An attractive feature of phase retrieval is that it is known to be locally convex around the global optimum under certain settings [55, 64].

The work in this paper develops a unified framework for analyzing the behavior of projected gradient descent in application to low-rank estimation problems, covering many of the models described above as well as various others. Our theory applies to matrices of arbitrary rank r, and is framed in the statistical setting of noisy observations, allowing for noiseless observations as a special case, When specialized to particular models, our framework yields a variety of corollaries providing guarantees for concrete statistical models that have not been studied in the work above. Notably, our general conditions *do not* depend on local convexity, and thus can be applied to models such as sparse PCA and clustering in which no form of local convexity holds. (In fact, our results apply even when the loss function is globally concave). In addition, we impose only a natural gradient smoothness condition that is much less restrictive than the vanishing gradient condition imposed in other work. Thus, one of the main contributions of this paper is to illuminate to weakest known conditions under which nonconvex gradient descent can succeed, and also allows for applications to several problems that lack local convexity and vanishing gradients.

It is also worth noting that other types of algorithms for nonconvex problems have also been analyzed, including alternating minimization [37, 32, 34], EM algorithms [4, 63] and power methods [33], various hard-thresholding and singular value projection [36, 53, 38, 9], gradient descent for nonconvex regression and spectrally sparse recovery problems [47, 61, 13], as well as gradient descent on Grassmannian manifolds [40, 65]. Finally, there is a large body of work on convex-optimization based approach to the concrete examples considered in this paper. We compare our statistical guarantees with results of these types after the statements of each of our corollaries.

Notation: The i-th row and j-th column of a matrix Z are denoted by $Z_{i\cdot}$ and $Z_{\cdot j}$, respectively. The spectral norm $|||Z|||_{\mathrm{op}}$ is the largest singular value of Z. The nuclear norm $|||Z|||_{\mathrm{nuc}}$ is the sum of the singular values of Z. For parameters $1 \leq a, b \leq \infty$ and a matrix Z, the ℓ_a/ℓ_b norm of Z is $|||Z|||_{b,a} = \left(\sum_i \|Z_{i\cdot}\|_b^a\right)^{\frac{1}{a}}$—that is, the ℓ_a norm of the vector of the ℓ_b norms of the rows. Special cases include the Frobenius norm $|||Z|||_{\mathrm{F}} = |||Z|||_{2,2}$, the elementwise ℓ_1 norm $\|Z\|_1 = |||Z|||_{1,1}$ and the elementwise ℓ_∞ norm $\|Z\|_\infty = |||Z|||_{\infty,\infty}$. For a convex set T, we use Π_T to denote the Euclidean projection onto T.

2 Background

We begin by setting up the class of matrix estimators to be studied in this paper, and then providing various concrete examples of specific models to which our general theory applies.

2.1 Matrix estimators in the factorized formulation

Letting $\mathcal{S}^{d\times d}$ denote the space of all symmetric d-dimensional matrices, this paper focuses on a class of matrix estimators that take the following general form. For a given sample size $n \geq 1$, let $\mathcal{L}_n : \mathcal{S}^{d\times d} \to \mathbb{R}$ be a cost function. It is a random function, since it depends (implicitly in our notation) on the observed data, and the function value $\mathcal{L}_n(M)$ provides some measure of fit of the matrix M to the given data. For a given convex set $\mathcal{M} \subseteq \mathcal{S}^{d\times d}$, we then consider a minimization problem of the form

$$\min_{M\in\mathcal{S}^{d\times d}} \mathcal{L}_n(M) \qquad \text{such that } M \succeq 0 \text{ and } M \in \mathcal{M}. \tag{1}$$

The goal of solving this optimization problem is to estimate some unknown target matrix M^*. Typically, the target matrix is a (near)-minimizer of the population version of the program—that is, a solution to the same constrained minimization problem with $\mathcal{L}_n$ replaced by its expectation $\bar{\mathcal{L}}(M) = \mathbb{E}[\mathcal{L}_n(M)]$. However, our theory does not require that M^* minimizes this quantity, nor that the gradient $\nabla\bar{\mathcal{L}}(M^*)$ vanish.

In many cases, the matrix M^* either has low rank, or can be well-approximated by a matrix of low rank. Concretely, if the target matrix M^* has rank $r < d$, then it can be written in the outer product form $M^* = F^*\otimes F^*$ for some other matrix $F^* \in \mathbb{R}^{d\times r}$ with orthogonal columns. This factorized representation motivates us to consider the function $\widetilde{\mathcal{L}}_n(F) := \mathcal{L}_n(F\otimes F)$, and the factorized formulation

$$\min_{F\in\mathbb{R}^{d\times r}} \widetilde{\mathcal{L}}_n(F) \qquad \text{such that } F \in \mathcal{F}, \tag{2}$$

where $\mathcal{F}$ is some convex set that contains F^*, and for which the set $\{F\otimes F \mid F \in \mathcal{F}\}$ acts as a surrogate for $\mathcal{M}$. Note that due to the factorized representation of the low-rank matrix, this factorized program is (in general) nonconvex, and is typically so even if the original program (1) is convex.

Nonetheless, we can apply a projected gradient descent method in order to compute an approximate minimizer. For this particular problem, the projected gradient descent updates take the form

$$F^{t+1} = \Pi_{\mathcal{F}}\Big(F^t - \eta^t \nabla\widetilde{\mathcal{L}}_n(F^t)\Big) \tag{3}$$

where $\eta^t > 0$ is a step size parameter, $\Pi_{\mathcal{F}}$ denotes the Euclidean projection onto the set $\mathcal{F}$, and the gradient[1] is given by $\nabla\widetilde{\mathcal{L}}_n(F) = \big[\nabla_M\mathcal{L}_n(F\otimes F) + (\nabla_M\mathcal{L}_n(F\otimes F))^\top\big]F$. The main goal of this paper is to provide a general set of sufficient conditions under which—up to a statistical tolerance term ε_n—the sequence $\{F^t\}_{t=0}^\infty$ converges to some F^* such that $F^*\otimes F^* = M^*$.

A significant challenge in the analysis is the fact that there are *many* possible factorizations of the form $M^* = F^*\otimes F^*$. In order to address this issue, it is convenient to define an equivalent class of valid solutions as follows

$$\mathcal{E}(M^*) := \{F^* \in \mathbb{R}^{d\times r} \mid F^*\otimes F^* = M^*, F^{*\top}_{\cdot i} F^*_{\cdot j} = 0, \forall i \neq j\}. \tag{4}$$

[1]This gradient takes the simpler form $\nabla\widetilde{\mathcal{L}}_n(F) = 2\nabla_M\mathcal{L}_n(F\otimes F)F$ whenever $\nabla\mathcal{L}_n(F\otimes F)$ is symmetric, which is the case in the concrete examples that we treat.

For the applications of interest here, the underlying goal is to obtain a good estimate of *any* matrix in the set $\mathcal{E}(M^*)$. In particular, such an estimate implies a good estimate of M^* itself as well as the column space and singular values of all the members of the class $\mathcal{E}(M^*)$. Accordingly, we define the pseudometric

$$\mathrm{d}(F, F^*) := \min_{F^* \in \mathcal{E}(M^*)} |||F - F^*|||_{\mathrm{F}}. \tag{5}$$

Note that all matrices $F^* \in \mathcal{E}(M^*)$ have the same singular values, so that we may write the singular values $\sigma_1(F^*) \geq \cdots \geq \sigma_r(F^*) > 0$ as well as $|||F^*|||_{\mathrm{op}}$ and $|||F^*|||_{\mathrm{F}}$ without any ambiguity. In fact, this invariant property holds more generally for any function of the sorted singular values and column space of F^* (e.g., any unitarially invariant norm).

2.2 Illustrative examples

Let us now consider a few specific models to illustrate the general set-up from the previous section. We return to demonstrate consequences of our general theory for these (and other) models in Section 4.

2.2.1 Matrix regression

We begin with a simple example, namely one in which we make noisy observations of linear projections of an unknown low-rank matrix $M^* \in \mathcal{S}^{d \times d}$. In particular, suppose that we are given n i.i.d. observations $\{(y_i, X_i)\}_{i=1}^n$ of the form

$$y_i = \mathrm{trace}(X^i M^*) + \epsilon_i \quad \text{for } i = 1, \ldots, n, \tag{6}$$

and $\{\epsilon_i\}_{i=1}^n$ is some i.i.d. sequence of zero-mean noise variables. The paper [50] provides various examples of such matrix regression problems, depending on the particular choice of the regression matrices $\{X^i\}_{i=1}^n$.

Original estimator: Without considering computational complexity, a reasonable estimate of M^* would be based on minimizing the least-squares cost

$$\mathcal{L}_n(M) := \frac{1}{2n} \sum_{i=1}^n \big(y_i - \mathrm{trace}(X^i M)\big)^2 \tag{7}$$

subject to a rank constraint. However, this problem is computationally intractable in general due to the nonconvexity of the rank function. A standard convex relaxation is based on the nuclear norm $|||M|||_{\mathrm{nuc}} := \sum_{j=1}^d \sigma_j(M)$, corresponding to the sum of the singular values of the matrix. In the symmetric PSD case, it is equivalent to the trace of the matrix. Using the nuclear norm as regularizer leads to the estimator

$$\min_{M \in \mathcal{S}^{d \times d}} \Big\{ \frac{1}{2n} \sum_{i=1}^n \big(y_i - \mathrm{trace}(X^i M)\big)^2 \Big\} \qquad \text{such that } M \succeq 0 \text{ and } |||M|||_{\mathrm{nuc}} \leq R,$$

where $R > 0$ is a radius to be chosen. This is a special case of our general estimator (1) with $\mathcal{L}_n$ being the least-squares cost (7), and the constraint set $\mathcal{M} = \{M \in \mathcal{S}^{d \times d} \mid |||M|||_{\mathrm{nuc}} \leq R\}$.

Population version: Suppose that the noise variables ϵ_i are i.i.d. zero-mean with variance σ^2, and the regression matrices $\{X_i\}_{i=1}^n$ are also i.i.d., zero-mean and such that $\mathbb{E}[\text{trace}(X_i M)^2] = |||M|||_{\text{F}}^2$ for any matrix M. Under these conditions, an easy calculation yields that the population cost function is given by $\bar{\mathcal{L}}(M) = \frac{1}{2}|||M - M^*|||_{\text{F}}^2 + \frac{1}{2}\sigma^2$. For this particular case, note that M^* is the unique minimizer of the population cost.

Projected gradient descent: The factorized cost function is given by

$$\widetilde{\mathcal{L}}_n(F) = \frac{1}{2n}\sum_{i=1}^{n}\Big\{y_i - \text{trace}\left(X^i(F\otimes F)\right\}^2, \tag{8}$$

and has gradient $\nabla\widetilde{\mathcal{L}}_n(F) = \frac{2}{n}\sum_{i=1}^n(y_i - \text{trace}\left(X^i(F\otimes F)\right)(X^i)^T F$ assuming each X^i is symmetric. Setting $\mathcal{F} = \mathbb{R}^{d\times r}$, the projected gradient descent updates (3) reduce to usual gradient descent—that is,

$$F^{t+1} = F^t - \eta^t\nabla\widetilde{\mathcal{L}}_n(F^t), \qquad \text{for } t = 0, 1, \ldots.$$

We return to analyze these updates in Section 4.2.

2.2.2 Rank-r PCA with row sparsity

Principal component analysis is a widely used method for dimensionality reduction. For high-dimensional problems in which $d \gg n$, it is well-known that classical PCA is inconsistent [39]. Moreover, minimax lower bounds show that consistent eigen-estimation is impossible in the absence of structure in the eigenvectors. Accordingly, a recent line of work (e.g., [39, 3, 14, 6, 59, 10]) has studied different forms of PCA with structured eigenvectors.

Here we consider one such form of structured PCA, namely a rank r model with row-wise sparsity. For a given signal-to-noise ratio $\gamma > 0$ and an orthonormal matrix $F^* \in \mathbb{R}^{d\times r}$, consider a covariance matrix of the form

$$\Sigma = \gamma\underbrace{\left(F^*\otimes F^*\right)}_{M^*} + I_d. \tag{9}$$

By construction, the columns of F^* span the top rank-r eigenspace of Σ with the corresponding maximal eigenvalues $\gamma + 1$. In the row-sparse version of this model [59], this leading eigenspace is assumed to be supported on k coordinates—that is, the matrix F^* has at most k non-zero rows. Given n i.i.d. samples $\{x_i\}_{i=1}^n$ from the Gaussian distribution $N(0, \Sigma)$, the goal of sparse PCA is to estimate the sparse eigenspace spanned by F^*.

Original estimator: A natural estimator is based on a semidefinite program, referred to as the Fantope relaxation in the paper [60], given by

$$\min_{\substack{M\in\mathcal{S}^{d\times d}\\ 0\preceq M\preceq I_d}}\Big\{-\text{trace}(\widehat{\Sigma}_n M)\Big\} \qquad \text{such that } \text{trace}(M) \le r \text{ and } \|M\|_1 \le R, \tag{10}$$

where $\widehat{\Sigma}_n$ is the empirical covariance matrix, and $R > 0$ is a radius to be chosen. This is a special case of our general set-up with $\mathcal{L}_n(M) = -\text{trace}(\widehat{\Sigma}_n M)$ and

$$\mathcal{M} := \Big\{M \in \mathcal{S}^{d\times d} \mid 0 \preceq M \preceq I_d,\ \text{trace}(M) \le r \text{ and } \|M\|_1 \le R\Big\}.$$

Population version: Since $\mathbb{E}[\widehat{\Sigma}_n] = \Sigma$, the population cost function is given by

$$\bar{\mathcal{L}}(M) = \mathbb{E}[\mathcal{L}_n(M)] = -\operatorname{trace}(\Sigma M).$$

Thus, by construction, for any radius $R \geq \|F^* \otimes F^*\|_1$, the matrix $M^* = F^* \otimes F^*$ is the unique minimizer of the population version of the problem (10), subject to the constraint $M \in \mathcal{M}$.

Projected gradient descent: For a radius $\widetilde{R}$ to be chosen, we consider a factorized version of the SDP

$$\widetilde{\mathcal{L}}_n(F) := -\langle\langle \widehat{\Sigma}_n, \, F \otimes F \rangle\rangle, \qquad \mathcal{F} := \big\{ F \in \mathbb{R}^{d \times r} \mid |||F|||_{\mathrm{op}} \leq 1, \, \|F\|_{2,1} \leq \widetilde{R} \big\}, \tag{11}$$

where we recall that $\|F\|_{2,1} = \sum_{i=1}^{d} \|F_i\|_2$. This norm is the appropriate choice for selecting matrices with sparse rows, as assumed in our initial set-up. We return in Section 4.3 to analyze the projected gradient updates (3) applied to pair $(\widetilde{\mathcal{L}}_n, \mathcal{F})$ in equation (11).

As a side-comment, this example illustrates that our theory does not depend on local convexity of the function $\widetilde{\mathcal{L}}_n$. In this case, even though the original function $\mathcal{L}_n$ is convex (in fact, linear) in the matrix $M \in \mathcal{S}^{d \times d}$, observe that that the function $\widetilde{\mathcal{L}}_n$ from equation (11) is never locally convex in the low-rank matrix $F \in \mathbb{R}^{d \times r}$; in fact, since $\widehat{\Sigma}_n$ is positive semidefinite, it is a globally concave function.

2.2.3 Low-rank and sparse matrix decomposition

There are various applications in which it is natural to model an unknown matrix as the sum of two matrices, one of which is low-rank and the other of which is sparse. Concretely, suppose that we make observations of the form $Y = M^* + S^* + E$ where M^* is low-rank, the matrix S^* is symmetric and elementwise-sparse, and E is a symmetric matrix of noise variables. Many problems can be cast in this form, including robust forms of PCA, factor analysis, and Gaussian graphical model estimation; see the papers [19, 1, 22, 20, 35] and references therein for further details on these and other applications.

Original estimator: Letting $S_j \in \mathbb{R}^d$ denote the j^{th} column of a matrix $S \in \mathbb{R}^{d \times d}$, define the set of matrices $\mathcal{S} := \{S \in \mathbb{R}^{d \times d} \mid \|S_j\|_1 \leq R_j \quad \text{for } j = 1, 2, \ldots, d\}$, where $(R_1, \ldots, R_d)$ are user-defined radii. Using the nuclear norm and ℓ_1 norm as surrogates for rank and sparsity respectively, a popular convex relaxation approach is based on the SDP

$$\min_{M \in \mathcal{S}^{d \times d}} \Big\{ \frac{1}{2} \min_{S \in \mathcal{S}} |||Y - (M + S)|||_{\mathrm{F}}^2 \Big\} \qquad \text{subject to } M \succeq 0 \text{ and } |||M|||_{\mathrm{nuc}} \leq R,$$

This is a special case of our general estimator with $\mathcal{L}_n(M) := \frac{1}{2} \min_{S \in \mathcal{S}} |||Y - (M + S)|||_{\mathrm{F}}^2$, and the constraint set $\mathcal{M} := \{M \in \mathcal{S}^{d \times d} \mid |||M|||_{\mathrm{nuc}} \leq R\}$.

Population version: In this case, the population function is given by

$$\bar{\mathcal{L}}(M) := \mathbb{E}\Big[\frac{1}{2} \min_{S \in \mathcal{S}} |||Y - (M + S)|||_{\mathrm{F}}^2 \Big\} \Big],$$

where the expectation is over the random noise matrix E. In general, we are not guaranteed that M^* is the unique minimizer of this objective, but our analysis shows that (under suitable conditions) it is a near-minimizer, and this is adequate for our theory.

Projected gradient descent: In this paper, we analyze a version of gradient descent that operates on the pair $(\widetilde{\mathcal{L}}_n, \mathcal{F})$ given by

$$\widetilde{\mathcal{L}}_n(F) = \frac{1}{2} \min_{S \in \mathcal{S}} |||Y - ((F \otimes F) + S)|||_{\mathrm{F}}^2, \quad \text{and} \quad \mathcal{F} := \Big\{ F \in \mathbb{R}^{d \times r} \mid \|F\|_{2,\infty} \leq \sqrt{\frac{2\mu}{d}} |||F^0|||_{\mathrm{F}} \Big\}. \quad (12)$$

Here F^0 is the initialization of the algorithm, and the parameter $\mu > 0$ controls the matrix incoherence. See Sections 4.1 and 4.6 for discussion of matrix incoherence parameters, and their necessity in such problems. The gradient of $\widetilde{\mathcal{L}}_n$ takes the form

$$\nabla \widetilde{\mathcal{L}}_n(F) = 2 \Big\{ \Pi_{\mathcal{S}}\big(Y - (F \otimes F)\big) - \big(Y - (F \otimes F)\big) \Big\} F,$$

where $\Pi_{\mathcal{S}}$ denotes projection onto the constraint set $\mathcal{S}$. This projection is easy to carry it, as it simply involves a soft-thresholding of the columns of the matrix. Likewise, the projection onto the set $\mathcal{F}$ from equation (12) is easy to carry out. We return to analyze these projected gradient updates in Section 4.6.

In addition to the three examples introduced so far, our theory also applies to various other low-rank estimation problems, including that of matrix completion with real-valued observations (Section 4.1) and binary observations (Section 4.5), as well as planted clustering problems (Section 4.4).

3 Main results

In this section, we turn to the set-up and statement of our main results on the convergence properties of projected gradient descent for low-rank factorizations. We begin in Section 3.1 by stating the conditions on the function $\widetilde{\mathcal{L}}_n$ and $\mathcal{F}$ that underlie our analysis. In Section 3.2, we state a result (Theorem 1) that guarantees sublinear convergence, whereas Section 3.3 is devoted to a result (Theorem 2) that guarantees faster linear convergence under slightly stronger assumptions. In Section 4 to follow, we derive various corollaries of these theorems for different concrete versions of low-rank estimation.

Given a radius $\rho > 0$, we define the ball $\mathbb{B}_2(\rho; F^*) := \big\{ F \in \mathbb{R}^{d \times r} \mid \mathrm{d}(F, F^*) \leq \rho \big\}$. At a high level, our goal is to provide conditions under which the projected gradient sequence $\{F^t\}_{t=0}^{\infty}$ converges some multiple of the ball $\mathbb{B}_2(\varepsilon_n; F^*)$, where $\varepsilon_n > 0$ is a *statistical tolerance*.

3.1 Conditions on the pair $(\widetilde{\mathcal{L}}_n, \mathcal{F})$

Recall the definition of the set $\mathcal{E}(M^*)$ of equivalent orthogonal factorizations of a given matrix M^*. We begin with a condition on $\mathcal{F}$ that guarantees that it respects the structure of this set.

M^*-faithfulness of $\mathcal{F}$: For a radius ρ, the constraint set $\mathcal{F}$ is said to be *M^*-faithful* if for each matrix $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$, we guaranteed that

$$\arg \min_{A \in \mathcal{E}(M^*)} |||A - F|||_{\mathrm{F}} \subseteq \mathcal{F}. \quad (13)$$

Of course, this condition is implied by the inclusion $\mathcal{E}(M^*) \subseteq \mathcal{F}$. The M^*-faithfulness condition is natural for our setting, as our goal is to estimate the eigen structure of M^*, and the set $\mathcal{F}$ should therefore represent prior knowledge of this structure and be independent of a specific factorization of M^*.

Local descent condition: Our next condition provides a guarantee on the cost improvement that can be obtained by taking a gradient step when starting from any matrix F that is "sufficiently" far away from the set $\mathcal{E}(M^*)$.

Definition 1 (Local descent condition)**.** For a given radius $\rho > 0$, curvature parameter $\alpha > 0$ and statistical tolerance $\varepsilon_n \geq 0$, a cost function $\widetilde{\mathcal{L}}_n$ satisfies a *local descent condition* with parameters $(\alpha, \beta, \varepsilon_n, \rho)$ over $\mathcal{F}$ if for each $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$, there is some $F_{\pi^*} \in \arg\min_{A \in \mathcal{E}(M^*)} |||A - F|||_{\mathrm{F}}$ such that

$$\langle\langle \nabla \widetilde{\mathcal{L}}_n(F),\ F - F^* \rangle\rangle \geq \alpha |||F - F_{\pi^*}|||_{\mathrm{F}}^2 - \frac{\beta^2}{\alpha} |||F_{\pi^*} - F^*|||_{\mathrm{F}}^2 - \alpha \varepsilon_n^2, \quad \forall F^* \in \mathcal{E}(M^*). \tag{14}$$

In order to gain intuition for this condition, note that by a first-order Taylor series expansion, we have $\widetilde{\mathcal{L}}_n(F) - \widetilde{\mathcal{L}}_n(F_{\pi^*}) \approx \langle\langle \nabla \widetilde{\mathcal{L}}_n(F),\ F - F_{\pi^*} \rangle\rangle$, so that this inner product measures the potential gains afforded by taking a gradient step. Now consider some matrix F such that $|||F - F_{\pi^*}|||_{\mathrm{F}} > \sqrt{2}\varepsilon_n$, so that its distance from $\mathcal{E}(M^*)$ is larger than the statistical precision. The lower bound (14) with $F^* = F_{\pi^*}$then implies that

$$\widetilde{\mathcal{L}}_n(F) - \widetilde{\mathcal{L}}_n(F_{\pi^*}) \approx \langle\langle \nabla \widetilde{\mathcal{L}}_n(F),\ F - F_{\pi^*} \rangle\rangle \geq \frac{\alpha}{2} |||F - F_{\pi^*}|||_{\mathrm{F}}^2,$$

which guarantees a quadratic descent condition. Note that the condition (14) actually allows for additional freedom in the choice of F^* so as to accommodate the non-uniqueness of the factorization.

One way in which to establish a bound of the form (14) is by requiring that $\widetilde{\mathcal{L}}_n$ be locally strongly convex, and that the gradient $\nabla \widetilde{\mathcal{L}}_n(F_{\pi^*})$ approximately vanishes. In particular, suppose $\widetilde{\mathcal{L}}_n$ is 2α-strongly convex over the set $\mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$, in the sense that

$$\langle\langle \nabla \widetilde{\mathcal{L}}_n(F) - \nabla \widetilde{\mathcal{L}}_n(F_{\pi^*}),\ F - F_{\pi^*} \rangle\rangle \geq 2\alpha |||F - F_{\pi^*}|||_{\mathrm{F}}^2 \qquad \text{for all } F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*).$$

If we assume that $|||\nabla \widetilde{\mathcal{L}}_n(F_{\pi^*})|||_{\mathrm{F}} \leq \alpha \varepsilon_n$, then some simple algebra yields that the lower bound (14) holds.

However, it is essential to note that our theory covers several examples in which a lower bound (14) of the form holds, even though $\widetilde{\mathcal{L}}_n$ fails to be locally convex, and/or the gradient $\nabla \widetilde{\mathcal{L}}_n(F_{\pi^*})$ does not approximately vanish.[2] Examples include the problem of sparse PCA, previously introduced in Section 2.2.2; in this case, the function $\widetilde{\mathcal{L}}_n$ is actually globally concave, but nonetheless our analysis in Section 4.3 shows that a local descent condition of the form (14) holds. Similarly, for the planted clustering model studied in Section 4.4, the same form of global concavity holds. In addition, for the matrix regression problem previously introduced in Section 2.2.1, we prove in Section 4.2 that the condition (14) holds over a set over which $\widetilde{\mathcal{L}}_n$ is nonconvex. The generality of our condition (14) is essential to accommodate these and other examples.

[2]We note that the vanishing gradient condition is needed in all existing work on nonconvex gradient descent [17, 57, 66].

Local Lipschitz condition: Our next requirement is a straightforward local Lipschitz property:

Definition 2 (Local Lipschitz)**.** The loss function $\widetilde{\mathcal{L}}_n$ is locally Lipschitz in the sense that

$$\|\!|\nabla \widetilde{\mathcal{L}}_n(F)|\!\|_{\mathrm{F}} \leq L |\!|\!| F^* |\!|\!|_{\mathrm{op}}. \tag{15}$$

for all $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$.

Local smoothness: Our last condition is *not* required to establish convergence of projected gradient descent, but rather to guarantee a faster geometric rate of convergence. It is a condition—complementary to the local descent condition—that upper bounds the behavior of the gradient map $\nabla \widetilde{\mathcal{L}}_n$.

Definition 3 (Local smoothness)**.** For some curvature and smoothness parameters α and β, statistical tolerance ε_n and radius ρ, we say that the loss function $\mathcal{L}_n$ satisfies a *local smoothness condition* with parameters $(\alpha, \beta, \varepsilon_n, \rho)$ over $\mathcal{F}$ if for each $F, F' \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$ and $F^* \in \mathcal{E}(M^*)$,

$$|\langle\!\langle \nabla_M \widetilde{\mathcal{L}}_n(F) - \nabla_M \widetilde{\mathcal{L}}_n(F'),\ F - F^* \rangle\!\rangle| \leq \big(\beta |\!|\!| F - F' |\!|\!|_{\mathrm{F}} + \alpha \varepsilon_n\big) |\!|\!| F - F^* |\!|\!|_{\mathrm{F}}. \tag{16}$$

The above conditions are stated in terms of the loss function $\widetilde{\mathcal{L}}_n$ for the factor matrix F. Alternatively, one may restate these conditions in terms of the loss function $\mathcal{L}_n$ on the original space, and we make use of this type of reformulation in parts of our proofs. For instance, see Section 6 for details.

3.2 Sublinear convergence under Lipschitz condition

With our basic conditions in place, we are now ready to state our first main result. It guarantees a sublinear rate of convergence under the M^*-faithfulness, local descent, and local Lipschitz conditions.

More precisely, for some descent and Lipschitz parameters $\alpha \leq L$, a statistical tolerance $\varepsilon_n \geq 0$, and a constant $\tau \in (0, \frac{1}{2})$, suppose that $\varepsilon_n \leq \frac{1-\tau}{2}\sigma_r(F^*)$, the cost functions $\widetilde{\mathcal{L}}_n$ satisfies the local descent and Lipschitz conditions (Definitions 1 and 2) with parameters α, L, ε_n and $\rho = (1-\tau)\sigma_r(F^*)$, and the constraint set $\mathcal{F}$ is M^*-faithful and convex. Let $\kappa = \kappa(F^*) := \frac{\sigma_1(F^*)}{\sigma_r(F^*)}$ be the condition number of F^*. We then have the following guarantee:

Theorem 1. *Under the previously stated conditions, given any initial point F^0 belonging to the set $\mathcal{F} \cap \mathbb{B}_2((1-\tau)\sigma_r(F^*); F^*)$, the projected gradient iterates $\{F^t\}_{t=1}^{\infty}$ with step size $\eta^t = \frac{1}{\alpha(t + 20\kappa^2 L^2/\alpha^2)}$ satisfy the bound*

$$\mathrm{d}^2(F^t, F^*) \leq \frac{20 L^2 |\!|\!| F^* |\!|\!|_{op}^2}{t\alpha^2} + 4\varepsilon_n^2 \qquad \textit{for all iterations } t = 1, 2, \ldots. \tag{17}$$

See Section 5.1 for the proof of this claim.

As a minor remark, we note that the assumption $\varepsilon_n \leq \frac{1-\tau}{2}\sigma_r(F^*)$ entails no loss of generality—if it fails to hold, then the initial solution F^0 already satisfies an error bound better than what is guaranteed for subsequent iterates.

Conceptually, Theorem 1 provides a minimal set of conditions for the convergence of projected gradient descent using the nonconvex factorization $M = F \otimes F$. The first term on the right hand side

of equation (17) corresponds to the *optimization error*, whereas the second ε_n^2 term is the *statistical error*. The bound (17) shows that the distance between F^t and F^* drops at the rate $\mathcal{O}(\frac{1}{t})$ up to the statistical limit ε_n^2 that is determined by the sample size and the signal-to-noise ratio (SNR) of the problem. We see concrete instances of this statistical error in the examples to follow.

3.3 Linear convergence under smoothness condition

Although Theorem 1 does guarantee convergence, the resulting rate is sublinear ($\mathcal{O}(1/t)$), and hence rather slow. In this section, we show that if in addition to the local Lipschitz and descent conditions, the function $\widetilde{\mathcal{L}}_n$ satisfies the local smoothness conditions in Definition 3, then much faster convergence can be guaranteed.

More precisely, suppose that for some numbers α, β, L, ε_n and τ with $0 < \alpha \leq \beta = L$, $0 < \tau < 1$ and $\varepsilon_n \leq \frac{1-\tau}{4}\sigma_r(F^*)$, the loss function $\widetilde{\mathcal{L}}_n$ satisfies the local descent, Lipschitz and smoothness conditions in Definitions 1–3 over $\mathcal{F}$ with parameters α, β, L, ε_n and $\rho = (1-\tau^2)\sigma_r(F^*)$, and that the set $\mathcal{F}$ is M^*-faithful and convex.

Theorem 2. *Under the previously stated conditions, there is a constant $0 < c_\tau < 1$ depending only on τ such that given an initial matrix F^0 in the set $\mathcal{F} \cap \mathbb{B}_2((1-\tau)\sigma_r(F^*); F^*)$, the projected gradient iterates $\{F^t\}_{t=1}^{\infty}$ with step size $\eta^t = c_\tau \frac{\alpha}{\kappa^6 \beta^2}$ satisfy the bound*

$$\mathrm{d}^2(F^t, F^*) \leq \Big(1 - c_\tau \frac{\alpha^2}{\kappa^6 \beta^2}\Big)^t \mathrm{d}^2(F^0, F^*) + 16\varepsilon_n^2 \qquad \text{for all iterations } t = 1, 2, \ldots. \tag{18}$$

See Section 5.2 for the proof of this claim.

The right hand side of the bound (18) again consists of an optimization error term and a statistical error term. The theorem guarantees that the optimization error converges linearly at the geometric rate $\mathcal{O}((1-c)^t)$ up to a statistical limit. Note that the theorem requires the initial solution F^0 to lie within a ball around F^* with radius $(1-\tau)\sigma_r(F^*)$, which is slightly smaller than the radius $\rho = (1-\tau^2)\sigma_r(F^*)$ for which the local descent, Lipschitz and smoothness conditions hold. Moreover, the step size and the convergence rate depend on the condition number of F^* as well as the quality of the initialization through τ. We did not make an attempt to optimize this dependence, but improvement in this direction, including adaptive choices of the step size, is certainly an interesting problem for future work.

4 Concrete results for specific models

In this section, we turn to the consequences of our general theory for specific models that arise in applications. Throughout this section, we focus on geometric convergence guaranteed by Theorem 2 using a constant step size. The main technical challenges are to verify the local descent, local Lipschitz and local smoothness assumptions that are needed to apply this result. Since Theorem 1 depends on weaker assumptions—it does not need the local smoothness property—it should be understood also that our analysis can be used to derive corollaries based on Theorem 1 as well.

Note: In all of the analysis to follow, we adopt the shorthand $\sigma_j = \sigma_j(F^*)$ for the singular values of F^*, and $\kappa = \frac{\sigma_1}{\sigma_r}$ for its condition number.

4.1 Noisy matrix completion

We begin by deriving a corollary for the problem of noisy matrix completion. Since we did not discuss this model in Section 2.2, let us provide some background here. There are a wide variety of matrix completion problems (e.g., [43]), and the variant of interest here arises when the unknown matrix has low rank. More precisely, for an unknown PSD and low-rank matrix $M^* \in \mathcal{S}^{d\times d}$, suppose that we are given noisy observations of a subset of its entries. In the so-called Bernoulli model, the random subset of observed entries is chosen uniformly at random—that is, each entry is observed with some probability p, independently of all other entries. We can represent these observations by a random symmetric matrix $Y \in \mathcal{S}^{d\times d}$ with entries of the form

$$Y_{ij} = \begin{cases} M^*_{ij} + E_{ij}, & \text{with probability } p, \text{ and} \\ * & \text{otherwise.} \end{cases}, \quad \text{for each } i \geq j. \tag{19}$$

Here the variables $\{E_{ij}, i \geq j\}$ represent a form of measurement noise.

A standard method for matrix completion is based on solving the semidefinite program

$$\min_{M\in\mathcal{S}^{d\times d}} \underbrace{\Big\{ \frac{1}{2p} \sum_{(i,j)\in\Omega} (M_{ij} - Y_{ij})^2 \Big\}}_{\mathcal{L}_n(M)} \qquad \text{such that } M \succeq 0 \quad \text{and} \quad |||M|||_{\text{nuc}} \leq R, \tag{20}$$

where $R > 0$ is a radius to be chosen. As noted above, the PSD constraint and nuclear norm bound are equivalent to the trace constraint $\text{trace}(M) \leq R$. In either case, this is a special case of our general estimator (1).

The SDP-based estimator (20) is known to have good performance when the underlying matrix M^* satisfies certain matrix incoherence conditions. These conditions involve its leverage scores, defined in the following way. Here we consider a simplified setting where the eigenvalues of M^* are equal. By performing an eigendecomposition, we can write $M^* = UDU^T$ where $D \in \mathbb{R}^{r\times r}$ is a diagonal matrix of eigenvalues (a constant multiple of the identity when they are constant), and take $F^* = UD^{1/2}$. With this notation, the *incoherence* parameter of $M^* = F^* \otimes F^*$ is given by

$$\mu := \frac{d \max_{i=1,\ldots,d} \|F^*_{i\cdot}\|_2^2}{r|||F^*|||_{\text{op}}^2} = \frac{d\|F^*\|_{2,\infty}^2}{r|||F^*|||_{\text{op}}^2}. \tag{21}$$

Since we already enforce low-rankness in the factorized formulation, we can drop nuclear norm constraint. The generalized projected gradient descent (3) is specified by letting $\widetilde{\mathcal{L}}_n$ and $\mathcal{F}$ set

$$\widetilde{\mathcal{L}}_n(F) := \frac{1}{2p} \sum_{(i,j)\in\Omega} \big((F\otimes F)_{ij} - Y_{ij}\big)^2 \quad \text{and} \quad \mathcal{F} := \Big\{ F \in \mathbb{R}^{d\times r} \;\mid\; \|F\|_{2,\infty} \leq \sqrt{\frac{2\mu}{d}} |||F^0|||_{\text{F}} \Big\}.$$

Note that $\mathcal{F}$ is convex, and depends on the initial solution M^0. The gradient of $\mathcal{L}_n$ is $\nabla_M \mathcal{L}_n(M) = \frac{1}{p}\Pi_\Omega(M - Y)$, and the projection $\Pi_{\mathcal{F}}$ is given by the row-wise "clipping" operation

$$[\Pi_{\mathcal{F}}(\theta)]_{i\cdot} = \begin{cases} F_{i\cdot}\,, & \|F_{i\cdot}\|_2 \leq \sqrt{\frac{2\mu r}{d}} |||F^0|||_{\text{op}}, \\ F_{i\cdot} \sqrt{\frac{2\mu r}{d}} \frac{|||F^0|||_{\text{op}}}{\|F_{i\cdot}\|_2}\,, & \|F_{i\cdot}\|_2 > \sqrt{\frac{2\mu r}{d}} |||F^0|||_{\text{op}}, \end{cases} \quad \text{for } i = 1, 2, \ldots, d.$$

This projection ensures that the iterates of gradient descent (3) remains incoherent.

Remark 1. Note that $\|F^*_{i\cdot}\|_2^2 = \|\Pi_{\mathrm{col}(F^*)}(e_i)\|_2^2$ (e_i is the i-th standard basis vector and $\mathrm{col}(F^*)$ is the column space of F^*), so the values of $\|F^*\|_{2,\infty}$ and μ depend only on $\mathrm{col}(F^*)$ and are the same for any F^* in $\mathcal{E}(M^*)$.

With this notation in place, we are now ready to apply Theorem 2 to the noisy matrix completion problem. As we show below, if the initial matrix F^0 satisfies the bound $\mathrm{d}(F^0, F^*) \leq \frac{1}{5}|||F^*|||_{\mathrm{op}}$, then the set $\mathcal{F}$ is M^*-faithful. Moreover, if the expected sample size satisfies $n = pd^2 \gtrsim \max\{\mu r d \log d, \mu^2 r^2 d\}$, then with probability at least $1 - 4d^{-3}$ the loss function $\widetilde{\mathcal{L}}_n$ satisfies the local descent, Lipschitz and smoothness conditions with parameters

$$\rho = \frac{3}{5}|||F^*|||_{\mathrm{op}}, \quad \alpha = \frac{2}{25}|||F^*|||_{\mathrm{op}}^2, \quad L = \beta = c_2 \mu r |||F^*|||_{\mathrm{op}}^2 \quad \text{and} \quad \varepsilon_n = 100 \frac{\sqrt{r}|||\Pi_\Omega(E)|||_{\mathrm{op}}}{p|||F^*|||_{\mathrm{op}}}.$$

Using this fact, we have the following consequence of Theorem 2, which holds when the sample size size n satisfies the bound above and is large enough to ensure that $\varepsilon_n \leq \frac{1}{10}|||F^*|||_{\mathrm{op}}$.

Corollary 1. *Under the previously stated conditions, if we are given an initial matrix F^0 satisfying the bound $\mathrm{d}(F^0, F^*) \leq \frac{1}{5}|||F^*|||_{op}$, then with probability at least $1 - 4d^{-3}$, the gradient iterates $\{F^t\}_{t=1}^{\infty}$ with step size $\eta^t = c_3 \frac{1}{\mu^2 r^2 |||F^*|||_{op}^2}$ satisfy the bound*

$$\mathrm{d}^2(F^t, F^*) \leq \Big(1 - c_4 \frac{1}{\mu^2 r^2}\Big)^t \mathrm{d}^2(F^0, F^*) + c_5 \frac{r|||\Pi_\Omega(E)|||_{op}^2}{p^2 |||F^*|||_{op}^2}. \tag{22}$$

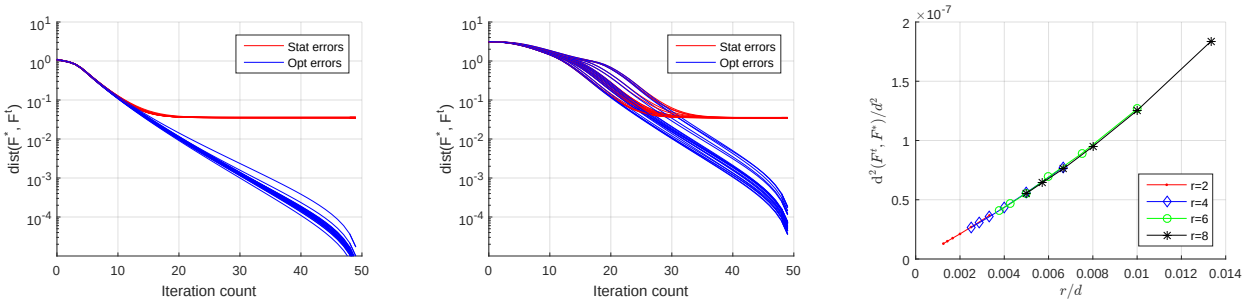
See Section 6.2 for the proof of this claim.

Even though Corollary 1 is a consequence of our general theory, it leads to results for exact/approximate recovery in the noiseless/noisy setting that are as good as or better than known results. In the noiseless setting ($E = 0$), our sample size requirement and contraction factor are sharper than those in the paper [57] by a polynomial factor in the rank r. Turning to the noisy setting, suppose the noise matrix E has independent sub-Gaussian entries with parameter σ^2. A technical result to be proved later (see Lemma 11) guarantees that given a sample size $n = pd^2 \gtrsim c_1 d \log^2 d$, we have the operator norm bound $|||\Pi_\Omega(E)|||_{\mathrm{op}} \precsim \sigma\sqrt{pd}$ with probability at least $1 - d^{-12}$. Together with the bound (22), we conclude that

$$\frac{1}{d^2}|||F^\infty \otimes F^\infty - F^* \otimes F^*|||_{\mathrm{F}}^2 \leq \frac{3}{d^2}|||F^*|||_{\mathrm{op}}^2 |||F^\infty - F^\infty_{\pi^*}|||_{\mathrm{F}}^2 \precsim \frac{\sigma^2 r}{pd} = \frac{\sigma^2 r d}{n}. \tag{23}$$

The scaling $\frac{\sigma^2 r d}{n}$ is better than the results in past work (e.g., [51, 42, 41]) on noisy matrix completion by a $\log d$ factor; in fact, it matches the minimax lower bounds established in the papers [51, 42]. Thus, Corollary 1 in fact establishes that the projected gradient descent method yields minimax-optimal estimates.

Initialization: Suppose the rank-r SVD of the matrix $\Pi_\Omega(M)$ is given by $USV^\top$. We can take $F^0 = \Pi_{\mathcal{F}}(US^{\frac{1}{2}})$. Under the previously stated condition on the sample size n, the matrix F^0 satisfies the requirement in Corollary 1 as shown in, e.g., [40] (combined with the above bound on $|||\Pi_\Omega(E)|||_{\mathrm{op}}$).



(a) (b) (c)

Figure 1. Simulation results for matrix completion. (a) Plots of optimization error $\mathrm{d}(F^t, F^T)$ and statistical error $\mathrm{d}(F^t, F^*)$ versus the iteration number t using SVD initialization. Panel (b): same plots using a random initialization. The simulation is performed using $d = 1000$, $r = 10$, $p = 0.1$ and $\sigma = 0.01 \cdot \frac{r}{d}$. Panel (c): plots of per-entry estimation error $\frac{1}{d^2}\mathrm{d}(\widehat{F}, F^*)$ versus $\frac{r}{d}$, for different values of (d, r) using SVD-based initialization. Each point represents the average over 20 random instances. The simulation is performed using $p = 0.1$ and $\sigma = 0.001$.

Computation: Computing the gradient $\nabla_F \widetilde{\mathcal{L}}_n(F) = \frac{2}{p}\Pi_\Omega(F \otimes F - Y)F$ takes time $\mathcal{O}(r^2|\Omega|)$. The projection $\Pi_{\mathcal{F}}(F)$ can be computed in time $\mathcal{O}(rd)$.

Simulations: In order to illustrate the predictions of Corollary 1, we performed a number of simulations. Since the distance measures $\mathrm{d}(F, F^*)$ and $\|F \otimes F - F^* \otimes F^*\|_{\mathrm{F}}$ are difficult to compute, so we instead use the subspace distance

$$\mathrm{d}(F, F^*) \approx \|\sin \angle(F, F^*)\|_{\mathrm{F}}^2, \tag{24}$$

as an approximation.[3] Here $\sin \angle(F, F^*)$ is the vector of principal angles between the column spaces of matrices F and F^*. For each example and given values of model parameters d, r, n, σ etc., we generate a random instance by sampling the true matrix F^* and the problem data randomly from the relevant model, and then run our projected gradient descent algorithm with $T = 50$ iterations.

In the matrix completion case, we sampled the true matrix F^* uniformly at random from all $d \times r$ orthonormal matrix uniformly at random, generated a noise matrix E with i.i.d. $N(0, \sigma^2)$ entries, and chose the observed entries randomly according to the Bernoulli model with probability p. We considered two approaches for obtaining the initial matrix F^0: (a) the SVD-based procedure described in Section 4.1, and (b) random initialization, where F^0 is a random $d \times r$ orthonormal matrix projected onto the associated constraint set $\mathcal{F}$. The step size for projected gradient descent is fixed at $\eta^t \equiv \frac{0.5}{p}$. Panels (a) and (b) in Figure 1 show the resulting convergence behavior of the algorithm, which confirm the geometric convergence (and threshold effect for the statistical error) that is predicted by our theory.

For these random ensembles, our theory predicts thatwith high probability the per-entry error of the output $\widehat{F}$ satisfies a bound of the form

$$\frac{1}{d^2}\mathrm{d}^2(\widehat{F}, F^*) \precsim \frac{\sigma^2 r}{pd};$$

cf. equation (23). Therefore, with p and σ fixed, the ratio $\frac{1}{d^2}\mathrm{d}^2(\widehat{F}, F^*)$ should be proportional to $\frac{r}{d}$.

[3]This approximation is valid up to a constant of 2 if both F and F^* are orthonormal (cf. Proposition 2.2 in [59]).

4.2 Matrix regression

Recall the matrix regression model previously introduced in Section 2.2.1. In order to simplify notation, it is convenient to introduce define a linear mapping $\mathfrak{X}_n : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^n$ via $[\mathfrak{X}_n(M)]_i := \langle\langle X^i,\ M\rangle\rangle$ for $i = 1, 2, \ldots, n$. Note that the adjoint operator $\mathfrak{X}_n^* : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}$ is given by $\mathfrak{X}_n^*(u) = \sum_{i=1}^n X^i u_i$. With this notation, we have the compact representation

$$\nabla\widetilde{\mathcal{L}}_n(F) = \frac{2}{n}\Big(\mathfrak{X}_n^*(\mathfrak{X}_n(F\otimes F) - y)\Big)F.$$

Since $\langle\langle X^i,\ F\otimes F\rangle\rangle = \langle\langle (X^i + X^{i\top})/2,\ F\otimes F\rangle\rangle$, we may assume without loss of generality that the matrices $\{X^i\}$ are symmetric.

In this case, projected gradient descent can be performed with $\mathcal{F} = \mathbb{R}^{d\times r}$, so that the M^*-faithfulness condition holds trivially. It remains to verify that the cost function $\widetilde{\mathcal{L}}_n$ from equation (8) satisfies the local descent, local Lipschitz and local smoothness properties, and these properties depend on the structure of the operator $\mathfrak{X}_n$. For instance, one way in which to certify the conditions of Theorem 2 is via a version of *restricted isometry property* (RIP) applied to the operator $\mathfrak{X}_n$.

Definition 4 (Restricted isometry property)**.** The operator $\mathfrak{X}_n : \mathbb{R}^{d\times d} \to \mathbb{R}^n$ is said to satisfied the restricted isometry property with parameter δ_k if

$$(1-\delta_k)|||M|||_{\mathrm{F}}^2 \leq \frac{1}{n}\|\mathfrak{X}_n(M)\|_2^2 \leq (1+\delta_k)|||M|||_{\mathrm{F}}^2, \quad \text{for all } d\text{-dimensional matrices with } \mathrm{rank}(M) \leq k.$$

It is well known [54, 50] that RIP holds for various random ensembles. For instance, suppose that the entries of $X^i_{j\ell}$ are i.i.d. zero-mean unit variance random variables, satisfying a sub-Gaussian tail bound. Examples of such ensembles include the standard Gaussian case ($X^i_{j\ell} \sim N(0,1)$) as well as Rademacher variables ($X^i_{j\ell} \in \{-1, 1\}$ equiprobably). For such ensembles, it is known that with high probability, a RIP condition of order r holds with a sample size $n \gtrsim rd$.

The RIP condition provides a straightforward way of verifying the conditions of Theorem 2. More precisely, as we show in the proof of Corollary 2, if the operator $\mathfrak{X}_n$ satisfies RIP with parameter $\delta_{4r} \in [0, \frac{1}{12})$, then the loss function $\widetilde{\mathcal{L}}_n$ satisfies the local descent, descent and smoothness conditions with parameters

$$\rho = (1 - 12\delta_{4r})\sigma_r, \quad \alpha = 6\delta_{4r}\sigma_r^2, \quad L = \beta = 64\kappa^2\sigma_r^2 \quad \text{and} \quad \varepsilon_n = \frac{2\sqrt{r}\kappa|||n^{-1}\mathfrak{X}_n^*(\epsilon)|||_{\mathrm{op}}}{\delta_{4r}\sigma_r}.$$

Using this fact, we have the following corollary of Theorem 2. We state it assuming that the operator $\mathfrak{X}_n$ satisfies RIP with parameter $\delta_{4r} \in [0, \frac{1}{12})$, and the sample size n is large enough to ensure that $\varepsilon_n \leq \frac{1-\sqrt{12\delta_{4r}}}{4}\sigma_r$.

Corollary 2. *Under the previously stated conditions, there is a universal function $\psi : [0, 1/12] \to (0, 1)$ such that given any initial matrix F^0 satisfying the bound $\mathrm{d}(F^0, F^*) \leq (1 - \sqrt{12\delta_{4r}})\sigma_r$, the projected gradient iterates $\{F^t\}_{t=1}^\infty$ with step size $\eta^t = \frac{\psi(\delta_{4r})}{\delta_{4r}\kappa^{10}\sigma_r^2}$ satisfy the bound*

$$\mathrm{d}^2(F^t, F^*) \leq \Big(1 - \frac{\psi(\delta_{4r})}{\kappa^{10}}\Big)^t \mathrm{d}^2(F^0, F^*) + c_0\frac{r\kappa^2|||\mathfrak{X}_n^*(\epsilon)|||_{op}^2}{n^2\delta_{4r}^2\sigma_r^2},$$

See Section 6.1 for the proof of this claim.

Note that the radius of the region of convergence $\mathbb{B}_2((1-\sqrt{12\delta_{4r}})\sigma_r; F^*)$ can be arbitrarily close to σ_r with δ_{4r} sufficiently small. Moreover, the function $\widetilde{\mathcal{L}}_n$ need not be convex in this region. As a simple example, consider the scalar case with $d = r = 1$, with noiseless observations ($\epsilon = 0$) of the target parameter $F^* = 1$. A simple calculation then yields that $\widetilde{\mathcal{L}}_n(F) = c\,(F^2-1)^2$ for some constant $c > 0$, which is nonconvex outside of the ball $\mathbb{B}_2(\frac{1}{\sqrt{3}}; F^*)$.

Specified to the noiseless setting with $\epsilon = 0$, Corollary 2 is similar to the results for nonconvex gradient descent in [58, 66]. In the more general noisy setting, our statistical error rate ε_n is consistent with the results in [51]. For a more concrete example, suppose $\kappa = \mathcal{O}(1)$, and each X^i and ϵ have i.i.d. Gaussian entries with $X^i_{j\ell} \sim \mathcal{N}(0,1)$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$. It can be shown that as long as $n \gtrsim rd\log d$, RIP holds with $\delta_{4r} < \frac{1}{192}$ and $|||\mathfrak{X}_n^*(\epsilon)|||_{\mathrm{op}} \precsim \sigma\sqrt{nd}$. The bound in Corollary 2 therefore implies a constant contraction factor and that

$$|||F^\infty \otimes F^\infty - F^* \otimes F^*|||_{\mathrm{F}}^2 \leq 3\mathrm{d}^2(F^\infty, F^*)|||F^*|||_{\mathrm{op}}^2 \precsim \sigma^2 \frac{rd}{n}.$$

Initialization: Suppose the rank-r SVD of the matrix $\frac{1}{n}\mathfrak{X}_n^*(y)$ is given by $USV^\top$. We can take $F^0 = US^{\frac{1}{2}}$. Under the above Gaussian example, it can be shown the condition on the initial solution is satisfied if $n \gtrsim dr^2\kappa^4 \log d$ and σ is small enough [36, 37].

The sample size required for this initialization scales quadratically in the rank r, as compared to the linear scaling that is the best possible [54, 50]. This looseness is a consequence of requiring the initialization error to satisfy a Frobenius norm bound instead of an operator norm one. It can be avoided by using a more sophisticated initialization procedures—for instance, one based on a few iterations of the singular value projection (SVP) algorithm [36]. In the current setting, since our primary focus is on understanding low-complexity algorithms via gradient descent, we do not pursue this direction further.

Computation: Let T_{mul} be the maximum time to multiply X^i with a vector in $\mathbb{R}^d$. Finding the initial solution as above requires computing the rank-r SVD of the $d \times d$ matrix $\frac{1}{n}\mathfrak{X}_n^*(y)$, which can be done in time $\mathcal{O}(nrT_{\mathrm{mul}} + dr^2)$; cf. [31]. The gradient $\frac{1}{n}\mathfrak{X}_n^*(\mathfrak{X}_n(M) - y)$ and can be computed in time $\mathcal{O}(nrT_{\mathrm{mul}} + dr)$. Therefore, the overall time complexity is $\mathcal{O}(nrT_{\mathrm{mul}} + dr^2)$ times the number of iterations.

4.3 Rank-r PCA with row sparsity

Recall the problem of sparse PCA previously introduced in Section 2.2.2. In this section, we analyze the projected gradient updates applied to this problem, in particular with the loss function $\widetilde{\mathcal{L}}_n$ from equation (11), and the constraint set

$$\mathcal{F} := \big\{F \in \mathbb{R}^{d\times r} \;\mid\; |||F|||_{\mathrm{op}} \leq 1,\; \|F\|_{2,1} \leq \|F^*\|_{2,1}\big\}.$$

To be clear, this choice of constraint set is somewhat unrealistic, since it assumes knowledge of the norm $\|F^*\|_{2,1}$. This condition could be removed by analyzing instead a penalized form of the estimator, but as our main goal is to illustrate the general theory, we remain with the constrained version here.

We now apply Theorem 2 to this problem. As we show in the proof of Corollary 3, the set $\mathcal{F}$ is M^*-faithful. Moreover, for each $0 < \tau < 1$, suppose that the SNR satisfies $\gamma > \frac{2}{\tau^2}$ in the row-sparse spiked covariance model, then with probability at least $1 - 2d^{-3}$, the loss function $\widetilde{\mathcal{L}}_n$ satisfies the local descent, smoothness conditions and the relaxed Lipschitz condition (50) with parameters

$$\rho = 1 - \tau^2, \quad \alpha = \frac{\gamma\tau^2}{4}, \quad L = \beta = 4(\gamma+1)\sqrt{r} \quad \text{and} \quad \varepsilon_n = c_1 \frac{\gamma+1}{\gamma\tau^2}\sqrt{r}\max\Big\{\sqrt{\frac{k\log d}{n}}, \frac{k\log d}{n}\Big\}. \tag{25}$$

Using these fact, we have the following corollary of Theorem 2. We state it assuming that the SNR obeys the bound $\gamma > \frac{2}{\tau^2}$ and the sample size n is large enough to ensure $\varepsilon_n \leq \frac{1-\tau}{20}$

Corollary 3. *Under the previously stated conditions, there is a function $\psi : (0,1) \to (0,1)$ such that given any initial matrix $F^0 \in \mathcal{F} \cap \mathbb{B}_2(1-\tau; F^*)$, with probability at least $1 - 2d^{-3}$, the projected gradient iterates $\{F^t\}_{t=1}^{\infty}$ with step size $\eta^t = \psi(\tau)\frac{\gamma}{(\gamma+1)^2 r}$ satisfy the bound*

$$\mathrm{d}^2(F^t, F^*)^2 \leq \Big(1 - \psi(\tau)\frac{\gamma^2}{(\gamma+1)^2 r}\Big)^t \mathrm{d}^2(F^0, F^*) + c_2 \cdot \frac{(\gamma+1)^2 r}{\gamma^2\tau^4}\max\Big\{\frac{k\log d}{n}, \frac{k^2\log^2 d}{n^2}\Big\}.$$

See Section 6.3 for the proof of this corollary.

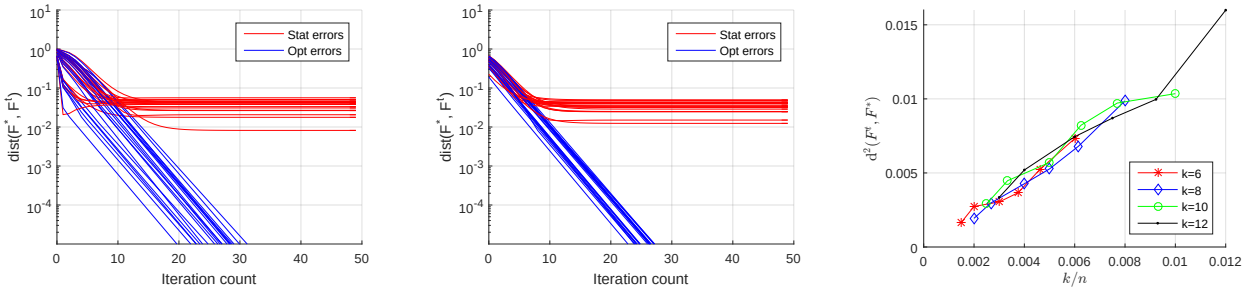
Remark 2. It is noteworthy that $\widetilde{\mathcal{L}}_n(F)$ is in fact *globally concave* in F. In order to see this fact, consider the scalar case with $d = r = 1$, where $\widetilde{\mathcal{L}}_n(F) = -CF^2$ for some $C > 0$.

The error rate ε_n is in fact minimax optimal (up to a logarithmic factor) with respect to n, d, k as well as the rank r; for instance, see the paper [59, 15], where the upper bound is achieved using computationally intractable estimators. Similar error rates are obtained in [14, 48, 62] using more sophisticated algorithms, but under a scaling of the sample size—in particular one that is quadratic in sparsity (see below)—that allows for a good initialization.

Initialization: The above results require an initial solution F^0 with $\mathrm{d}(F^0, F^*) < 1$. Under the spiked covariance model and given a sample size $n \gtrsim k^2 \log d$, such solution can be found by the diagonal thresholding method [39, 48]. Here the quadratic dependence on k is related to a computational barrier [6], and thus may not improvable using a polynomial-time algorithm.

Computation: The algorithm requires projection onto the intersection of the spectral norm and ℓ_1/ℓ_2 norm balls. In the rank one case ($r = 1$), it reduces to projecting to the intersection of the vector ℓ_2 and ℓ_1 balls, which can be done efficiently [56]. In the general case with $r > 1$, it can be done by alternating projection. The speed of convergence depends on the eigengap γ, exhibiting similarity to the standard power method for finding eigenvectors.

Simulations: We performed experiments under the same general set-up as the matrix completion (see the discussion surrounding equation (24)). For sparse PCA, we generated random ensembles of problems by fixing the rank $r = 1$, and choosing a random unit-norm $F^* \in \mathbb{R}^d$ supported on k randomly chosen coordinates. Using this random vector, we formed the spiked covariance matrix Σ with top eigenvector F^* and SNR γ. We considered two approaches for initialization: (a) diagonal thresholding as described in the papers [39, 48], and (b) choosing F^0 to be the perturbed version



(a) (b) (c)

Figure 2. Simulation results for sparse PCA. Panel (a); plots of optimization error $\mathrm{d}(F^t, F^T)$ and statistical error $\mathrm{d}(F^t, F^*)$ versus the iteration number t, using diagonal thresholding initialization. Panel (b): same plots using perturbation initialization. For both panels (a) and (b), simulations are performed using $d = 5000$, $r = 1$, $k = 5$, $\gamma = 4$ and $n = 4000$. Panel (c): plot of estimation error $\mathrm{d}(\widehat{F}, F^*)$ versus $\frac{k}{n}$, for different values of (k, n) using diagonal thresholding initialization. Each point represents the average over 20 random instances. The simulation is performed using $d = 5000$, $r = 1$ and $\gamma = 4$.

$F^0_R = F^*_R + \frac{1}{\sqrt{2}}E_1$ and $F^0_{R^c} = F^*_{R^c} + \frac{1}{\sqrt{2}}E_2$, where $R = \mathrm{support}(F^*)$ and E_1 and E_2 are random unit norm vectors with the appropriate dimensions. The step size is fixed at $\eta^t \equiv \frac{0.5\gamma}{(\gamma+1)^2}$.

Panels (a) and (b) of Figure 2 show the convergence rates of the optimization and statistical error using these two different types of initializations. Consistent with our theory, we witness an initially geometric convergence in terms of statistical error followed by an error floor at the statistical precision. In panel (c), we study the scaling of the estimation error. Our theory predicts that given a suitable initialization and sample size $n \gtrsim k \log d$, then with high probability the output $\widehat{F}$ satisfies

$$\mathrm{d}^2(\widehat{F}, F^*) \lesssim \frac{(\gamma+1)^2 r}{\gamma^2} \cdot \frac{k \log d}{n}.$$

Therefore, with the triplet of parameters (d, r, γ) fixed, the error $\mathrm{d}^2(\widehat{F}, F^*)$ should grow proportionally with the ratio $\frac{k}{n}$, a prediction that is confirmed in Figure 2(c).

4.4 Planted densest subgraph

The planted densest subgraph problem is a generalization of the planted clique problem; it can be viewed as a single cluster (or rank one) version of the more general planted partition problem. For a collection of d vertices, there is an unknown subset of size k which forms a cluster. Based on this cluster and two probabilities $p > q$, a random symmetric matrix $A \in \{0,1\}^{d \times d}$, which we think of as the adjacency matrix of the observed graph, is generated in the following way:

- for each pair of vertices i, j in the cluster, $A_{ij} = 1$ with probability p, and zero otherwise.
- for all other pairs of vertices, $A_{ij} = 1$ with probability q, and zero otherwise.

Let $F^* \in \{0,1\}^d$ be the cluster membership vector: i.e., $F^*_j = 1$ if and only if vertex j belongs to the cluster.

A previous approach is to recover the cluster matrix $M^* = F^* \otimes F^*$ by solving a particular SDP, derived as a relaxation of the MLE. Let $S := A - \frac{p+q}{2} J_d$ be a shifted version of the adjacency matrix, where J_d is the $d \times d$ all one matrix. Consider the semidefinite program

$$\min_{M \in \mathcal{S}^{d \times d}} \Big\{ - \langle\!\langle S,\ M \rangle\!\rangle \Big\} \qquad \text{such that } M \succeq 0,\ \textstyle\sum_{i,j} M_{ij} = k^2 \text{ and } M \in [0,1]^{d \times d}. \tag{26}$$

It is known [21] that with probability at least $1 - d^2$, the true cluster matrix M^* is the unique optimal solution to this program when

$$\frac{(p-q)^2}{p} \geq c_1 \Big(\frac{\log d}{k} + \frac{d}{k^2} \Big), \tag{27}$$

for some universal constant $c_1 > 0$. When $p = 1 = 2q$, this condition reduces to the well-known $k \gtrsim \sqrt{d}$ tractability region for the planted clique problem [2].

Alternatively, we may solve the factorized formulation by projected gradient decent (3), as applied to the problem

$$\widetilde{\mathcal{L}}_n(F) = \langle\!\langle -S,\ F \otimes F \rangle\!\rangle, \qquad \mathcal{F} = \{F \mid F \in [0,1]^d,\ \textstyle\sum_{i=1}^d F_i = k\}.$$

This setting is a $r = 1$ special case of our general framework. In this case $\mathcal{E}(M^*) = \{\pm F^*\}$ contains only two elements, and can be verified to be M^*-faithful.

We now ready to apply our general theory to this problem. As we show in proof of Corollary 4, if the model parameters satisfy the condition (27), then with probability at least $1 - d^{-3}$, the loss function $\widetilde{\mathcal{L}}_n$ satisfies the local descent and smoothness conditions and the relaxed Lipschitz condition (40) with parameters

$$\rho = \frac{2}{5}\sqrt{k}, \quad \alpha = \frac{1}{20}(p-q)k, \quad \beta = 12(p-q)k \quad \text{and} \quad \varepsilon_n = 0.$$

Using this fact, we have the following corollary of Theorem 2. We state it assuming that the condition condition (27) holds

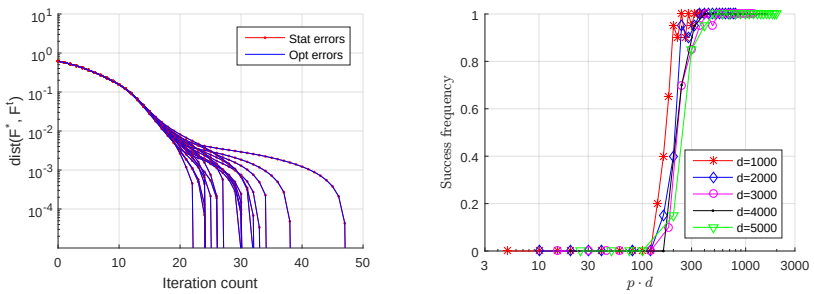
Corollary 4. *Under the previously stated conditions, given an initial vector $F^0 \in \mathcal{F} \cap \mathbb{B}_2(\frac{1}{5}\sqrt{k}; F^*)$, the projected gradient iterates $\{F^t\}_{t=1}^{\infty}$ with step size $\eta^t = c_2 \frac{1}{(p-q)k}$ satisfy the bound*

$$\mathrm{d}^2(F^t, F^*) \leq \Big(1 - c_3\Big)^t \mathrm{d}^2(F^0, F^*).$$

See Section 6.4 for the proof of this claim.

The corollary guarantees exact recovery of M^* when $t \to \infty$. The condition (27) matches the best existing results; see e.g., [21] and the references therein.

Initialization Set F^0 to be the top left singular vector of $A - qJ_d$ projected onto the set $\mathcal{F}$. Note that F^* is a left singular vector of the matrix $\mathbb{E}[A - qJ_d]$ corresponding to the only non-zero singular value $(p-q)k$. Under the condition (27), Proposition 1 ensures that $\|\!|(A - qJ_d) - \mathbb{E}[A - qJ_d]|\!\|_{\mathrm{op}} \leq \frac{1}{4}(p-q)k$ with probability at least $1 - d^{-3}$. On this event, applying Wedin's $\sin\Theta$ theorem [30] guarantees that F^0 satisfies the requirement in Corollary 4.



(a) (b)

Figure 3. Simulations for planted densest subgraph. Panel (a): plots of optimization error $\mathrm{d}(F^t, F^T)$ and statistical error $\mathrm{d}(F^t, F^*)$ versus the iteration number t, using SVD-based initialization. The simulation is performed using $d = 8000$, $k = 2000$, $p = 0.13$ and $q = 0.05$. Panel (b): plot of the probability of successful exact recovery of F^* versus pd, for different values of (d, p) using SVD-based initialization. We declare exact recovery if $\mathrm{d}(\widehat{F}, F^*) \leq 2 \times 10^{-3}$, and each point represents frequency of exact recovery over 20 random instances. The simulation is performed with $q = \frac{p}{4}$ and $k = \frac{d}{2}$.

Computation: The set $\mathcal{F}$ is the intersection of a hyperplane and a box in $\mathbb{R}^d$, so the associated projection $\Pi_{\mathcal{F}}$ can be computed in time $\mathcal{O}(d)$ [49]. Computing the gradient $\nabla \widetilde{\mathcal{L}}_n(F) = -2SF$ only requires matrix-vector multiplication with the matrix S, which is the sum of a rank-1 matrix and the (usually sparse) graph adjacency matrix A. In contrast, solving the SDP in equation (26) using ADMM requires multiple full SVD of dense matrices even when the graph is sparse.

Simulations: We performed experiments under the same general set-up as the matrix completion (see the discussion surrounding equation (24)). The 0-1 cluster indicator matrix $F^* \in \mathbb{R}^{d \times 1}$ is supported on k coordinates, and we sampled the graph adjacency matrix A from the planted densest subgraph model with edge probabilities (p, q) and cluster size k. The initial matrix F^0 is obtained using the SVD-based procedure described in Section 4.4. The step size is fixed at $\eta^t \equiv \frac{0.1}{(p-q)k}$. Panel (a) of Figure 3 shows plots of the optimization and statistical errors versus the iteration number; consistent with Corollary 4, these iterates converge at least geometrically.

In terms of the scaling of the sample size required for exact recovery, we know that if $k \precsim \frac{d}{\log d}$, then the fixed point of the algorithm $\bar{F}$ will be equal to F^* with high probability provided that $\frac{(p-q)^2}{p} \succsim \frac{d}{k^2}$. In particular, see equation (27). Therefore, with $q = \frac{p}{3}$ and $k = \frac{d}{2}$, exact recovery of F^* can be achieved with probability close to one as soon as pd is above a constant threshold. This theoretical prediction is confirmed in panel (b) of Figure 3.

4.5 One-bit matrix completion

Let us now turn to an extension of the standard (linear) matrix completion model studied in Section 4.1. It provides a more challenging problem to analyze, and our general theory provides (to the best of our knowledge) the first known polynomial-time algorithm for achieving the minimax rate in the case of rank r matrices.

In order to set up the problem, suppose that $F^* \in \mathbb{R}^{d \times r}$ is an orthonormal matrix and has incoherence parameter μ as previsouly defined in equation (21). Given a set $\Omega \subseteq [d] \times [d]$ of observed elements, a noise parameter $\sigma > 0$ and a differentiable function $f : \mathbb{R} \mapsto [0, 1]$ with Lipschitz derivative, we observe a binary symmetric matrix $Y \in \{-1, 1\}^{d \times d}$ such that for each

$(i,j) \in \Omega$ with $i \geq j$,

$$Y_{ij} = \begin{cases} 1, & \text{with probability } f(M^*_{ij}/\sigma), \\ -1, & \text{with probability } 1 - f(M^*_{ij}/\sigma). \end{cases}$$

We further assume that the observation set Ω is symmetric and generated by the Bernoulli model with parameter p, that is, $\mathbb{P}\big((i,j),(j,i) \in \Omega\big) = p$ independently for each (i,j) with $i \geq j$ The goal is to estimate M^* given the binary observations Y. Examples of the function f include the *logistic model* with $f(x) = \frac{\exp(x)}{1+\exp(x)}$; the *probit model* with $f(x) = \Phi(x)$, where $\Phi(x)$ is the cumulative distribution function of a standard Gaussian; and the *Laplacian model* in which f is the cumulative distribution function of Laplacian$(0,1)$: variable. See the papers [16, 24] for more details on these choices.

For a given f, consider thee negative log-likelihood of M, given by

$$\begin{aligned} \mathcal{L}_n(M) &= -2 \sum_{(i,j)\in\Omega} \Big[\frac{1+Y_{ij}}{2} \log f(M_{ij}/\sigma) + \frac{1-Y_{ij}}{2} \log\big(1 - f(M_{ij}/\sigma)\big) \Big] \\ &= -\langle\!\langle \Pi_\Omega(J_d + Y),\ \log f(M/\sigma) \rangle\!\rangle - \langle\!\langle \Pi_\Omega(J_d - Y),\ \log\big(1 - f(M/\sigma)\big) \rangle\!\rangle, \end{aligned} \tag{28}$$

where J_d is the $d \times d$ all one matrix, $\circ$ denotes the Hadamard product, and functions are applied to a matrix element-wise. As in matrix completion, we use the set $\mathcal{F} = \big\{F \mid \|F\|_{2,\infty} \leq \sqrt{\frac{2\mu}{d}} |\!|\!| F^0 |\!|\!|_{\mathrm{F}} \big\}$ Note that the gradient of the loss function is given by

$$\nabla_M \mathcal{L}_n(M) = -\frac{1}{\sigma} \Pi_\Omega \Big[\frac{f'(M/\sigma) \circ (Y - 2f(M/\sigma) + J_d)}{f(M/\sigma) \circ (1 - f(M/\sigma))} \Big],$$

where the fraction are also element-wise.

Since the function f is differentiable with a Lipschitz derivative f', Rademacher's theorem guarantees that the second derivative f'' is defined almost everywhere. Our corollary depends function f through the following two quantities, defined for each $a > 0$:

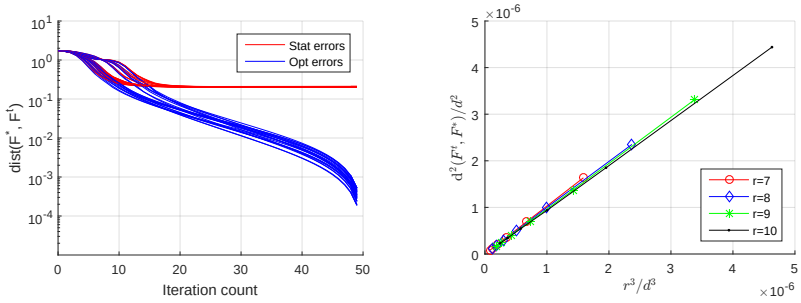
$$\begin{aligned} L_a &:= \max\Big\{ \sup_{|x|<a} \frac{|f'(x)|}{f(x)(1-f(x))},\ \sup_{|x|<a} \frac{f'(x)^2}{f(x)^2(1-f(x))^2},\ \sup_{|x|<a} \frac{|f''(x)|}{f(x)(1-f(x))} \Big\}, \quad \text{and} \\ \ell_a &:= \sup_{|x|<a} \frac{f(x)(1-f(x))}{f'(x)^2}. \end{aligned} \tag{29}$$

These quantities are similar to those in the paper [24], along with the additional control over the second derivative f'' required for proving (fast) geometric convergence. We introduce the shorthand $\nu := \frac{\mu r}{d\sigma} = \frac{\|M^*\|_\infty}{\sigma}$, which we think of as a measure of SNR. In the constant SNR setting $\nu = \Theta(1)$, the quantities $L_{4\nu}$ and $\ell_{4\nu}$ are positive universal constants independent of the other model parameters d, r, p etc.

We now apply Theorem 2 to the one-bit matrix completion problem. Set

$$\rho = c_1 \max\Big\{1, \frac{1}{\ell_{4\nu} L_{4\nu}}\Big\}, \quad \alpha = c_2 \frac{p}{\ell_{4\nu}\sigma^2}, \quad L = \beta = c_3 \frac{L_{4\nu} p \mu r}{\sigma^2} \quad \text{and} \quad \varepsilon_n = c_4 \sigma L_{4\nu} \ell_{4\nu} (1+\nu) \sqrt{\tfrac{dr}{p}}.$$

As we shown in the proof of Corollary 5, if the initial matrix satisfies the condition $\mathrm{d}(F^0, F^*) \leq 1 - \sqrt{1-\rho}$, then the set $\mathcal{F}$ is M^*-faithful. Moreover, if the expected sample size satisfies the bound



(a) (b)

Figure 4. Simulation results for one-bit matrix completion. Panel (a); plots of OB optimization error $\mathrm{d}(F^t, F^T)$ and statistical error $\mathrm{d}(F^t, F^*)$ versus the iteration number t, using random initialization. The simulation is performed using $d = 1000$, $r = 3$ and $p = 0.5$. Panel (b) plot of per-entry estimation error $\frac{1}{d^2}\mathrm{d}(\widehat{F}, F^*)$ versus $\frac{r^3}{d^3}$, for different values of (d, r) using random initialization. Each point represents the average over 20 random instances. The simulation is performed using $p = 0.5$ and $\sigma = \frac{0.5r}{d}$.

$n = pd^2 \geq c_5 \max\{\mu r d \log d, d\log^2 d, \mu^2 r^2 d\}$ and is large enough to ensure $\varepsilon_n < \frac{1}{20}(1 - \sqrt{1-\rho})$, then with probability at least $1 - c_6 d^{-3}$, the loss function $\widetilde{\mathcal{L}}_n$ associated with (28) satisfies the local descent, Lipschitz and smoothness conditions with parameters ρ, α, L, β and ε_n given above. Using these facts, we obtain the following guarantee, which we state assuming that the sample size n satisfies the above conditions.

Corollary 5. *Under the previously stated conditions, if we are given an initial matrix F^0 with $\mathrm{d}(F^0, F^*) \leq 1 - \sqrt{1-\rho}$, then with probability at least $1 - c_6 d^{-3}$, the gradient descent iterates $\{F^t\}_{t=1}^{\infty}$ with step size $\eta^t = c_7 \frac{\sigma^2}{\ell_{4\nu} L_{4\nu}^2 p \mu r}$ satisfy the bound*

$$\mathrm{d}^2(F^t, F^*) \leq \left(1 - c_8 \frac{1}{\ell_{4\nu}^2 L_{4\nu}^2 \mu r}\right)^t \mathrm{d}^2(F^0, F^*) + c_9 \sigma^2 L_{4\nu}^2 \ell_{4\nu}^2 (1+\nu)^2 \frac{dr}{p}.$$

See Section 6.5 for the proof of this claim.

In order to interpret the above result, let us consider the setting with a constant SNR $\nu = \Theta(1)$, in which case $\sigma = \frac{\mu r}{d\nu} \asymp \|F^* \otimes F^*\|_\infty$. Corollary 5 guarantees that given an initial matrix F^0 within a constant radius of F^*, the projected gradient descent converges geometrically and has per-entry error

$$\frac{1}{d^2}\|F^\infty \otimes F^\infty - F^* \otimes F^*\|_{\mathrm{F}}^2 \leq \frac{3}{d^2}\|F^*\|_{\mathrm{op}}^2 \mathrm{d}^2(F^\infty, F^*) \precsim \frac{dr}{n}\sigma^2 \asymp \frac{dr}{n}\|F^* \otimes F^*\|_\infty^2. \tag{30}$$

This bound has the same form as that in Section 4.1 for standard matrix completion, with the important difference that it is an essentially *multiplicative* bound where the pre-factor depends on the SNR ν.

It is worth comparing our error bounds with previous results under the setting $\nu = \Theta(1)$. One body of past work [24, 16] has studied the recovery of *approximately low-rank* matrices with bounded nuclear norm—that is, matrices whose vectors of singular values are in the ℓ_q ball with $q = 1$. This is a milder sparsity assumption, and so leads to the slower error rate $O\left(\sqrt{\frac{dr}{n}}\right)$. The result here applies to exactly low-rank matrices ($q = 0$), and so leads to the faster rate $\frac{dr}{n}$. Both of these scalings are be

minimax-optimal in the simpler linear setting [51]. On the other hand, Bhaskar et al. [8] also analyze the case of exactly low-rank matrices, but their algorithm relies on rank-constrained optimization and does not have convergence guarantees in polynomial time. Moreover, their error rate scales as $\frac{d^7 r^3}{n^4}$, and thus has a worse dependence on r, d and n as compared to ours.

Initialization and time complexity: In theory, we can obtain a good initial solution F^0 by solving one of the convex programs in the papers [24, 16] followed by a projection onto the set $\mathcal{F}$. Since we only need the initial error to be a constant, it suffices to have $n \gtrsim dr + d \log d$ observations. In fact, in our simulations, we find that a randomly chosen initial matrix F^0 is often good enough (see Figure 4(a)). Given such an initial solution, the projected gradient iterates converges geometrically with a contraction factor $1 - \frac{c_8}{\mu r}$, so we need $\mathcal{O}(\mu r \log(1/\delta))$ iterations to compute a δ-accurate solution. Therefore, we can achieve the $\mathcal{O}(\frac{dr}{n})$ error rate in polynomial time; to the best of our knowledge, this polynomial-time guarantee for achieving the minimax-rate in the exact low-rank case is the first such result in the literature.

Simulations: We performed experiments under the same general set-up as the matrix completion (see the discussion surrounding equation (24)). The matrix F^* is random orthonormal, and the observations are generated using the Bernoulli model with observation probability p and the standard Gaussian CDF as the link function f with noise magnitude $\sigma = \frac{2r}{d}$. The initial matrix F^0 is obtained by random initialization. The step size is fixed at $\eta^t \equiv \frac{0.5\sigma^2}{p}$. Panel (a) of Figure 4 illustrates the geometric convergence of the algorithm.

In terms of the scaling of the estimation error, with $\sigma = \frac{2r}{d}$, $n = pd^2$ and p fixed, the per-entry error of the output $\widehat{F}$ satisfies

$$\frac{1}{d^2} \mathrm{d}^2(\widehat{F}, F^*) \precsim \frac{dr}{n} \propto \frac{r^3}{d^3}$$

with high probability; cf. equation (30). Therefore, we should expect that the squared error $\frac{1}{d^2} \mathrm{d}^2(\widehat{F}, F^*)$ scales proportionally with the ratio $\frac{r^3}{d^3}$, a prediction that is confirmed in Figure 4(b).

4.6 Low-rank and sparse matrix decomposition

Recall from Section 2.2.3 the problem of noisy matrix decomposition, in which we observe a noisy sum of the form $Y = F^* \otimes F^* + S^* + E$, where E is a symmetric noise matrix. Our goal is to estimate F^*, and in this section, we analyze a version of this model in which the factor matrix $F^* \in \mathbb{R}^{d \times r}$ has equal eigenvalues and incoherence parameter μ as defined in equation (21), and the perturbing matrix $S^* \in \mathcal{S}^{d \times d}$ is element-wise sparse.

One line of work concerns the setting where the non-zero entries of S^* are randomly located [19, 22], whereas another line of work focuses on deterministic models [20, 35, 1]. We focus on one version of the deterministic setting, in which each row/column of the matrix S^* has at most k non-zero entries, whose locations and values are otherwise arbitrary. In light of keeping the presentation as simple as possible, we assume here the values of $\|S^*_i\|_1$, the ℓ_1 norm of each row of S^* are known.[4]

[4]This is unrealistic and could be relaxed, albeit at the price of more involved analysis of the Lagrangian version instead of the constrained version.

Using the nuclear norm and ℓ_1 norms as surrogates for rank and sparsity (respectively), the constrained version of a popular convex relaxation approach is based on the SDP

$$\min_{M \in \mathcal{S}^{d \times d}} \Big\{ \frac{1}{2} \big(\min_{S \in \mathcal{S}} |||Y - (M+S)|||_{\mathrm{F}}^2 \big) + \lambda |||M|||_{\mathrm{nuc}} \Big\},$$

where $\mathcal{S} := \{S \in \mathbb{R}^{d \times d} \mid \|S_i\|_1 \leq \|S_i^*\|_1, i = 1, 2, \ldots, d\}$. Alternatively, we may drop the nuclear norm regularizer and solve the factorized formulation by projected gradient descent, as applied to the problem

$$\mathcal{L}_n(M) = \frac{1}{2} \min_{S \in \mathcal{S}} |||M + S - Y|||_{\mathrm{F}}^2, \qquad \mathcal{F} = \Big\{ F \mid \|F\|_{2,\infty} \leq \sqrt{\frac{2\mu}{d}} |||F^0|||_{\mathrm{F}} \Big\}.$$

Note that $\mathcal{L}_n(M)$ is the squared Euclidean distance between the point $Y - M$ and the closed convex set $\mathcal{S}$. Therefore, the function $\mathcal{L}_n$ is convex and has gradient

$$\nabla_M \mathcal{L}_n(M) = M + \Pi_{\mathcal{S}}(Y - M) - Y.$$

We now derive a guarantee for this problem using our Theorem 2. As we show in the proof of Corollary 6, if the initial matrix F^0 satisfies $\mathrm{d}(F^0, F^*) \leq \frac{1}{5}$, then the constraint set $\mathcal{F}$ is M^*-faithful. Moreover, if the sparsity of the matrix S^* satisfies $\frac{\mu r k}{d} \leq c_1$ and the noise matrix satisfies $|||E|||_{\mathrm{F}} \leq c_2 |||F^*|||_{\mathrm{op}}$, then the loss function and the feasible set satisfy the local descent, Lipschitz and smoothness conditions with parameters

$$\rho = \frac{3}{5} |||F^*|||_{\mathrm{op}}, \quad \alpha = \frac{1}{10} |||F^*|||_{\mathrm{op}}^2, \quad L = \beta = 48 |||F^*|||_{\mathrm{op}}^2 \quad \text{and} \quad \varepsilon_n = 128 \frac{|||E|||_{\mathrm{F}}}{|||F^*|||_{\mathrm{op}}}.$$

Using these facts, we have the following guarantee, which is stated assuming that the matrices S^* and σ satisfy the assumptions above.

Corollary 6. *Under the previously stated conditions, given any initial matrix F^0 satisfying the bound $\mathrm{d}(F^0, F^*) \leq \frac{1}{5}$, the gradient iterates $\{F^t\}_{t=1}^{\infty}$ with step size $\eta^t = c_3 \frac{1}{|||F^*|||_{op}^2}$ satisfy the bound*

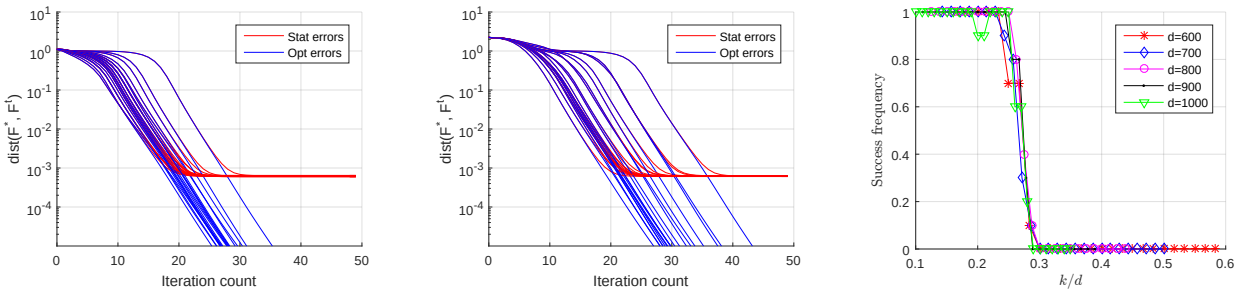
$$\mathrm{d}^2(F^t, F^*) \leq \big(1 - c_4\big)^t \mathrm{d}^2(F^0, F^*) + c_5 \frac{|||E|||_F^2}{|||F^*|||_{op}^2}. \tag{31}$$

See Section 6.6 for the proof.

The condition $\frac{\mu r k}{d} \leq c_1$ matches the best existing results in [35, 22] for the deterministic setting of matrix decomposition. As a passing observation, the above results can be applied to matrix completion with *adversarial* missing entries—by arbitrarily filling in the missing entries and treating them as sparse corruption, Corollary 6 guarantees recovery when each row/column has at most $k \leq c_1 \frac{d}{\mu r}$ missing entries, whose locations can be arbitrary.

Initialization: We describe how to get a good initial matrix F^0 in the noiseless setting $E = 0$. Suppose $|||F^*|||_{\mathrm{op}} = 1$. Let $\overline{Y}$ be obtained from hard-thresholding Y at the level $\frac{\mu r}{d}$; that is, for each element (i, j),

$$\overline{Y}_{ij} = \begin{cases} Y_{ij}, & \text{if } |Y_{ij}| \leq \frac{\mu r}{d}, \\ \frac{\mu r}{d} \operatorname{sign}(Y_{ij}), & \text{if } |Y_{ij}| > \frac{\mu r}{d}. \end{cases}$$



(a) (b) (c)

Figure 5. Matrix decomposition: plots of optimization error $\mathrm{d}(F^t, F^T)$ and statistical error $\mathrm{d}(F^t, F^*)$ versus the iteration number t, using (a) SVD-based initialization and (b) random initialization. The simulation is performed using $d = 600$, $r = 5$, $k = 100$ and $\sigma = 0.1 \cdot \frac{r}{d}$. Panel (c): plots of the probability of successful exact recovery of F^* versus $\frac{k}{d}$, for different values of (d, k) using SVD-based initialization. We declare exact recovery if $\mathrm{d}(\widehat{F}, F^*) \leq 2 \times 10^{-3}$, and each point represents frequency of exact recovery over 20 random instances. The simulation is performed using $r = 6$ and $\sigma = 0$.

We then set F^0 to be the $d \times r$ matrix with columns being the top-r singular vector of $\bar{Y}$ projected onto the set $\mathcal{F}$. In Appendix F, we prove that under these conditions, we have

$$\mathrm{d}(F^0, F^*) \leq \frac{4\mu r\sqrt{r}k}{d}. \tag{32}$$

Therefore, the requirement in Corollary 6 is satisfied if $\frac{\mu r\sqrt{r}k}{d} \leq c_1$ for a universal constant c_1 that is sufficiently small. The condition $\frac{\mu r\sqrt{r}k}{d} \leq c_1$ is sub-optimal by a factor of $\sqrt{r}$.

Computation: To compute the gradient $\nabla_M \mathcal{L}_n(M) = M + \Pi_{\mathcal{S}}(Y - M) - Y$, we need to project each row of $Y - M$ to the ℓ_1 balls, which can be done efficiently [27]. As discussed in Section 4.1, the projection $\Pi_{\mathcal{F}}$ can be computed by row-wise clipping.

Simulations: We performed experiments under the same general set-up as the matrix completion (see the discussion surrounding equation (24)). The matrix F^* is random orthonormal; the sparse matrix S^* has $k \times d$ non-zero entries whose locations are sampled uniformly without replacement and whose values are independently and uniformly sampled from the interval $[0, 10 \cdot \frac{r}{d}]$; the noise matrix E has i.i.d. zero-mean entries with standard deviation σ. The initial matrix F^0 is obtained using the SVD-based procedure described in Section 4.6. The step size is fixed at $\eta^t \equiv 1$. Panels (a) and (b) of Figure 5 confirms the predicted geometric convergence.

In terms of the estimation error, in the noiseless setting with $\sigma = 0$, the output $\bar{F}$ equals F^* with high probability provided that $\frac{rk}{d} \precsim 1$. Therefore, with r fixed, exact recovery of F^* can be achieved with probability close to one as soon as $\frac{k}{d}$ is below a constant threshold. Panel (c) of Figure 5 confirms this prediction.

5 Proofs of general theorems

This section is devoted to the proofs of our general theorems on projected gradient convergence, namely Theorem 1 in the Lipschitz case, and Theorem 2 under stronger smoothness conditions. Throughout these proofs, we make use of the convenient shorthand $\widetilde{G}^t := \nabla \widetilde{\mathcal{L}}_n(F^t)$ for the gradient of $\widetilde{\mathcal{L}}_n$ at step t. We also define the difference matrix $\Lambda^t := F^{t+1} - F^t$, as well as the parameters $\psi := |||F^*|||_{\mathrm{op}}$ and $\sigma_r := \sigma_r(F^*)$.

5.1 Proof of Theorem 1

Our proof proceeds via induction on the event

$$\mathcal{E}_t := \big\{ \mathrm{d}(F^s, F^*) \leq \underbrace{(1-\tau)\sigma_r}_{\rho} \quad \text{for all } s \in \{0, 1, \ldots, t\} \big\}. \tag{33}$$

For the base case $t = 0$, note that $\mathcal{E}_0$ holds by the assumptions of the theorem. Assuming that $\mathcal{E}_t$ holds, it suffices to show $\mathrm{d}(F^{t+1}, F^*) \leq \rho$, which then implies that $\mathcal{E}_{t+1}$ holds.

We require the following auxiliary result:

Lemma 1. *For any matrix $F \in \mathbb{R}^{d \times r}$ such that $\mathrm{d}(F, F^*) < \sigma_r(F^*)$, the optimization problem $\min_{A \in \mathcal{E}(M^*)} |||A - F|||_F$ has a unique optimum F_{π^*} such that (i) the matrix $F^\top F_{\pi^*} \in \mathbb{R}^{r \times r}$ is positive definite; and (ii) the matrix $(F - F_{\pi^*})^\top F_{\pi^*}$ is symmetric.*

See Section 5.1.1 for the proof of this claim. In view of Lemma 1, the matrix $F^s_{\pi^*} := \arg\min_{A \in \mathcal{E}(M^*)} |||A - F^s|||_{\mathrm{F}}$ is uniquely defined for each time step $s \in \{0, 1, \ldots, t\}$.

The projected gradient descent update can be decomposed into the two steps

$$\widetilde{F}^{s+1} = F^s - \eta^s \nabla \widetilde{\mathcal{L}}_n(F^s) \quad \text{and} \quad F^{s+1} = \Pi_{\mathcal{F}}\big(\widetilde{F}^{s+1}\big). \tag{34}$$

For each $s \in \{0, 1, \ldots, t\}$, the local descent condition (14) implies that

$$\langle\!\langle \nabla \widetilde{\mathcal{L}}_n(F^s),\ F^s - F^s_{\pi^*} \rangle\!\rangle \geq \alpha |||F^s - F^s_{\pi^*}|||_{\mathrm{F}}^2 - \alpha \varepsilon_n^2.$$

On the other hand, from the decomposition (34), we have $\nabla \widetilde{\mathcal{L}}_n(F^s) = \frac{F^s - \widetilde{F}^{s+1}}{\eta^s}$, and hence the above inequality implies that

$$\begin{aligned}
\alpha |||F^s - F^s_{\pi^*}|||_{\mathrm{F}}^2 - \alpha \varepsilon_n^2 &\leq \frac{1}{\eta^s} \langle\!\langle F^s - \widetilde{F}^{s+1},\ F^s - F^s_{\pi^*} \rangle\!\rangle \\
&= \frac{1}{2\eta^s} \big(|||F^s - F^s_{\pi^*}|||_{\mathrm{F}}^2 + |||F^s - \widetilde{F}^{s+1}|||_{\mathrm{F}}^2 - |||\widetilde{F}^{s+1} - F^s_{\pi^*}|||_{\mathrm{F}}^2 \big) \\
&= \frac{1}{2\eta^s} \big(|||F^s - F^s_{\pi^*}|||_{\mathrm{F}}^2 - |||\widetilde{F}^{s+1} - F^s_{\pi^*}|||_{\mathrm{F}}^2 \big) + \eta^s |||\nabla \widetilde{\mathcal{L}}_n(F^s)|||_{\mathrm{F}}^2.
\end{aligned}$$

Due to the M^*-faithfulness and convexity assumption on $\mathcal{F}$, we are guaranteed that $F^s_{\pi^*} \in \mathcal{F}$, and hence

$$|||F^{s+1} - F^s_{\pi^*}|||_{\mathrm{F}} \leq |||\widetilde{F}^{s+1} - F^s_{\pi^*}|||_{\mathrm{F}}$$

since Euclidean projection onto a convex set is non-expansive [7]. Moreover, by the Lipschitz condition (15) on $\widetilde{\mathcal{L}}_n$, we have $\|\nabla\widetilde{\mathcal{L}}_n(F^s)\|_{\mathrm{F}}^2 \leq L^2\psi^2$. Combining the pieces, we find that

$$\alpha\|F^s - F^s_{\pi^*}\|_{\mathrm{F}}^2 - \alpha\varepsilon_n^2 \leq \frac{1}{2\eta^s}\big(\|F^s - F^s_{\pi^*}\|_{\mathrm{F}}^2 - \|F^{s+1} - F^s_{\pi^*}\|_{\mathrm{F}}^2\big) + \eta^s L^2\psi^2. \tag{35}$$

Introducing the shorthand $\gamma := \frac{20\kappa^2L^2}{\alpha^2} - 1$, we then make the step size choice $\eta^s = \frac{1}{\alpha(s+1+\gamma)}$. Substituting into our bound (35) and rearranging yields

$$\frac{\alpha(s+1+\gamma)}{2}\|F^{s+1} - F^s_{\pi^*}\|_{\mathrm{F}}^2 \leq \frac{\alpha(s-1+\gamma)}{2}\|F^s - F^s_{\pi^*}\|_{\mathrm{F}}^2 + \frac{1}{\alpha(s+1+\gamma)}L^2\psi^2 + \alpha\varepsilon_n^2.$$

Multiplying both sides by $s+\gamma$ and using the fact that $\|F^{s+1} - F^{s+1}_{\pi^*}\|_{\mathrm{F}} \leq \|F^{s+1} - F^s_{\pi^*}\|_{\mathrm{F}}$ yields

$$\begin{aligned}\frac{\alpha(s+\gamma)(s+1+\gamma)}{2}\|F^{s+1} - F^{s+1}_{\pi^*}\|_{\mathrm{F}}^2 &\leq \frac{\alpha(s+\gamma)(s-1+\gamma)}{2}\|F^s - F^s_{\pi^*}\|_{\mathrm{F}}^2 + \frac{(s+\gamma)}{\alpha(s+1+\gamma)}L^2\psi^2 + \alpha(s+\gamma)\varepsilon_n^2. \\ &\leq \frac{\alpha(s+\gamma)(s-1+\gamma)}{2}\|F^s - F^s_{\pi^*}\|_{\mathrm{F}}^2 + \frac{1}{\alpha}L^2\psi^2 + \alpha(s+\gamma)\varepsilon_n^2.\end{aligned}$$

Summing the above inequality over $s = 0, \ldots t$ yields

$$\frac{\alpha(t+\gamma)(t+1+\gamma)}{2}\|F^{t+1} - F^{t+1}_{\pi^*}\|_{\mathrm{F}}^2 \leq \frac{\alpha\gamma^2}{2}\|F^0 - F^0_{\pi^*}\|_{\mathrm{F}}^2 + \frac{t+1}{\alpha}L^2\psi^2 + \alpha(t+1)(t+\gamma)\varepsilon_n^2. \tag{36}$$

Now observe that the assumptions $\tau \leq \frac{1}{2}$ and $\alpha \leq L$ imply that $\gamma \geq \frac{4\kappa^2L^2}{(1-\tau)^2\alpha^2} \geq 1$. These inequalities, combined with the facts that $\|F^0 - F^0_{\pi^*}\|_{\mathrm{F}} \leq (1-\tau)\sigma_r$ by assumption and $\psi/\kappa = \sigma_r$, when applied to the bound (36), yield

$$\begin{aligned}\|F^{t+1} - F^{t+1}_{\pi^*}\|_{\mathrm{F}}^2 &\leq \frac{\gamma^2}{(t+\gamma)(t+1+\gamma)}(1-\tau)^2\sigma_r^2 + \frac{(t+1)\gamma/2}{(t+\gamma)(t+1+\gamma)}(1-\tau)^2\frac{\psi^2}{\kappa^2} + \frac{2(t+1)}{t+1+\gamma}\varepsilon_n^2 \\ &= \frac{\gamma^2 + (t+1)\gamma/2}{(t+\gamma)(t+1+\gamma)}(1-\tau)^2\sigma_r^2 + \frac{2(t+1)}{t+1+\gamma}\varepsilon_n^2. \end{aligned}\tag{37}$$

This bound, together with the assumed bound $\varepsilon_n \leq \frac{(1-\tau)\sigma_r}{2}$ yields

$$\|F^{t+1} - F^{t+1}_{\pi^*}\|_{\mathrm{F}}^2 \leq \frac{\gamma^2 + (t+1)\gamma/2 + (t+1)(t+\gamma)/2}{(t+\gamma)(t+1+\gamma)}(1-\tau)^2\sigma_r^2 \leq (1-\tau)^2\sigma_r^2 \;=\; \rho^2$$

whence $\mathrm{d}(F^{t+1}, F^*) \leq \rho$, thereby proving the induction hypothesis for $t+1$.

Moreover, since $\gamma \geq 1$, the inequality (37) implies that

$$\|F^{t+1} - F^*_{t+1}\|_{\mathrm{F}}^2 \leq \frac{\gamma}{t+\gamma}(1-\tau)^2\sigma_r^2 + 2\varepsilon_n^2 \leq \frac{20L^2\psi^2}{(t+1)\alpha^2} + 4\varepsilon_n^2,$$

thereby establishing the bound (17) stated in the theorem.

5.1.1 Proof of Lemma 1

We use the shorthand $\sigma_k = \sigma_k(F^*)$ for $k = 1, \ldots, r$. Since $\mathrm{d}(F, F^*) = \min_{A \in \mathcal{E}(M^*)} |||F - A|||_{\mathrm{F}} < \sigma_r$, there must exist a matrix $F_0^* \in \mathcal{E}(M^*)$ such that

$$|||F - F_0^*|||_{\mathrm{op}} \leq |||F - F_0^*|||_{\mathrm{F}} < \sigma_r. \tag{38}$$

It follows that the matrix F must have full column rank with all its singular values contained in the interval $\big[\sigma_r - \mathrm{d}(F, F^*), \sigma_1 + \mathrm{d}(F, F^*)\big]$. Let I_r denote the r-dimensional identity matrix, and the rank-r SVD of F_0^* be $F_0^* = VSR^\top$, where $V \in \mathbb{R}^{d \times r}$, $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ and $R \in \mathbb{R}^{r \times r}$ is orthonormal. Any unit vector in $\mathbb{R}^r$ has the form Rw for some unit vector $w \in \mathbb{R}^r$, and $F^\top F_0^* Rw = (F_0^* + F - F_0^*)^\top VSw = (RS + (F - F_0^*)^\top V)Sw$. We therefore have the bound

$$\|F^\top F_0^* Rw\|_2 \geq \sigma_{\min}(RS + (F - F_0^*)^\top V)\|Sw\|_2 \geq (\sigma_{\min}(RS) - \mathrm{d}(F, F^*))\sigma_r = (\sigma_r - \mathrm{d}(F, F^*))\sigma_r.$$

It follows that the r-dimensional matrix $U := F^\top F_0^*$ satisfies $\sigma_r(U) \geq (\sigma_r - \mathrm{d}(F, F^*))\sigma_r > 0$ and is thus invertible. A similar argument shows that $\sigma_1(U) \leq (\sigma_1 + \mathrm{d}(F, F^*))\sigma_1$.

Defining the matrix $F_{\pi^*} := F_0^* U^\top (UU^\top)^{-1/2}$, it is easy to verify $F_{\pi^*} \in \mathcal{E}(M^*)$. Observe that

$$F^\top F_{\pi^*} = F^\top F_0^* U^\top (UU^\top)^{-1/2} = UU^\top (UU^\top)^{-1/2} = (UU^\top)^{1/2},$$

which is symmetric and positive definite since U has strictly positive singular values. It is then clear that the matrix $(F - F_{\pi^*})^\top F_{\pi^*}$ is symmetric.

Any matrix $A \in \mathcal{E}(M^*)$ can be written as $A = F_{\pi^*}\Xi$ for some orthonormal matrix $\Xi \in \mathbb{R}^{r \times r}$, whence

$$\mathrm{trace}(F^\top A) = \mathrm{trace}\left((UU^\top)^{1/2}\Xi\right) \leq |||\Xi|||_{\mathrm{op}} \sum_{i=1}^r \sigma_i(U) = \sum_{i=1}^r \sigma_i(U). \tag{39}$$

Noting that

$$|||F - F_{\pi^*}|||_{\mathrm{F}}^2 \;=\; |||F|||_{\mathrm{F}}^2 + |||F_{\pi^*}|||_{\mathrm{F}}^2 - 2\,\mathrm{trace}\left((UU^\top)^{1/2}\right) = |||F|||_{\mathrm{F}}^2 + |||A|||_{\mathrm{F}}^2 - 2\sum_{i=1}^r \sigma_i(U),$$

we thus have the bound

$$|||F - F_{\pi^*}|||_{\mathrm{F}}^2 \leq |||F|||_{\mathrm{F}}^2 + |||A|||_{\mathrm{F}}^2 - 2\,\mathrm{trace}\left(F^\top A\right) \;=\; |||F - A|||_{\mathrm{F}}^2.$$

Since A was arbitrary in $\mathcal{E}(M^*)$. we conclude that F_{π^*} is a constrained minimizer of $|||F - F^*|||_{\mathrm{F}}$ over $\mathcal{E}(M^*)$.

Finally, we claim that the inequality in (39) is strict if $\Xi \neq I_r$, so that F_{π^*} is in fact the unique minimizer, as claimed. To establish this strictness, suppose that the SVD of $(UU^\top)^{1/2}$ is given by $(UU^\top)^{1/2} = R'\Sigma R'^\top$, where $\Sigma = \mathrm{diag}\big(\sigma_1(U), \ldots, \sigma_r(U)\big)$ and R' is an $r \times r$ orthonormal matrix. Then $\mathrm{trace}\left((UU^\top)^{1/2}\Xi\right) = \mathrm{trace}(\Sigma R'^\top \Xi R') = \mathrm{trace}(\Sigma\Xi')$, where $\Xi' := R'^\top \Xi R'$ is also orthonormal. If $\Xi \neq I_r$, then $\Xi' \neq I_r$ and therefore

$$\mathrm{trace}\left((UU^\top)^{1/2}\Xi\right) \;=\; \sum_{i=1}^r \sigma_i(U)\Xi'_{ii} \;<\; \sum_{i=1}^r \sigma_i(U)\|\Xi'_{i\cdot}\|_2 \;=\; \sum_{i=1}^r \sigma_i(U),$$

where the inequality follows from the facts that $\sigma_i(U) > 0$ for each index $i \in [r]$, and since $\Xi' \neq I_r$, we must have $|\Xi'_{ii}| < 1 = \|\Xi'_{i\cdot}\|_2$ for some index $i \in [r]$.

5.2 Proof of Theorem 2

We will in fact prove a slightly stronger form of the theorem, where the L-Lipschitz condition is replaced by the following relaxed condition:

$$|\langle\langle \nabla_M \widetilde{\mathcal{L}}_n(F \otimes F),\ F - F' \rangle\rangle| \leq L\big(|||F^*|||_{\mathrm{op}}^2 + |||F^*|||_{\mathrm{op}} |||F - F'|||_{\mathrm{F}}\big), \tag{40}$$

valid for all $F \in \mathcal{F} \cap \mathbb{B}_2((1-\tau^2)\sigma_r(F^*); F^*)$ and $F' \in \mathcal{F}$. We use the step size choice $\eta^t := c_\tau \frac{\alpha}{\beta} = \frac{\tau^4(1-\tau)^8}{C} \frac{\alpha}{\kappa^6 \beta^2}$, where $C = 272$.

As in the proof of Theorem 1, we will show that $\mathrm{d}(F^t, F^*) \leq (1-\tau)\psi$ for all iterates $t = 0, 1, 2, \ldots$. We do so via induction on the event $\mathcal{E}_t$ previously defined in equation (33). As before, the base event $\mathcal{E}_0$ holds by the theorem's conditions. The induction step is based on assuming that $\mathcal{E}_t$ holds, and then showing that $\mathcal{E}_{t+1}$ also holds. As before, it suffices to establish the bound $\mathrm{d}(F^{t+1}, F^*) \leq (1-\tau)\psi$. The proof is divided into several steps below.

Showing that $\mathbf{d}(F^{t+1}, F^*) \leq \rho = (1-\tau^2)\psi$: In the first step, we establish a slightly weaker bound on $\mathrm{d}(F^{t+1}, F^*)$.

In view of Lemma 1, the matrix $F^s_{\pi^*} := \arg\min_{A \in \mathcal{E}(M^*)} |||A - F^s|||_{\mathrm{F}}$ is uniquely defined for each time step $s \in \{0, 1, \ldots, t\}$. Recall that the iterate F^{t+1} is an optimal solution to the convex optimization problem (3). Consequently, the first-order conditions for optimality imply that

$$\langle\langle \widetilde{G}^t + \frac{1}{\eta^t}\Lambda^t,\ -F^{t+1} + F \rangle\rangle \geq 0, \quad \forall F \in \mathcal{F}. \tag{41}$$

Applying this condition with $F = F^t$ and using the relaxed Lipschitz condition (50) with the assumption $L = \beta$, we obtain

$$|||\Lambda^t|||_{\mathrm{F}}^2 \leq \eta^t \langle\langle \widetilde{G}^t,\ -\Lambda^t \rangle\rangle \leq \eta^t \beta(\psi^2 + \psi |||\Lambda^t|||_{\mathrm{F}}) = \frac{\tau^8(1-\tau)^2\alpha}{C\kappa^6\beta}(\psi^2 + \psi |||\Lambda^t|||_{\mathrm{F}}).$$

With the constant $C = 272$ and the fact that $\max\{\tau, 1/\kappa, \alpha/\beta\} \leq 1$, this inequality implies that $|||\Lambda^t|||_{\mathrm{F}} \leq \tau(1-\tau)\psi$ and hence that

$$\mathrm{d}(F^{t+1}, F^*) \leq |||F^{t+1} - F^t_{\pi^*}|||_{\mathrm{F}} \leq |||F^t - F^t_{\pi^*}|||_{\mathrm{F}} + |||\Lambda^t|||_{\mathrm{F}} \leq (1-\tau^2)\psi. \tag{42}$$

Note that we have not yet completed the induction step, but the bound (42) is useful below.

Establishing a recursive bound: With F^{t+1} satisfying the bound (42), Lemma 1 guarantees that the matrix $F^{t+1}_{\pi^*} := \arg\min_{A \in \mathcal{E}(M^*)} |||F^* - F^{t+1}|||_{\mathrm{F}}$ is well-defined. Our analysis involves the matrix differences $\Delta^s_{s'} := F^s - F^{s'}_{\pi^*}$, defined by pairs $s, s' \in \{0, 1, \ldots, t+1\}$. The following inequality is central to our analysis:

$$\frac{1}{\eta^t}\langle\langle \Lambda^t,\ -\Delta^{t+1}_t \rangle\rangle \geq \frac{\alpha}{2}|||\Delta^{t+1}_{t+1}|||_{\mathrm{F}}^2 - \frac{136\kappa^6\beta^2|||\Lambda^t|||_{\mathrm{F}}^2}{\alpha\tau^8} - 3\alpha\varepsilon_n^2. \tag{43}$$

We return to prove it shortly; taking it as given for the moment, it follows that

$$\begin{aligned}
\|\Delta_{t+1}^{t+1}\|_{\mathrm{F}}^2 \leq \|\Delta_t^{t+1}\|_{\mathrm{F}}^2 &= \|F^{t+1} - F_t^*\|_{\mathrm{F}}^2 \\
&= \|F^t - F^{t+1} + F^{t+1} - F_t^*\|_{\mathrm{F}}^2 - \|\Lambda^t\|_{\mathrm{F}}^2 + 2\langle\langle \Lambda^t,\ F^{t+1} - F_t^* \rangle\rangle \\
&\leq \|\Delta_t^t\|_{\mathrm{F}}^2 - \|\Lambda^t\|_{\mathrm{F}}^2 + \eta^t \cdot \Big(-\alpha\|\Delta_{t+1}^{t+1}\|_{\mathrm{F}}^2 + \frac{272\kappa^6\beta^2}{\alpha\tau^8}\|\Lambda^t\|_{\mathrm{F}}^2 + 6\alpha\varepsilon_n^2\Big) \\
&= \|\Delta_t^t\|_{\mathrm{F}}^2 - \|\Lambda^t\|_{\mathrm{F}}^2 - \frac{\alpha^2\tau^8(1-\tau)^2}{C\kappa^6\beta^2}\|\Delta_{t+1}^{t+1}\|_{\mathrm{F}}^2 + \frac{272(1-\tau)^2}{C}\|\Lambda^t\|_{\mathrm{F}}^2 + \frac{6\alpha^2\tau^8(1-\tau)^2}{C\kappa^6\beta^2}\varepsilon_n^2 \\
&\leq \|\Delta_t^t\|_{\mathrm{F}}^2 - \frac{\alpha^2\tau^8}{C\kappa^6\beta^2}\|\Delta_{t+1}^{t+1}\|_{\mathrm{F}}^2 + \frac{6\alpha^2\tau^8}{C\kappa^6\beta^2}\varepsilon_n^2.
\end{aligned}$$

where we used the step size choice $\eta^t = \frac{\alpha\tau^8(1-\tau)^2}{C\kappa^6\beta^2}$ and the assumption $C \geq 272$ in the last two lines. Rearranging this inequality yields the recursive bound

$$\|\Delta_{t+1}^{t+1}\|_{\mathrm{F}}^2 \leq \Big(1 + \frac{\alpha^2\tau^8}{C\kappa^6\beta^2}\Big)^{-1}\Big(\|\Delta_t^t\|_{\mathrm{F}}^2 + \frac{6\alpha^2\tau^8}{C\kappa^6\beta^2}\varepsilon_n^2\Big) \leq \Big(1 - \frac{\alpha^2\tau^8}{2C\kappa^6\beta^2}\Big)\Big(\|\Delta_t^t\|_{\mathrm{F}}^2 + \frac{6\alpha^2\tau^8}{C\kappa^6\beta^2}\varepsilon_n^2\Big). \tag{44}$$

Completing the induction and proof: Since $\|\Delta_t^t\|_{\mathrm{F}} \leq (1-\tau)\psi$ by induction hypothesis and $\varepsilon_n \leq \frac{1-\tau}{4}\psi$ by assumption, the inequality (44) above implies

$$\|\Delta_{t+1}^{t+1}\|_{\mathrm{F}}^2 \leq \Big(1 - \frac{\alpha^2\tau^8}{2C\kappa^6\beta^2}\Big)\Big(1 + \frac{\alpha^2\tau^8}{2C\kappa^6\beta^2}\Big)(1-\tau)^2\psi^2 \leq (1-\tau)^2\psi^2,$$

which completes the induction step. Moreover, by applying the inequality (44) recursively, we find that

$$\|\Delta_t^t\|_{\mathrm{F}}^2 \leq \Big(1 - \frac{\alpha^2\tau^8}{2C\kappa^6\beta^2}\Big)^t\|\Delta_0^0\|_{\mathrm{F}}^2 + \frac{2C\kappa^6\beta^2}{\alpha^2\tau^8} \cdot \frac{6\alpha^2\tau^8}{C\kappa^6\beta^2}\varepsilon_n^2 \leq \Big(1 - \frac{\alpha^2\tau^8}{2C\kappa^6\beta^2}\Big)^t\|\Delta_0^0\|_{\mathrm{F}}^2 + (4\varepsilon_n)^2,$$

thereby completing the proof of the theorem.

Proof of inequality (43)**:** It remains to prove the intermediate claim (43). With $\mathrm{d}(F^{t+1}, F^*) \leq \rho = (1-\tau^2)\psi$ as established in (42), the local descent condition (14) yields

$$\langle\langle \widetilde{G}^{t+1},\ \Delta_t^{t+1} \rangle\rangle \geq \alpha\|\Delta_{t+1}^{t+1}\|_{\mathrm{F}}^2 - \frac{\beta^2}{\alpha}\|F_{\pi^*}^{t+1} - F_{\pi^*}^t\|_{\mathrm{F}}^2 - \alpha\varepsilon_n^2.$$

In order to proceed, we need a second technical lemma. Recall that $\kappa = \frac{\sigma_1(F^*)}{\sigma_r(F^*)}$ is the condition number of F^*.

Lemma 2. *Under the conditions of Theorem 2, we have*

$$\|F_{\pi^*}^{t+1} - F_{\pi^*}^t\|_F \leq \frac{10\kappa^3\|\Lambda^t\|_F}{\tau^4}.$$

See Section 5.2.1 for the proof of this claim.

Applying Lemma 2, we find that

$$\langle\!\langle \widetilde{G}^{t+1},\ \Delta_t^{t+1} \rangle\!\rangle \geq \alpha |\!|\!| \Delta_{t+1}^{t+1} |\!|\!|_{\mathrm{F}}^2 - \frac{100\kappa^6\beta^2}{\alpha\tau^8} |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}^2 - \alpha\varepsilon_n^2. \tag{45}$$

On the other hand, the smoothness condition (16) yields that

$$\langle\!\langle \widetilde{G}^t - \widetilde{G}^{t+1},\ \Delta_t^{t+1} \rangle\!\rangle \geq -\beta |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}} |\!|\!| \Delta_t^{t+1} |\!|\!|_{\mathrm{F}} - \alpha\varepsilon_n |\!|\!| \Delta_t^{t+1} |\!|\!|_{\mathrm{F}}.$$

Together with Lemma 2, we obtain

$$\begin{aligned}
&\langle\!\langle \widetilde{G}^t - \widetilde{G}^{t+1},\ \Delta_t^{t+1} \rangle\!\rangle \qquad (46)\\
&\geq -\Big(\beta |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}} + \alpha\varepsilon_n\Big)\Big(|\!|\!| \Delta_{t+1}^{t+1} |\!|\!|_{\mathrm{F}} + \frac{10\kappa^3 |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}}{\tau^4}\Big).\\
&= -\beta |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}} |\!|\!| \Delta_{t+1}^{t+1} |\!|\!|_{\mathrm{F}} - \alpha\varepsilon_n |\!|\!| \Delta_{t+1}^{t+1} |\!|\!|_{\mathrm{F}} - \frac{10\kappa^3\beta}{\tau^4} |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}^2 - \frac{10\kappa^3\alpha}{\tau^4}\varepsilon_n |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}\\
&\overset{(i)}{\geq} -\Big(\frac{\alpha}{4} |\!|\!| \Delta_{t+1}^{t+1} |\!|\!|_{\mathrm{F}}^2 + \frac{\beta^2}{\alpha} |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}^2\Big) - \Big(\frac{\alpha}{4} |\!|\!| \Delta_{t+1}^{t+1} |\!|\!|_{\mathrm{F}}^2 + \alpha\varepsilon_n^2\Big) - \frac{10\kappa^3\beta}{\tau^4} |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}^2 - \Big(\frac{25\kappa^6\alpha}{\tau^8} |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}^2 + \alpha\varepsilon_n^2\Big)\\
&\geq -\frac{\alpha}{2} |\!|\!| \Delta_{t+1}^{t+1} |\!|\!|_{\mathrm{F}}^2 - 2\alpha\varepsilon_n^2 - \frac{36\kappa^6\beta^2}{\alpha\tau^8} |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}^2, \qquad (47)
\end{aligned}$$

where the step (i) follows from the AM-GM inequality.

Finally, the M^*-faithfulness of $\mathcal{F}$ ensures that $F_{\pi^*}^t \in \mathcal{F}$, so that we may apply the bound (41) with $F = F_{\pi^*}^t$, thereby obtaining

$$\langle\!\langle \widetilde{G}^t + \frac{1}{\eta^t}\Lambda^t,\ -\Delta_t^{t+1} \rangle\!\rangle \geq 0. \tag{48}$$

Adding together inequalities (45), (47) and (48) yields the claim (43).

5.2.1 Proof of Lemma 2

By dividing through by ψ we may assume that $\psi = 1$, so $\sigma_1(F^*) = \kappa$, where we recall that κ is the condition number of F^*. Define the matrices $U_s := (F^s)^\top F_{\pi^*}^t$ for $s \in \{t, t+1\}$, and recall that we have shown

$$\max\Big\{ |\!|\!| F^t - F_{\pi^*}^t |\!|\!|_{\mathrm{F}}, |\!|\!| F^{t+1} - F_{\pi^*}^t |\!|\!|_{\mathrm{F}} \Big\} \leq 1 - \tau^2.$$

The same argument as in the proof of Lemma 1 from the previous section show that the singular values of U_s are in the interval $[\tau^2, \kappa + (1-\tau^2)]$, and we have the expression $F_{\pi^*}^s := F_{\pi^*}^t U_s^\top (U_s U_s^\top)^{-1/2}$ for $s \in \{t, t+1\}$. Since $U_t = U_t^\top = (U_t U_t^\top)^{1/2}$, we have

$$|\!|\!| U_{t+1} - U_t |\!|\!|_{\mathrm{F}} = |\!|\!| (F^{t+1} - F^t)^\top F_{\pi^*}^t |\!|\!|_{\mathrm{F}} \leq \sigma_1 |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}.$$

By applying a known perturbation bound for matrix square roots ([29, Lemma 15]), we find that

$$\begin{aligned}
|\!|\!| (U_t U_t^\top)^{1/2} - (U_{t+1} U_{t+1}^\top)^{1/2} |\!|\!|_{\mathrm{F}} &\leq \frac{|\!|\!| U_t U_t^\top - U_{t+1} U_{t+1}^\top |\!|\!|_{\mathrm{F}}}{\sigma_{\min}((U_t U_t^\top)^{1/2}) + \sigma_{\min}((U_{t+1} U_{t+1}^\top)^{1/2})}\\
&\leq \frac{1}{2\tau^2} |\!|\!| U_t U_t^\top - U_{t+1} U_{t+1}^\top |\!|\!|_{\mathrm{F}}\\
&= \frac{1}{2\tau^2} |\!|\!| (U_t - U_{t+1}) U_t^\top + U_{t+1} (U_t - U_{t+1})^\top |\!|\!|_{\mathrm{F}}\\
&\leq \frac{2\kappa^2}{\tau^2} |\!|\!| \Lambda^t |\!|\!|_{\mathrm{F}}.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}\|U_t^\top \left[(U_{t+1}U_{t+1}^\top)^{-1/2} - (U_tU_t^\top)^{-1/2}\right]\|_{\mathrm{F}} &= \|(U_{t+1}U_{t+1}^\top)^{-1/2}\left[(U_tU_t^\top)^{1/2} - (U_{t+1}U_{t+1}^\top)^{1/2}\right]\|_{\mathrm{F}} \\ &\le \frac{1}{\tau^2}\|(U_tU_t^\top)^{1/2} - (U_{t+1}U_{t+1}^\top)^{1/2}\|_{\mathrm{F}} \\ &\le \frac{2\kappa^2\|\Lambda^t\|_{\mathrm{F}}}{\tau^4}.\end{aligned}$$

Putting together the pieces, it follows that

$$\begin{aligned}\|F_{t+1}^* - F_t^*\|_{\mathrm{F}} &= \|F_{\pi^*}^t U_{t+1}^\top (U_{t+1}U_{t+1}^\top)^{-1/2} - F_{\pi^*}^t U_t^\top (U_tU_t^\top)^{-1/2}\|_{\mathrm{F}} \\ &= \|F_{\pi^*}^t (U_{t+1}-U_t)^\top (U_{t+1}U_{t+1}^\top)^{-1/2} + F_{\pi^*}^t U_t^\top \left[(U_{t+1}U_{t+1}^\top)^{-1/2} - (U_tU_t^\top)^{-1/2}\right]\|_{\mathrm{F}} \\ &\le \|F_{\pi^*}^t\|_{\mathrm{op}}\|U_{t+1}-U_t\|_{\mathrm{F}}\|(U_{t+1}U_{t+1}^\top)^{-1/2}\|_{\mathrm{op}} + \|F_{\pi^*}^t\|_{\mathrm{op}}\|U_t^\top\left[(U_{t+1}U_{t+1}^\top)^{-1/2} - (U_tU_t^\top)^{-1/2}\right]\|_{\mathrm{F}} \\ &\le \kappa\cdot\kappa\|\Lambda^t\|_{\mathrm{F}}\cdot\frac{1}{\tau^2} + \kappa\cdot\frac{2\kappa^2\|\Lambda^t\|_{\mathrm{F}}}{\tau^4} \\ &\le \frac{10\kappa^3\|\Lambda^t\|_{\mathrm{F}}}{\tau^4},\end{aligned}$$

as claimed.

6 Proofs of corollaries

In this section, we prove the corollaries by applying our general theory.

The general theorems in Section 3 are stated in terms of the loss function $\widetilde{\mathcal{L}}_n$ of the factor variable F. Sometimes it is convenient to work with the original loss function $\mathcal{L}_n$ of the $d \times d$ variable M. These two loss functions are related by $\widetilde{\mathcal{L}}_n(F) = \mathcal{L}_n(F\otimes F)$ and $\nabla_F\widetilde{\mathcal{L}}_n(F) = \left[\nabla_M\mathcal{L}_n(F\otimes F) + (\nabla_M\mathcal{L}_n(F\otimes F))^\top\right]F$, and the convergence results can be restated in terms of $\mathcal{L}_n$. We do so below for the result in Theorem 2.

The following conditions for $\mathcal{L}_n$ are the counterparts of the corresponding conditions for $\widetilde{\mathcal{L}}_n$.

Definition 5 (Local descent condition for $\mathcal{L}_n$)**.** For some curvature parameter α, statistical tolerance ε_n and radius ρ, we say that the cost function $\mathcal{L}_n$ satisfies a *local descent condition* with parameters $(\alpha, \varepsilon_n, \rho)$ over $\mathcal{F}$ if for each $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$, there exists $F_{\pi^*} \in \arg\min_{F^*\in\mathcal{E}(M^*)}\|F^* - F\|_{\mathrm{F}}$ such that

$$\langle\!\langle \nabla_M\mathcal{L}_n(F\otimes F),\ F\otimes F - F_{\pi^*}\otimes F_{\pi^*} + (F-F_{\pi^*})\otimes(F-F_{\pi^*})\rangle\!\rangle \ge 2\alpha\|F-F_{\pi^*}\|_{\mathrm{F}}^2 - \frac{1}{4}\alpha\varepsilon_n\|F-F_{\pi^*}\|_{\mathrm{F}}. \tag{49}$$

Definition 6 (Relaxed Local Lipschitz condition for $\mathcal{L}_n$)**.** For some Lipschitz constant L and radius ρ, we say that the $\mathcal{L}_n$ satisfies a *relaxed local Lipschitz condition* with parameter (L, ρ) over $\mathcal{F}$ if for each $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$ and $F' \in \mathcal{F}$,

$$|\langle\!\langle \nabla_M\mathcal{L}_n(F\otimes F),\ (F-F')\otimes F\rangle\!\rangle| \le \frac{1}{2}L\big(\|F^*\|_{\mathrm{op}}^2 + \|F^*\|_{\mathrm{op}}\|F-F'\|_{\mathrm{F}}\big). \tag{50}$$

Of course, this relaxed Lipschitz condition for $\mathcal{L}_n$ is implied by a Lipschitz condition of the form

$$\|\nabla_M \mathcal{L}_n(F \otimes F)F\|_{\mathrm{F}} \leq \frac{1}{2} L \|F^*\|_{\mathrm{op}}, \quad \forall F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*). \tag{51}$$

Definition 7 (Local smoothness condition for $\mathcal{L}_n$)**.** For some curvature and smoothness parameters α and β, statistical tolerance ε_n and radius ρ, we say that the loss function $\mathcal{L}_n$ satisfies a *local smoothness condition* with parameters $(\alpha, \beta, \varepsilon_n, \rho)$ over $\mathcal{F}$ if for each $F, F', F'' \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$ and any $F^* \in \mathcal{E}(M^*)$,

$$|\langle\langle \nabla_M \mathcal{L}_n(F \otimes F) - \nabla_M \mathcal{L}_n(F' \otimes F'),\ F' \otimes (F - F^*) \rangle\rangle| \leq \frac{1}{4}\big(\beta \|F - F'\|_{\mathrm{F}} + \alpha\varepsilon_n\big)\|F - F^*\|_{\mathrm{F}}, \tag{52a}$$

$$|\langle\langle \nabla_M \mathcal{L}_n(F \otimes F),\ (F - F^*) \otimes (F' - F'') \rangle\rangle| \leq \frac{1}{4}\big(\beta \|F' - F''\|_{\mathrm{F}} + \alpha\varepsilon_n\big)\|F - F^*\|_{\mathrm{F}}. \tag{52b}$$

Now suppose that for some numbers α, β, L, ε_n and τ with $0 < \alpha \leq \beta = L$, $0 < \tau < 1$ and $\varepsilon_n \leq \frac{1-\tau}{4}\sigma_r(F^*)$, the empirical loss $\mathcal{L}_n$ satisfies the local descent, relaxed Lipschitz and smoothness conditions in Definitions 5–7 over $\mathcal{F}$ with parameters α, β, L, ε_n and $\rho = (1-\tau^2)\sigma_r(F^*)$, that the set $\mathcal{F}$ is M^*-faithful and convex, and that the matrix $\nabla \mathcal{L}_n(M)$ is symmetric for any symmetric matrix M. As we show in the proof of Theorem 3, the loss function $\widetilde{\mathcal{L}}_n$ then satisfies the corresponding conditions with the same parameters. Consequently, we have the following result:

Theorem 3. *Under the previously stated conditions, the conclusion in Theorem 2 holds.*

See Section A for the proof of this claim.

In remainder of this section, we verify the above conditions for each of our examples. It is easy to see that in these examples the matrix $\nabla \mathcal{L}_n(M)$ is indeed symmetric for any symmetric M, so it remains to verify the conditions in Definitions 5–7 for $\mathcal{L}_n$ and the M^*-faithfulness of $\mathcal{F}$.

Recall that σ_i and κ are the i-th singular value and the condition number of F^*, respectively. Throughout this section, we let F be an arbitrary matrix in $\mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$ and $M = F \otimes F$, where $\mathcal{F}$ and ρ are specified for each of our examples. In all these examples $\rho < \sigma_r$, so Lemma 1 guarantees that we can write $F_{\pi^*} := \arg\min_{A \in \mathcal{E}(M^*)} \|A - F\|_{\mathrm{F}}$ and $\Delta := F - F_{\pi^*}$, and $\Delta^\top F_{\pi^*}$ is a symmetric matrix. Let F^* be an arbitrary matrix in $\mathcal{E}(M^*)$, and recall that $M^* = F^* \otimes F^* = F_{\pi^*} \otimes F_{\pi^*}$. Denote by C, c, c_1 etc. positive universal constants, whose values could change from line to line.

6.1 Proof of Corollary 2

We begin by proving our claims for the matrix sensing observation model. By dividing through by σ_r, we may assume without loss of generality $\sigma_r = 1$, so $\kappa = \sigma_1$. Recall that $\alpha = 6\delta_{4r}$, $L = \beta = 64\kappa^2$, $\varepsilon_n = \frac{2\sqrt{r}\kappa \|n^{-1}\mathfrak{X}_n^*(\epsilon)\|_{\mathrm{op}}}{\delta_{4r}}$ and $\rho = 1 - 12\delta_{4r}$. It is a standard result that RIP implies preservation of inner products between low rank matrices, as summarized in the lemma below:

Lemma 3. *If $\mathfrak{X}_n$ satisfies a RIP-δ_{4r} condition, then*

$$\frac{1}{n}\Big|\langle \mathfrak{X}_n(A),\, \mathfrak{X}_n(B)\rangle - \langle A,\, B\rangle\Big| \leq \delta_{4r}\|A\|_F\|B\|_F \quad \textit{for all matrices } A, B \in \mathbb{R}^{d\times d} \textit{ of rank at most } 2r.$$

For completeness, we provide a proof in Appendix B.1.

Under the matrix sensing observation model (6), the gradient of $\mathcal{L}_n$ takes the form

$$\nabla \mathcal{L}_n(F \otimes F) = \frac{1}{n} \mathfrak{X}_n^* \mathfrak{X}_n(F \otimes F - F^* \otimes F^*) - \frac{1}{n} \mathfrak{X}_n^*(\epsilon), \tag{53}$$

Below we verify the local descent, Lipschitz and smoothness conditions.

Local descent: We have the decomposition $\langle\!\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle = T_1 + T_2$, where

$$T_1 := \frac{1}{n} \langle\!\langle \mathfrak{X}_n^* \mathfrak{X}_n(M - M^*),\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle \qquad \text{and} \quad T_2 := -\frac{1}{n} \langle\!\langle \mathfrak{X}_n^*(\epsilon),\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle.$$

Lemma 3 implies that

$$\begin{aligned} T_1 &= \frac{1}{n} \langle\!\langle \mathfrak{X}_n(M - M^*),\ \mathfrak{X}_n(M - M^* + \Delta \otimes \Delta) \rangle\!\rangle \\ &\geq \langle\!\langle M - M^*,\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle - \delta_{4r} |\!|\!| M - M^* |\!|\!|_{\mathrm{F}} |\!|\!| M - M^* + \Delta \otimes \Delta |\!|\!|_{\mathrm{F}}. \end{aligned}$$

Since the matrix $\Delta^\top F_{\pi^*}$ is symmetric, some algebra shows that

$$\begin{aligned} \langle\!\langle M - M^*,\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle &= \langle\!\langle F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*} + \Delta \otimes \Delta,\ F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*} + 2\Delta \otimes \Delta \rangle\!\rangle \\ &= 2 |\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}}^2 + 2 \langle\!\langle F_{\pi^*} \otimes \Delta,\ \Delta \otimes \Delta \rangle\!\rangle + 2 |\!|\!| (F_{\pi^*})^\top \Delta + \Delta^\top \Delta |\!|\!|_{\mathrm{F}}^2 \\ &\geq 2 |\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} \big(|\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} - |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2 \big) \end{aligned}$$

In addition, we have

$$\begin{aligned} |\!|\!| M - M^* |\!|\!|_{\mathrm{F}} |\!|\!| M - M^* + \Delta \otimes \Delta |\!|\!|_{\mathrm{F}} &= |\!|\!| F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*} + \Delta \otimes \Delta |\!|\!|_{\mathrm{F}} |\!|\!| F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*} + 2\Delta \otimes \Delta |\!|\!|_{\mathrm{F}} \\ &\leq \big(2 |\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} + |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2 \big) \big(2 |\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} + 2 |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2 \big). \end{aligned}$$

It follows that

$$\begin{aligned} T_1 &\geq |\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} \Big((2 - 4\delta_{4r}) |\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} - (2 + 6\delta_{4r}) |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2 \Big) - 2\delta_{4r} |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^4 \\ &\geq |\!|\!| \Delta |\!|\!|_{\mathrm{F}} \Big((2 - 4\delta_{4r}) |\!|\!| \Delta |\!|\!|_{\mathrm{F}} - (2 + 6\delta_{4r})(1 - 12\delta_{4r}) |\!|\!| \Delta |\!|\!|_{\mathrm{F}} \Big) - 2\delta_{4r} |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2 \geq 12\delta_{4r} |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2, \end{aligned}$$

where the second step uses the inequalities $|\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} \geq \sigma_r |\!|\!| \Delta |\!|\!|_{\mathrm{F}} = |\!|\!| \Delta |\!|\!|_{\mathrm{F}}$ and $|\!|\!| \Delta |\!|\!|_{\mathrm{F}} \leq \rho = 1 - 12\delta_{4r}$.

On the other hand, we have

$$\begin{aligned} |T_2| &\leq |\!|\!| n^{-1} \mathfrak{X}_n^*(\epsilon) |\!|\!|_{\mathrm{op}} \cdot \sqrt{2r} |\!|\!| M - M^* + \Delta \otimes \Delta |\!|\!|_{\mathrm{F}} \\ &\leq |\!|\!| n^{-1} \mathfrak{X}_n^*(\epsilon) |\!|\!|_{\mathrm{op}} \cdot \sqrt{2r} \big(2 |\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} + 2 |\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2 \big) \leq |\!|\!| n^{-1} \mathfrak{X}_n^*(\epsilon) |\!|\!|_{\mathrm{op}} \cdot 6\sqrt{r}\kappa |\!|\!| \Delta |\!|\!|_{\mathrm{F}}, \end{aligned}$$

where the last step uses the inequalities $|\!|\!| F_{\pi^*} \otimes \Delta |\!|\!|_{\mathrm{F}} \leq \sigma_1(F_{\pi^*}) |\!|\!| \Delta |\!|\!|_{\mathrm{F}} = \kappa |\!|\!| \Delta |\!|\!|_{\mathrm{F}}$ and $|\!|\!| \Delta |\!|\!|_{\mathrm{F}} \leq 1$. Combining this upper bound with our lower bound on T_1, we find that

$$\langle\!\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle \geq 12\delta_{4r} \Big(|\!|\!| \Delta |\!|\!|_{\mathrm{F}}^2 - \frac{\sqrt{r}\kappa}{2\delta_{4r}} |\!|\!| n^{-1} \mathfrak{X}_n^*(\epsilon) |\!|\!|_{\mathrm{op}} |\!|\!| \Delta |\!|\!|_{\mathrm{F}} \Big),$$

thereby establishing the local descent condition (49) for $\mathcal{L}_n$.

Local Lipschitz and smoothness: We have the following variational representation:

$$\|\nabla\widetilde{\mathcal{L}}_n(F)\|_{\mathrm{F}} = \sup_{\substack{H\in\mathbb{R}^{d\times r}\\ \|H\|_{\mathrm{F}}=1}} \langle\!\langle \nabla\mathcal{L}_n(F\otimes F),\ H\otimes F\rangle\!\rangle.$$

Using the form of the gradient $\nabla\mathcal{L}_n$ given in equation (53), we have

$$\nabla\mathcal{L}_n(F\otimes F) - \nabla\mathcal{L}_n(F'\otimes F') = \frac{1}{n}\mathfrak{X}_n^*\mathfrak{X}_n\Big(F\otimes F - F'\otimes F'\Big).$$

Note moreover that $0\in\mathbb{B}_2(1;F^*)$, $\alpha\leq L=\beta$ and $\varepsilon_n\leq 1$. Using these facts, it can be verified that the local Lipschitz and smoothness conditions (51)–(52b) for $\mathcal{L}_n$ are implied by a bound of the form

$$|\langle\!\langle n^{-1}\mathfrak{X}_n^*\mathfrak{X}_n(F\otimes F - F'\otimes F'),\ H\otimes G\rangle\!\rangle| + |\langle\!\langle n^{-1}\mathfrak{X}_n^*(\epsilon),\ H\otimes G\rangle\!\rangle| \leq \frac{1}{8\kappa}\Big(\beta\|F-F'\|_{\mathrm{F}} + \alpha\varepsilon_n\Big)\|H\|_{\mathrm{F}}\|G\|_{\mathrm{op}}, \tag{54}$$

valid for all $F,F'\in\mathbb{B}_2(1;F^*)$ and for all $H,G\in\mathbb{R}^{d\times r}$.

Let us prove the bound (54). Lemma 3 guarantees that

$$\begin{aligned}
|\langle\!\langle n^{-1}\mathfrak{X}_n^*\mathfrak{X}_n(F\otimes F - F'\otimes F'),\ H\otimes G\rangle\!\rangle| &\leq (1+\delta_{4r})\|F\otimes F - F'\otimes F'\|_{\mathrm{F}}\cdot\|H\|_{\mathrm{F}}\|G\|_{\mathrm{op}}\\
&\leq (1+\delta_{4r})\big(\|F\|_{\mathrm{op}} + \|F'\|_{\mathrm{op}}\big)\|F-F'\|_{\mathrm{F}}\cdot\|H\|_{\mathrm{F}}\|G\|_{\mathrm{op}}\\
&\leq 8\kappa\|F-F'\|_{\mathrm{F}}\|H\|_{\mathrm{F}}\|G\|_{\mathrm{op}},
\end{aligned}$$

where the last step follows from the facts that $\|F\|_{\mathrm{op}}\leq\|F^*\|_{\mathrm{op}}+\mathrm{d}(F,F^*)\leq 2\kappa$ for all $F\in\mathbb{B}_2(1;F^*)$ and $\delta_{4r}\leq 1$. We also have

$$|\langle\!\langle n^{-1}\mathfrak{X}_n^*(\epsilon),\ H\otimes G\rangle\!\rangle| \leq \|n^{-1}\mathfrak{X}_n^*(\epsilon)\|_{\mathrm{op}}\cdot\|H\otimes G\|_{\mathrm{nuc}} \leq \|n^{-1}\mathfrak{X}_n^*(\epsilon)\|_{\mathrm{op}}\cdot\sqrt{r}\|H\|_{\mathrm{F}}\|G\|_{\mathrm{op}}.$$

Combining these inequalities and recalling the values of $\alpha,\beta,\varepsilon_n$ yields the claim (54).

6.2 Proof of Corollary 1

We now turn to the proof of our claims for the matrix completion model. By dividing through by $\|F^*\|_{\mathrm{op}}$ and using the equal eigenvalue assumption, we may assume without loss of generality $\|F^*\|_{\mathrm{op}} = \sigma_r(F^*) = 1$. We first show that $\mathcal{F}$ is M^*-faithful. Note that $\mathcal{F}$ is the set of matrices with each row in the ℓ_2 ball of radius $\gamma := \sqrt{\frac{2\mu}{dr}}\|F^0\|_{\mathrm{op}}$. Because $F^0\in\mathbb{B}_2(\frac{1}{5};F^*)$, we have $\|F^0\|_{\mathrm{op}}\geq\frac{4}{5}\|F^*\|_{\mathrm{op}}$, whence $\gamma\geq\sqrt{\frac{\mu}{dr}}\|F^*\|_{\mathrm{op}}$. Combined with the definition of the incoherence parameter μ, we see that any matrix $F^*\in\mathcal{E}(M^*)$ satisfies the maximum row norm bound $\|F^*\|_{2,\infty}\leq\gamma$, so that $F^*\in\mathcal{F}$ as desired.

For future reference, we make note of a useful property satisfied by any matrices $F\in\mathcal{F}\cap\mathbb{B}_2(\rho;F^*)$ and $F^*\in\mathcal{E}(M^*)$. As a consequence of the clipping operation $\Pi_{\mathcal{F}}$, the row norms of the matrices F and $F-F^*$ satisfy the bounds

$$\|F\|_{2,\infty}\leq\sqrt{\frac{2\mu r}{d}}\|F^0\|_{\mathrm{op}}\leq 2\sqrt{\frac{\mu r}{d}}\|F^*\|_{\mathrm{op}}\leq 2\sqrt{\frac{\mu r}{d}}\quad\text{and} \tag{55a}$$

$$\|F-F^*\|_{2,\infty}\leq\|F\|_{2,\infty}+\|F^*\|_{2,\infty}\leq 3\sqrt{\frac{\mu r}{d}} \tag{55b}$$

where we use the inequality $\|F^0\|_{\rm op} \leq \|F^*\|_{\rm op} + {\rm d}(F^0, F^*) \leq \frac{6}{5}\|F^*\|_{\rm op}$ and the normalization assumption $\|F^*\|_{\rm op} = 1$. Inequality (55b) applies in particular to the difference matrix $\Delta := F - F_{\pi^*}$, where we recall that the matrix $F_{\pi^*} := \arg\min_{A \in \mathcal{E}(M^*)} \|A - F\|_{\rm F}$ is uniquely defined thanks to Lemma 1.

It remains to verify that the local descent, Lipschitz and smoothness conditions are satisfied with high probability. Under the matrix completion observation model (19), the gradient takes the form $\nabla \mathcal{L}_n(M) = \frac{1}{p}\Pi_\Omega(M - M^* - E)$, where Π_Ω is the projection operator. We need two technical lemmas. The first lemma shows that the projection operator Π_Ω approximately preserves inner products between matrices whose column or row spaces are equal to the column space of F^*.

Lemma 4. *There is a universal constant c such that for any $0 < \epsilon < 1$ and $p \geq c\frac{\mu r \log d}{\epsilon^2 d}$, uniformly for all $H, G \in \mathbb{R}^{d \times r}$, we have*

$$|p^{-1}\langle\langle \Pi_\Omega\,(F^* \otimes H),\ \Pi_\Omega(G \otimes F^*)\rangle\rangle - \langle\langle F^* \otimes H,\ G \otimes F^*\rangle\rangle| \leq \epsilon \|F^*\|_{op}^2 \|H\|_F \|G\|_F, \tag{56a}$$

$$|p^{-1}\langle\langle \Pi_\Omega\,(F^* \otimes H),\ \Pi_\Omega(F^* \otimes G)\rangle\rangle - \langle\langle F^* \otimes H,\ F^* \otimes G\rangle\rangle| \leq \epsilon \|F^*\|_{op}^2 \|H\|_F \|G\|_F, \tag{56b}$$

with probability at least $1 - 2d^{-3}$.

See Appendix C.1 for the proof of this claim.

Our second lemma is useful for controlling the projection of "small" matrices to Ω.

Lemma 5. *There is a universal constant $c > 0$ such that for any $\epsilon \in (0, 1)$ and $p \geq \frac{C}{\epsilon^2}\big(\frac{\mu^2 r^2}{d} + \frac{\log d}{d}\big)$, then unformly for all matrices $Z \in \mathbb{R}^{d \times d}$, $G \in \mathbb{R}^{d \times r}$ and H with $\|H\|_{2,\infty} \leq 6\sqrt{\frac{\mu r}{d}}$, we have*

$$p^{-1}\|\Pi_\Omega(H \otimes H)\|_F^2 \leq (1 + \epsilon)\|H\|_F^4 + \epsilon\|H\|_F^2, \tag{57a}$$

$$p^{-1}\|\Pi_\Omega(Z)H\|_F^2 \leq 72\mu r\|\Pi_\Omega(Z)\|_F^2, \tag{57b}$$

$$p^{-1}\|\Pi_\Omega(\phi \otimes \omega)\|_F^2 \leq 72\mu r\|G\|_F^2 \tag{57c}$$

with probability at least $1 - 2d^{-4}$.

See Section C.2 for the proof of this claim.

For the remainder of the proof, we condition on the intersection of the events in Lemmas 4 and 5.

Local descent: We have the decomposition $\langle\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta\rangle\rangle = T_1 - T_2$, where

$$T_1 := \frac{1}{p}\langle\langle \Pi_\Omega(M - M^*),\ \Pi_\Omega(M - M^* + \Delta \otimes \Delta)\rangle\rangle, \qquad \text{and} \quad T_2 := \frac{1}{p}\langle\langle \Pi_\Omega(E),\ M - M^* + \Delta \otimes \Delta\rangle\rangle.$$

Our strategy is to lower bound T_1 and upper bound $|T_2|$. Beginning with T_1, we have

$$\begin{aligned} T_1 &= p^{-1}\langle\langle \Pi_\Omega(F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*} + \Delta \otimes \Delta),\ \Pi_\Omega(F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*} + 2\Delta \otimes \Delta)\rangle\rangle \\ &\geq p^{-1}\Big(\|\Pi_\Omega(F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*})\|_{\rm F} - 2\|\Pi_\Omega(\Delta \otimes \Delta)\|_{\rm F}\Big)\Big(\|\Pi_\Omega(F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*})\|_{\rm F} - \|\Pi_\Omega(\Delta \otimes \Delta)\|_{\rm F}\Big). \end{aligned}$$

Recall that $p \geq \frac{C}{\epsilon^2}\Big(\frac{\mu^2 r^2}{d} + \frac{\mu r d}{d}\Big)$ by assumption of the corollary. By Lemma 4, we find that

$$p^{-1}\|\Pi_\Omega(F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*})\|_{\mathrm{F}}^2 \geq (1-\epsilon)\|F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*}\|_{\mathrm{F}}^2 \geq 2(1-\epsilon)\|\Delta\|_{\mathrm{F}}^2,$$

where the last step follow from the inequality

$$\|F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*}\|_{\mathrm{F}}^2 = 2\|F_{\pi^*} \otimes \Delta\|_{\mathrm{F}}^2 + 2\langle\!\langle F_{\pi^*} \otimes \Delta,\ \Delta \otimes F_{\pi^*}\rangle\!\rangle = 2\|F_{\pi^*} \otimes \Delta\|_{\mathrm{F}}^2 + 2\|\Delta^\top F_{\pi^*}\|_{\mathrm{F}}^2 \geq 2\|\Delta\|_{\mathrm{F}}^2$$

thanks to the symmetry of the matrix $\Delta^\top F_{\pi^*}$ (cf. Lemma 1). Since $\|\Delta\|_{2,\infty} \leq 3\sqrt{\frac{\mu r}{d}}$ and $\|\Delta\|_{\mathrm{F}} \leq \frac{3}{5}$, we can use the inequality (57a) from Lemma 5 to get

$$p^{-1}\|\Pi_\Omega(\Delta \otimes \Delta)\|_{\mathrm{F}}^2 \leq (1+\epsilon)\|\Delta\|_{\mathrm{F}}^4 + \epsilon\|\Delta\|_{\mathrm{F}}^2 \leq \frac{9}{25}(1+4\epsilon)\|\Delta\|_{\mathrm{F}}^2. \tag{58}$$

With the constant ϵ sufficiently small, we get that $2\|\Pi_\Omega(\Delta \otimes \Delta)\|_{\mathrm{F}} \leq \|\Pi_\Omega(F^* \otimes \Delta + \Delta \otimes F^*)\|_{\mathrm{F}}$ and

$$T_1 \geq \|\Delta\|_{\mathrm{F}}^2\big(\sqrt{2(1-\epsilon)} - \frac{6}{5}\sqrt{1+4\epsilon}\big)\big(\sqrt{2(1-\epsilon)} - \frac{3}{5}\sqrt{1+4\epsilon}\big) \geq \frac{4}{25}\|\Delta\|_{\mathrm{F}}^2. \tag{59a}$$

On the other hand, we have

$$|T_2| \leq \frac{1}{p}\|\Pi_\Omega(E)\|_{\mathrm{op}} \cdot \sqrt{r}\|M - M^* + \Delta \otimes \Delta\|_{\mathrm{F}} \leq \frac{4\sqrt{r}}{p}\|\Pi_\Omega(E)\|_{\mathrm{op}} \cdot \|\Delta\|_{\mathrm{F}}. \tag{59b}$$

Combining inequalities (59a) and (59b) with our original decomposition yields

$$\langle\!\langle \nabla\mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta\rangle\!\rangle \geq \frac{4}{25}\|\Delta\|_{\mathrm{F}}^2 - \frac{4\sqrt{r}}{p}\|\Pi_\Omega(E)\|_{\mathrm{op}} \cdot \|\Delta\|_{\mathrm{F}},$$

showing that the local descent (49) for $\mathcal{L}_n$ holds with $\alpha = \frac{2}{25}$ and $\varepsilon_n = \frac{100\sqrt{r}}{p}\|\Pi_\Omega(E)\|_{\mathrm{op}}$.

Local Lipschitz and smoothness: Observe that $\alpha \leq L = \beta$ and $\max\{\rho, \varepsilon_n\} \leq 1$, $\|F - F'\|_{2,\infty} \leq 6\sqrt{\frac{\mu r}{d}}$ for all $F, F' \in \mathcal{F}$, and

$$\|\nabla\mathcal{L}_n(F)F\|_{\mathrm{F}} = \sup_{G \in \mathbb{R}^{d\times r}, \|G\|_{\mathrm{F}} \leq 1} \langle\!\langle \nabla\mathcal{L}_n(F \otimes F),\ G \otimes F\rangle\!\rangle.$$

Using these facts, it follows that the Lipschitz and smoothness conditions (51)–(52b) for $\mathcal{L}_n$ can be verified by showing that

$$\left|\frac{\langle\!\langle \Pi_\Omega(F \otimes F - F' \otimes F'),\ G \otimes H\rangle\!\rangle}{p}\right| + \left|\frac{\langle\!\langle \Pi_\Omega(E),\ G \otimes H\rangle\!\rangle}{p}\right| \leq \frac{1}{8}\big(\beta\|F - F'\|_{\mathrm{F}} + \alpha\varepsilon_n\|H\|_{\mathrm{op}}\big)\|G\|_{\mathrm{F}}, \tag{60}$$

valid for all $F, F' \in \mathcal{F} \cap \mathbb{B}_2(1; F^*)$, and for all matrices $H, G \in \mathbb{R}^{d\times r}$ such that $\|H\|_{2,\infty} \leq 6\sqrt{\frac{\mu r}{d}}$.

Let us now verify the bound (60). For an arbitrary $F^* \in \mathcal{E}(M^*)$, define the matrices $\Lambda = F' - F$, $\Delta_1 = F - F^*$ and $\Delta_2 = F' - F^*$, and observe that

$$\begin{aligned}\frac{\|\Pi_\Omega(F \otimes F - F' \otimes F')\|_{\mathrm{F}}}{\sqrt{p}} &= \frac{\|\Pi_\Omega(F^* \otimes \Lambda + \Lambda \otimes F^* + \Lambda \otimes \Delta_2 + \Delta_1 \otimes \Lambda)\|_{\mathrm{F}}}{\sqrt{p}} \\ &\leq \underbrace{\frac{\|\Pi_\Omega(F^* \otimes \Lambda)\|_{\mathrm{F}}}{\sqrt{p}}}_{T_1} + \underbrace{\frac{\|\Pi_\Omega(\Lambda \otimes \Delta_2)\|_{\mathrm{F}}}{\sqrt{p}}}_{T_2} + \underbrace{\frac{\|\Pi_\Omega(\Delta_1 \otimes \Lambda)\|_{\mathrm{F}}}{\sqrt{p}}}_{T_3}.\end{aligned}$$

Lemma 4 implies that $T_1 \leq (1+\epsilon)\|F^* \otimes \Lambda + \Lambda \otimes F^*\|_{\mathrm{F}} \leq 2(1+\epsilon)\|\Lambda\|_{\mathrm{F}}$, whereas inequality (57c) from Lemma 5 ensures that $\max\{T_2, T_3\} \leq 6\sqrt{2\mu r}\|\Lambda\|_{\mathrm{F}}$. Combining these bounds yields

$$\frac{\|\Pi_\Omega(F\otimes F - F'\otimes F')\|_{\mathrm{F}}}{\sqrt{p}} \leq 14\sqrt{2\mu r}\|F - F'\|_{\mathrm{F}}. \tag{61}$$

On the other hand, using the inequality (57b) from Lemma 5, we have

$$\begin{aligned}|p^{-1}\langle\langle \Pi_\Omega(F\otimes F - F\otimes F'),\, G\otimes H\rangle\rangle| &\leq p^{-1/2}\cdot\|p^{-1/2}\Pi_\Omega(F\otimes F - F'\otimes F')H\|_{\mathrm{F}}\|G\|_{\mathrm{F}} \\ &\leq 6\sqrt{\frac{2\mu r}{p}}\|\Pi_\Omega(F\otimes F - F'\otimes F')\|_{\mathrm{F}}\|G\|_{\mathrm{F}}.\end{aligned}$$

Combining with the earlier inequality (61) yields

$$|p^{-1}\langle\langle \Pi_\Omega(F\otimes F - F\otimes F'),\, G\otimes H\rangle\rangle| \leq 168\mu r\|F - F'\|_{\mathrm{F}}\|G\|_{\mathrm{F}}.$$

Finally, observe that

$$|p^{-1}\langle\langle \Pi_\Omega(E),\, G\otimes H\rangle\rangle| \leq \frac{1}{p}\|\Pi_\Omega(E)\|_{\mathrm{op}}\|G\otimes H\|_{\mathrm{nuc}} \leq \frac{\sqrt{r}}{p}\|\Pi_\Omega(E)\|_{\mathrm{op}}\|G\|_{\mathrm{F}}\|H\|_{\mathrm{op}}.$$

Combining the last two inequalities establishes the claim (60), thereby completing the proof of Corollary 1.

6.3 Proof of Corollary 3

We now prove our claims for the sparse PCA model. Define the sampling noise matrix $W := \widehat{\Sigma}_n - \Sigma$, corresponding to the deviation between the sample and population covariance matrices. Recall that $\Sigma = \gamma\left(F^*\otimes F^*\right) + I_d$ with $\|F^*\|_{\mathrm{op}} = 1$. Under the spiked covariance model (9), we have $\nabla\mathcal{L}_n(\Theta) = -\widehat{\Sigma}_n = -(\Sigma + W)$. Let R index the non-zero rows of F^*; observe that the choice of R does not depend on the choice of F^* in $\mathcal{E}(M^*)$.

In light of Remark 1, we have $\mathcal{E}(M^*) \subseteq \mathcal{F}$, which guarantees the M^*-faithfulness condition. It remains to verify the local descent, Lipschitz and smoothness conditions.

Local descent: For a given matrix F, let F_{π^*} be its projection onto $\mathcal{E}(M^*)$, and deifne $\Delta = F - F_{\pi^*}$. Since $F_{\pi^*}\otimes F_{\pi^*} = F^*\otimes F^*$, we have

$$\langle\langle \nabla\mathcal{L}_n(M),\, M - M^* + \Delta\otimes\Delta\rangle\rangle = \underbrace{-\langle\langle \Sigma,\, M - M^* + \Delta\otimes\Delta\rangle\rangle}_{T_1}\underbrace{-\langle\langle W,\, M - M^* + \Delta\otimes\Delta\rangle\rangle}_{T_2}.$$

The remainder of the proof consists of lower bounding T_1 and T_2.

Beginning with T_1, observe that

$$\begin{aligned}T_1 &= -\langle\langle \gamma F_{\pi^*}\otimes F_{\pi^*} + I,\, F_{\pi^*}\otimes\Delta + \Delta\otimes F_{\pi^*} + 2\Delta\otimes\Delta\rangle\rangle \\ &= -2\gamma\big(\langle F_{\pi^*},\, \Delta\rangle + \|\Delta^\top F_{\pi^*}\|_{\mathrm{F}}^2\big) - 2\langle\langle F_{\pi^*},\, \Delta\rangle\rangle - 2\|\Delta\|_{\mathrm{F}}^2.\end{aligned}$$

By Lemma 1, the matrix $F^\top F_{\pi^*}$ is positive semidefinite, and has operator norm bounded as $\|F^\top F_{\pi^*}\|_{\rm op} \leq \|F\|_{\rm op}\|F_{\pi^*}\|_{\rm op} \leq 1$, so that the matrix $-\Delta^\top F_{\pi^*} = I_r - F^\top F_{\pi^*}$ is also positive semidefinite. We therefore have the bound

$$\|\Delta^\top F_{\pi^*}\|_{\rm F}^2 \leq \|\Delta^\top F_{\pi^*}\|_{\rm nuc}\|\Delta^\top F_{\pi^*}\|_{\rm op} = -\operatorname{trace}(\Delta^\top F_{\pi^*}) \cdot \|\Delta^\top F_{\pi^*}\|_{\rm op}.$$

Combined with the bound $\|F\|_{\rm op} \leq 1$ and the orthonormality of F_{π^*}, we find that

$$-\langle F_{\pi^*}, \Delta\rangle = -\langle F_{\pi^*}, F\rangle + \|F_{\pi^*}\|_{\rm F}^2 \geq -\langle F_{\pi^*}, F\rangle + \frac{1}{2}\|F_{\pi^*}\|_{\rm F}^2 + \frac{1}{2}\|F\|_{\rm F}^2 = \frac{1}{2}\|\Delta\|_{\rm F}^2.$$

It follows that

$$T_1 \geq \gamma\|\Delta\|_{\rm F}^2\big(1 - \|\Delta^\top F_{\pi^*}\|_{\rm op}\big) + \|\Delta\|_{\rm F}^2 - 2\|\Delta\|_{\rm F}^2 \geq \|\Delta\|_{\rm F}^2\big(\gamma\tau^2 - 1\big),$$

where in the last inequality we use $\|\Delta^\top F_{\pi^*}\|_{\rm op} \leq \|\Delta\|_{\rm op} \leq 1 - \tau^2$. Combined with the assumption $\gamma \geq \frac{2}{\tau^2}$, it thus follows that $T_1 \geq \frac{\gamma\tau^2}{2}\|\Delta\|_{\rm F}^2$.

In order to bound T_2, we require control on how the matrix W behaves when acting on matrices in the set $\mathbb{C}(k) := \{U \in \mathbb{R}^{d\times r} \mid \|U\|_{2,1} \leq \sqrt{k}\|U\|_{\rm F}\}$.

Lemma 6. *There is a universal constant $c > 0$ such that*

$$|\langle\!\langle W, U \otimes V\rangle\!\rangle| \leq c\ \max\Big\{\sqrt{\frac{k\log d}{n}}, \frac{k\log d}{n}\Big\}(\gamma+1)\|U\|_F\|V\|_F, \quad \text{for all } U, V \in \mathbb{C}(k) \tag{62}$$

with probability at least $1 - 2d^{-4}$.

See Appendix D for the proof of this lemma; it is based on variants of techniques from Lemma 12 of Loh and Wainwright [46].

Now observe that the row sparsity of F^* implies $F^* \in \mathbb{C}(k)$, $\forall F^* \in \mathcal{E}(M^*)$. Recall that R is the row support set of F^*, with R^c denoting its complement. Since $F \in \mathcal{F}$, we are guaranteed that $\|F\|_{2,1} \leq \|F^*\|_{2,1}$, which implies the cone inequality $\|\Delta_{R^c}\|_{2,1} \leq \|\Delta_R\|_{2,1}$. By assumption $|R| \leq k$, whence $\|\Delta\|_{2,1} \leq \sqrt{k}\|\Delta\|_{\rm F}$. It follows that $\Delta \in \mathbb{C}(k)$. Applying Lemma 6, we find that with probability at least $1 - 8d^{-4}$,

$$\begin{aligned}|T_2| \ &= \ \big|2\langle\!\langle W,\ F_{\pi^*} \otimes \Delta + 2\Delta \otimes \Delta\rangle\!\rangle\big| \leq 2|(F_{\pi^*})^\top W\Delta| + 2|\Delta^\top W\Delta| \\ &\leq 4c\max\Big\{\sqrt{\frac{k\log d}{n}}, \frac{k\log d}{n}\Big\}(\gamma+1)\sqrt{r}\|\Delta\|_{\rm F}.\end{aligned}$$

Combining the bounds for T_1 and T_2 proves that the local descent condition (49) for $\mathcal{L}_n$ is satisfied.

Local Lipschitz Let us verify the relaxed Lipschitz condition (50). Observe that for all matrices $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$, $F' \in \mathcal{F}$ and $F^* \in \mathcal{E}(M^*)$, we have $F = F^* + (F - F^*)$ and

$$F - F' = (F - F^*) - (F' - F^*).$$

Following the argument above one can show that the three matrices $F^*, F - F^*, F' - F^*$ all belong to the set $\mathbb{C}(k)$. Consequently, Lemma 6 guarantees that with probability at least $1 - 8d^{-4}$,

$$\begin{aligned}|\langle\!\langle W,\ (F - F') \otimes F\rangle\!\rangle| &\leq 2c\max\Big\{\sqrt{\frac{k\log d}{n}}, \frac{k\log d}{n}\Big\}(\gamma+1) \cdot \big(\|F^*\|_{\rm F} + \|F - F^*\|_{\rm F}\big)\big(\|F - F^*\|_{\rm F} + \|F' - F^*\|_{\rm F}\big) \\ &\leq \underbrace{12c\max\Big\{\sqrt{\frac{k\log d}{n}}, \frac{k\log d}{n}\Big\}(\gamma+1) \cdot \sqrt{r}}_{\frac{3}{4}\alpha\varepsilon_n},\end{aligned}$$

where in the last inequality we use $\|F' - F^*\|_{\mathrm{F}} \leq \sqrt{r}(\|F'\|_{\mathrm{op}} + \|F^*\|_{\mathrm{op}}) \leq 2\sqrt{r}$. It follows that

$$\begin{aligned}
|\langle\langle \nabla \mathcal{L}_n(F),\ (F - F') \otimes F \rangle\rangle| &\leq |\langle\langle \Sigma,\ (F - F') \otimes F \rangle\rangle| + |\langle\langle W,\ (F - F') \otimes F \rangle\rangle| \\
&\leq \|\Sigma\|_{\mathrm{op}} \sqrt{r} \|F - F'\|_{\mathrm{F}} \|F\|_{\mathrm{op}} + \frac{3}{4}\alpha\varepsilon_n \\
&\leq 2(\gamma+1)\sqrt{r}\|F - F'\|_{\mathrm{F}} + \frac{3}{4}\alpha\varepsilon_n \\
&\leq 2(\gamma+1)\sqrt{r}(1 + \|F - F'\|_{\mathrm{F}}),
\end{aligned}$$

where the last inequality follows from $\varepsilon_n \leq 1$ and $\alpha \leq 2(\gamma+1)\sqrt{r}$. Thus, we have established the relaxed Lipschitz condition (50) for $\mathcal{L}_n$.

Local smoothness: Since $\nabla\mathcal{L}_n(F \otimes F) - \nabla\mathcal{L}_n(F' \otimes F') = 0$, the first smoothness condition (52a) for $\mathcal{L}_n$is satisfied trivially. On the other hand, we have

$$\begin{aligned}
|\langle\langle \nabla \mathcal{L}_n(M),\ (F - F^*) \otimes (F' - F'') \rangle\rangle| &\leq |\langle\langle \Sigma,\ (F - F^*) \otimes (F' - F'') \rangle\rangle| + |\langle\langle W,\ (F - F^*) \otimes (F' - F'') \rangle\rangle| \\
&\leq (\gamma+1)\sqrt{r}\|F - F^*\|_{\mathrm{F}}\|F' - F''\|_{\mathrm{F}} + |\langle\langle W,\ (F - F^*) \otimes (F' - F'') \rangle\rangle|.
\end{aligned}$$

Following the same argument above, we can show that $F - F^*, F' - F^*, F'' - F^* \in \mathbb{C}(k)$, whence Lemma 6 guarantees that

$$\begin{aligned}
|\langle\langle W,\ (F - F^*) \otimes (F' - F'') \rangle\rangle| \leq& c(\gamma+1)\max\Big\{\sqrt{\frac{k\log d}{n}}, \frac{k\log d}{n}\Big\} \cdot \|F - F^*\|_{\mathrm{F}}\big(\|F' - F^*\|_{\mathrm{F}} + \|F'' - F^*\|_{\mathrm{F}}\big) \\
\overset{(i)}{\leq}& 2c(\gamma+1)\sqrt{r}\max\Big\{\sqrt{\frac{k\log d}{n}}, \frac{k\log d}{n}\Big\} \cdot \|F - F^*\|_{\mathrm{F}} \\
=& \frac{1}{8}\alpha\varepsilon_n,
\end{aligned}$$

with probability at least $1 - 8d^{-4}$. Here step (i) follows from the inequality $\max\{\|F' - F^*\|_{\mathrm{F}}, \|F'' - F^*\|_{\mathrm{F}}\} \leq \sqrt{r}$, valid for any pair $F', F'' \in \mathbb{B}_2(\rho; F^*)$. We conclude that

$$|\langle\langle \nabla \mathcal{L}_n(M),\ (F - F^*) \otimes (F' - F'') \rangle\rangle| \leq (\gamma+1)\sqrt{r}\|F - F^*\|_{\mathrm{F}}\|F' - F''\|_{\mathrm{F}} + \frac{1}{8}\alpha\varepsilon_n\|F - F^*\|_{\mathrm{F}},$$

thereby establishing the second smoothness condition (52b) for $\mathcal{L}_n$.

6.4 Proof of Corollary 4

We now prove our claims for the planted densest subgraph model. Since $\mathcal{E}(M^*)$ is a two-element set, it follows that for any vector $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$, the projection $F_{\pi^*} := \arg\min_{F^* \in \mathcal{E}(M^*)} \|F^* - F\|_{\mathrm{F}}$ is always equal to the cluster membership vector. The M^*-invariance of the set $\mathcal{F}$ thus follows. Under the planted densest subgraph model, the expectation of the shifted adjacency matrix S has the expression

$$\bar{S} := \mathbb{E}[S] = \frac{p-q}{2}\Big\{2F^* \otimes F^* - \mathbf{1} \otimes \mathbf{1}\Big\},$$

where $\mathbf{1} \in \mathbb{R}^d$ denotes a vector of all ones. The noise matrix $W := S - \bar{S}$ has i.i.d. zero mean entries with variance bounded by p. The gradient of $\mathcal{L}_n$ is given by $\nabla\mathcal{L}_n(F \otimes F) = -2SF$. Below we verify the local descent, Lipschitz and smoothness conditions.

local descent: We have the decomposition

$$\langle\!\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta\otimes\Delta \rangle\!\rangle = \underbrace{-2\langle\!\langle \bar{S},\ F_{\pi^*} \otimes \Delta + \Delta \otimes \Delta \rangle\!\rangle}_{T_1} \quad \underbrace{-2\langle\!\langle W,\ F_{\pi^*} \otimes \Delta + \Delta \otimes \Delta \rangle\!\rangle}_{T_2}.$$

We proceed by lower bounding T_1 and upper bounding $|T_2|$.

Beginning with the term T_1, for any feasible deviation Δ, the two matrices $-\bar{S}$ and $F^*\Delta^\top$ have the same sign on each entry, whence

$$-2\langle\!\langle \bar{S},\ F^* \otimes \Delta \rangle\!\rangle = (p-q)\|F^* \otimes \Delta\|_1 \geq (p-q)k\|\Delta\|_1.$$

On the other hand, the bounds $|\!|\!|\Delta|\!|\!|_{\mathrm{F}} \leq \frac{2}{5}|\!|\!|F^*|\!|\!|_{\mathrm{op}} = \frac{2\sqrt{k}}{5}$ and $\|F\|_1 \leq k = \|F^*\|_1$ imply that $\|\Delta\|_1 \leq 2\sqrt{k}|\!|\!|\Delta|\!|\!|_{\mathrm{F}} \leq \frac{4}{5}k$, from which it follows that

$$|2\langle\!\langle \bar{S},\ \Delta\otimes\Delta \rangle\!\rangle| \leq 2\|\bar{S}\|_\infty \|\Delta\otimes\Delta\|_1 = (p-q)\|\Delta\|_1^2 \leq \frac{4}{5}(p-q)k\|\Delta\|_1.$$

Putting together the piecesr, we obtain the lower bound $T_1 \geq \frac{1}{5}(p-q)k\|\Delta\|_1$.

Now turning to term T_2, by Bernstein's inequality and Proposition 1, there is a universal constant $c_0 > 0$ such that

$$\|WF^*\|_\infty \leq c_0\sqrt{pk\log d} \qquad \text{and} \qquad |\!|\!|W|\!|\!|_{\mathrm{op}} \leq c_0\sqrt{pd + pk\log d},$$

with probability at least $1 - d^{-3}$. On this event, the term T_2 can be bounded as

$$\begin{aligned} |T_2| &\leq 2\big(\|WF^*\|_\infty\|\Delta\|_1 + |\!|\!|W|\!|\!|_{\mathrm{op}}|\!|\!|\Delta|\!|\!|_{\mathrm{F}}^2\big) \\ &\leq 2c_0\big(\sqrt{pk\log d}\|\Delta\|_1 + \sqrt{pd+pk\log d}|\!|\!|\Delta|\!|\!|_{\mathrm{F}}^2\big) \\ &\overset{(ii)}{\leq} \frac{1}{10}k(p-q)\|\Delta\|_1, \end{aligned}$$

where the step (ii) follows from the clustering condition (27), as well as the upper bound $|\!|\!|\Delta|\!|\!|_{\mathrm{F}}^2 \leq \|d\|_\infty\|\Delta\|_1 \leq \|d\|_1$, using the fact that $\|\Delta\|_\infty \leq 1$. Combining the bounds for T_1 and T_2, we conclude that

$$\langle\!\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta\otimes\Delta \rangle\!\rangle \geq \frac{1}{10}k(p-q)\|\Delta\|_1 \geq \frac{1}{10}k(p-q)|\!|\!|\Delta|\!|\!|_{\mathrm{F}}^2,$$

thereby establishing the local descent condition (49) for $\mathcal{L}_n$.

Local Lipschitz and smoothness: Since $\nabla\mathcal{L}_n(M) - \nabla\mathcal{L}_n(M') = 0$, the first smoothness condition (52a) is satisfied trivially. It remains to verify the second smoothness condition (52b) and the relaxed Lipschitz condition (50). For each $F, F', F'' \in \mathcal{F}$ and $G \in \mathbb{R}^{d\times r}$, observe that

$$|\langle\!\langle \nabla\mathcal{L}_n(M),\ (F-G)\otimes(F'-F'') \rangle\!\rangle| \leq \underbrace{|\langle\!\langle \bar{S},\ (F-G)\otimes(F'-F'') \rangle\!\rangle|}_{T_1} + \underbrace{|\langle\!\langle W,\ (F-G)\otimes(F'-F'') \rangle\!\rangle|}_{T_2}.$$

Note that the matrices $F', F'' \in \mathcal{F}$ satisfy the constraint $\sum_i F' = \sum_i F'' = k$, which implies that $(\mathbf{1}\otimes\mathbf{1})(F' - F'') = 0$. It follows that T_1 can be upper bounded as

$$\begin{aligned} T_1 &= \frac{p-q}{2}\left|\langle\!\langle 2F^*\otimes F^* - \mathbf{1}\otimes\mathbf{1},\ (F-G)\otimes(F'-F'')\rangle\!\rangle\right| \\ &= (p-q)\left|\langle\!\langle F^*\otimes F^*,\ (F-G)\otimes(F'-F'')\rangle\!\rangle\right| \\ &\leq (p-q)|\!|\!|F^*|\!|\!|_{\mathrm{F}}^2 |\!|\!|F-G|\!|\!|_{\mathrm{F}} |\!|\!|F'-F''|\!|\!|_{\mathrm{F}} \\ &= 2(p-q)k|\!|\!|F-G|\!|\!|_{\mathrm{F}} |\!|\!|F'-F''|\!|\!|_{\mathrm{F}}. \end{aligned}$$

Similarly, the second term can be upper bounded as

$$\begin{aligned} T_2 &\leq |\!|\!|W|\!|\!|_{\mathrm{op}} |\!|\!|F-G|\!|\!|_{\mathrm{F}} |\!|\!|F'-F''|\!|\!|_{\mathrm{F}} \\ &\overset{(i)}{\leq} c_0\sqrt{pd+pk\log d}\,|\!|\!|F-G|\!|\!|_{\mathrm{F}} |\!|\!|F'-F''|\!|\!|_{\mathrm{F}} \\ &\overset{(ii)}{\leq} \frac{1}{64}k(p-q)|\!|\!|F-G|\!|\!|_{\mathrm{F}} |\!|\!|F'-F''|\!|\!|_{\mathrm{F}}, \end{aligned}$$

where inequality (i) holds with probability at least $1 - d^{-3}$ as proved above, and inequality (ii) holds under the clustering condition (27). Combining the bounds for T_1 and T_2 with the choice $\beta = 12(p-q)k$, we conclude that

$$|\langle\!\langle \nabla\mathcal{L}_n(M),\ (F-G)\otimes(F'-F'')\rangle\!\rangle| \leq \frac{\beta}{4}|\!|\!|F-G|\!|\!|_{\mathrm{F}} |\!|\!|F'-F''|\!|\!|_{\mathrm{F}}, \qquad \forall F, F', F'' \in \mathcal{F},\ G \in \mathbb{R}^{d\times r}.$$

For an arbitrary $F^* \in \mathcal{E}(M^*)$, setting $G = F^*$ in this inequality establishes the smoothness condition (52b). On the other hand, setting $G = 0$ and noting that $|\!|\!|F|\!|\!|_{\mathrm{F}}^2 \leq \|F\|_1\|F\|_\infty \leq k = |\!|\!|F^*|\!|\!|_{\mathrm{op}}^2$, we obtain the relaxed Lipschitz condition (50) for $\mathcal{L}_n$.

6.5 Proof of Corollary 5

We now prove our claims for the one-bit matrix completion model. By assumption, the initial matrix F^0 belongs to the set $\mathbb{B}_2(\frac{1}{5}; F^*) \cap \mathcal{F}$, where the set $\mathcal{F}$ is was previously involved in our analysis of ordinary matrix completion (see Section 6.2). Therefore, following the argument in Remark 1, we can show that $\|F\|_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}}$, $\|F - F^*\|_{2,\infty} \leq 3\sqrt{\frac{\mu r}{d}}$ for all $F^* \in \mathcal{E}(M^*)$ and $F \in \mathbb{B}_2(\epsilon; F^*) \cap \mathcal{F}$, and that $\mathcal{F}$ is M^*-faithful. Consequently, we have for all relevant matrices M^* and M, we are guaranteed that

$$\max\left\{\|M^*/\sigma\|_\infty,\ \|M/\sigma\|_\infty\right\} \leq 4\ \underbrace{\frac{\mu r}{d\sigma}}_{\nu}.$$

Now define a (random function) $H : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ with entries $[H(x)]_{ij} := \frac{f'(x)(-Y_{ij}+2f(x)-1)}{f(x)(1-f(x))}$. With this notation, we have

$$\nabla\mathcal{L}_n(M) = \frac{1}{\sigma}\Pi_\Omega\left[H(M/\sigma)\right]. \tag{63}$$

For future reference, we claim that each component $[H(\cdot)]_{ij}$ is bounded and $4L_{4\nu}$-Lipschitz over $[-4\nu, 4\nu]$. This property follows because f satisfies the bounds in equation (29) and $|Y_{ij}| \leq 1$, so that $|[H(x)]_{ij}| \leq 2L_{4\nu}$ surely. Moreover, the derivative can be bounded as

$$|[H'(x)]_{ij}| = \left| \frac{f''(x)(Y_{ij} - 2f(x) + 1) - 2f'(x)^2}{f(x)(1-f(x))} - \frac{f'(x)(Y_{ij} - 2f(x) + 1)}{f^2(x)(1-f(x))^2} f'(x)(1 - 2f(x)) \right| \leq 4L_{4\nu}$$

which certifies the Lipschitz property.

With this set-up, we are now prepared to establish the local descent, Lipschitz and smoothness conditions for $\mathcal{L}_n$.

Local descent: Let us introduce the shorthand $\bar{G}(M) := \mathbb{E}[\nabla \mathcal{L}_n(M)] = \nabla \mathbb{E}[\mathcal{L}_n(M)]$. We begin by splitting the gradient into two terms, corresponding to the expectation and the zero-mean deviation, thereby obtaining

$$\langle\!\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle = \underbrace{\langle\!\langle \bar{G}(M),\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle}_{T_1} + \underbrace{\langle\!\langle \nabla \mathcal{L}_n(M) - \bar{G}(M),\ M - M^* + \Delta \otimes \Delta \rangle\!\rangle}_{T_2}.$$

Controlling T_1: For the expectation term T_1, we first note that $\mathbb{E}[\mathcal{L}_n(M)]$ is convex in M and has the form of expected negative log likelihood, whence

$$\begin{aligned} \langle\!\langle \bar{G}(M),\ M - M^* \rangle\!\rangle &\geq \mathbb{E}[\mathcal{L}_n(M)] - \mathbb{E}[\mathcal{L}_n(M^*)] \\ &= p\, D\big(f(M^*/\sigma) \| f(M/\sigma)\big) \\ &\geq p\, d_H^2\big(f(M^*/\sigma) \| f(M/\sigma)\big), \end{aligned}$$

where $D(\cdot)$ and $d_H^2(\cdot)$ denote the KL and Hellinger distances, respectively. To proceed, we use a known lower bound (Lemma 2 in the paper [24]) on the Hellinger distance: for matrices $X, X' \in \mathbb{R}^{d \times d}$ such that $\|X\|_\infty, \|X'\|_\infty \leq a$, we have

$$d_H^2\big(f(X), f(X')\big) \geq \frac{\|X - X'\|_{\mathrm{F}}^2}{8\ell_a}. \tag{64}$$

Since $\|M/\sigma\|_\infty, \|M^*/\sigma\|_\infty \leq 4\nu$, applying the lower bound (64) with $a = 4\nu$ yields the lower bound $\langle\!\langle \bar{G}(M),\ M - M^* \rangle\!\rangle \geq \frac{p}{8\sigma^2 \ell_{4\nu}} \|M - M^*\|_{\mathrm{F}}^2$. Furthermore, since F^* is orthonormal, $\|\Delta\|_{\mathrm{op}} \leq \frac{1}{16}$ and $\Delta^\top F_{\pi^*}$ is symmetric, we have

$$\langle\!\langle \bar{G}(M),\ M - M^* \rangle\!\rangle \geq \frac{p}{16\sigma^2 \ell_{4\nu}} \|\Delta\|_{\mathrm{F}}^2,$$

On the other hand, using the expression (63) for the gradient and the assumptions in equation (29), we find that

$$\|\bar{G}(M)\|_{\mathrm{F}} = \frac{p}{\sigma} \cdot \Big\| \frac{f'(M/\sigma) \circ (f(M/\sigma) - f(M^*/\sigma))}{f(M/\sigma) \circ (1 - f(M/\sigma))} \Big\|_{\mathrm{F}} \leq \frac{p\sqrt{L_{4\nu}}}{\sigma} \cdot \|f(M/\sigma) - f(M^*/\sigma)\|_{\mathrm{F}}.$$

Inequality (29) combined with the bound $\sup_{z\in\mathbb{R}} |f(z)(1-f(z))| \leq 1$ ensures that f is $\sqrt{L_{4\nu}}$-Lipschitz over the interval $[-4\nu, 4\nu]$. It follows that

$$\|\!|\overline{G}(M)|\!\|_{\mathrm{F}} \leq \frac{pL_{4\nu}}{\sigma^2}\|\!|M - M^*|\!\|_{\mathrm{F}} \leq \frac{3pL_{4\nu}}{\sigma^2}\|\!|\Delta|\!\|_{\mathrm{F}}, \tag{65}$$

where the last step uses the upper bound $\|\!|\Delta|\!\|_{\mathrm{op}} \leq 1$. Combining these bounds, we obtain the following lower bound on the expectation:

$$\begin{aligned} T_1 = \langle\!\langle \overline{G}(M),\ M - M^* + \Delta\otimes\Delta \rangle\!\rangle &\geq \langle\!\langle \overline{G}(M),\ M - M^* \rangle\!\rangle - \|\!|\overline{G}(M)|\!\|_{\mathrm{F}}\|\!|\Delta|\!\|_{\mathrm{F}}^2 \\ &\geq \frac{p}{16\sigma^2\ell_{4\nu}}\|\!|\Delta|\!\|_{\mathrm{F}}^2 - \frac{3pL_{4\nu}}{\sigma^2}\|\!|\Delta|\!\|_{\mathrm{op}} \cdot \|\!|\Delta|\!\|_{\mathrm{F}}^2 \\ &\geq \frac{p}{32\sigma^2\ell_{4\nu}}\|\!|\Delta|\!\|_{\mathrm{F}}^2 \end{aligned}$$

where the last step uses the bound $\|\!|\Delta|\!\|_{\mathrm{op}} \leq \frac{1}{96\ell_{4\nu}L_{4\nu}}$.

Controlling T_2: We now turn to analysis of the deviation term T_2. Using the symmetry of the matrix $\nabla\mathcal{L}_n(M) - \overline{G}(M)$, we may rewrite it as

$$T_2 = 2\langle \nabla\mathcal{L}_n(M) - \overline{G}(M),\, \Delta \otimes F\rangle$$

We control this quantity via two auxiliary lemmas. For each $B \in (0,1)$, define the annular set

$$\begin{aligned} \Gamma(B) &:= \{\Delta \mid \frac{B}{2} < \|\!|\Delta|\!\|_{\mathrm{F}} \leq B, \|\Delta\|_{2,\infty}^2 \leq 4\nu\sigma\}, \qquad \text{and the event} \\ \mathcal{E}(B) &:= \Big\{ \sup_{\Delta\in\Gamma(B)} \frac{|2\langle \nabla\mathcal{L}_n(F\otimes F) - \overline{G}(F\otimes F),\, \Delta\otimes F\rangle|}{\|\!|\Delta|\!\|_{\mathrm{F}}} > \frac{1}{4}\alpha\varepsilon_n \Big\}. \end{aligned}$$

Our first lemma controls the probability of this "bad event":

Lemma 7. *For any $B \in (0,1)$, we have $\mathbb{P}[\mathcal{E}(B)] \leq 2d^{-14}$.*

Our next lemma gives controls over the small ball Γ_0 around the origin. In particular, we define the set

$$\Gamma_0 := \{\Delta \mid \|\!|\Delta|\!\|_{\mathrm{F}} \leq 2^{-d^2}, \|\Delta\|_{2,\infty}^2 \leq 4\nu\sigma\}, \text{ and } \mathcal{E}_0 := \Big\{ \sup_{\Delta\in\Gamma_0} \frac{|2\langle \nabla\mathcal{L}_n(F\otimes F) - \overline{G}(F\otimes F),\, \Delta\otimes F\rangle|}{\|\!|\Delta|\!\|_{\mathrm{F}}} > \frac{1}{4}\alpha\varepsilon_n \Big\}.$$

Lemma 8. *We have $\mathbb{P}[\mathcal{E}_0] \leq d^{-12}$.*

See Appendices E.1 and E.2 for the proofs of these two auxiliary results.

Taking the lemmas as given, the union bound then guarantees that

$$\begin{aligned} \mathbb{P}\left[\sup_{\|\!|\Delta|\!\|_{\mathrm{F}}\leq 1, \|\Delta\|_{2,\infty}^2 \leq 4\nu\sigma} \frac{|2\langle \nabla\mathcal{L}_n(F\otimes F) - \overline{G}(F\otimes F),\, \Delta\otimes F\rangle|}{\|\!|\Delta|\!\|_{\mathrm{F}}} > \frac{1}{4}\alpha\varepsilon_n \right] &\leq \mathbb{P}[\mathcal{E}_0] + \sum_{i=0}^{d^2} \mathbb{P}[\mathcal{E}(2^{-i})] \\ &\leq d^2 \cdot 2d^{-14} + d^{-12} = 3d^{-12}, \end{aligned}$$

which implies that $|T_2| \leq \frac{1}{4}\alpha\varepsilon_n \|\Delta\|_{\mathrm{F}}$ with probability at least $1 - 3d^{-12}$, Conditioned on this event, we can combine the bounds for T_1 and T_2 to conclude that

$$\langle\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta \rangle\rangle \geq T_1 - |T_2| \geq 2\alpha \|\Delta\|_{\mathrm{F}}^2 - \frac{1}{4}\alpha\varepsilon_n \|\Delta\|_{\mathrm{F}},$$

valid for all matrices Δ such that $\|\Delta\|_{\mathrm{F}} \leq \rho = \max\{\frac{1}{16}, \frac{1}{96\ell_{4\nu} L_{4\nu}}\}$, thereby establishing the local descent condition (49) for $\mathcal{L}_n$.

Local Lipschitz condition (51) **and smoothness condition** (52b)**:** We begin by making note of the bounds

$$\|F' - F''\|_{2,\infty} \leq 4\sqrt{\frac{\mu r}{d}}, \quad \text{and} \quad \|F' - F''\|_{\mathrm{F}} \leq 2,$$

valid for all matrices $F', F'' \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$. Moreover, we have $\alpha \leq L = \beta$ and $\varepsilon_n \leq 1$. Using these facts, it follows that the local Lipschitz and smoothness conditions (51) and (52b) for $\mathcal{L}_n$ are implied a bound of the form

$$\|\nabla \mathcal{L}_n(M) H\|_{\mathrm{F}} \leq \frac{1}{8}\big(\beta \|F - F^*\|_{\mathrm{F}} + \alpha\varepsilon_n \|H\|_{\mathrm{F}}\big), \tag{66}$$

valid for all matrices $F \in \mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$, and $F^* \in \mathcal{E}(M^*)$, and matrices H such that $\|H\|_{2,\infty} \leq 4\sqrt{\frac{\mu r}{d}}$. Accordingly, the remainder of our effort is devoted to establishing the bound (66). We first condition on the event in Lemma 5, which occurs with probability at least $1 - 2d^{-4}$. We then make note of the decomposition

$$\|\nabla \mathcal{L}_n(M) H\|_{\mathrm{F}} = \frac{1}{\sigma}\|\Pi_\Omega[h(M/\sigma)]H\|_{\mathrm{F}} \ \leq \underbrace{\frac{2}{\sigma}\|\Pi_\Omega\big[h(M/\sigma) - h(M^*/\sigma)\big]H\|_{\mathrm{F}}}_{T_1} + \underbrace{\frac{2}{\sigma}\|\Pi_\Omega[h(M^*/\sigma)]H\|_{\mathrm{F}}}_{T_2},$$

and we bound each of these two terms separately.
Bounding term T_1: We have

$$\begin{aligned} T_1 &= \frac{2}{\sigma} \sup_{G \in \mathbb{R}^{d\times r}: \|G\|_{\mathrm{F}}=1} |\langle\langle \Pi_\Omega\big[H(M/\sigma) - H(M^*/\sigma)\big],\ G \otimes H \rangle\rangle| \\ &\leq \frac{2}{\sigma}\|\Pi_\Omega\big(H(M/\sigma) - H(M^*/\sigma)\big)\|_{\mathrm{F}} \cdot \sup_{G \in \mathbb{R}^{d\times r}: \|G\|_{\mathrm{F}}=1} \|\Pi_\Omega(G \otimes H)\|_{\mathrm{F}}. \end{aligned}$$

Recall that each component $[H(\cdot)]_{ij}$ is almostly surely $4L_{4\nu}$-Lipschitz over the interval $[-4\nu, 4\nu]$, and that $\|M/\sigma\|_\infty, \|M^*/\sigma\|_\infty \leq 4\nu$. Combining these facts yields the bound

$$\begin{aligned} \|\Pi_\Omega\big(H(M/\sigma) - H(M^*/\sigma)\big)\|_{\mathrm{F}} &\leq \frac{4L_{4\nu}}{\sigma}\|\Pi_\Omega\big(M - M^*\big)\|_{\mathrm{F}} \\ &\leq \frac{4L_{4\nu}}{\sigma}\Big(2\|\Pi_\Omega\big(F^* \otimes (F - F^*)\big)\|_{\mathrm{F}} + \|\Pi_\Omega\big((F - F^*) \otimes (F - F^*)\big)\|_{\mathrm{F}}\Big) \\ &\leq \frac{72L_{4\nu}\sqrt{2p\mu r}}{\sigma}\|\Delta\|_{\mathrm{F}}, \end{aligned}$$

where the last inequality follows from the inequality (57c) in Lemma 5 combined with the fact that $\max\big\{\|F^*\|_{2,\infty}, \|F - F^*\|_{2,\infty}\big\} \leq 4\sqrt{\frac{\mu r}{d}}$. Applying inequality (57c) a second time yields

$$\sup_{G \in \mathbb{R}^{d \times r} : \|G\|_{\mathrm{F}} = 1} \|\Pi_\Omega(G \otimes H)\|_{\mathrm{F}} \leq 6\sqrt{2p\mu r} \sup_{G \in \mathbb{R}^{d \times r} : \|G\|_{\mathrm{F}} = 1} \|G\|_{\mathrm{F}} = 6\sqrt{2p\mu r}.$$

Putting together the pieces yields the upper bound

$$T_1 \leq \frac{48 L_{4\nu} p\mu r}{\sigma^2} \|F - F^*\|_{\mathrm{F}} \leq \frac{\beta}{8} \|F - F^*\|_{\mathrm{F}}. \tag{67a}$$

Bounding term T_2: Turning to the term T_2, we observe that conditioned on Y, the matrix $H(M^*/\sigma)$ is a deterministic quantity with entries bounded uniformly as $|[H(M^*/\sigma)]_{oj}| \leq L_{4\nu}$ for all indices $i, j \in [d]$. By applying Lemma 11 and integrating out the conditioning, we find that the second term is bounded as

$$T_2 \leq \frac{2}{\sigma} \|\Pi_\Omega[h(M^*/\sigma)]\|_{\mathrm{op}} \cdot \sqrt{r} \|H\|_{\mathrm{F}} \leq \frac{4\sqrt{prd} L_{4\nu}}{\sigma} \cdot \|H\|_{\mathrm{F}} \leq \frac{1}{8} \alpha \varepsilon_n \|H\|_{\mathrm{F}}, \tag{67b}$$

where these bounds hold with probability at least $1 - d^{-4}$.

Finally, combining our bounds for T_1 and T_2 in equations (67a) and (67b) yields the desired bound (66).

Local smoothness condition (52a): We begin by conditioning on the event in Lemma 5, which holds with probability at least $1 - 2d^{-4}$. For each pair of matrices F', F'' in the set $\mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$ and any matrix $F^* \in \mathcal{E}(M^*)$, we then have

$$\begin{aligned}
|\langle\!\langle \nabla \mathcal{L}_n(M) - \nabla \mathcal{L}_n(M'),\ F' \otimes (F - F^*) \rangle\!\rangle| &= \frac{1}{\sigma} |\langle \Pi_\Omega\big(H(M/\sigma) - h(M'/\sigma)\big),\ F' \otimes (F - F^*) \rangle| \\
&\leq \frac{1}{\sigma} \|\Pi_\Omega\big(H(M/\sigma) - h(M'/\sigma)\big)\|_{\mathrm{F}} \|\Pi_\Omega(F' \otimes (F - F^*))\|_{\mathrm{F}} \\
&\leq \frac{4L_{4\nu}}{\sigma^2} \|\Pi_\Omega\big(M - M'\big)\|_{\mathrm{F}} \|\Pi_\Omega(F' \otimes (F - F^*))\|_{\mathrm{F}},
\end{aligned}$$

where the last inequality follows from the fact that the function H is element-wise $4L_{4\nu}$-Lipschitz. Recall equation (61) from Section 6.2, which ensures that

$$\|\Pi_\Omega\big(M - M'\big)\|_{\mathrm{F}} \leq 14\sqrt{2p\mu r} \|F - F'\|_{\mathrm{F}}.$$

Combined with inequality (57c) from Lemma 5, we find that $\|\Pi_\Omega(F' \otimes (F - F^*))\|_{\mathrm{F}} \leq 6\sqrt{2p\mu r} \|F - F^*\|_{\mathrm{F}}$, from which it follows that

$$\begin{aligned}
|\langle\!\langle \nabla \mathcal{L}_n(M) - \nabla \mathcal{L}_n(M'),\ F' \otimes (F - F^*) \rangle\!\rangle| &\leq \frac{672 L_{4\nu}}{\sigma^2} \cdot p\mu r \|F - F'\|_{\mathrm{F}} \|F - F^*\|_{\mathrm{F}} \\
&\leq \frac{1}{4} \beta \|F - F'\|_{\mathrm{F}} \|F - F^*\|_{\mathrm{F}},
\end{aligned}$$

thereby establishing the local smoothness condition (52a).

6.6 Proof of Corollary 6

We now prove our claims for the matrix decomposition problem. By dividing through $|||F^*|||_{\text{op}}$, we may assume without loss of generality that $|||F^*|||_{\text{op}} = 1$. The set $\mathcal{F}$ and the values of ρ and $\mathrm{d}(F^0, F^*)$ are the same as used in the proof of matrix completion in Section 6.2, so we make use of the results therein. In particular, we showed there that the set $\mathcal{F}$ is M^*-faithful.

Given the observation matrix $Y = M^* + S^* + E$, the gradient takes the form

$$\nabla \mathcal{L}_n(M) = (M - M^*) + (S(M) - S^*) - E,$$

where $S(M) := \Pi_{\mathcal{S}}(Y - M)$. Below we verify the local descent, Lipschitz and smoothness conditions.

local descent: Expanding $\nabla \mathcal{L}_n(M)$, the quantity $\langle\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta \rangle\rangle$ can be decomposed into the sum

$$\underbrace{|||M - M^*|||_{\text{F}}^2}_{T_1} + \underbrace{\langle\langle M - M^*,\ \Delta \otimes \Delta \rangle\rangle}_{T_2} + \underbrace{\langle\langle S(M) - S^*,\ M - M^* + \Delta \otimes \Delta \rangle\rangle}_{T_3} + \underbrace{\langle\langle -E,\ M - M^* + \Delta \otimes \Delta \rangle\rangle}_{T_4}.$$

Note that $M - M^* = F_{\pi^*} \otimes \Delta + \Delta \otimes F_{\pi^*} + \Delta \otimes \Delta$. By Lemma 1, the matrix $\Delta^\top F_{\pi^*}$ is symmetric, so expanding the Frobenius norm shows that $T_1 \geq 2|||\Delta|||_{\text{F}}^2$. Since $|||\Delta|||_{\text{op}} \leq \frac{3}{5}$, we have

$$|T_2| \leq 2|||\Delta|||_{\text{op}}|||\Delta|||_{\text{F}}^2 + |||\Delta|||_{\text{op}}^2|||\Delta|||_{\text{F}}^2 \leq \frac{39}{25}|||\Delta|||_{\text{F}}^2.$$

With $\Delta_S := S(M) - S^*$ and e_j being the j-th standard basis, we find that

$$\begin{aligned} |T_3| = |2\langle\langle \Delta_S,\ \Delta \otimes F \rangle\rangle| \leq & 2|||F^\top \Delta_S|||_{\text{F}}|||\Delta|||_{\text{F}} \\ = & 2\sqrt{\sum_{j=1}^d \|F^\top \Delta_S e_j\|_2^2}|||\Delta|||_{\text{F}} \leq 2\sqrt{\sum_{j=1}^d |||F|||_{2,\infty}^2 \|\Delta_S e_j\|_1^2}|||\Delta|||_{\text{F}}, \end{aligned}$$

where we use the symmetry of Δ_S in the first equality. Inequality (55a) ensures that $|||F|||_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}}$. Moreover, for each $S \in \mathcal{S}$, we have the inequalities $\|\Delta_S e_j\|_1 \leq 2\sqrt{k}\|\Delta_S e_j\|_2, j \in [d]$ thanks to the row-wise ℓ_1 constraints and the k-sparsity of the columns of S^*. It follows that

$$|T_3| \leq 4\sqrt{\frac{\mu r k}{d}}\sqrt{\sum_{j=1}^d \|\Delta_S e_j\|_2^2}|||\Delta|||_{\text{F}} = 4\sqrt{\frac{\mu r k}{d}}|||\Delta_S|||_{\text{F}}|||\Delta|||_{\text{F}}.$$

Under the assumption $\frac{\mu r k}{d} \leq c_1$ of the corollary, we obtain $|T_3| \leq \frac{2}{25}|||\Delta_S|||_{\text{F}}|||\Delta|||_{\text{F}}$. But $S^* \in \mathcal{S}$ and the projection $\Pi_{\mathcal{S}}$ is non-expansive, whence

$$|||\Delta_S|||_{\text{F}} = |||\Pi_{\mathcal{S}}(Y - M) - S^*|||_{\text{F}} \leq |||(Y - M) - S^*|||_{\text{F}} = |||M - M^*|||_{\text{F}} \leq 3|||\Delta|||_{\text{F}}, \tag{68}$$

and so we have shown that $|T_3| \leq \frac{6}{25}|||\Delta|||_{\text{F}}^2$. Finally, we have

$$|T_4| \leq |||E|||_{\text{F}}|||M - M^* + \Delta \otimes \Delta|||_{\text{F}} \leq \frac{16}{5}|||E|||_{\text{F}}|||\Delta|||_{\text{F}}.$$

Putting together the bounds for T_1, T_2 and T_3 and T_4, we conclude that

$$\langle\langle \nabla \mathcal{L}_n(M),\ M - M^* + \Delta \otimes \Delta \rangle\rangle \geq \frac{1}{5}|||\Delta|||_{\text{F}}^2 - \frac{16}{5}|||E|||_{\text{F}}|||\Delta|||_{\text{F}},$$

thereby proving the local descent condition (49) for $\mathcal{L}_n$.

Local Lipschitz condition: Using the inequality (68) above and the assumption $\|E\|_{\mathrm{F}} \leq \frac{2}{5}$, we have

$$\begin{aligned}\|\nabla \mathcal{L}_n(M)F\|_{\mathrm{F}} &= \|(M - M^* + S(M) - S^* - E)F\|_{\mathrm{F}} \\ &\leq \big(\|M - M^*\|_{\mathrm{F}} + \|\Delta_S\|_{\mathrm{F}} + \|E\|_{\mathrm{F}}\big)\|F\|_{\mathrm{op}} \\ &\leq (6\|\Delta\|_{\mathrm{F}} + 1)\|F\|_{\mathrm{op}} \\ &\leq 8,\end{aligned}$$

where the last inequality follows from $\|\Delta\|_{\mathrm{op}} \leq \|\Delta\|_{\mathrm{F}} \leq \frac{3}{5}$. Therefore, $\mathcal{L}_n$ satisfies the local Lipschitz condition in (51).

Local smoothness: Observe that

$$\begin{aligned}|\langle\!\langle \nabla \mathcal{L}_n(M) - \nabla \mathcal{L}_n(M'),\ F' \otimes (F - F^*)\rangle\!\rangle| &= |\langle\!\langle M - M' + S(M) - S(M'),\ F' \otimes (F - F^*)\rangle\!\rangle| \\ &\leq \big(\|M - M'\|_{\mathrm{F}} + \|S(M) - S(M')\|_{\mathrm{F}}\big)\|F - F^*\|_{\mathrm{F}}\|F'\|_{\mathrm{op}}.\end{aligned}$$

The non-expansiveness of the projection $\Pi_{\mathcal{S}}$ ensures that

$$\|S(M) - S(M')\|_{\mathrm{F}} = \|\Pi_{\mathcal{S}}(Y - M) - \Pi_{\mathcal{S}}(Y - M')\|_{\mathrm{F}} \leq \|M - M'\|_{\mathrm{F}} \leq \frac{16}{5}\|F - F'\|_{\mathrm{F}},.$$

where we use $F, F' \in \mathbb{B}_2(\frac{3}{5}; F^*)$. It follows that

$$|\langle\!\langle \nabla \mathcal{L}_n(M) - \nabla \mathcal{L}_n(M'),\ F' \otimes (F - F^*)\rangle\!\rangle| \leq 12\|F - F'\|_{\mathrm{F}}\|F - F^*\|_{\mathrm{F}},$$

proving the first smoothness condition (52a).

Similarly, combining inequality (68) with the bound $\|E\|_{\mathrm{F}} \leq \frac{2}{5}$ implies that

$$\begin{aligned}|\langle\!\langle \nabla \mathcal{L}_n(M),\ (F - F^*) \otimes (F' - F'')\rangle\!\rangle| &= |\langle\!\langle M - M^* + S(M) - S^* - E,\ (F - F^*) \otimes (F' - F'')\rangle\!\rangle| \\ &\leq \big(\|M - M^*\|_{\mathrm{F}} + \|S(M) - S^*\|_{\mathrm{F}} + \|E\|_{\mathrm{F}}\big)\|F - F^*\|_{\mathrm{F}}\|F' - F''\|_{\mathrm{F}} \\ &\leq 7\|F - F^*\|_{\mathrm{F}}\|F' - F''\|_{\mathrm{F}},\end{aligned}$$

thereby verifying the second smoothness condition (52b).

7 Discussion

In this paper, we have laid out a general framework for analyzing the behavior of projected gradient descent for solving low-rank optimization problems in the factorized space. We have illustrated the consequences of our general theory for a number of concrete models, including matrix regression, structured PCA, matrix completion, matrix decomposition and graph clustering.

Acknowledgements

This work was partially supported by ONR-MURI grant DOD-002888, AFOSR grant FA9550-14-1-0016, NSF grant CIF-31712-23800, and ONR MURI grant N00014-11-1-0688.

A Proof of Theorem 3

Recall that in Section 5.2 we proved Theorem 2 under the assumption that $\widetilde{\mathcal{L}}_n$ satisfies the *relaxed* local Lipschitz condition (40) as well as the local descent condition (14) and smoothness conditions (16). We establish Theorem 3 by showing that these conditions for $\widetilde{\mathcal{L}}_n$ are implied by the corresponding conditions for $\mathcal{L}_n$ in Definitions 5–7 with the same parameters α, β, L, ε_n and ρ with $\rho < \sigma_r(F^*)$.

Let F be an arbitrary matrix in $\mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$ and F^* an arbitrary member in $\mathcal{E}(M^*)$. Since $\rho < \sigma_r(F^*)$, Lemma 1 guarantees that $F_{\pi^*} = \arg\min_{A \in \mathcal{E}(M^*)} \|A - F\|_{\mathrm{F}}$ is uniquely defined. We use the shorthand $G := \nabla_M \mathcal{L}_n$, $\widetilde{G} := \nabla_F \widetilde{\mathcal{L}}_n$, $\Delta_{\pi^*} := F - F_{\pi^*}$ and $\Delta = F - F^*$.

Local descent condition: Observe that

$$\begin{aligned}
\langle\!\langle \widetilde{G}(F),\ F - F^* \rangle\!\rangle =& \langle\!\langle G(F \otimes F),\ \Delta \otimes F + F \otimes \Delta \rangle\!\rangle \\
=& \langle\!\langle G(F \otimes F),\ F \otimes F - F^* \otimes F^* \rangle\!\rangle + \langle\!\langle G(F \otimes F),\ \Delta \otimes \Delta \rangle\!\rangle \\
=& \langle\!\langle G(F \otimes F),\ F \otimes F - F_{\pi^*} \otimes F_{\pi^*} \rangle\!\rangle + \langle\!\langle G(F \otimes F),\ \Delta_{\pi^*} \otimes \Delta_{\pi^*} \rangle\!\rangle \\
&+ \langle\!\langle G(F \otimes F),\ (F_{\pi^*} - F^*) \otimes \Delta \rangle\!\rangle + \langle\!\langle G(F \otimes F),\ \Delta_{\pi^*} \otimes (F_{\pi^*} - F^*) \rangle\!\rangle,
\end{aligned}$$

where the last step follows from $F^* \otimes F^* = F_{\pi^*} \otimes F_{\pi^*}$ and $\Delta = \Delta_{\pi^*} + F_{\pi^*} - F^*$. We then apply the local descent condition (49) for $\mathcal{L}_n$ to the first two terms above, and the local smoothness condition (52b) for $\mathcal{L}_n$ to the last two terms. Doing so yields

$$\begin{aligned}
\langle\!\langle \widetilde{G}(F),\ F - F^* \rangle\!\rangle \geq& 2\alpha \|\Delta_{\pi^*}\|_{\mathrm{F}}^2 - \frac{\beta}{4} \|F_{\pi^*} - F^*\|_{\mathrm{F}} \big(\|\Delta\|_{\mathrm{F}} + \|\Delta_{\pi^*}\|_{\mathrm{F}} \big) - \frac{\alpha}{2} \varepsilon_n \|\Delta_{\pi^*}\|_{\mathrm{F}} - \frac{\alpha}{2} \varepsilon_n \|\Delta\|_{\mathrm{F}} \\
\geq& 2\alpha \|\Delta_{\pi^*}\|_{\mathrm{F}}^2 - \frac{\beta}{2} \|F_{\pi^*} - F^*\|_{\mathrm{F}} \|\Delta_{\pi^*}\|_{\mathrm{F}} - \frac{\beta}{2} \|F_{\pi^*} - F^*\|_{\mathrm{F}}^2 - \alpha \varepsilon_n \|\Delta_{\pi^*}\|_{\mathrm{F}} - \frac{\alpha}{2} \varepsilon_n \|F_{\pi^*} - F^*\|_{\mathrm{F}} \\
\geq& 2\alpha \|\Delta_{\pi^*}\|_{\mathrm{F}}^2 - \Big(\frac{1}{2} \alpha \|\Delta_{\pi^*}\|_{\mathrm{F}}^2 + \frac{\beta^2}{4\alpha} \|F_{\pi^*} - F^*\|_{\mathrm{F}}^2 \Big) - \frac{\beta}{2} \|F_{\pi^*} - F^*\|_{\mathrm{F}}^2 \\
&- \Big(\frac{1}{2} \alpha \|\Delta_{\pi^*}\|_{\mathrm{F}}^2 + \frac{1}{2} \alpha \varepsilon_n^2 \Big) - \Big(\frac{1}{2} \alpha \varepsilon_n^2 + \frac{\alpha}{4} \|F_{\pi^*} - F^*\|_{\mathrm{F}}^2 \Big) \\
\geq& \alpha \|\Delta_{\pi^*}\|_{\mathrm{F}}^2 - \frac{\beta^2}{\alpha} \|F_{\pi^*} - F^*\|_{\mathrm{F}} - \alpha \varepsilon_n^2,
\end{aligned}$$

where the last three steps follow from $\|\Delta\|_{\mathrm{F}} \leq \|\Delta_{\pi^*}\|_{\mathrm{F}} + \|F_{\pi^*} - F^*\|_{\mathrm{F}}$, the AM-GM inequality and the upper bound $\alpha \leq \beta$, respectively. This proves the local descent condition for $\widetilde{\mathcal{L}}_n$ in (14).

Relaxed local Lipschitz condition: Let F' be an arbitrary matrix in $\mathcal{F}$. With $G(F \otimes F)$ symmetric and $\mathcal{L}_n$ satisfying the relaxed Lipschitz condition in (50), we have

$$|\langle\!\langle \widetilde{G}(F),\ F - F' \rangle\!\rangle| = |\langle\!\langle 2G(F \otimes F),\ (F - F') \otimes F \rangle\!\rangle| \leq L \big(\|F^*\|_{\mathrm{op}}^2 + \|F^*\|_{\mathrm{F}} \|F - F'\|_{\mathrm{F}} \big),$$

which proves the relaxed local Lipschitz condition for $\widetilde{\mathcal{L}}_n$ in (40).

Local smoothness condition: Let F' be an arbitrary matrix in $\mathcal{F} \cap \mathbb{B}_2(\rho; F^*)$. The smoothness conditions (52a) and (52b) yields that

$$\begin{aligned}|\langle\!\langle \widetilde{G}(F) - \widetilde{G}(F'),\ F - F^* \rangle\!\rangle| &= |2\langle\!\langle G(F \otimes F)F - G(F' \otimes F')F',\ F - F^* \rangle\!\rangle| \\ &= |2\langle\!\langle G(F \otimes F)F' - G(F' \otimes F')F',\ F - F^* \rangle\!\rangle + 2\langle\!\langle G(F \otimes F)(F - F'),\ F - F^* \rangle\!\rangle| \\ &\leq \beta |\!|\!| F - F' |\!|\!|_{\mathrm{F}} |\!|\!| F - F^* |\!|\!|_{\mathrm{F}} + \alpha \varepsilon_n |\!|\!| F - F^* |\!|\!|_{\mathrm{F}},\end{aligned}$$

which establishes the smoothness condition for $\widetilde{\mathcal{L}}_n$ in (16).

B Technical lemmas for Corollary 2

In this appendix, we prove the technical lemmas involved in the proof of Corollary 2 on the matrix sensing model.

B.1 Proof of Lemma 3

By the bilinearity of the inner product, we may assume without loss of generality that $|\!|\!| A |\!|\!|_{\mathrm{F}} = |\!|\!| B |\!|\!|_{\mathrm{F}} = 1$. Since the matrices $A \pm B$ have rank at most $4r$, the RIP with δ_{4r} ensures that

$$(1 - \delta_{4r}) |\!|\!| A \pm B |\!|\!|_{\mathrm{F}}^2 \leq \frac{1}{n} |\!|\!| \mathfrak{X}_n(A \pm B) |\!|\!|_{\mathrm{F}}^2 \leq (1 + \delta_{4r}) |\!|\!| A \pm B |\!|\!|_{\mathrm{F}}^2.$$

It follows that

$$\begin{aligned}\frac{1}{n} \langle\!\langle \mathfrak{X}_n(A),\ \mathfrak{X}_n(B) \rangle\!\rangle &= \frac{1}{4n} \Big(|\!|\!| \mathfrak{X}_n(A + B) |\!|\!|_{\mathrm{F}}^2 - |\!|\!| \mathfrak{X}_n(A - B) |\!|\!|_{\mathrm{F}}^2 \Big) \\ &\leq \frac{1}{4} \Big((1 + \delta_{4r}) |\!|\!| A + B |\!|\!|_{\mathrm{F}}^2 - (1 - \delta_{4r}) |\!|\!| A - B |\!|\!|_{\mathrm{F}}^2 \Big) \\ &= \langle\!\langle A,\ B \rangle\!\rangle + \frac{1}{2} \delta_{4r} \big(|\!|\!| A |\!|\!|_{\mathrm{F}}^2 + |\!|\!| B |\!|\!|_{\mathrm{F}}^2 \big) \\ &= \langle\!\langle A,\ B \rangle\!\rangle + \delta_{4r} |\!|\!| A |\!|\!|_{\mathrm{F}} |\!|\!| B |\!|\!|_{\mathrm{F}}.\end{aligned}$$

It follows from a similar argument that $\frac{1}{n} \langle\!\langle \mathfrak{X}_n(A),\ \mathfrak{X}_n(B) \rangle\!\rangle \geq \langle\!\langle A,\ B \rangle\!\rangle - \delta_{4r} |\!|\!| A |\!|\!|_{\mathrm{F}} |\!|\!| B |\!|\!|_{\mathrm{F}}$.

C Technical lemmas for Corollary 1

In this appendix, we prove the technical lemmas involved in the proof of Corollary 1 on the matrix completion model.

C.1 Proof of Lemma 4

Define a subspace $\mathcal{T} \subseteq \mathbb{R}^{d \times d}$ of d-dimensional matrices as follows

$$\mathcal{T} := \Big\{ X \mid X = (F^* \otimes U) + (V \otimes F^*) \quad \text{for some } U, V \in \mathbb{R}^{d \times r} \Big\},$$

and let $\Pi_{\mathcal{T}}$ be the Euclidean projection onto $\mathcal{T}$. Since F^* is 4μ-incoherent, a known result in exact matrix completion [18] guarantees that as long as $p \geq c\frac{\mu r \log d}{\epsilon^2 d}$ for a suficiently large universal constant c, then

$$\|\!|(\Pi_{\mathcal{T}}\Pi_{\Omega}\Pi_{\mathcal{T}} - p\Pi_{\mathcal{T}})X|\!\|_{\mathrm{F}} \leq \epsilon p \|\!|X|\!\|_{\mathrm{F}}, \quad \text{for all } X \in \mathcal{T}$$

with probability at least $1-2d^{-3}$. Noting that the matrices $F^* \otimes H \pm G \otimes F^*$ belong to the subspace $\mathcal{T}$, we can apply the above inequality to obtain

$$(1-\epsilon)p\|\!|F^* \otimes H \pm G \otimes F^*|\!\|_{\mathrm{F}} \leq \|\!|\Pi_{\mathcal{T}}\Pi_{\Omega}\Pi_{\mathcal{T}}\big(F^* \otimes H \pm G \otimes F^*\big)|\!\|_{\mathrm{F}} \leq (1+\epsilon)p\|\!|F^* \otimes H \pm G \otimes F^*|\!\|_{\mathrm{F}}.$$

The rest of the proof is similar to that of Lemma 3. In particular, by the bilinearity of the inner product, we may assume $\|\!|F^* \otimes H|\!\|_{\mathrm{F}} = \|\!|G \otimes F^*|\!\|_{\mathrm{F}} = 1$. Using the above inequalities, we find that

$$\begin{aligned}
&\langle\!\langle \Pi_{\Omega}\left(F^* \otimes H\right),\, \Pi_{\Omega}\left(G \otimes F^*\right)\rangle\!\rangle = \langle\!\langle \Pi_{\Omega}\Pi_{\mathcal{T}}(F^* \otimes H),\, \Pi_{\Omega}\Pi_{\mathcal{T}}(G \otimes F^*)\rangle\!\rangle \\
=&\frac{1}{4}\left(\|\!|\Pi_{\Omega}\Pi_{\mathcal{T}}(F^* \otimes H + G \otimes F^*)|\!\|_{\mathrm{F}}^2 - \|\!|\Pi_{\Omega}\Pi_{\mathcal{T}}(F^* \otimes H - G \otimes F^*)|\!\|_{\mathrm{F}}^2\right) \\
\leq&\frac{1}{4}\big(\|\!|F^* \otimes H + G \otimes F^*|\!\|_{\mathrm{F}}\|\!|\Pi_{\mathcal{T}}\Pi_{\Omega}\Pi_{\mathcal{T}}(F^* \otimes H + G \otimes F^*)|\!\|_{\mathrm{F}} - \|\!|F^* \otimes H - G \otimes F^*|\!\|_{\mathrm{F}}\|\!|\Pi_{\mathcal{T}}\Pi_{\Omega}\Pi_{\mathcal{T}}(F^* \otimes H + G \otimes F^*)|\!\| \\
\leq&\frac{1}{4}\Big((1+\epsilon)p\|\!|F^* \otimes H + G \otimes F^*|\!\|_{\mathrm{F}}^2 - (1-\epsilon)p\|\!|F^* \otimes H - G \otimes F^*|\!\|_{\mathrm{F}}^2\Big) \\
=&p\langle\!\langle F^* \otimes H,\, G \otimes F^*\rangle\!\rangle + \epsilon p = p\langle\!\langle F^* \otimes H,\, G \otimes F^*\rangle\!\rangle + \epsilon p\|\!|F^* \otimes H|\!\|_{\mathrm{F}}\|\!|G \otimes F^*|\!\|_{\mathrm{F}}.
\end{aligned}$$

This proves the first inequality in the lemma. The second inequality can be proved in the same fashion by noting that the matrices $(F^* \otimes H) \pm (F^* \otimes G)$ also belong to the subspace $\mathcal{T}$.

C.2 Proof of Lemma 5

We need the following result on random graphs [28], which involves some universal constants c_1 and c_2.

Lemma 9. *If $p \geq c_1\frac{\log d}{\epsilon^2 d}$, then with probability at least $1 - \frac{1}{2}d^{-4}$,*

$$\sum_{(i,j)\in\Omega} U_i V_j \leq (1+\epsilon)p\|U\|_1\|V\|_1 + c_2\sqrt{pd}\|U\|_2\|V\|_2, \quad \forall U, V \in \mathbb{R}^d. \tag{69}$$

From the bound (69) and the assumption $p \geq \frac{C}{\epsilon^2}\big(\frac{\mu^2 r^2}{d} + \frac{\log d}{d}\big)$, we find that with probability at least $1 - \frac{1}{2}d^{-4}$,

$$\begin{aligned}
p^{-1}\|\!|\Pi_{\Omega}(H \otimes H)|\!\|_{\mathrm{F}}^2 &\leq p^{-1}\sum_{(i,j)\in\Omega}\|H_{i\cdot}\|_2^2\|H_{j\cdot}\|_2^2 \leq (1+\epsilon)\Big(\sum_i \|H_{i\cdot}\|_2^2\Big)^2 + C_2\sqrt{\frac{d}{p}}\sum_i \|H_{i\cdot}\|_2^4 \\
&= (1+\epsilon)\|\!|H|\!\|_{\mathrm{F}}^4 + C_2\sqrt{\frac{d}{p}}\|\!|H|\!\|_{\mathrm{F}}^2\|H\|_{2,\infty}^2 \\
&\leq (1+\epsilon)\|\!|H|\!\|_{\mathrm{F}}^4 + \epsilon\|\!|H|\!\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last step follows from the bound $\|H\|_{2,\infty} \leq 6\sqrt{\frac{\mu r}{d}}$. We have thus established the first inequality (57a) in the lemma statement.

Let $\Omega_i := \{j \mid (i,j) \in \Omega\}$. When $p \geq C\frac{\log d}{d}$ for C sufficiently large, the event $\max_i |\Omega_i| \leq 2pd$ holds with probability at least $1 - d^{-4}$. On this event, we have

$$\begin{aligned}
p^{-1}\|\Pi_\Omega(Z)H\|_{\mathrm{F}}^2 &= p^{-1}\sum_{i=1}^{d}\sum_{k=1}^{r}\Big(\sum_{j\in\Omega_i}(\Pi_\Omega(Z^\top)e_i)_j \cdot H_{jk}\Big)^2 \\
&\leq p^{-1}\sum_{i=1}^{d}\sum_{k=1}^{r}\|\Pi_\Omega(Z^\top)e_i\|_2^2\sum_{j\in\Omega_i}H_{jk}^2 \\
&= p^{-1}\sum_{i=1}^{d}\|\Pi_\Omega(Z^\top)e_i\|_2^2\sum_{j\in\Omega_i}\|H_{j\cdot}\|_2^2 \\
&\leq p^{-1}\sum_{i=1}^{d}\|\Pi_\Omega(Z^\top)e_i\|_2^2\cdot\big(\max_i|\Omega_i|\big)\|H\|_{2,\infty}^2 \leq \|\Pi_\Omega(Z)\|_{\mathrm{F}}^2\cdot 2d\|H\|_{2,\infty}^2.
\end{aligned}$$

But $\|H\|_{2,\infty} \leq 6\sqrt{\frac{\mu r}{d}}$ by assumption, so the second inequality (57b) in the lemma follows.

To establish the third inequality (57c) in the lemma, observe that conditioned on the event $\{\max_i |\Omega_i| \leq 2pd\}$, we have

$$\begin{aligned}
p^{-1}\|\Pi_\Omega(H\otimes G)\|_{\mathrm{F}}^2 \;=\; p^{-1}\sum_{j=1}^{d}\|\Pi_\Omega(H\otimes G)e_j\|_2^2 &\leq p^{-1}\sum_{j=1}^{d}\|G_{j\cdot}\|_2^2\sum_{\{i\,|\,(i,j)\in\Omega\}}\|H_{i\cdot}\|_2^2 \\
&\leq p^{-1}\sum_{j=1}^{d}\|G_{j\cdot}\|_2^2\big(\max_i|\Omega_i|\big)\|H\|_{2,\infty}^2 \\
&\leq p^{-1}\sum_{i=1}^{d}\|G_{j\cdot}\|_2^2\cdot 2pd\cdot 36\frac{\mu r}{d} \\
&= 72\mu r\|G\|_{\mathrm{F}}^2.
\end{aligned}$$

D Proof of Lemma 6

By rescaling, it suffices to consider matrices U and V with $\|U\|_{\mathrm{F}} = \|V\|_{\mathrm{F}} = 1$ and $U, V \in \mathbb{B}_{2,1}(\sqrt{k})$. We need the following geometric result, which is a simple generalization of Lemma 11 in the paper [46]. For completeness, we provide the proof in Section D.1 to follow.

Lemma 10. *For each integer $1 \leq k \leq d$, we have*

$$\mathbb{B}_{2,1}(\sqrt{k}) \cap \mathbb{B}_F(1) \subseteq 3\,\mathrm{cl}\big\{\mathrm{conv}\{\mathbb{B}_{2,0}(k) \cap \mathbb{B}_F(1)\}\big\}. \tag{70}$$

Based on this this lemma and continuity, it suffices to prove the bound (62) for pairs of matrices $U, V \in \mathrm{conv}\{\mathbb{B}_{2,0}(k) \cap \mathbb{B}_F(3)\}$. Any such pair can be written as a weighted combination of the form $U = \sum_i \alpha_i U_i$ and $V = \sum_j \beta_j V_j$, with weights $\alpha_i, \beta_j \geq 0$ such that $\sum_i \alpha_i = \sum_j \beta_j = 1$, and constituent matrices $U_i, V_j \in \mathbb{B}_{2,0}(k) \cap \mathbb{B}_F(3)$ for each i, j. With this notation, observe that

$$|\langle\!\langle W,\, U\otimes V\rangle\!\rangle| \leq \sum_{i,j}\alpha_i\beta_j|\langle\!\langle W,\, U_i\otimes V_j\rangle\!\rangle| \leq \Big(\sum_{i,j}\alpha_i\beta_j\Big)\max_{i,j}|\langle\!\langle W,\, U_i\otimes V_j\rangle\!\rangle| = \max_{i,j}|\langle\!\langle W,\, U_i\otimes V_j\rangle\!\rangle|.$$

If we use $(U_i)_{\cdot\ell}$ and $(V_j)_{\cdot\ell}$ to denote the ℓ-th column of U_i and V_j, respectively, then

$$|\langle\!\langle W,\, U_i \otimes V_j \rangle\!\rangle| \leq \sum_{\ell=1}^{r} |\langle\!\langle W,\, (U_i)_{\cdot\ell} \otimes (V_j)_{\cdot\ell} \rangle\!\rangle| \leq \Big(\sup_{x,y \in \mathbb{B}_0(k)} \frac{|x^\top W y|}{\|x\|_2 \|y\|_2} \Big) \sum_{\ell=1}^{r} \|(U_i)_{\cdot\ell}\|_2 \|(V_j)_{\cdot\ell}\|_2$$
$$\overset{(i)}{\leq} 9 \Big(\sup_{x,z \in \mathbb{B}_0(k)} \frac{|x^\top W y|}{\|x\|_2 \|y\|_2} \Big) = 9 \Big(\sup_{x,y \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)} |x^\top W y| \Big),$$

where step (i) follows from the Cauchy-Schwarz inequality, and $|\!|\!| U_i |\!|\!|_{\mathrm{F}} = |\!|\!| V_j |\!|\!|_{\mathrm{F}} \leq 3$. It suffices to bound the supremum in the last RHS by $18t$, where

$$t := c' \, (\gamma + 1) \max \Big\{ \sqrt{\frac{k \log d}{n}}, \frac{k \log d}{n} \Big\} \tag{71}$$

for a universal constant c' to be specified later.

To proceed, we make use of a standard concentration result. Recall that $X \in \mathbb{R}^{n \times d}$ is the matrix of independent samples from a d-dimensional Gaussian distribution with zero mean and covariance $\Sigma = \gamma(F^* \otimes F^*) + I_d$. By Lemma 15 in Loh and Wainwright [46], there is a universal constant $c > 0$ such that

$$\mathbb{P}\Big[\sup_{z \in \mathbb{B}_0(2k) \cap \mathbb{B}_2(1)} |\|Xz\|_2^2/n - z^\top \Sigma z| \geq t \Big] \leq 2 \exp\Big(- cn \min\big\{ \frac{t^2}{(\gamma+1)^2}, \frac{t}{\gamma+1} \big\} + 2k \log d \Big).$$

Applying this inequality with $z = \frac{1}{6}(x \pm y)$ and our previously specified (71) of t with $c' = \frac{8}{c}$, we find that with probability $1 - 2d^{-4}$, we have

$$\frac{1}{36} |\frac{1}{n} \|X(x+y)\|_2^2 - (x+y)^\top \Sigma (x+y)| \leq t \quad \text{and} \quad \frac{1}{36} |\frac{1}{n} \|X(x-y)\|_2^2 - (x-y)^\top \Sigma (x-y)| \leq t$$

for all $x, y \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)$. On this event, we have

$$\begin{aligned} \frac{4}{n} x^\top X^\top X y &= \frac{1}{n} \|X(x+y)\|_2^2 - \frac{1}{n} \|X(x-y)\|_2^2 \\ &\leq (x+y)^\top \Sigma (x+y) - (x-y)^\top \Sigma (x-y) + 72t \\ &= 4 x^\top \Sigma y + 72t \end{aligned}$$

and similarly $\frac{4}{n} x^\top X^\top X y \geq 4 x \Sigma y - 72t$, whence $|x^\top W y| = |\frac{1}{n} x^\top X^\top X y - x^\top \Sigma y| \leq 18t$.

D.1 Proof of Lemma 10

Let $A, B \subseteq \mathbb{R}^{d \times r}$ be closed convex sets, with support function given by $\phi_A(U) = \sup_{F \in A} \langle\!\langle F,\, U \rangle\!\rangle$ and ϕ_B similarly defined. It is well-known that $\phi_A(U) \leq \phi_B(U)$ for all $U \in \mathbb{R}^{d \times r}$ if and only if $A \subseteq B$. Accordingly, let us verify the first condition for the sets $A = \mathbb{B}_{2,1}(\sqrt{k}) \cap \mathbb{B}_F(1)$ and $B = 3\mathrm{cl}\big\{\mathrm{conv}\{\mathbb{B}_{2,0}(k) \cap \mathbb{B}_F(1)\}\big\}$.

For any $U \in \mathbb{R}^{d \times r}$, let $S \subseteq \{1, 2, \ldots, d\}$ be the subset that indexes the top $\lfloor k \rfloor$ rows of U in ℓ_2 norm. Then $\|U_{S^c}\|_{2,\infty} \leq \|U_{j\cdot}\|_2$ for all $j \in S$, whence

$$\|U_{S^c}\|_{2,\infty} \leq \frac{1}{\lfloor k \rfloor} \|U_S\|_{2,1} \leq \frac{1}{\sqrt{\lfloor k \rfloor}} |\!|\!| U_S |\!|\!|_{\mathrm{F}}. \tag{72}$$

Therefore, we obtain

$$\begin{aligned}\phi_A(U) = \sup_{F \in A} \langle\!\langle F_S,\, U_S \rangle\!\rangle + \langle\!\langle F_{S^c},\, U_{S^c} \rangle\!\rangle &\leq \sup_{\|\!| F_S |\!\|_{\mathrm{F}} \leq 1} \langle\!\langle F_S,\, U_S \rangle\!\rangle + \sup_{\|F_{S^c}\|_{2,1} \leq \sqrt{k}} \langle\!\langle F_{S^c},\, U_{S^c} \rangle\!\rangle \\ &\leq |\!\| U_S |\!\|_{\mathrm{F}} + \sqrt{k} \|U_{S^c}\|_{2,\infty} \\ &\stackrel{(i)}{\leq} \Big(1 + \sqrt{\frac{k}{\lfloor k \rfloor}}\Big) |\!\| U_S |\!\|_{\mathrm{F}} \leq 3 |\!\| U_S |\!\|_{\mathrm{F}},\end{aligned}$$

where inequality (i) follows from the earlier bound (72). The claim then follows from the observation that $\phi_B(U) = \sup_{F \in B} \langle\!\langle F,\, U \rangle\!\rangle = 3 \max_{|T| = \lfloor k \rfloor} \sup_{|\!| F_T |\!\|_{\mathrm{F}} \leq 1} \langle\!\langle F_T,\, U_T \rangle\!\rangle = 3 |\!\| U_S |\!\|_{\mathrm{F}}$.

E Technical lemmas for Corollary 5

In this appendix, we provide the proofs of the technical lemmas required for Corollary 5 on one-bit matrix completion.

E.1 Proof of Lemma 7

In order to prove the lemma, we need to establish an upper tail bound on the random variable

$$Z := \sup_{\Delta \in \Gamma(B)} \frac{\langle\!\langle \nabla \mathcal{L}_n(F \otimes F) - \bar{G}(F \otimes F),\, \Delta \otimes F \rangle\!\rangle}{|\!\| \Delta |\!\|_{\mathrm{F}}}.$$

Define the indicator variable $Q_{ij} := \mathbb{I}\{(i,j) \in \Omega\}$ for each (i,j). Using the expression (63) for $\nabla \mathcal{L}_n(\cdot)$ and the definition of $\Gamma(B)$, we observe that

$$Z \leq \frac{2}{B\sigma} \sup_{\Delta \in \Gamma(B)} \sum_{i,j} [Q_{ij} h_{ij}(M_{ij}/\sigma) - \mathbb{E}_{Q,Y} Q_{ij} h_{ij}(M_{ij}/\sigma)] \cdot (\Delta \otimes \theta)_{ij}. \tag{73}$$

By the usual symmetrization argument [45], the expectation $\mathbb{E}_{Q,Y}[Z]$ is at most a factor of two times the Rademacher-symmetrized version. This is the supremum of a sub-Gaussian process in terms of Rademacher variables, and so is majorized by the expected supremum of the corresponding Gaussian process [45] up to a universal constant c. In conjunction, these two steps yield the bound

$$\mathbb{E}_{Q,Y} Z \leq c \mathbb{E}_{Q,Y} \mathbb{E}_g \sup_{\Delta \in \Gamma(B)} \{Z(\Delta)\} := c \mathbb{E}_{Q,Y} \mathbb{E}_g \sup_{\Delta \in \Gamma(B)} \left\{ \frac{4}{B\sigma} \sum_{i,j} g_{ij} Q_{ij} h_{ij}(M_{ij}/\sigma) \cdot (\Delta \otimes \theta)_{ij} \right\},$$

where $\{g_{ij}\}$ are independent standard Gaussian variables.

Our next step is to bound $\mathbb{E}_g \sup_{\Delta \in \Gamma(B)} Z(\Delta)$ using the Sudakov-Fernique comparison inequality.

For any $\Delta, \Delta' \in \Gamma(B)$ and $F' := F^* + \Delta'$, $M' := F' \otimes F'$, we have

$$\begin{aligned}
\gamma(\Delta, \Delta') &:= \mathbb{E}_g (Z(\Delta) - Z(\Delta'))^2 \\
&= \frac{16}{(B\sigma)^2} \mathbb{E}_g \left[\sum_{i,j} g_{ij} Q_{ij} \left(h_{ij}(M_{ij}/\sigma) \cdot (\Delta \otimes F)_{ij} - h_{ij}(M'_{ij}/\sigma) \cdot (\Delta' \otimes F')_{ij} \right) \right]^2 \\
&= \frac{16}{(B\sigma)^2} \sum_{i,j} Q_{ij}^2 \left\{ h_{ij}(M_{ij}/\sigma) \cdot (\Delta \otimes F)_{ij} - h_{ij}(M'_{ij}/\sigma) \cdot (\Delta' \otimes F')_{ij} \right\}^2 \\
&\leq \frac{32}{(B\sigma)^2} \sum_{i,j} Q_{ij}^2 \left\{ h_{ij}(\Theta_{ij}/\sigma)^2 \cdot (\Delta \otimes F - \Delta' \otimes F')_{ij}^2 + \left(h_{ij}(\Theta_{ij}/\sigma) - h_{ij}(M'_{ij}/\sigma) \right)^2 \cdot (\Delta' \otimes F')_{ij}^2 \right\}.
\end{aligned}$$

Recall that for each (i, j) and over the interval $[-4\nu, 4\nu]$, the function $h_{ij}(\cdot)$ is surely bounded by $4L_{4\nu}$ and $4L_{4\nu}$-Lipschitz. Moreover, the Cauchy-Schwarz inequality implies that

$$|\Theta_{ij}| = |(F \otimes F)_{ij}| \leq \|F_{i\cdot}\|_2 \|F_{j\cdot}\|_2 \leq 2\sqrt{\frac{\mu r}{d}} \cdot 2\sqrt{\frac{\mu r}{d}} = 4\sigma\nu.$$

Note that the same bound holds for $|M'_{ij}|$, $|(\Delta \otimes F)_{ij}|$ and $|(\Delta' \otimes F')_{ij}|$. It follows that

$$\begin{aligned}
\gamma(\Delta, \Delta') &\leq C^2 \frac{L_{4\nu}^2}{(B\sigma)^2} \sum_{i,j} Q_{ij}^2 \left\{ (\Delta \otimes F - \Delta' \otimes F')_{ij}^2 + (M_{ij}/\sigma - M'_{ij}/\sigma)^2 \cdot 4\sigma^2\nu^2 \right\} \\
&= C^2 \frac{L_{4\nu}^2}{(B\sigma)^2} \sum_{i,j} Q_{ij}^2 \left\{ (\Delta \otimes F - \Delta' \otimes F')_{ij}^2 + \left(2\Delta \otimes F_{ij} + \Delta \otimes \Delta_{ij} - 2\Delta' \otimes F'_{ij} - \Delta' \otimes \Delta'_{ij} \right)^2 4\nu^2 \right\} \\
&\leq C^2 (1+\nu)^2 \frac{L_{4\nu}^2}{(B\sigma)^2} \sum_{i,j} Q_{ij}^2 \left\{ (\Delta \otimes F - \Delta' \otimes F')_{ij}^2 + (\Delta \otimes \Delta - \Delta' \otimes \Delta')_{ij}^2 \right\}.
\end{aligned}$$

We compare $Z(\Delta)$ with an alternative stochastic process given by

$$\bar{Z}(\Delta) := C(1+\nu) \frac{L_{4\nu}}{B\sigma} \sum_{i,j} \left[g_{ij} Q_{ij} (\Delta \otimes \Delta)_{ij} + g'_{ij} Q_{ij} (\Delta \otimes F)_{ij} \right],$$

where $\{g_{ij}, g'_{ij}\}$ are independent standard Gaussian variables. Both $Z(\Delta)$ and $\bar{Z}(\Delta)$ are surely continuous in Δ. Observe that by independence, we have

$$\begin{aligned}
\bar{\gamma}(\Delta, \Delta') &:= \mathbb{E}_{g,g'} (\bar{Z}(\Delta) - \bar{Z}(\Delta'))^2 \\
&= C^2 (1+\nu)^2 \frac{L_{4\nu}^2}{(B\sigma)^2} \sum_{i,j} Q_{ij}^2 \left\{ (\Delta \otimes \Delta - \Delta' \otimes \Delta')_{ij}^2 + (\Delta \otimes F - \Delta' \otimes F')_{ij}^2 \right\},
\end{aligned}$$

so we have $\gamma(\Delta, \Delta') \leq \bar{\gamma}(\Delta, \Delta'), \forall \Delta, \Delta' \in \Gamma(B)$. By the Sudakov-Fernique comparison [45], we find that

$$\begin{aligned}
\mathbb{E}_{Q,Y} \mathbb{E}_g \sup_{\Delta \in \Gamma(B)} Z(\Delta) &\leq \mathbb{E}_{Q,Y} \mathbb{E}_{g,g'} \sup_{\Delta \in \Gamma(B)} \bar{Z}(\Delta) \\
&= C(1+\nu) \frac{L_{4\nu}}{B\sigma} \cdot \mathbb{E}_{Q,g,g'} \sup_{\Delta \in \Gamma(B)} \sum_{i,j} \left[g_{ij} Q_{ij} (\Delta \otimes \Delta)_{ij} + g'_{ij} Q_{ij} (\Delta \otimes F)_{ij} \right] \\
&\leq C(1+\nu) \frac{L_{4\nu}}{B\sigma} \left(\sup_{\Delta \in \Gamma(B)} |\!|\!| \Delta \otimes \Delta |\!|\!|_{\text{nuc}} + \sup_{\Delta \in \Gamma(B)} |\!|\!| \Delta \otimes F |\!|\!|_{\text{nuc}} \right) \mathbb{E}_{Q,g} |\!|\!| g \circ Q |\!|\!|_{\text{op}},
\end{aligned}$$

where the last inequality follows from the generalized Holder's inequality and that g and g' are identically distributed. To proceed, we use Lemma 11 to get that $\mathbb{E}_{Q,g}\|g \circ Q\|_{\mathrm{op}} \leq c(\sqrt{pd} + \log d) \leq 2c\sqrt{pd}$, where the last inequality follows from the assumption $p \geq \frac{\log^2 d}{d}$. Moreover, for each $\Delta \in \Gamma(B)$, the matrices $\Delta \otimes \Delta$ and $\Delta \otimes F$ have rank at most r, so

$$\max\{\|\Delta \otimes \Delta\|_{\mathrm{nuc}}, \|\Delta \otimes F\|_{\mathrm{nuc}}\} \leq \sqrt{r} \cdot \|\Delta\|_{\mathrm{F}} \max\{\|\Delta\|_{\mathrm{op}}, \|F\|_{\mathrm{op}}\} \leq 2\sqrt{r}B.$$

Putting together the pieces yields

$$\mathbb{E}_{Q,Y} Z \leq \mathbb{E}_{Q,Y}\mathbb{E}_g \sup_{\Delta \in \Gamma(B)} Z(\Delta) \leq \frac{4C'}{\sigma} L_{4\nu}(1+\nu)\sqrt{pdr} \leq \frac{1}{8}\alpha\varepsilon_n.$$

In order establish concentration of Z around $\mathbb{E}_{Q,Y}Z$, we use a standard functional Hoeffding inequality [44]. In particular, letting $\{X_i\}_{i=1}^n$ be independent random variables such that X_i takes values in $\mathcal{X}_i$, consider a random variable of the form $Y := \sup_{g \in \mathcal{G}} \sum_{i=1}^n g(X_i)$ where for each $g \in \mathcal{G}$, we have $\sup_{x \in \mathcal{X}_i} |g(x)| \leq b_i$. Then we are guaranteed that

$$\mathbb{P}\left[Y \geq \mathbb{E}[Y] + \tau\right] \leq e^{-\tau^2/16D^2} \qquad \text{for all } \tau \geq 0, \text{ and } D^2 := \textstyle\sum_{i=1}^n b_i^2. \tag{74}$$

Setting $\tau = \frac{1}{8}\alpha\varepsilon_n$, we have

$$\begin{aligned}\sup_{\Delta \in \Gamma(B)} \sum_{i,j} \frac{4[\nabla \mathcal{L}_n(M) - \bar{G}(M)]_{ij}^2 (\Delta \otimes F)_{ij}^2}{\|\Delta\|_{\mathrm{F}}^2} &\leq C \frac{L_{4\nu}^2}{B^2\sigma^2} \sup_{\Delta \in \Gamma(B)} \|\Delta \otimes F\|_{\mathrm{F}}^2 \\ &\leq C \frac{L_{4\nu}^2}{\sigma^2 B^2} \cdot \sup_{\Delta \in \Gamma(B)} \|\Delta\|_{\mathrm{F}}^2 \|F\|_{\mathrm{op}}^2 \leq \frac{\alpha^2\varepsilon_n^2}{128 \times 14 \log d} \to D^2.\end{aligned}$$

Consequently, applying the bound (74) with these choices of (τ, D^2), [5] we obtain $\mathbb{P}[Z \geq \mathbb{E}_{Q,Y}[Z] + \frac{1}{8}\alpha\varepsilon_n] \leq d^{-14}$. Combining with the the expectation bound $\mathbb{E}_{Q,Y}Z \leq \frac{1}{8}\alpha\varepsilon_n$, we find that

$$\mathbb{P}\Big[Z = \sup_{\Delta \in \Gamma(B)} \langle\!\langle \nabla \mathcal{L}_n(F \otimes F) - \bar{G}(F \otimes F),\ \Delta \rangle\!\rangle / \|\Delta\|_{\mathrm{F}} \geq \frac{1}{4}\alpha\varepsilon_n\Big] \leq d^{-14}.$$

Following the same lines of argument we obtain a similar bound on the lower tail:

$$\mathbb{P}\Big[\inf_{\Delta \in \Gamma(B)} \langle\!\langle \nabla \mathcal{L}_n(F \otimes F) - \bar{G}(F \otimes F),\ \Delta \rangle\!\rangle / \|\Delta\|_{\mathrm{F}} \leq -\frac{1}{4}\alpha\varepsilon_n\Big] \leq d^{-14}.$$

The proof the lemma is completed by applying the union bound.

[5] In particular, we apply it with the following setup: $X_{ij} = (e_i \otimes e_j, Q_{ij}, Y_{ij})$ and $\mathcal{X}_{ij} = \{e_i \otimes e_j\} \times \{0,1\} \times \{-1,1\}$, where e_i is the i-th standard basis vector in $\mathbb{R}^d$; $\mathcal{F} = \{\zeta_\Delta : \Delta \in \Gamma(B)\}$ with

$$\zeta_\Delta(X_{ij}) = \frac{2}{\sigma\|\Delta\|_{\mathrm{F}}} \langle\!\langle \frac{f'(M/\sigma)\,(Y_{ij}(e_i \otimes e_j) - 2f(M/\sigma) + 1)}{f(M/\sigma)\,(1 - f(M/\sigma))} - p\frac{f'(M/\sigma)\,(2f(M^*/\sigma) - 2f(M/\sigma))}{f(M/\sigma)\,(1 - f(M/\sigma))},\ Q_{ij}(e_i \otimes e_j) \circ (\Delta \otimes F) \rangle\!\rangle.$$

E.2 Proof of Lemma 8

Using the Cauchy-Schwarz inequality and the expression (63) for the gradient $\nabla \mathcal{L}_n$, we have

$$\sup_{\Delta \in \Gamma_0} \frac{|\langle\!\langle \nabla \mathcal{L}_n(F \otimes F) - \overline{G}(F \otimes F),\, \Delta \otimes F \rangle\!\rangle|}{|\!|\!| \Delta |\!|\!|_{\mathrm{F}}} \leq \frac{1}{\sigma} \sup_{\Delta \in \Gamma_0} |\!|\!| \left[\Pi_\Omega h(M/\sigma) - \mathbb{E}\Pi_\Omega h(M/\sigma)\right] F |\!|\!|_{\mathrm{F}} \leq \sum_{j=1}^{3} T_3$$

where $T_1 := \frac{2}{\sigma} \sup_{\Delta \in \Gamma_0} |\!|\!| \left[\Pi_\Omega h(M/\sigma) - \Pi_\Omega h(M^*/\sigma)\right] F |\!|\!|_{\mathrm{F}}$, and

$$T_2 := \frac{2}{\sigma} \sup_{\Delta \in \Gamma_0} |\!|\!| \left[\Pi_\Omega h(M^*/\sigma)\right] F |\!|\!|_{\mathrm{F}}, \quad \text{and} \quad T_3 := \frac{2}{\sigma} \sup_{\Delta \in \Gamma_0} |\!|\!| [\mathbb{E}\Pi_\Omega h(M/\sigma)] F |\!|\!|_{\mathrm{F}}.$$

To prove the lemma, it suffices to show that with probability at least $1 - d^{-12}$, each of $\{T_1, T_2, T_3\}$ is bounded from above by $\frac{1}{12}\alpha\varepsilon_n$. For T_1, we have

$$\begin{aligned} T_1 \leq \frac{2}{\sigma} \sup_{\Delta \in \Gamma_0} d \|H(M/\sigma) - H(M^*/\sigma)\|_\infty \cdot |\!|\!| F |\!|\!|_{\mathrm{op}} &\overset{(i)}{\leq} \frac{16d}{\sigma} \cdot L_{4\nu} \cdot \sup_{\Delta \in \Gamma_0} \|M/\sigma - M^*/\sigma\|_\infty \\ &\leq \frac{16 d 2^{-d^2} L_{4\nu}}{\sigma^2} \overset{(ii)}{\leq} \frac{\sqrt{r \log^2 d} L_{4\nu}}{\sigma} \nu \leq \frac{1}{12}\alpha\varepsilon_n, \end{aligned}$$

where the step (i) follows from the fact that h is element-wise $4L_{4\nu}$-Lipschitz over $[-4\nu, 4\nu]$ and that $|\!|\!| F |\!|\!|_{\mathrm{op}} \leq 2$ for $\Delta \in \Gamma_0$, and in step (ii) from the definition $\nu := \frac{\mu r}{d\sigma}$. Since $|\!|\!| F |\!|\!|_{\mathrm{F}} \leq 2\sqrt{r}$ for $\Delta \in \Gamma_0$, we have

$$T_2 \leq \frac{2}{\sigma} |\!|\!| \Pi_\Omega h(M^*/\sigma) |\!|\!|_{\mathrm{op}} \sup_{\Delta \in \Gamma_0} |\!|\!| F |\!|\!|_{\mathrm{F}} \leq \frac{4\sqrt{r}}{\sigma} |\!|\!| \Pi_\Omega h(M^*/\sigma) |\!|\!|_{\mathrm{op}}$$

Note that for each index pair (i, j), $\mathbb{E}_Y h_{ij}(M^*_{ij}/\sigma) = 0$, and that $|h_{ij}(M^*_{ij}/\sigma)| \leq 4L_{4\nu}$ since $\|M^*/\sigma\|_\infty \leq 4\nu$. Therefore, the matrix $\frac{1}{4L_{4\nu}} \Pi_\Omega h(M^*/\sigma)$ is a censored sub-Gaussian random matrix satisfying the assumptions in Lemma 11, by which we obtain $|\!|\!| \Pi_\Omega h(M^*/\sigma) |\!|\!|_{\mathrm{op}} \leq C L_{4\nu} \sqrt{pd}$ with probability at least $1 - d^{-12}$. It follows that with the same probability, the second term is bounded as $T_2 \leq \frac{4C L_{4\nu}\sqrt{pdr}}{\sigma} \leq \frac{1}{12}\alpha\varepsilon_n$. Finally, the third term T_3 can be bounded as

$$T_3 \leq \frac{2}{\sigma} \sup_{\Delta \in \Gamma_0} |\!|\!| [\mathbb{E}\Pi_\Omega h(M/\sigma)] |\!|\!|_{\mathrm{F}} |\!|\!| F |\!|\!|_{\mathrm{op}} = \frac{2}{\sigma} \sup_{\Delta \in \Gamma_0} |\!|\!| p \cdot \frac{2 f'(M/\sigma)(f(M^*/\sigma) - f(M/\sigma))}{f(M/\sigma)(1 - f(M/\sigma))} |\!|\!|_{\mathrm{F}} |\!|\!| F |\!|\!|_{\mathrm{op}}.$$

Note that $|\!|\!| F |\!|\!|_{\mathrm{op}} \leq 2$ for $\Delta \in \Gamma_0$. Moreover, because f satisfies (29), we know that $|\frac{f'(x)}{f(x)(1-f(x))}| \leq \sqrt{L_{4\nu}}$ and $|f(x) - f(x')| \leq \sqrt{L_{4\nu}}|x - x'|$ for all $x, x' \in [-4\nu, 4\nu]$. It follows that

$$T_3 \leq \frac{8pL_{4\nu}}{\sigma^2} \sup_{\Delta \in \Gamma_0} |\!|\!| M^* - M |\!|\!|_{\mathrm{F}} \leq \frac{8pL_{4\nu}}{\sigma^2} \cdot \frac{3}{2^{d^2}} \overset{(i)}{\leq} \frac{24\sqrt{pdr} L_{4\nu} \nu}{\sigma} \leq \frac{1}{12}\alpha\varepsilon_n,$$

where the step (i) follows from the definition $\nu := \frac{\mu r}{d\sigma}$. This completes the proof of the lemma.

F Proof of inequality (32)

Recalling that the matrix F^* is orthonormal and μ-incoherent, we have $\|F^* \otimes F^*\|_\infty \leq \frac{\mu r}{d}$, and hence $\|\overline{Y} - F^* \otimes F^*\|_\infty \leq 2\frac{\mu r}{d}$. On the other hand, we claim that each row and column of the matrix $\overline{Y} - F^* \otimes F^*$ has at most k non-zero elements. To see this, let Φ^* be the set of the non-zero element of S^*. If $(i,j) \notin \Phi^*$, then $|Y_{ij}| = |(F^* \otimes F^*)_{ij}| \leq \frac{\mu r}{d}$, so $\overline{Y}_{ij} = Y_{ij} = (F^* \otimes F^*)_{ij}$ and thus $(\overline{Y} - F^* \otimes F^*)_{ij} = 0$. Therefore, we find that $\overline{Y} - F^* \otimes F^*$ is supported on the elements in Φ^*, hence the claim. With the above two facts, we apply Proposition 3 in the paper [20] to obtain that

$$\|\!|\overline{Y} - F^* \otimes F^*|\!\|_{\mathrm{op}} \leq k\|\overline{Y} - F^* \otimes F^*\|_\infty \leq 2\frac{\mu r k}{d}.$$

On the other hand, the gap between the r-th and $(r+1)$-th singular values of the matrix $F^* \otimes F^*$ is 1. Letting $\overline{U}$ be the matrix of the top-r singular vectors of $\overline{Y}$ and using $\Theta[\cdot,\cdot]$ to denote the principal angles between two subspaces, we find that

$$\min_{F^* \in \mathcal{E}(M^*)} \|\!|\overline{U} - F^*|\!\|_{\mathrm{op}} \leq \sqrt{2}\|\!|\sin\Theta\big[\mathrm{col}(\overline{U}), \mathrm{col}(F^*)\big]|\!\|_{\mathrm{op}} \leq 2\|\!|\overline{Y} - F^* \otimes F^*|\!\|_{\mathrm{op}} \leq \frac{4\mu r k}{d},$$

where the first step follows from Proposition 2.2 in the paper [59] and the second step follows from Wedin's $\sin\Theta$ theorem [30]. It follows that

$$\mathrm{d}(F^0, F^*) \leq \mathrm{d}(\overline{U}, F^*) \leq \sqrt{r} \min_{F^* \in \mathcal{E}(M^*)} \|\!|\overline{U} - F^*|\!\|_{\mathrm{op}} \leq \frac{4\mu r\sqrt{r} k}{d},$$

where the first step holds because $F^0 = \Pi_{\mathcal{F}}(\overline{U})$ and projection onto the convex set $\mathcal{F}$ is non-expansive, which completes the proof of the claim.

G Spectral norms of censored sub-Gaussian random matrices

In this appendix, we state and prove a useful bound on the spectral norm sub-Gaussian random matrices with censored entries.

Lemma 11. *Suppose $X \in \mathbb{R}^{d \times d}$ is a symmetric random matrix with $X_{ij} = g_{ij}Q_{ij}$, where $\{g_{ij} \mid i \geq j\}$ are independent zero-mean sub-Gaussian random variables with parameter 1, $\{Q_{ij} \mid i \geq j\}$ are independent Bernoulli variables with parameter p, and they are mutually independent. Then there exists a universal constant $c > 0$ such that*

$$\mathbb{E}\big[\|\!|X|\!\|_{op}\big] \leq c\big(\sqrt{pd} + \log d\big), \quad \text{and} \tag{75a}$$

$$\mathbb{P}\Big[\|\!|X|\!\|_{op} \geq c\big(\sqrt{pd} + \log d\big)\Big] \leq d^{-12}. \tag{75b}$$

Let us now prove this lemma. By a standard symmetrization argument, we can assume without loss of generality that each g_{ij} is a symmetric random variable. To proceed, we need the following result from Bandeira and van Handel [5]:

Proposition 1 (Corollaries 3.6 and 3.12 in [5])**.** *Let $\tilde{X}$ be the $d \times d$ symmetric random matrix whose entries $\tilde{X}_{ij}$ are independent symmetric random variables bounded by $\tilde{\sigma}_*$, and define*

$\tilde{\sigma} := \max_i \sqrt{\sum_j \mathbb{E}[\tilde{X}_{ij}^2]}$. *Then there exist universal constants* $\tilde{c}_1$ *and* $\tilde{c}_2$ *such that*

$$\mathbb{E}\|\tilde{X}\|_{op} \le 3\tilde{\sigma} + \tilde{c}_1\tilde{\sigma}_*\sqrt{\log d}, \quad \textit{and} \tag{76a}$$

$$\mathbb{P}\big[\|\tilde{X}\|_{op} \ge 3\tilde{\sigma} + t\big] \le d \cdot \exp\Big(-\frac{t^2}{\tilde{c}_2\tilde{\sigma}_*^2}\Big) \qquad \textit{for each } t \ge 0. \tag{76b}$$

To apply the proposition with unbounded entries in X, we use a standard truncation argument. For some constant b to be specified later, let $\tilde{X}$ be the matrix with $\tilde{X}_{ij} = X_{ij}1_{|X_{ij}| \le b\sqrt{\log d}}$. Observe that $\tilde{X}$ satisfies the assumption in Proposition 1 with $\tilde{\sigma}_* \le b\sqrt{\log d}$ and $\tilde{\sigma} \le \sqrt{pd}$. Applying Proposition 1 with $t = \sqrt{12\tilde{c}_2 b^2}\log d$, we obtain the bounds

$$\mathbb{E}\|\tilde{X}\|_{\text{op}} \le 3\sqrt{pd} + \tilde{c}_1 b \log d, \quad \text{and} \tag{77a}$$

$$\mathbb{P}\big[\|\tilde{X}\|_{\text{op}} \ge 3\sqrt{pd} + \sqrt{12\tilde{c}_2 b^2}\log d\big] \le d\exp\Big(-\frac{t^2}{\tilde{c}_2\tilde{\sigma}_*^2}\Big) \le d^{-13}, \tag{77b}$$

where the last inequality follows from $t \ge \tilde{\sigma}_*\sqrt{12\tilde{c}\log d}$. On the other hand, by choosing the constant b sufficiently large and using a standard bound on the maximum of sub-Gaussian variables, we know that

$$\mathbb{P}\big[\tilde{X} \ne X\big] \le \mathbb{P}\Big[\max_{i,j}|g_{ij}| > b\sqrt{\log d}\Big] \le d^{-13}.$$

Combining with the tail bound (77b) yields

$$\mathbb{P}\left[\|X\|_{\text{op}} \ge 3\sqrt{pd} + \sqrt{12\tilde{c}b^2}\log d\right] \le \mathbb{P}\left[X \ne \tilde{X}\right] + \mathbb{P}\left[\|\tilde{X}\|_{\text{op}} \ge 3\sqrt{pd} + \sqrt{12\tilde{c}b^2}\log d\right] \le d^{-12},$$

which proves the second inequality in Lemma 11.

Turning to the first inequality in the lemma, we let $\breve{X}$ be the matrix with $\breve{X}_{ij} = X_{ij} - \tilde{X}_{ij} = X_{ij}1_{|X_{ij}| > b\sqrt{\log d}}$, and observe that by definition, $\mathbb{P}(0 < \max_{i,j}|\breve{X}_{ij}| \le b\sqrt{\log d}) = 0$. Moreover, by choosing the constant b sufficiently large and using a standard concentration inequality for convex Lipschitz functions [44], we find that for each $t \ge 0$,

$$\begin{aligned}\mathbb{P}\left[\max_{i,j}|\breve{X}_{ij}| > b\sqrt{\log d} + t\right] &\le \mathbb{P}\left[\max_{i,j}|g_{ij}| \ge \mathbb{E}\max_{i,j}|g_{ij}| + t + 4\sqrt{\log d}\right] \\ &\le 2e^{-(t+4\sqrt{\log d})^2/5} \le \frac{2}{d^2}e^{-t^2/5}.\end{aligned}$$

Integrating these tail bounds gives $\mathbb{E}\big[\max_{i,j}|\breve{X}_{ij}|\big] \le \frac{\breve{c}\sqrt{\log d}}{d^2}$ Combining with equation (77a) yields the upper bound

$$\mathbb{E}\|X\|_{\text{op}} \le \mathbb{E}\|\tilde{X}\|_{\text{op}} + \mathbb{E}\|\breve{X}\|_{\text{op}} \le \mathbb{E}\|\tilde{X}\|_{\text{op}} + d\mathbb{E}\max_{i,j}|\breve{X}_{ij}| \le 3\sqrt{pd} + \tilde{c}_1 b\log d + \frac{\breve{c}\sqrt{\log d}}{d},$$

which completes the proof of Lemma 11.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions". In: *The Annals of Statistics* 40.2 (2012), pp. 1171–1197.

[2] N. Alon, M. Krivelevich, and B. Sudakov. "Finding a large hidden clique in a random graph". In: *Random Structures and Algorithms* 13.3-4 (1998), pp. 457–466.

[3] A. A. Amini and M. J. Wainwright. "High-dimensional analysis of semdefinite relaxations for sparse principal component analysis". In: *Annals of Statistics* 5B (2009), pp. 2877–2921.

[4] S. Balakrishnan, M. J. Wainwright, and B. Yu. "Statistical guarantees for the EM algorithm: From population to sample-based analysis". In: *arXiv preprint arXiv:1408.2156* (2014).

[5] A. S. Bandeira and R. van Handel. "Sharp nonasymptotic bounds on the norm of random matrices with independent entries". In: *arXiv preprint arXiv:1408.6185* (2014).

[6] Q. Berthet and P. Rigollet. "Complexity theoretic lower bounds for sparse principal component detection". In: *Journal of Machine Learning Research: Workshop and Conference Proceedings* 30 (2013), pp. 1046–1066.

[7] D. Bertsekas. *Nonlinear programming*. Belmont, MA: Athena Scientific, 1995.

[8] S. A. Bhaskar and A. Javanmard. "1-bit matrix completion under exact low-rank constraint". In: *49th Annual Conference on Information Sciences and Systems (CISS)*. IEEE. 2015, pp. 1–6.

[9] K. Bhatia, P. Jain, and P. Kar. "Robust Regression via Hard Thresholding". In: *arXiv preprint arXiv:1506.02428* (2015).

[10] A. Birnbaum et al. "Minimax bounds for sparse PCA with noisy high-dimensional data". In: *Annals of Statistics* 41.3 (2012), pp. 1055–1084.

[11] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge, UK: Cambridge University Press, 2004.

[12] S. Burer and R. D. C. Monteiro. "Local minima and convergence in low-rank semidefinite programming". In: *Mathematical Programming* 103.3 (2005), pp. 427–444.

[13] J.-F. Cai, S. Liu, and W. Xu. "Projected Wirtinger Gradient Descent for Low-Rank Hankel Matrix Completion in Spectral Compressed Sensing". In: *arXiv preprint arXiv:1507.03707* (2015).

[14] T. T. Cai, Z. Ma, Y. Wu, et al. "Sparse PCA: Optimal rates and adaptive estimation". In: *The Annals of Statistics* 41.6 (2013), pp. 3074–3110.

[15] T. Cai, Z. Ma, and Y. Wu. "Optimal estimation and rank detection for sparse spiked covariance matrices". In: *Probability Theory and Related Fields* (2013), pp. 1–35.

[16] T. Cai and W.-X. Zhou. "A max-norm constrained minimization approach to 1-bit matrix completion". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3619–3647.

[17] E. Candès, X. Li, and M. Soltanolkotabi. "Phase Retrieval via Wirtinger Flow: Theory and Algorithms". In: *arXiv preprint arXiv:1407.1065* (2014). URL: `http://arxiv.org/abs/1407.1065`.

[18] E. J. Candès and B. Recht. "Exact matrix completion via convex optimization". In: *Foundations of Computational Mathematics* 9.6 (2009), pp. 717–772.

[19] E. J. Candès et al. "Robust principal component analysis?" In: *Journal of the ACM* 58.3 (2011), p. 11.

[20] V. Chandrasekaran et al. "Rank-Sparsity Incoherence for Matrix Decomposition". In: *SIAM Journal on Optimization* 21.2 (2011), pp. 572–596.

[21] Y. Chen, S. Sanghavi, and H. Xu. "Improved Graph Clustering". In: *IEEE Transactions on Information Theory* 60.10 (2014), pp. 6440–6455.

[22] Y. Chen et al. "Low-rank Matrix Recovery from Errors and Erasures". In: *IEEE Transactions on Information Theory* 59.7 (2013), pp. 4324–4337.

[23] Y. Chen and E. J. Candes. "Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems". In: *arXiv preprint arXiv:1505.05114* (2015).

[24] M. A. Davenport et al. "1-bit matrix completion". In: *Information and Inference* 3.3 (2014), pp. 189–223.

[25] C. De Sa, K. Olukotun, and C. Ré. "Global Convergence of Stochastic Gradient Descent for Some Nonconvex Matrix Problems". In: *arXiv preprint arXiv:1411.1134* (2014).

[26] C. De Sa et al. "Taming the Wild: A Unified Analysis of Hogwild!-Style Algorithms". In: *arXiv preprint arXiv:1506.06438* (2015).

[27] J. Duchi et al. "Efficient projections onto the l 1-ball for learning in high dimensions". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 272–279.

[28] U. Feige and E. Ofek. "Spectral techniques applied to sparse random graphs". In: *Random Structures & Algorithms* 27.2 (2005), pp. 251–275.

[29] C. Gao et al. "Minimax Estimation in Sparse Canonical Correlation Analysis". In: *The Annals of Statistics, to appear. arXiv preprint arXiv:1405.1595* (2014).

[30] G. H. Golub and C. F. Van Loan. *Matrix computations*. 3rd ed. The Johns Hopkins University Press, 1996.

[31] N. Halko, P.-G. Martinsson, and J. A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM review* 53.2 (2011), pp. 217–288.

[32] M. Hardt. "Understanding alternating minimization for matrix completion". In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*. 2014, pp. 651–660.

[33] M. Hardt and E. Price. "The Noisy Power Method: A Meta Algorithm with Applications". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2861–2869.

[34] M. Hardt and M. Wootters. "Fast Matrix Completion Without the Condition Number". In: *Proceedings of The 27th Conference on Learning Theory*. 2014, pp. 638–678.

[35] D. Hsu, S. M. Kakade, and T. Zhang. "Robust matrix decomposition with sparse corruptions". In: *IEEE Transactions on Information Theory* 57.11 (2011), pp. 7221–7234.

[36] P. Jain, R. Meka, and I. S. Dhillon. "Guaranteed rank minimization via singular value projection". In: *Advances in Neural Information Processing Systems*. 2010, pp. 937–945.

[37] P. Jain, P. Netrapalli, and S. Sanghavi. "Low-rank matrix completion using alternating minimization". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM. 2013, pp. 665–674.

[38] P. Jain, A. Tewari, and P. Kar. "On iterative hard thresholding methods for high-dimensional M-estimation". In: *Advances in Neural Information Processing Systems*. 2014, pp. 685–693.

[39] I. M. Johnstone and A. Y. Lu. "On consistency and sparsity for principal components analysis in high dimensions". In: *Journal of the American Statistical Association* 104 (2009), pp. 682–693.

[40] R. H. Keshavan, A. Montanari, and S. Oh. "Matrix completion from a few entries". In: *IEEE Transactions on Information Theory* 56.6 (2010), pp. 2980–2998.

[41] R. H. Keshavan, A. Montanari, and S. Oh. "Matrix completion from noisy entries". In: *The Journal of Machine Learning Research* 99 (2010), pp. 2057–2078.

[42] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion". In: *The Annals of Statistics* 39.5 (2011), pp. 2302–2329.

[43] M. Laurent. "Matrix Completion Problems". In: *The Encyclopedia of Optimization*. Kluwer Academic, 2001, pp. 221–229.

[44] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2001.

[45] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. New York, NY: Springer-Verlag, 1991.

[46] P.-L. Loh and M. J. Wainwright. "High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity". In: *The Annals of Statistics* 40.3 (2012), pp. 1637–1664.

[47] P.-L. Loh and M. J. Wainwright. "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima". In: *Advances in Neural Information Processing Systems*. 2013, pp. 476–484.

[48] Z. Ma. "Sparse principal component analysis and iterative thresholding". In: *The Annals of Statistics* 41.2 (2013), pp. 772–801.

[49] N. Maculan et al. "An $O(n)$ algorithm for projecting a vector on the intersection of a hyperplane and a box in R^n". In: *Journal of optimization theory and applications* 117.3 (2003), pp. 553–574.

[50] S. Negahban and M. J. Wainwright. "Estimation of (near) low-rank matrices with noise and high-dimensional scaling". In: *Arxiv preprint arXiv:0912.5100* (2009).

[51] S. Negahban and M. J. Wainwright. "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise". In: *The Journal of Machine Learning Research* 13 (2012), pp. 1665–1697.

[52] Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM studies in applied and numerical mathematics. Society for Industrial and Applied Mathematics, 1987.

[53] P. Netrapalli et al. "Non-convex robust PCA". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1107–1115.

[54] B. Recht, M. Fazel, and P. A. Parrilo. "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization". In: *SIAM Review* 52.471 (2010).

[55] M. Soltanolkotabi. "Algorithms and Theory for Clustering and Nonconvex Quadratic Programming". PhD thesis. Stanford University, 2014.

[56] H. Su, A. W. Yu, and L. Fei-Fei. "Efficient euclidean projections onto the intersection of norm balls". In: *Proceedings of the 29th International Conference on Machine Learning*. Vol. 1. International Machine Learning Society. 2012, pp. 433–440.

[57] R. Sun and Z.-Q. Luo. "Guaranteed Matrix Completion via Non-convex Factorization". In: *arXiv preprint arXiv:1411.8003* (2014).

[58] S. Tu et al. "Low-rank Solutions of Linear Matrix Equations via Procrustes Flow". In: *arXiv preprint arXiv:1507.03566* (2015).

[59] V. Q. Vu, J. Lei, et al. "Minimax sparse principal subspace estimation in high dimensions". In: *The Annals of Statistics* 41.6 (2013), pp. 2905–2947.

[60] V. Q. Vu et al. "Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2670–2678.

[61] Z. Wang, H. Liu, and T. Zhang. "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems". In: *arXiv preprint arXiv:1306.4960* (2013).

[62] Z. Wang, H. Lu, and H. Liu. "Nonconvex statistical optimization: Minimax-optimal Sparse PCA in polynomial time". In: *arXiv preprint arXiv:1408.5352* (2014).

[63] Z. Wang et al. "High Dimensional Expectation-Maximization Algorithm: Statistical Optimization and Asymptotic Normality". In: *arXiv preprint arXiv:1412.8729* (2014).

[64] C. D. White, R. Ward, and S. Sanghavi. "The Local Convexity of Solving Quadratic Equations". In: *arXiv preprint arXiv:1506.07868* (2015).

[65] D. Zhang and L. Balzano. "Global Convergence of a Grassmannian Gradient Descent Algorithm for Subspace Estimation". In: *arXiv preprint arXiv:1506.07405* (2015).

[66] Q. Zheng and J. Lafferty. "A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming from Random Linear Measurements". In: *arXiv preprint arXiv:1506.06081* (2015).