Matrix Completion and Related Problems via Strong Duality*†

Maria-Florina Balcan[1], Yingyu Liang[2], David P. Woodruff[3], and Hongyang Zhang[4]

1 **Carnegie Mellon University, Pittsburgh, USA**
ninamf@cs.cmu.edu
2 **University of Wisconsin-Madison, Madison, USA**
yliang@cs.wisc.edu
3 **Carnegie Mellon University, Pittsburgh, USA**
dwoodruf@cs.cmu.edu
4 **Carnegie Mellon University, Pittsburgh, USA**
hongyanz@cs.cmu.edu

Abstract

This work studies the *strong duality of non-convex matrix factorization problems*: we show that under certain dual conditions, these problems and its dual have the same optimum. This has been well understood for convex optimization, but little was known for non-convex problems. We propose a novel analytical framework and show that under certain dual conditions, the optimal solution of the matrix factorization program is the same as its bi-dual and thus the global optimality of the non-convex program can be achieved by solving its bi-dual which is convex. These dual conditions are satisfied by a wide class of matrix factorization problems, although matrix factorization problems are hard to solve in full generality. This analytical framework may be of independent interest to non-convex optimization more broadly.

We apply our framework to two prototypical matrix factorization problems: matrix completion and robust Principal Component Analysis (PCA). These are examples of efficiently recovering a hidden matrix given limited reliable observations of it. Our framework shows that exact recoverability and strong duality hold with nearly-optimal sample complexity guarantees for matrix completion and robust PCA.

1998 ACM Subject Classification G.1.6 Optimization

Keywords and phrases Non-Convex Optimization, Strong Duality, Matrix Completion, Robust PCA, Sample Complexity

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.5

1 Introduction

Non-convex matrix factorization problems have been an emerging object of study in theoretical computer science [37, 30, 53, 45], optimization [58, 50], machine learning [11, 23, 21, 36, 42, 57], and many other domains. In theoretical computer science and optimization, the study of such models has led to significant advances in provable algorithms that converge to local

* A full version of the paper is available at https://arxiv.org/abs/1704.08683

† This work was supported in part by NSF grants NSF CCF-1422910, NSF CCF-1535967, NSF CCF-1451177, NSF IIS-1618714, NSF CCF-1527371, a Sloan Research Fellowship, a Microsoft Research Faculty Fellowship, DMS-1317308, Simons Investigator Award, Simons Collaboration Grant, and ONR-N00014-16-1-2329.

minima in linear time [37, 30, 53, 2, 3]. In machine learning, matrix factorization serves as a building block for large-scale prediction and recommendation systems, e.g., the winning submission for the Netflix prize [41]. Two prototypical examples are matrix completion and robust Principal Component Analysis (PCA).

This work develops a novel framework to analyze a class of non-convex matrix factorization problems with strong duality, which leads to exact recoverability for matrix completion and robust Principal Component Analysis (PCA) via the solution to a convex problem. The matrix factorization problems can be stated as finding a target matrix $\mathbf{X}^*$ in the form of $\mathbf{X}^* = \mathbf{AB}$, by minimizing the objective function $H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2$ over factor matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times n_2}$ with a known value of $r \ll \min\{n_1, n_2\}$, where $H(\cdot)$ is some function that characterizes the desired properties of $\mathbf{X}^*$.

Our work is motivated by several promising areas where our analytical framework for non-convex matrix factorizations is applicable. The first area is low-rank matrix completion, where it has been shown that a low-rank matrix can be exactly recovered by finding a solution of the form $\mathbf{AB}$ that is consistent with the observed entries (assuming that it is incoherent) [37, 53, 23]. This problem has received a tremendous amount of attention due to its important role in optimization and its wide applicability in many areas such as quantum information theory and collaborative filtering [30, 61, 7]. The second area is robust PCA, a fundamental problem of interest in data processing that aims at recovering both the low-rank and the sparse components exactly from their superposition [13, 43, 27, 62, 61, 59], where the low-rank component corresponds to the product of $\mathbf{A}$ and $\mathbf{B}$ while the sparse component is captured by a proper choice of function $H(\cdot)$, e.g., the ℓ_1 norm [13, 6]. We believe our analytical framework can be potentially applied to other non-convex problems more broadly, e.g., matrix sensing [54], dictionary learning [52], weighted low-rank approximation [45, 42], and deep linear neural network [39], which may be of independent interest.

Without assumptions on the structure of the objective function, direct formulations of matrix factorization problems are NP-hard to optimize in general [31, 60]. With standard assumptions on the structure of the problem and with sufficiently many samples, these optimization problems can be solved efficiently, e.g., by convex relaxation [14, 18]. Some other methods run local search algorithms given an initialization close enough to the global solution in the basin of attraction [37, 30, 53, 21, 38]. However, these methods have sample complexity significantly larger than the information-theoretic lower bound; see Table 1 for a comparison. The problem becomes more challenging when the number of samples is small enough that the sample-based initialization is far from the desired solution, in which case the algorithm can run into a local minimum or a saddle point.

Another line of work has focused on studying the loss surface of matrix factorization problems, providing positive results for approximately achieving global optimality. One nice property in this line of research is that there is no spurious local minima for specific applications such as matrix completion [23], matrix sensing [11], dictionary learning [52], phase retrieval [51], linear deep neural networks [39], etc. However, these results are based on concrete forms of objective functions. Also, even when any local minimum is guaranteed to be globally optimal, in general it remains NP-hard to escape high-order saddle points [5], and additional arguments are needed to show the achievement of a local minimum. Most importantly, all existing results rely on strong assumptions on the sample size.

1.1 Our Results

Our work studies the exact recoverability problem for a variety of non-convex matrix factorization problems. The goal is to provide a unified framework to analyze a large class

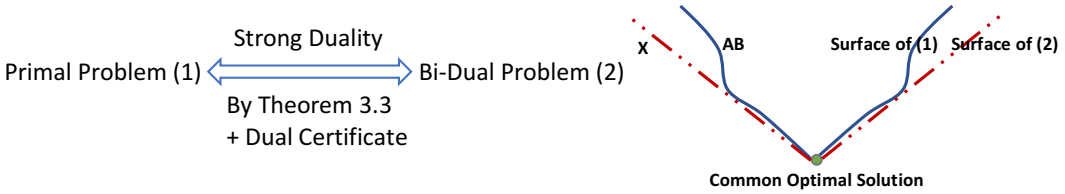


Figure 1 Strong duality of matrix factorizations.

of matrix factorization problems, and to achieve efficient algorithms. Our main results show that although matrix factorization problems are hard to optimize in general, *under certain dual conditions the duality gap is zero*, and thus the problem can be converted to an equivalent convex program. The main theorem of our framework is the following.

Theorem 4. (Strong Duality. Informal.) *Under certain dual conditions, strong duality holds for the non-convex optimization problem*

$$(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \operatorname*{argmin}_{\mathbf{A}\in\mathbb{R}^{n_1\times r}, \mathbf{B}\in\mathbb{R}^{r\times n_2}} F(\mathbf{A},\mathbf{B}) = H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2, \quad H(\cdot) \text{ is convex and closed}, \tag{1}$$

where "the function $H(\cdot)$ is closed" means that for each $\alpha \in \mathbb{R}$, the sub-level set $\{\mathbf{X} \in \mathbb{R}^{n_1\times n_2} : H(\mathbf{X}) \le \alpha\}$ is a closed set. In other words, problem (1) and its bi-dual problem

$$\widetilde{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} H(\mathbf{X}) + \|\mathbf{X}\|_{r*}, \tag{2}$$

have exactly the same optimal solutions in the sense that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$, where $\|\mathbf{X}\|_{r*}$ is a convex function defined by $\|\mathbf{X}\|_{r*} = \max_{\mathbf{M}} \langle \mathbf{M}, \mathbf{X}\rangle - \frac{1}{2}\|\mathbf{M}\|_r^2$ and $\|\mathbf{M}\|_r^2 = \sum_{i=1}^r \sigma_i^2(\mathbf{M})$ is the sum of the first r largest squared singular values.

Theorem 4 connects the non-convex program (1) to its convex counterpart via strong duality; see Figure 1. We mention that strong duality rarely happens in the non-convex optimization region: low-rank matrix approximation [44] and quadratic optimization with two quadratic constraints [10] are among the few paradigms that enjoy such a nice property. Given strong duality, the computational issues of the original problem can be overcome by solving the convex bi-dual problem (2).

The positive result of our framework is complemented by a lower bound to formalize the hardness of the above problem in general. Assuming that the random 4-SAT problem is hard [45], we give a strong negative result for deterministic algorithms. If also $\mathsf{BPP} = \mathsf{P}$ (see Section 6 for a discussion), then the same conclusion holds for randomized algorithms succeeding with probability at least $2/3$.

Theorem 9. (Hardness Statement. Informal.) *Assuming that random 4-SAT is hard on average, there is a problem in the form of (1) such that any deterministic algorithm achieving $(1+\epsilon)\mathsf{OPT}$ in the objective function value with $\epsilon \le \epsilon_0$ requires $2^{\Omega(n_1+n_2)}$ time, where OPT is the optimum and $\epsilon_0 > 0$ is an absolute constant. If $\mathsf{BPP} = \mathsf{P}$, then the same conclusion holds for randomized algorithms succeeding with probability at least $2/3$.*

Our framework only requires the dual conditions in Theorem 4 to be verified. We will show that two prototypical problems, matrix completion and robust PCA, obey the conditions. They belong to the linear inverse problems of form (1) with a proper choice of function $H(\cdot)$, which aim at exactly recovering a hidden matrix $\mathbf{X}^*$ with $\mathsf{rank}(\mathbf{X}^*) \le r$ given a limited number of linear observations of it.

For matrix completion, the linear measurements are of the form $\{\mathbf{X}^*_{ij} : (i,j) \in \Omega\}$, where Ω is the support set which is uniformly distributed among all subsets of $[n_1] \times [n_2]$

Table 1 Comparison of matrix completion methods. Here $\kappa = \sigma_1(\mathbf{X}^*)/\sigma_r(\mathbf{X}^*)$ is the condition number of $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$, ϵ is the accuracy such that the output $\widetilde{\mathbf{X}}$ obeys $\|\widetilde{\mathbf{X}} - \mathbf{X}^*\|_F \leq \epsilon$, $n_{(1)} = \max\{n_1, n_2\}$ and $n_{(2)} = \min\{n_1, n_2\}$. *The first line of ours is an information-theoretic upper bound and the second line is a polynomial-time approach.*

Work	Sample Complexity	μ-Incoherence
[37]	$\mathcal{O}\left(\kappa^4\mu^2 r^{4.5} n_{(1)} \log n_{(1)} \log\left(\frac{r\|\mathbf{X}^*\|_F}{\epsilon}\right)\right)$	Condition (3)
[30]	$\mathcal{O}\left(\mu r n_{(1)}(r + \log\left(\frac{n_{(1)}\|\mathbf{X}^*\|_F}{\epsilon}\right)\frac{\|\mathbf{X}^*\|_F^2}{\sigma_r^2}\right)$	Condition (3)
[23]	$\mathcal{O}(\max\{\mu^6\kappa^{16}r^4, \mu^4\kappa^4 r^6\} n_{(1)} \log^2 n_{(1)})$	$\|\mathbf{X}^*_{i\cdot}\|_2 \leq \frac{\mu}{\sqrt{n_{(2)}}}\|\mathbf{X}^*\|_F$
[53]	$\mathcal{O}(r n_{(1)}\kappa^2 \max\left\{\mu \log n_{(2)}, \sqrt{\frac{n_{(1)}}{n_{(2)}}}\mu^2 r^6 \kappa^4\right\}$	Condition (3)
[65]	$\mathcal{O}(\mu r^2 n_{(1)}\kappa^2 \max(\mu, \log n_{(1)}))$	Condition (3)
[20]	$\mathcal{O}\left(\left(\mu^2 r^4\kappa^2 + \mu r \log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\right) n_{(1)} \log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\right)$	Condition (3)
[64]	$\mathcal{O}\left(\mu r^3 n_{(1)} \log n_{(1)} \log\left(\frac{1}{\epsilon}\right)\right)$	Condition (3)
[40]	$\mathcal{O}\left(n_{(2)} r \sqrt{\frac{n_{(1)}}{n_{(2)}}}\kappa^2 \max\left\{\mu \log n_{(2)}, \mu^2 r \sqrt{\frac{n_{(1)}}{n_{(2)}}}\kappa^4\right\}\right)$	Similar to (3) and (12)
[17]	$\mathcal{O}(\max\{\mu\kappa n_{(1)} r \log n, \mu^2 r^2 \kappa^2 n_{(1)}\})$	Condition (3)
[25]	$\mathcal{O}(\mu r n_{(1)} \log^2 n_{(1)})$	Conditions (3) and (12)
[18]	$\mathcal{O}(\mu r n_{(1)} \log^2 n_{(1)})$	Condition (3)
Ours	$\mathcal{O}(\mu r n_{(1)} \log n_{(1)})$	Condition (3)
	$\mathcal{O}(\kappa^2 \mu r n_{(1)} \log(n_{(1)}) \log_{2\kappa}(n_{(1)}))$	Condition (3)
Lower Bound[1] [15]	$\Omega(\mu r n_{(1)} \log n_{(1)})$	Condition (3)

of cardinality m. With strong duality, we can either study the exact recoverability of the primal problem (1), or investigate the validity of its convex dual (or bi-dual) problem (2). Here we study the former with tools from geometric functional analysis. Recall that in the analysis of matrix completion, one typically requires a μ-incoherence condition for a given rank-r matrix $\mathbf{X}^*$ with skinny SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ [46, 15]:

$$\|\mathbf{U}^T\mathbf{e}_i\|_2 \leq \sqrt{\frac{\mu r}{n_1}}, \qquad \text{and} \qquad \|\mathbf{V}^T\mathbf{e}_i\|_2 \leq \sqrt{\frac{\mu r}{n_2}}, \qquad \text{for all } i \tag{3}$$

where $\mathbf{e}_i$'s are vectors with i-th entry equal to 1 and other entries equal to 0. The incoherence condition claims that information spreads throughout the left and right singular vectors and is quite standard in the matrix completion literature. Under this standard condition, we have the following results.

Theorems 5, 7, and 6. (Matrix Completion. Informal.) *$\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ is the unique matrix of rank at most r that is consistent with the m measurements with high probability, provided $m = \mathcal{O}(\mu(n_1+n_2)r\log(n_1+n_2))$ and $\mathbf{X}^*$ satisfies incoherence (3). In addition, there exists a convex optimization for matrix completion in the form of (2) that exactly recovers $\mathbf{X}^*$ with high probability, provided that $m = \mathcal{O}(\kappa^2\mu(n_1+n_2)r\log(n_1+n_2)\log_{2\kappa}(n_1+n_2))$, where κ is the condition number of $\mathbf{X}^*$.*

To the best of our knowledge, our result is the first to connect convex matrix completion to non-convex matrix completion, two parallel lines of research that have received significant attention in the past few years. Table 1 compares our result with prior results.

[1] This lower bound is information-theoretic.

For robust PCA, instead of studying exact recoverability of problem (1) as for matrix completion, we investigate problem (2) directly. The robust PCA problem is to decompose a given matrix $\mathbf{D} = \mathbf{X}^* + \mathbf{S}^*$ into the sum of a low-rank component $\mathbf{X}^*$ and a sparse component $\mathbf{S}^*$ [1]. We obtain the following theorem for robust PCA.

Theorem 8. (Robust PCA. Informal.) *There exists a convex optimization formulation for robust PCA in the form of problem* (2) *that exactly recovers the incoherent matrix* $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ *and* $\mathbf{S}^* \in \mathbb{R}^{n_1 \times n_2}$ *with high probability, even if* $\mathsf{rank}(\mathbf{X}^*) = \Theta\left(\frac{\min\{n_1,n_2\}}{\mu \log^2 \max\{n_1,n_2\}}\right)$ *and the size of the support of* $\mathbf{S}^*$ *is* $m = \Theta(n_1 n_2)$*, where the support set of* $\mathbf{S}^*$ *is uniformly distributed among all sets of cardinality* m*, and the incoherence parameter* μ *satisfies constraints* (3) *and* $\|\mathbf{X}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}\sigma_r(\mathbf{X}^*)$.

The bounds in Theorem 8 match the best known results in the robust PCA literature when the supports of $\mathbf{S}^*$ are uniformly sampled [13], while our assumption is arguably more intuitive; see Section 5. Note that our results hold even when $\mathbf{X}^*$ is close to full rank and a constant fraction of the entries have noise. Independently of our work, Ge et al. [22] developed a framework to analyze the loss surface of low-rank problems, and applied the framework to matrix completion and robust PCA. Their bounds are: for matrix completion, the sample complexity is $\mathcal{O}(\kappa^6 \mu^4 r^6 (n_1+n_2)\log(n_1+n_2))$; for robust PCA, the outlier entries are deterministic and the number that the method can tolerate is $\mathcal{O}\left(\frac{n_1 n_2}{\mu r \kappa^5}\right)$. Zhang et al. [63] also studied the robust PCA problem using non-convex optimization, where the outlier entries are deterministic and the number of outliers that their algorithm can tolerate is $\mathcal{O}\left(\frac{n_1 n_2}{r\kappa}\right)$. The strong duality approach is unique to our work.

1.2 Our Techniques

Reduction to Low-Rank Approximation. Our results are inspired by the low-rank approximation problem:

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r},\mathbf{B}\in\mathbb{R}^{r\times n_2}} \frac{1}{2}\|-\widetilde{\mathbf{\Lambda}} - \mathbf{AB}\|_F^2. \tag{4}$$

We know that all local solutions of (4) are globally optimal (see Lemma 1) and that strong duality holds for any given matrix $-\widetilde{\mathbf{\Lambda}} \in \mathbb{R}^{n_1 \times n_2}$ [26]. To extend this property to our more general problem (1), our main insight is to reduce problem (1) to the form of (4) using the ℓ_2-regularization term. While some prior work attempted to apply a similar reduction, their conclusions either depended on unrealistic conditions on local solutions, e.g., all local solutions are rank-deficient [28, 26], or their conclusions relied on strong assumptions on the objective functions, e.g., that the objective functions are twice-differentiable [29]. Instead, our general results formulate strong duality via the existence of a dual certificate $\widetilde{\mathbf{\Lambda}}$. For concrete applications, the existence of a dual certificate is then converted to mild assumptions, e.g., that the number of measurements is sufficiently large and the positions of measurements are randomly distributed. We will illustrate the importance of randomness below.

The Blessing of Randomness. The desired dual certificate $\widetilde{\mathbf{\Lambda}}$ may not exist in the deterministic world. A hardness result [45] shows that for the problem of weighted low-rank approximation, which can be cast in the form of (1), without some randomization in the measurements made on the underlying low rank matrix, it is NP-hard to achieve a good objective value, not to mention to achieve strong duality. A similar phenomenon was observed for deterministic matrix completion [32]. Thus we should utilize such randomness to analyze the

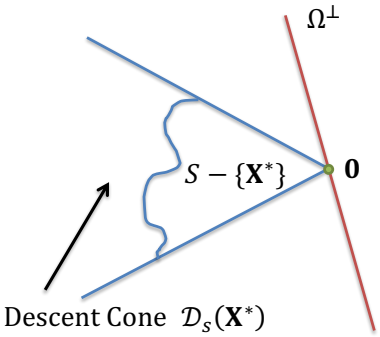


Figure 2 Feasibility.

existence of a dual certificate. For matrix completion, the assumption that the measurements are random is standard, under which, the angle between the space Ω (the space of matrices which are consistent with observations) and the space $\mathcal{T}$ (the space of matrices which are low-rank) is small with high probability, namely, $\mathbf{X}^*$ is almost the unique low-rank matrix that is consistent with the measurements. Thus, our dual certificate can be represented as another form of a convergent Neumann series concerning the projection operators on the spaces Ω and $\mathcal{T}$. The remainder of the proof is to show that such a construction obeys the dual conditions.

To prove the dual conditions for matrix completion, we use the fact that the subspace Ω and the complement space $\mathcal{T}^\perp$ are almost orthogonal when the sample size is sufficiently large. This implies the projection of our dual certificate on the space $\mathcal{T}^\perp$ has a very small norm, which exactly matches the dual conditions.

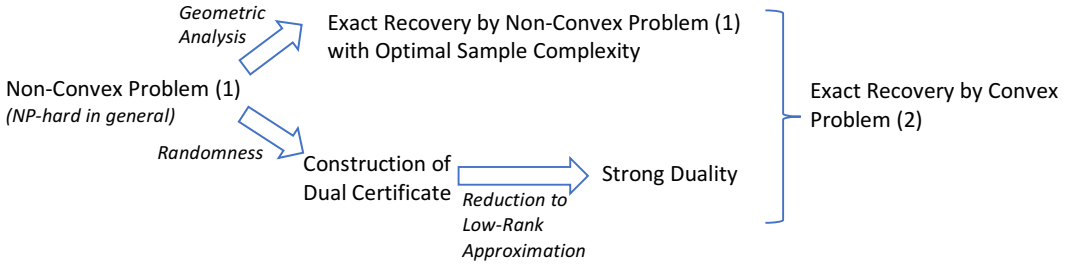
Non-Convex Geometric Analysis. Strong duality implies that the primal problem (1) and its bi-dual problem (2) have exactly the same solutions in the sense that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$. Thus, to show exact recoverability of linear inverse problems such as matrix completion and robust PCA, it suffices to study either the non-convex primal problem (1) or its convex counterpart (2). Here we do the former analysis for matrix completion. We mention that traditional techniques [15, 46, 16] for convex optimization break down for our non-convex problem, since the subgradient of a non-convex objective function may not even exist [12]. Instead, we apply tools from geometric functional analysis [55] to analyze the geometry of problem (1). Our non-convex geometric analysis is in stark contrast to prior techniques of convex geometric analysis [56] where convex combinations of non-convex constraints were used to define the Minkowski functional (e.g., in the definition of atomic norm) while our method uses the non-convex constraint itself.

For matrix completion, problem (1) has two hard constraints: a) the rank of the output matrix should be no larger than r, as implied by the form of $\mathbf{AB}$; b) the output matrix should be consistent with the sampled measurements, i.e., $\mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*)$. We study the feasibility condition of problem (1) from a geometric perspective: $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$ is the unique feasible solution to problem (1) if and only if starting from $\mathbf{X}^*$, the rank of $\mathbf{X}^* + \mathbf{D}$ increases for all directions $\mathbf{D}$'s in the constraint set $\Omega^\perp = \{\mathbf{D} \in \mathbb{R}^{n_1 \times n_2} : \mathcal{P}_\Omega(\mathbf{X}^* + \mathbf{D}) = \mathcal{P}_\Omega(\mathbf{X}^*)\}$ (a.k.a. the feasibility condition). This can be geometrically interpreted as the requirement that the descent cone $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) = \{t(\mathbf{X} - \mathbf{X}^*) \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \leq r,\ t \geq 0\}$ and the constraint set $\Omega^\perp$ must intersect uniquely at $\mathbf{0}$ (see Figure 2), which means $\mathbf{X}^*$ is the unique matrix that satisfies the constraints a) and b). This is shown by the following tangent cone argument.

Let $\mathcal{S}$ be the set of all matrices with rank at most r around the underlying matrix $\mathbf{X}^*$. In the tangent cone argument, by definition, $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*)$ is a subset of the tangent cone of $\mathcal{S}$ at $\mathbf{X}^*$. The latter cone of interest has a very nice form, namely, it is just the space $\mathcal{T}$ mentioned above (the space of matrices which are low-rank). Now leverage results from prior work which imply $\mathcal{T} \cap \Omega^{\perp} = \{\mathbf{0}\}$ with a large enough sample size. Namely, among all matrices of the form $\mathbf{X}^* + \mathbf{D}$, $\mathbf{D} = \mathbf{0}$ is the only matrix such that $\mathsf{rank}(\mathbf{X}^* + \mathbf{D}) \leq r$ and $\mathbf{X}^* + \mathbf{D}$ is consistent with the observations.

Using this argument, we can show that the sample size needed for exact recovery in matrix completion matches the known lower bound up to a constant factor.

Putting Things Together. We summarize our new analytical framework with the following figure.



Other Techniques. An alternative method is to investigate the exact recoverability of problem (2) via standard convex analysis. We find that the sub-differential of our induced function $\|\cdot\|_{r*}$ is very similar to that of the nuclear norm. With this observation, we prove the validity of robust PCA in the form of (2) by combining this property of $\|\cdot\|_{r*}$ with standard techniques from [13].

2 Preliminaries

We will use calligraphy to represent a set, bold capital letters to represent a matrix, bold lower-case letters to represent a vector, and lower-case letters to represent scalars. Specifically, we denote by $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ the underlying matrix. We use $\mathbf{X}_{:t} \in \mathbb{R}^{n_1 \times 1}$ ($\mathbf{X}_{t:} \in \mathbb{R}^{1 \times n_2}$) to indicate the t-th column (row) of $\mathbf{X}$. The entry in the i-th row, j-th column of $\mathbf{X}$ is represented by $\mathbf{X}_{ij}$. The condition number of $\mathbf{X}$ is $\kappa = \sigma_1(\mathbf{X})/\sigma_r(\mathbf{X})$. We let $n_{(1)} = \max\{n_1, n_2\}$ and $n_{(2)} = \min\{n_1, n_2\}$. For a function $H(\mathbf{M})$ on an input matrix $\mathbf{M}$, its conjugate function H^* is defined by $H^*(\mathbf{\Lambda}) = \max_{\mathbf{M}} \langle \mathbf{\Lambda}, \mathbf{M} \rangle - H(\mathbf{M})$. Furthermore, let H^{**} denote the conjugate function of H^*.

We will frequently use $\mathsf{rank}(\mathbf{X}) \leq r$ to constrain the rank of $\mathbf{X}$. This can be equivalently represented as $\mathbf{X} = \mathbf{AB}$, by restricting the number of columns of $\mathbf{A}$ and rows of $\mathbf{B}$ to be r. For norms, we denote by $\|\mathbf{X}\|_F = \sqrt{\sum_{ij} \mathbf{X}_{ij}^2}$ the Frobenius norm of matrix $\mathbf{X}$. Let $\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq ... \geq \sigma_r(\mathbf{X})$ be the non-zero singular values of $\mathbf{X}$. The nuclear norm (a.k.a. trace norm) of $\mathbf{X}$ is defined by $\|\mathbf{X}\|_* = \sum_{i=1}^{r} \sigma_i(\mathbf{X})$, and the operator norm of $\mathbf{X}$ is $\|\mathbf{X}\| = \sigma_1(\mathbf{X})$. Denote by $\|\mathbf{X}\|_\infty = \max_{ij} |\mathbf{X}_{ij}|$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of equal dimensions, we denote by $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} \mathbf{A}_{ij} \mathbf{B}_{ij}$. We denote by $\partial H(\mathbf{X}) = \{\mathbf{\Lambda} \in \mathbb{R}^{n_1 \times n_2} : H(\mathbf{Y}) \geq H(\mathbf{X}) + \langle \mathbf{\Lambda}, \mathbf{Y} - \mathbf{X} \rangle \text{ for any } \mathbf{Y}\}$ the sub-differential of function H evaluated at $\mathbf{X}$.

We define the indicator function of convex set $\mathcal{C}$ by $\mathbf{I}_{\mathcal{C}}(\mathbf{X}) = \begin{cases} 0, & \text{if } \mathbf{X} \in \mathcal{C}; \\ +\infty, & \text{otherwise.} \end{cases}$ For any non-empty set $\mathcal{C}$, denote by $\mathsf{cone}(\mathcal{C}) = \{t\mathbf{X} : \mathbf{X} \in \mathcal{C},\ t \geq 0\}$.

We denote by Ω the set of indices of observed entries, and $\Omega^{\perp}$ its complement. Without confusion, Ω also indicates the linear subspace formed by matrices with entries in $\Omega^{\perp}$ being 0. We denote by $\mathcal{P}_{\Omega} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ the orthogonal projector of subspace Ω. We will consider a single norm for these operators, namely, the operator norm denoted by $\|\mathcal{A}\|$ and defined by $\|\mathcal{A}\| = \sup_{\|\mathbf{X}\|_F = 1} \|\mathcal{A}(\mathbf{X})\|_F$. For any orthogonal projection operator $\mathcal{P}_{\mathcal{T}}$ to any subspace $\mathcal{T}$, we know that $\|\mathcal{P}_{\mathcal{T}}\| = 1$ whenever $\mathsf{dim}(\mathcal{T}) \neq 0$. For distributions, denote by $\mathcal{N}(0,1)$ a standard Gaussian random variable, $\mathsf{Uniform}(m)$ the uniform distribution of cardinality m, and $\mathsf{Ber}(p)$ the Bernoulli distribution with success probability p.

3 ℓ_2-Regularized Matrix Factorizations: A New Analytical Framework

In this section, we develop a novel framework to analyze a general class of ℓ_2-regularized matrix factorization problems. Our framework can be applied to different specific problems and leads to nearly optimal sample complexity guarantees. In particular, we study the ℓ_2-regularized matrix factorization problem

$$(\mathbf{P}) \qquad \min_{\mathbf{A} \in \mathbb{R}^{n_1 \times r}, \mathbf{B} \in \mathbb{R}^{r \times n_2}} F(\mathbf{A}, \mathbf{B}) = H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2,\ H(\cdot) \text{ is convex and closed.}$$

We show that under suitable conditions the duality gap between $(\mathbf{P})$ and its dual (bi-dual) problem is zero, so problem $(\mathbf{P})$ can be converted to an equivalent convex problem.

3.1 Strong Duality

We first consider an easy case where $H(\mathbf{AB}) = \frac{1}{2}\|\widehat{\mathbf{Y}}\|_F^2 - \langle \widehat{\mathbf{Y}}, \mathbf{AB} \rangle$ for a fixed $\widehat{\mathbf{Y}}$, leading to the objective function $\frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{AB}\|_F^2$. For this case, we establish the following lemma.

▶ **Lemma 1.** *For any given matrix $\widehat{\mathbf{Y}}$, any local minimum of $f(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{AB}\|_F^2$ is globally optimal, given by $\mathsf{svd}_r(\widehat{\mathbf{Y}})$. The objective function $f(\mathbf{A}, \mathbf{B})$ around any saddle point has a negative second-order directional curvature. Moreover, $f(\mathbf{A}, \mathbf{B})$ has no local maximum.*[2]

The proof of Lemma 1 is basically to calculate the gradient of $f(\mathbf{A}, \mathbf{B})$ and let it equal to zero. Given this lemma, we can reduce $F(\mathbf{A}, \mathbf{B})$ to the form $\frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{AB}\|_F^2$ for some $\widehat{\mathbf{Y}}$ plus an extra term:

$$\begin{aligned} F(\mathbf{A}, \mathbf{B}) &= \frac{1}{2}\|\mathbf{AB}\|_F^2 + H(\mathbf{AB}) = \frac{1}{2}\|\mathbf{AB}\|_F^2 + H^{**}(\mathbf{AB}) \\ &= \max_{\mathbf{\Lambda}} \frac{1}{2}\|\mathbf{AB}\|_F^2 + \langle \mathbf{\Lambda}, \mathbf{AB} \rangle - H^*(\mathbf{\Lambda}) \\ &= \max_{\mathbf{\Lambda}} \frac{1}{2}\| - \mathbf{\Lambda} - \mathbf{AB}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - H^*(\mathbf{\Lambda}) \triangleq \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}), \end{aligned} \tag{5}$$

where we define $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) \triangleq \frac{1}{2}\| - \mathbf{\Lambda} - \mathbf{AB}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - H^*(\mathbf{\Lambda})$ as the Lagrangian of problem $(\mathbf{P})$,[3] and the second equality holds because H is closed and convex w.r.t. the

[2] Prior work studying the loss surface of low-rank matrix approximation assumes that the matrix $\widetilde{\mathbf{\Lambda}}$ is of full rank and does not have the same singular values [8]. In this work, we generalize this result by removing these two assumptions.

[3] One can easily check that $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = \min_{\mathbf{M}} L'(\mathbf{A}, \mathbf{B}, \mathbf{M}, \mathbf{\Lambda})$, where $L'(\mathbf{A}, \mathbf{B}, \mathbf{M}, \mathbf{\Lambda})$ is the Lagrangian of the constraint optimization problem $\min_{\mathbf{A}, \mathbf{B}, \mathbf{M}} \frac{1}{2}\|\mathbf{AB}\|_F^2 + H(\mathbf{M})$, s.t. $\mathbf{M} = \mathbf{AB}$. With a little abuse of notation, we call $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$ the Lagrangian of the unconstrained problem $(\mathbf{P})$ as well.

argument $\mathbf{AB}$. For any fixed value of $\mathbf{\Lambda}$, by Lemma 1, any local minimum of $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ is globally optimal, because minimizing $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ is equivalent to minimizing $\frac{1}{2}\|-\mathbf{\Lambda}-\mathbf{AB}\|_F^2$ for a fixed $\mathbf{\Lambda}$.

The remaining part of our analysis is to choose a proper $\widetilde{\mathbf{\Lambda}}$ such that $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$, so that $\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$ and problem $(\mathbf{P})$ have the same optimal solution $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$. For this, we introduce the following condition, and later we will show that the condition holds with high probability.

▶ **Condition 2.** *For a solution $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ to problem $(\mathbf{P})$, there exists an $\widetilde{\mathbf{\Lambda}}\in\partial_{\mathbf{X}}H(\mathbf{X})|_{\mathbf{X}=\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}}$ such that*

$$-\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T=\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{B}}^T \quad \textit{and} \quad \widetilde{\mathbf{A}}^T(-\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})=\widetilde{\mathbf{A}}^T\widetilde{\mathbf{\Lambda}}. \tag{6}$$

Explanation of Condition 2. We note that $\nabla_{\mathbf{A}}L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})=\mathbf{ABB}^T+\mathbf{\Lambda B}^T$ and $\nabla_{\mathbf{B}}L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})=\mathbf{A}^T\mathbf{AB}+\mathbf{A}^T\mathbf{\Lambda}$ for a fixed $\mathbf{\Lambda}$. In particular, if we set $\mathbf{\Lambda}$ to be the $\widetilde{\mathbf{\Lambda}}$ in (6), then $\nabla_{\mathbf{A}}L(\mathbf{A},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})|_{\mathbf{A}=\widetilde{\mathbf{A}}}=\mathbf{0}$ and $\nabla_{\mathbf{B}}L(\widetilde{\mathbf{A}},\mathbf{B},\widetilde{\mathbf{\Lambda}})|_{\mathbf{B}=\widetilde{\mathbf{B}}}=\mathbf{0}$. So Condition 2 implies that $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ is either a saddle point or a local minimizer of $L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$ as a function of $(\mathbf{A},\mathbf{B})$ for the fixed $\widetilde{\mathbf{\Lambda}}$.

The following lemma states that if it is a local minimizer, then strong duality holds.

▶ **Lemma 3** (Dual Certificate). *Let $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ be a global minimizer of $F(\mathbf{A},\mathbf{B})$. If there exists a dual certificate $\widetilde{\mathbf{\Lambda}}$ satisfying Condition 2 and the pair $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ is a local minimizer of $L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$ for the fixed $\widetilde{\mathbf{\Lambda}}$, then strong duality holds. Moreover, we have the relation $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}=\mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}})$.*

Proof. By the assumption of the lemma, $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ is a local minimizer of $L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})=\frac{1}{2}\|-\widetilde{\mathbf{\Lambda}}-\mathbf{AB}\|_F^2+c(\widetilde{\mathbf{\Lambda}})$, where $c(\widetilde{\mathbf{\Lambda}})$ is a function that is independent of $\mathbf{A}$ and $\mathbf{B}$. So according to Lemma 1, $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})=\operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$, namely, $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ globally minimizes $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ when $\mathbf{\Lambda}$ is fixed to $\widetilde{\mathbf{\Lambda}}$. Furthermore, $\widetilde{\mathbf{\Lambda}}\in\partial_{\mathbf{X}}H(\mathbf{X})|_{\mathbf{X}=\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}}$ implies that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\in\partial_{\mathbf{\Lambda}}H^*(\mathbf{\Lambda})|_{\mathbf{\Lambda}=\widetilde{\mathbf{\Lambda}}}$ by the convexity of function H, meaning that $\mathbf{0}\in\partial_{\mathbf{\Lambda}}L(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\mathbf{\Lambda})$. So $\widetilde{\mathbf{\Lambda}}=\operatorname{argmax}_{\mathbf{\Lambda}} L(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\mathbf{\Lambda})$ due to the concavity of $L(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\mathbf{\Lambda})$ w.r.t. variable $\mathbf{\Lambda}$. Thus $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$.

We now prove the strong duality. By the fact that $F(\mathbf{A},\mathbf{B})=\max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ and that $\widetilde{\mathbf{\Lambda}}=\operatorname{argmax}_{\mathbf{\Lambda}} L(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\mathbf{\Lambda})$, we have $F(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})=L(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})\le L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}}),\ \ \forall\mathbf{A},\mathbf{B}$, where the inequality holds because $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of L. So on the one hand, $\min_{\mathbf{A},\mathbf{B}}\max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})=F(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})\le\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})\le\max_{\mathbf{\Lambda}}\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$. On the other hand, by weak duality, $\min_{\mathbf{A},\mathbf{B}}\max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})\ge\max_{\mathbf{\Lambda}}\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$. Therefore, we have $\min_{\mathbf{A},\mathbf{B}}\max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})=\max_{\mathbf{\Lambda}}\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$, i.e., strong duality holds. Hence,

$$\begin{aligned}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}&=\operatorname*{argmin}_{\mathbf{AB}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})=\operatorname*{argmin}_{\mathbf{AB}}\frac{1}{2}\|-\widetilde{\mathbf{\Lambda}}-\mathbf{AB}\|_F^2-\frac{1}{2}\|\widetilde{\mathbf{\Lambda}}\|_F^2-H^*(\widetilde{\mathbf{\Lambda}})\\&=\operatorname*{argmin}_{\mathbf{AB}}\frac{1}{2}\|-\widetilde{\mathbf{\Lambda}}-\mathbf{AB}\|_F^2=\mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}}).\end{aligned}$$ ◀

This lemma then leads to the following theorem.

▶ **Theorem 4.** *Denote by $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ the optimal solution of problem $(\mathbf{P})$. Define a matrix space $\mathcal{T}\triangleq\{\widetilde{\mathbf{A}}\mathbf{X}^T+\mathbf{Y}\widetilde{\mathbf{B}},\ \mathbf{X}\in\mathbb{R}^{n_2\times r},\ \mathbf{Y}\in\mathbb{R}^{n_1\times r}\}$. Then strong duality holds for problem $(\mathbf{P})$, provided that*

$$(1)\ \widetilde{\mathbf{\Lambda}}\in\partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})\triangleq\Psi,\qquad (2)\ \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}})=\widetilde{\mathbf{A}}\widetilde{\mathbf{B}},\qquad (3)\ \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\|<\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \tag{7}$$

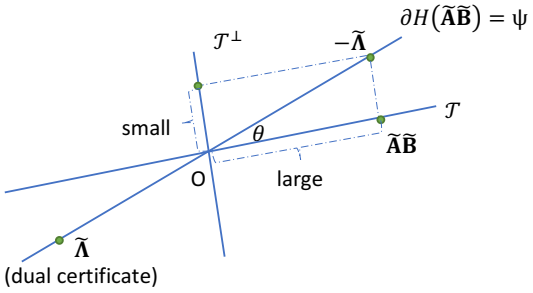


Figure 3 Geometry of dual condition (7) for general matrix factorization problems.

Proof. The proof idea is to construct a dual certificate $\widetilde{\mathbf{\Lambda}}$ so that the conditions in Lemma 3 hold. We note that $\widetilde{\mathbf{\Lambda}}$ should satisfy the following:

$$\begin{aligned}
&\text{(a)} \quad \widetilde{\mathbf{\Lambda}} \in \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \qquad \text{(by Condition 2)} \\
&\text{(b)} \quad (\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}})\widetilde{\mathbf{B}}^T = \mathbf{0} \quad \text{and} \quad \widetilde{\mathbf{A}}^T(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}}) = \mathbf{0}, \qquad \text{(by Condition 2)} \\
&\text{(c)} \quad \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}}). \qquad \text{(by the local minimizer assumption and Lemma 1)}
\end{aligned} \tag{8}$$

By the definition of $\mathcal{T}$ in the theorem statement, it turns out that for any matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, we have $\mathcal{P}_{\mathcal{T}^\perp}\mathbf{M} = (\mathbf{I} - \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger)\mathbf{M}(\mathbf{I} - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\dagger)$ and so $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{M}\| \le \|\mathbf{M}\|$, a fact that we will frequently use in the subsequent parts of the paper. Denote by $\mathcal{U}$ the left singular space of $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ and $\mathcal{V}$ the right singular space. Then the linear space $\mathcal{T}$ can be equivalently represented as $\mathcal{T} = \mathcal{U} + \mathcal{V}$ by the definition. Therefore, we have $\mathcal{T}^\perp = (\mathcal{U}+\mathcal{V})^\perp = \mathcal{U}^\perp \cap \mathcal{V}^\perp$. With this, we note that: (b) $(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}+\widetilde{\mathbf{\Lambda}})\widetilde{\mathbf{B}}^T = \mathbf{0}$ and $\widetilde{\mathbf{A}}^T(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}+\widetilde{\mathbf{\Lambda}}) = \mathbf{0}$ imply $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}+\widetilde{\mathbf{\Lambda}} \in \mathsf{Null}(\widetilde{\mathbf{A}}^T) = \mathsf{Col}(\widetilde{\mathbf{A}})^\perp$ and $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}} \in \mathsf{Row}(\widetilde{\mathbf{B}})^\perp$ (so $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}} \in \mathcal{T}^\perp$), and vice versa, where $\mathsf{Null}(\mathbf{Y})$, $\mathsf{Col}(\mathbf{Y})$, $\mathsf{Row}(\mathbf{Y})$ represent the null space, the row space, the column space of any given matrix $\mathbf{Y}$, respectively. And (c) $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}})$ implies that for an orthogonal decomposition $-\widetilde{\mathbf{\Lambda}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \mathbf{E}$, where $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} \in \mathcal{T}$, and $\mathbf{E} \in \mathcal{T}^\perp$, we have $\|\mathbf{E}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. Conversely, $\|\mathbf{E}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$ and condition (b) imply $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}})$. Therefore, the dual conditions (a), (b), and (c) in (8) are equivalent to (1) $\widetilde{\mathbf{\Lambda}} \in \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \triangleq \Psi$; (2) $\mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$; (3) $\|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$, as desired. ◀

To show the dual condition in Theorem 4, intuitively, we need to show that the angle θ between subspace $\mathcal{T}$ and Ψ is small (see Figure 3) for a specific function $H(\cdot)$. In the following (see Section B), we will demonstrate applications that, with randomness, obey this dual condition with high probability.

4 Matrix Completion

In matrix completion, there is a hidden matrix $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ with rank r. We are given measurements $\{\mathbf{X}^*_{ij} : (i,j) \in \Omega\}$, where $\Omega \sim \mathsf{Uniform}(m)$, i.e., Ω is sampled uniformly at random from all subsets of $[n_1] \times [n_2]$ of cardinality m. The goal is to exactly recover $\mathbf{X}^*$ with high probability. Here we apply our unified framework in Section 3 to matrix completion, by setting $H(\cdot) = \mathbf{I}_{\{\mathbf{M}:\mathcal{P}_\Omega(\mathbf{M})=\mathcal{P}_\Omega(\mathbf{X}^*)\}}(\cdot)$.

A quantity governing the difficulties of matrix completion is the incoherence parameter μ. Intuitively, matrix completion is possible only if the information spreads evenly throughout the low-rank matrix. This intuition is captured by the incoherence conditions. Formally, denote by $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ the skinny SVD of a fixed $n_1 \times n_2$ matrix $\mathbf{X}$ of rank r. Candès et

al. [13, 14, 46, 61] introduced the μ-incoherence condition (3) to the low-rank matrix $\mathbf{X}$. For conditions (3), it can be shown that $1 \leq \mu \leq \frac{n_{(1)}}{r}$. The condition holds for many random matrices with incoherence parameter μ about $\sqrt{r \log n_{(1)}}$ [40].

We have two positive results. The first result is an information-theoretic upper bound: with the standard incoherence condition (3), $\mathbf{X}^*$ is the unique matrix of rank at most r that is consistent with the observations. The proof is deferred to Appendix A.

▶ **Theorem 5** (Information-Theoretic Upper Bound). *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Suppose that $m \geq c\mu n_{(1)} r \log n_{(1)}$ for an absolute constant c. Then $\mathbf{X}^*$ is the unique $n_1 \times n_2$ matrix of rank at most r with μ-incoherence condition (3) such that $\mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*)$, with probability at least $1 - n_{(1)}^{-10}$.*

Proof Sketch. We consider the feasibility of the matrix completion problem:

$$\text{Find a matrix } \mathbf{X} \in \mathbb{R}^{n_1 \times n_2} \text{ such that } \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*), \quad \mathsf{rank}(\mathbf{X}) \leq r. \tag{9}$$

Our proof first identifies a feasibility condition for problem (9), and then shows that $\mathbf{X}^*$ is the only matrix which obeys this feasibility condition when the sample size is large enough. More specifically, we note that $\mathbf{X}^*$ obeys the conditions in problem (9). Therefore, $\mathbf{X}^*$ is the only matrix which obeys condition (9) if and only if $\mathbf{X}^* + \mathbf{D}$ does not follow the condition for all $\mathbf{D}$, i.e., $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$, where $\mathcal{D}_\mathcal{S}(\mathbf{X}^*)$ is the descent cone of all low-rank matrices. We note that the descent cone $\mathcal{D}_\mathcal{S}(\mathbf{X}^*)$ is contained in the subspace $\mathcal{T}$ by the tool of geometry functional analysis. Thus by a well-known fact that $\mathcal{T} \cap \Omega^\perp = \{\mathbf{0}\}$ when the sample size is large, the proof is completed. ◀

We describe a simple finite-time inefficient algorithm given Theorem 5 in Section C. This positive result matches a lower bound from prior work, which claims that the sample complexity in Theorem 5 is optimal.

▶ **Theorem 6** (Information-Theoretic Lower Bound. [15], Theorem 1.7). *Denote by $\Omega \sim \mathsf{Uniform}(m)$ the support set uniformly distributed among all sets of cardinality m. Suppose that $m \leq c\mu n_{(1)} r \log n_{(1)}$ for an absolute constant c. Then there exist infinitely many $n_1 \times n_2$ matrices $\mathbf{X}'$ of rank at most r obeying μ-incoherence (3) such that $\mathcal{P}_\Omega(\mathbf{X}') = \mathcal{P}_\Omega(\mathbf{X}^*)$, with probability at least $1 - n_{(1)}^{-10}$.*

Our second positive result converts the feasibility problem in Theorem 5 to a convex optimization problem, which can be *efficiently* solved.

▶ **Theorem 7** (Efficient Matrix Completion). *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Suppose $\mathbf{X}^*$ has condition number $\kappa = \sigma_1(\mathbf{X}^*)/\sigma_r(\mathbf{X}^*)$. Then there are absolute constants c and c_0 such that with probability at least $1 - c_0 n_{(1)}^{-10}$, the output of the convex problem*

$$\widetilde{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{X}} \|\mathbf{X}\|_{r*}, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*), \tag{10}$$

is unique and exact, i.e., $\widetilde{\mathbf{X}} = \mathbf{X}^*$, provided that $m \geq c\kappa^2 \mu r n_{(1)} \log_{2\kappa}(n_{(1)}) \log(n_{(1)})$ and $\mathbf{X}^*$ obeys μ-incoherence (3).

Proof Sketch. We have shown in Theorem 5 that $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{AB}\|_F^2$, s.t. $\mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*)$ exactly recovers $\mathbf{X}^*$, i.e., $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$, with the optimal sample complexity. So if strong duality holds, this non-convex optimization problem can be equivalently

converted to the convex program (10). Then Theorem 7 is straightforward from strong duality.

It now suffices to apply our unified framework in Section 3 to prove the strong duality. We show that the dual condition in Theorem 4 holds with high probability by the following arguments. Let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ be a global solution to problem (10). For $H(\mathbf{X}) = \mathbf{I}_{\{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}:\ \mathcal{P}_\Omega\mathbf{M}=\mathcal{P}_\Omega\mathbf{X}^*\}}(\mathbf{X})$, we have

$$\begin{aligned}\Psi &= \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \\ &= \{\mathbf{G}\in\mathbb{R}^{n_1\times n_2} : \langle \mathbf{G}, \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\rangle \geq \langle \mathbf{G}, \mathbf{Y}\rangle, \text{ for any } \mathbf{Y}\in\mathbb{R}^{n_1\times n_2} \text{ s.t. } \mathcal{P}_\Omega\mathbf{Y} = \mathcal{P}_\Omega\mathbf{X}^*\} \\ &= \{\mathbf{G}\in\mathbb{R}^{n_1\times n_2} : \langle \mathbf{G}, \mathbf{X}^*\rangle \geq \langle \mathbf{G}, \mathbf{Y}\rangle, \text{ for any } \mathbf{Y}\in\mathbb{R}^{n_1\times n_2} \text{ s.t. } \mathcal{P}_\Omega\mathbf{Y} = \mathcal{P}_\Omega\mathbf{X}^*\} = \Omega,\end{aligned}$$

where the third equality holds since $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$. Then we only need to show

$$(1)\ \widetilde{\mathbf{\Lambda}}\in\Omega, \qquad (2)\ \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \qquad (3)\ \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{2}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \tag{11}$$

It is interesting to see that dual condition (11) can be satisfied if the angle θ between subspace Ω and subspace $\mathcal{T}$ is very small; see Figure 3. When the sample size $|\Omega|$ becomes larger and larger, the angle θ becomes smaller and smaller (e.g., when $|\Omega| = n_1n_2$, the angle θ is zero as $\Omega = \mathbb{R}^{n_1\times n_2}$). We show that the sample size $m \geq \Omega(\kappa^2\mu r n_{(1)}\log_{2\kappa}(n_{(1)})\log(n_{(1)}))$ is a sufficient condition for condition (11) to hold. ◀

5 Robust Principal Component Analysis

In this section, we develop our theory for robust PCA based on our framework. In the problem of robust PCA, we are given an observed matrix of the form $\mathbf{D} = \mathbf{X}^* + \mathbf{S}^*$, where $\mathbf{X}^*$ is the ground-truth matrix and $\mathbf{S}^*$ is the corruption matrix which is sparse. The goal is to recover the hidden matrices $\mathbf{X}^*$ and $\mathbf{S}^*$ from the observation $\mathbf{D}$. We set $H(\mathbf{X}) = \lambda\|\mathbf{D}-\mathbf{X}\|_1$.

To make the information spread evenly throughout the matrix, the matrix cannot have one entry whose absolute value is significantly larger than other entries. In this work, we make the following incoherence assumption for robust PCA:

$$\|\mathbf{X}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1n_2}}\sigma_r(\mathbf{X}^*). \tag{12}$$

Note that condition (12) has an intuitive explanation, namely, that the entries must scatter almost uniformly across the low-rank matrix.

We have the following results for robust PCA.

▶ **Theorem 8** (Robust PCA). *Suppose $\mathbf{X}^*$ is an $n_1\times n_2$ matrix of rank r, and obeys incoherence (3) and (12). Assume that the support set Ω of $\mathbf{S}^*$ is uniformly distributed among all sets of cardinality m. Then with probability at least $1 - cn_{(1)}^{-10}$, the output of the optimization problem*

$$(\widetilde{\mathbf{X}}, \widetilde{\mathbf{S}}) = \underset{\mathbf{X},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{X}\|_{r*} + \lambda\|\mathbf{S}\|_1, \quad \text{s.t.} \quad \mathbf{D} = \mathbf{X} + \mathbf{S},$$

with $\lambda = \frac{\sigma_r(\mathbf{X}^*)}{\sqrt{n_{(1)}}}$ is exact, namely, $\widetilde{\mathbf{X}} = \mathbf{X}^*$ and $\widetilde{\mathbf{S}} = \mathbf{S}^*$, if $\mathsf{rank}(\mathbf{X}^*) \leq \rho_r\frac{n_{(2)}}{\mu\log^2 n_{(1)}}$ and $m \leq \rho_s n_1n_2$, where c, ρ_r, and ρ_s are all positive absolute constants, and function $\|\cdot\|_{r*}$ is given by (13).

The bounds on the rank of $\mathbf{X}^*$ and the sparsity of $\mathbf{S}^*$ in Theorem 8 match the best known results for robust PCA in prior work when we assume the support set of $\mathbf{S}^*$ is sampled uniformly [13].

6 Computational Aspects

Computational Efficiency. We discuss our computational efficiency given that we have strong duality. We note that the dual and bi-dual of primal problem (**P**) are given by

$$\begin{aligned} &(\textbf{Dual, D1}) \qquad \max_{\mathbf{\Lambda}\in\mathbb{R}^{n_1\times n_2}} -H^*(\mathbf{\Lambda}) - \frac{1}{2}\|\mathbf{\Lambda}\|_r^2, \quad \text{where } \|\mathbf{\Lambda}\|_r^2 = \sum_{i=1}^{r}\sigma_i^2(\mathbf{\Lambda}), \\ &(\textbf{Bi-Dual, D2}) \qquad \min_{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}} H(\mathbf{M}) + \|\mathbf{M}\|_{r*}, \quad \text{where } \|\mathbf{M}\|_{r*} = \max_{\mathbf{X}}\langle \mathbf{M}, \mathbf{X}\rangle - \frac{1}{2}\|\mathbf{X}\|_r^2. \end{aligned} \tag{13}$$

Problems (**D1**) and (**D2**) can be solved efficiently due to their convexity. In particular, Grussler et al. [26] provided a computationally efficient algorithm to compute the proximal operators of functions $\frac{1}{2}\|\cdot\|_r^2$ and $\|\cdot\|_{r*}$. Hence, the Douglas-Rachford algorithm can find the global minimum up to an ϵ error in function value in time $\mathsf{poly}(1/\epsilon)$ [33].

Computational Lower Bounds. Unfortunately, strong duality does not always hold for general non-convex problems (**P**). Here we present a very strong lower bound based on the random 4-SAT hypothesis. This is by now a fairly standard conjecture in complexity theory [19] and gives us constant factor inapproximability of problem (**P**) for deterministic algorithms, even those running in exponential time.

If we additionally assume that $\mathsf{BPP} = \mathsf{P}$, where BPP is the class of problems which can be solved in probabilistic polynomial time, and P is the class of problems which can be solved in deterministic polynomial time, then the same conclusion holds for randomized algorithms. This is also a standard conjecture in complexity theory, as it is implied by the existence of certain strong pseudorandom generators or if any problem in deterministic exponential time has exponential size circuits [34]. Therefore, any subexponential time algorithm achieving a sufficiently small constant factor approximation to problem (**P**) in general would imply a major breakthrough in complexity theory.

The lower bound is proved by a reduction from the Maximum Edge Biclique problem [4].

▶ **Theorem 9** (Computational Lower Bound). *Assume Conjecture 20 (the hardness of Random 4-SAT). Then there exists an absolute constant $\epsilon_0 > 0$ for which any deterministic algorithm achieving $(1+\epsilon)\mathsf{OPT}$ in the objective function value for problem (**P**) with $\epsilon \le \epsilon_0$, requires $2^{\Omega(n_1+n_2)}$ time, where OPT is the optimum. If in addition, $\mathsf{BPP} = \mathsf{P}$, then the same conclusion holds for randomized algorithms succeeding with probability at least $2/3$.*

Proof Sketch. Theorem 9 is proved by using the hypothesis that random 4-SAT is hard, in order to show hardness of the Maximum Edge Biclique problem for deterministic algorithms. We then do a reduction from the Maximum Edge Biclique problem to our problem. ◀

Due to space constraints, we defer the proofs of Lemma 1, Theorem 8, some synthetic experiments, and other related work to our full version on arXiv. The proofs of other theorems/lemmas can be found in the appendices.

Acknowledgments. We thank Rong Ge, Zhouchen Lin, and Benjamin Recht for useful discussions. We would like to thank Rina Foygel for finding a bug in the proof of Theorem 7 in a previous version.

References

1. Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, pages 1171–1197, 2012.
2. Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. In *ACM Symposium on Theory of Computing*, pages 1195–1199, 2017.
3. Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *ACM Symposium on Theory of Computing*, 2017.
4. Christoph Ambühl, Monaldo Mastrolilli, and Ola Svensson. Inapproximability results for maximum edge biclique, minimum linear arrangement, and sparsest cut. *SIAM Journal on Computing*, 40(2):567–596, 2011.
5. Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *Annual Conference on Learning Theory*, pages 81–102, 2016.
6. Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Annual Conference on Learning Theory*, pages 152–192, 2016.
7. Maria-Florina Balcan and Hongyang Zhang. Noise-tolerant life-long matrix completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 2955–2963, 2016.
8. Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.
9. Saugata Basu, Richard Pollack, and Marie Francoise Roy. On the combinatorial and algebraic complexity of quantifier elimination. *Journal of the ACM*, 43(6):1002–1045, 1996.
10. Amir Beck and Yonina C Eldar. Strong duality in nonconvex quadratic optimization with two quadratic constraints. *SIAM Journal on Optimization*, 17(3):844–860, 2006.
11. Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.
12. Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
13. Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.
14. Emmanuel J. Candès and Ben Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
15. Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
16. Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.
17. Ji Chen and Xiaodong Li. Memory-efficient kernel PCA via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. *arXiv preprint arXiv:1711.01742*, 2017.
18. Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.
19. Uriel Feige. Relations between average case complexity and approximation complexity. In *Annual IEEE Conference on Computational Complexity*, page 5, 2002.
20. David Gamarnik, Quan Li, and Hongyi Zhang. Matrix completion from $O(n)$ samples in linear time. In *Annual Conference on Learning Theory*, 2017.

21 Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Annual Conference on Learning Theory*, pages 797–842, 2015.

22 Rong Ge, Chi Jin, and Zheng Yi. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *International Conference on Machine Learning*, 2017.

23 Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

24 Andreas Goerdt and André Lanka. An approximation hardness result for bipartite clique. In *Electronic Colloquium on Computational Complexity, Report*, volume 48, 2004.

25 D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

26 Christian Grussler, Anders Rantzer, and Pontus Giselsson. Low-rank optimization with convex constraints. *arXiv preprint arXiv:1606.01793*, 2016.

27 Quanquan Gu, Zhaoran Wang, and Han Liu. Low-rank and sparse structure pursuit via alternating minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 600–609, 2016.

28 Benjamin Haeffele, Eric Young, and Rene Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International Conference on Machine Learning*, pages 2007–2015, 2014.

29 Benjamin D Haeffele and René Vidal. Global optimality in neural network training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339, 2017.

30 Moritz Hardt. Understanding alternating minimization for matrix completion. In *IEEE Symposium on Foundations of Computer Science*, pages 651–660, 2014.

31 Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Annual Conference on Learning Theory*, pages 703–725, 2014.

32 Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. *Annual Conference on Learning Theory*, 2013.

33 Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

34 Russell Impagliazzo and Avi Wigderson. $P = BPP$ if E requires exponential circuits: Derandomizing the XOR lemma. In *ACM Symposium on the Theory of Computing*, pages 220–229, 1997.

35 Johannes Jahn. *Introduction to the theory of nonlinear optimization*. Springer Berlin Heidelberg, 2007.

36 Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

37 Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM Symposium on Theory of Computing*, pages 665–674, 2013.

38 Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *International Conference on Machine Learning*, 2017.

39 Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

40 Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

41 Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

42 Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *International Conference on Machine Learning*, pages 2358–2367, 2016.

43 Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

44 Michael L Overton and Robert S Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.

45 Ilya Razenshteyn, Zhao Song, and David P. Woodruff. Weighted low rank approximations with provable guarantees. In *ACM Symposium on Theory of Computing*, pages 250–263, 2016.

46 Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

47 James Renegar. On the computational complexity and geometry of the first-order theory of the reals, part I: introduction. preliminaries. the geometry of semi-algebraic sets. the decision problem for the existential theory of the reals. *Journal of symbolic computation*, 13(3):255–300, 1992.

48 James Renegar. On the computational complexity and geometry of the first-order theory of the reals, part II: the general decision problem. preliminaries for quantifier elimination. *Journal of Symbolic Computation*, 13(3):301–328, 1992.

49 Reinhold Schneider and André Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Lojasiewicz inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015.

50 Yuan Shen, Zaiwen Wen, and Yin Zhang. Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, 29(2):239–263, 2014.

51 Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *IEEE International Symposium on Information Theory*, pages 2379–2383, 2016.

52 Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

53 Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via nonconvex factorization. In *IEEE Symposium on Foundations of Computer Science*, pages 270–289, 2015.

54 Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *International Conference on Machine Learning*, 2016.

55 Roman Vershynin. Lectures in geometric functional analysis, 2009. URL: `https://www.math.uci.edu/~rvershyn/papers/GFA-book.pdf`.

56 Roman Vershynin. Estimation in high dimensions: A geometric perspective. In *Sampling theory, a renaissance*, pages 3–66. Springer, 2015.

57 Yu-Xiang Wang and Huan Xu. Stability of matrix factorization for collaborative filtering. In *International Conference on Machine Learning*, pages 417–424, 2012.

58 Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

59 Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

60 Hongyang Zhang, Zhouchen Lin, and Chao Zhang. A counterexample for the validity of using nuclear norm as a convex surrogate of rank. In *European Conference on Machine*

Learning and Principles and Practice of Knowledge Discovery in Databases, volume 8189, pages 226–241, 2013.

61 Hongyang Zhang, Zhouchen Lin, and Chao Zhang. Completing low-rank matrices with corrupted samples from few coefficients in general basis. *IEEE Transactions on Information Theory*, 62(8):4748–4768, 2016.

62 Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Edward Chang. Exact recoverability of robust PCA via outlier pursuit with tight recovery bounds. In *AAAI Conference on Artificial Intelligence*, pages 3143–3149, 2015.

63 Xiao Zhang, Lingxiao Wang, and Quanquan Gu. A nonconvex free lunch for low-rank plus sparse matrix recovery. *arXiv preprint arXiv:1702.06525*, 2017.

64 Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

65 Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

A Proof of Theorem 5

Theorem 5. (Information-Theoretic Upper Bound. Restated.) *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set, which is uniformly distributed among all sets of cardinality m. Suppose that $m \geq c\mu n_{(1)} r \log n_{(1)}$ for an absolute constant c. Then $\mathbf{X}^*$ is the unique $n_1 \times n_2$ matrix of rank at most r with μ-incoherence (3) such that $\mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*)$, with probability at least $1 - n_{(1)}^{-10}$.*

Proof. We note that the sampling model $\mathsf{Uniform}(m)$ is equivalent to the sampling model $\mathsf{Ber}(p)$ with $p = \Theta\left(\frac{m}{n_1 n_2}\right)$, which we will frequently use in the sequel. We consider the feasibility of the matrix completion problem:

$$\text{Find a matrix } \mathbf{X} \in \mathbb{R}^{n_1 \times n_2} \text{ such that } \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*), \quad \mathsf{rank}(\mathbf{X}) \leq r. \tag{14}$$

Our proof first identifies a feasibility condition for problem (14), and then shows that $\mathbf{X}^*$ is the only matrix that obeys this feasibility condition when the sample size is large enough. We denote by $\mathcal{S} = \{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \leq r\}$, and define $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) = \{t(\mathbf{X} - \mathbf{X}^*) \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \leq r,\ t \geq 0\}$. We have the following proposition for the feasibility of problem (14).

▶ **Proposition 10** (Feasibility Condition). *$\mathbf{X}^*$ is the unique feasible solution to problem (14) if $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$.*

Proof. Notice that problem (14) is equivalent to another feasibility problem

$$\text{Find a matrix } \mathbf{D} \in \mathbb{R}^{n_1 \times n_2} \text{ such that } \quad \mathsf{rank}(\mathbf{X}^* + \mathbf{D}) \leq r, \quad \mathbf{D} \in \Omega^\perp.$$

Suppose that $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$. Since $\mathsf{rank}(\mathbf{X}^* + \mathbf{D}) \leq r$ implies $\mathbf{D} \in \mathcal{D}_\mathcal{S}(\mathbf{X}^*)$, and note that $\mathbf{D} \in \Omega^\perp$, we have $\mathbf{D} = \mathbf{0}$, which means $\mathbf{X}^*$ is the unique feasible solution to problem (14). ◀

The remainder of the proof is to show $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$. To proceed, we note that the "escaping through a mesh" techniques for matrix sensing do not work for matrix completion since Ω is not drawn from the Grassmanian according to the Haar measure. To address this issue, we instead need the following lemmas. The first lemma claims that the tangent cone of the set $\mathcal{S}$ evaluated at $\mathbf{X}^*$ is slightly larger than the cone $\mathsf{cone}(\mathcal{S} - \{\mathbf{X}^*\})$.

▶ **Lemma 11** ([35], Theorem 4.8). *Let $\mathcal{S}$ be a non-empty subset of a real normed space. If $\mathcal{S}$ is star-shaped w.r.t. some $\mathbf{X}^* \in \mathcal{S}$, i.e., $t(\mathcal{S} - \{\mathbf{X}^*\}) \subseteq \mathcal{S} - \{\mathbf{X}^*\}$ for all $t \in [0,1]$, then it follows $\mathsf{cone}(\mathcal{S} - \{\mathbf{X}^*\}) \subseteq T(\mathcal{S}, \mathbf{X}^*)$, where $T(\mathcal{S}, \mathbf{X}^*)$ is the tangent cone of the set $\mathcal{S}$ at point $\mathbf{X}^*$ defined by $T(\mathcal{S}, \mathbf{X}^*) = \{\mathbf{\Xi} \in \mathbb{R}^{n_1 \times n_2} : \exists \mathbf{X}_n \subseteq \mathcal{S},\ (a_n) \subseteq \mathbb{R}^+ \text{ s.t. } \mathbf{X}_n \to \mathbf{X}^*,\ a_n(\mathbf{X}_n - \mathbf{X}^*) \to \mathbf{\Xi}\}$.*

The second lemma states that the tangent cone of $\mathcal{S}$ evaluated at $\mathbf{X}^*$ can be represented in a closed form.

▶ **Lemma 12** ([49], Theorem 3.2). *Let $\mathbf{X}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the skinny SVD of matrix $\mathbf{X}^*$. The tangent cone $T(\mathcal{S}, \mathbf{X}^*)$ of the set $\mathcal{S} = \{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \le r\}$ at $\mathbf{X}^*$ is a linear subspace given by $T(\mathcal{S}, \mathbf{X}^*) = \{\mathbf{U}\mathbf{L}^T + \mathbf{M}\mathbf{V}^T :\ \mathbf{L} \in \mathbb{R}^{n_2 \times r}, \mathbf{M} \in \mathbb{R}^{n_1 \times r}\} \triangleq \mathcal{T}$.*

Now we are ready to prove Theorem 5. By Lemma 11 and 12, we have $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) = \mathsf{cone}(\mathcal{S} - \{\mathbf{X}^*\}) \subseteq T(\mathcal{S}, \mathbf{X}^*) = \mathcal{T}$, where the first equality holds by the definition of $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*)$. So if $\mathcal{T} \cap \Omega^\perp = \{\mathbf{0}\}$, then $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$, meaning that $\mathbf{X}^*$ is the unique feasible solution to the problem (14). Thus the rest of proof is to find a sufficient condition for $\mathcal{T} \cap \Omega^\perp = \{\mathbf{0}\}$. We have the following lemma.

▶ **Lemma 13.** *Assume that $\Omega \sim \mathsf{Ber}(p)$ and the incoherence condition (3) holds. Then with probability at least $1 - n_{(1)}^{-10}$, we have $\|\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}}\| \le \sqrt{1 - p + \epsilon p}$, provided that $p \ge C_0 \epsilon^{-2}(\mu r \log n_{(1)})/n_{(2)}$, where C_0 is an absolute constant.*

Proof. If $\Omega \sim \mathsf{Ber}(p)$, we have, by Theorem 15, that with high probability $\|\mathcal{P}_{\mathcal{T}} - p^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}\| \le \epsilon$, provided that $p \ge C_0 \epsilon^{-2} \frac{\mu r \log n_{(1)}}{n_{(2)}}$. Note, however, that since $\mathcal{I} = \mathcal{P}_{\Omega} + \mathcal{P}_{\Omega^\perp}$, $\mathcal{P}_{\mathcal{T}} - p^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}} = p^{-1}(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}} - (1-p)\mathcal{P}_{\mathcal{T}})$ and, therefore, by the triangle inequality $\|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}}\| \le \epsilon p + (1-p)$. Since $\|\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}}\|^2 \le \|\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}}\|$, the proof is completed. ◀

We note that $\|\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}}\| < 1$ implies $\Omega^\perp \cap \mathcal{T} = \{\mathbf{0}\}$. The proof is completed. ◀

B Proof of Theorem 7

We have shown in Theorem 5 that the problem $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \mathrm{argmin}_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{A}\mathbf{B}\|_F^2, \text{s.t.} \mathcal{P}_{\Omega}(\mathbf{A}\mathbf{B}) = \mathcal{P}_{\Omega}(\mathbf{X}^*)$, exactly recovers $\mathbf{X}^*$, i.e., $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$, with the optimal sample complexity. So if strong duality holds, this non-convex optimization problem can be equivalently converted to the convex program (10). Then Theorem 7 is straightforward from strong duality.

It now suffices to apply our unified framework in Section 3 to prove the strong duality. We show that the dual condition in Theorem 4 holds with high probability. Let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ be a global solution to problem (10). For $H(\mathbf{X}) = \mathbf{I}_{\{\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}:\ \mathcal{P}_{\Omega}\mathbf{M} = \mathcal{P}_{\Omega}\mathbf{X}^*\}}(\mathbf{X})$, we have

$$\begin{aligned}
\Psi &= \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \\
&= \{\mathbf{G} \in \mathbb{R}^{n_1 \times n_2} : \langle \mathbf{G}, \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} \rangle \ge \langle \mathbf{G}, \mathbf{Y} \rangle,\ \text{for any } \mathbf{Y} \in \mathbb{R}^{n_1 \times n_2} \text{ s.t. } \mathcal{P}_{\Omega}\mathbf{Y} = \mathcal{P}_{\Omega}\mathbf{X}^*\} \\
&= \{\mathbf{G} \in \mathbb{R}^{n_1 \times n_2} : \langle \mathbf{G}, \mathbf{X}^* \rangle \ge \langle \mathbf{G}, \mathbf{Y} \rangle,\ \text{for any } \mathbf{Y} \in \mathbb{R}^{n_1 \times n_2} \text{ s.t. } \mathcal{P}_{\Omega}\mathbf{Y} = \mathcal{P}_{\Omega}\mathbf{X}^*\} = \Omega,
\end{aligned}$$

where the third equality holds since $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$. Then we only need to show

$$(1)\ \widetilde{\mathbf{\Lambda}} \in \Omega, \quad (2)\ \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \quad (3)\ \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{2}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \tag{15}$$

We have the following lemma.

▶ **Lemma 14.** *If we can construct an* $\mathbf{\Lambda}$ *such that*

$$\text{(a) } \mathbf{\Lambda} \in \Omega, \quad \text{(b) } \|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_F \le \sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \quad \text{(c) } \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| < \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \tag{16}$$

then we can construct an $\widetilde{\mathbf{\Lambda}}$ *such that Eqn.* (15) *holds with probability at least* $1 - n_{(1)}^{-10}$.

Proof. To prove the lemma, we first claim the following theorem.

▶ **Theorem 15** ([14], Theorem 4.1)**.** *Assume that* Ω *is sampled according to the Bernoulli model with success probability* $p = \Theta(\frac{m}{n_1 n_2})$*, and incoherence condition* (3) *holds. Then there is an absolute constant* C_R *such that for* $\beta > 1$*, we have*

$$\|p^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\| \le C_R\sqrt{\frac{\beta\mu n_{(1)} r\log n_{(1)}}{m}} \triangleq \epsilon,$$

with probability at least $1 - 3n^{-\beta}$ *provided that* $C_R\sqrt{\frac{\beta\mu n_{(1)} r\log n_{(1)}}{m}} < 1$.

Suppose that Condition (16) holds. Let $\mathbf{Y} = \widetilde{\mathbf{\Lambda}} - \mathbf{\Lambda} \in \Omega$ be the perturbation matrix between $\mathbf{\Lambda}$ and $\widetilde{\mathbf{\Lambda}}$ such that $\mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$. Such a $\mathbf{Y}$ exists by setting $\mathbf{Y} = \mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1}(\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. So $\|\mathcal{P}_{\mathcal{T}}\mathbf{Y}\|_F \le \sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. We now prove Condition (3) in Eqn. (15). Observe that

$$\|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| \le \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| \le \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\|. \tag{17}$$

So we only need to show $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| \le \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$.

Before proceeding, we begin by introducing a normalized version $\mathcal{Q}_{\Omega} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ of $\mathcal{P}_{\Omega}$: $\mathcal{Q}_{\Omega} = p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}$. With this, we have $\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}} = p\mathcal{P}_{\mathcal{T}}(\mathcal{I} + \mathcal{Q}_{\Omega})\mathcal{P}_{\mathcal{T}}$. Note that for any operator $\mathcal{P} : \mathcal{T} \to \mathcal{T}$, we have $\mathcal{P}^{-1} = \sum_{k\ge 0}(\mathcal{P}_{\mathcal{T}} - \mathcal{P})^k$ whenever $\|\mathcal{P}_{\mathcal{T}} - \mathcal{P}\| < 1$. So according to Theorem 15, the operator $p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1}$ can be represented as a *convergent* Neumann series $p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1} = \sum_{k\ge 0}(-1)^k(\mathcal{P}_{\mathcal{T}}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}})^k$, because $\|\mathcal{P}_{\mathcal{T}}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}}\| \le \epsilon < \frac{1}{2}$ once $m \ge C\mu n_{(1)} r\log n_{(1)}$ for a sufficiently large absolute constant C. We also note that $p(\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}}) = \mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}$, because $\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\mathcal{T}} = 0$. Thus

$$\begin{aligned}
\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| &= \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1}(\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\| \\
&= \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}}p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1}((\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\| \\
&= \|\sum_{k\ge 0}(-1)^k\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_{\Omega}(\mathcal{P}_{\mathcal{T}}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}})^k((\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\| \\
&\le \sum_{k\ge 0}\|(-1)^k\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_{\Omega}(\mathcal{P}_{\mathcal{T}}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}})^k((\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\|_F \\
&\le \|\mathcal{Q}_{\Omega}\|\sum_{k\ge 0}\|\mathcal{P}_{\mathcal{T}}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}}\|^k\|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\|_F \\
&\le \frac{4}{p}\|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}})\|_F \le \Theta\left(\frac{n_1 n_2}{m}\right)\sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \le \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})
\end{aligned}$$

with high probability. The proof is completed. ◀

It thus suffices to construct a dual certificate $\mathbf{\Lambda}$ such that all conditions in (16) hold. To this end, partition $\Omega = \Omega_1 \cup \Omega_2 \cup ... \cup \Omega_b$ into b partitions of size q. By assumption, we may choose

$$q \geq \frac{128}{3} C\beta\kappa^2 \mu r n_{(1)} \log n_{(1)} \quad \text{and} \quad b \geq \frac{1}{2} \log_{2\kappa} \left(24^2 n_{(1)}^2 \kappa^2\right)$$

for a sufficiently large constant C. Let $\Omega_j \sim \mathsf{Ber}(q)$ denote the set of indices corresponding to the j-th partitions. Define $\mathbf{W}_0 = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ and set $\mathbf{\Lambda}_k = \frac{n_1 n_2}{q} \sum_{j=1}^{k} \mathcal{P}_{\Omega_j}(\mathbf{W}_{j-1})$, $\mathbf{W}_k = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} - \mathcal{P}_{\mathcal{T}}(\mathbf{\Lambda}_k)$ for $k = 1, 2, ..., b$. Then by Theorem 15,

$$\begin{aligned} \|\mathbf{W}_k\|_F &= \left\| \mathbf{W}_{k-1} - \frac{n_1 n_2}{q} \mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega_k}(\mathbf{W}_{k-1}) \right\|_F \\ &= \left\| \left(\mathcal{P}_{\mathcal{T}} - \frac{n_1 n_2}{q} \mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega_k} \mathcal{P}_{\mathcal{T}} \right) (\mathbf{W}_{k-1}) \right\|_F \leq \frac{1}{2\kappa} \|\mathbf{W}_{k-1}\|_F. \end{aligned}$$

So it follows that $\|\mathbf{W}_b\|_F \leq (2\kappa)^{-b} \|\mathbf{W}_0\|_F \leq (2\kappa)^{-b} \sqrt{r} \sigma_1(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \leq \sqrt{\frac{r}{24^2 n_{(1)}^2}} \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$.

The following lemma together implies the strong duality of (10) straightforwardly.

▶ **Lemma 16.** *Under the assumptions of Theorem 7, the dual certification $\mathbf{W}_b$ obeys the dual condition* (16) *with probability at least $1 - n_{(1)}^{-10}$.*

Proof. It is well known that for matrix completion, the Uniform model $\Omega \sim \mathsf{Uniform}(m)$ is equivalent to the Bernoulli model $\Omega \sim \mathsf{Ber}(p)$, where each element in $[n_1] \times [n_2]$ is included with probability $p = \Theta(m/(n_1 n_2))$ independently. By the equivalence, we can suppose $\Omega \sim \mathsf{Ber}(p)$.

To prove Lemma 16, as a preliminary, we need the following lemmas.

▶ **Lemma 17** ([18], Lemma 2)**.** *Suppose $\mathbf{Z}$ is a fixed matrix. Suppose $\Omega \sim \mathsf{Ber}(p)$. Then with high probability, $\|(\mathcal{I} - p^{-1}\mathcal{P}_\Omega)\mathbf{Z}\| \leq C_0' \left(\frac{\log n_{(1)}}{p} \|\mathbf{Z}\|_\infty + \sqrt{\frac{\log n_{(1)}}{p}} \|\mathbf{Z}\|_{\infty,2} \right)$, where $C_0' > 0$ is an absolute constant and $\|\mathbf{Z}\|_{\infty,2} = \max \left\{ \max_i \sqrt{\sum_b \mathbf{Z}_{ib}^2}, \max_j \sqrt{\sum_a \mathbf{Z}_{aj}^2} \right\}$.*

▶ **Lemma 18** ([13], Lemma 3.1)**.** *Suppose $\Omega \sim \mathsf{Ber}(p)$ and $\mathbf{Z}$ is a fixed matrix. Then with high probability, $\|\mathbf{Z} - p^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega \mathbf{Z}\|_\infty \leq \epsilon \|\mathbf{Z}\|_\infty$, provided that $p \geq C_0 \epsilon^{-2} (\mu r \log n_{(1)})/n_{(2)}$ for some absolute constant $C_0 > 0$.*

▶ **Lemma 19** ([18], Lemma 3)**.** *Suppose that $\mathbf{Z}$ is a fixed matrix and $\Omega \sim \mathsf{Ber}(p)$. If $p \geq c_0 \mu r \log n_{(1)}/n_{(2)}$ for some c_0 sufficiently large, then by high probability, $\|(p^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega - \mathcal{P}_{\mathcal{T}})\mathbf{Z}\|_{\infty,2} \leq \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}} \|\mathbf{Z}\|_\infty + \frac{1}{2}\|\mathbf{Z}\|_{\infty,2}$.*

Observe that by Lemma 18, $\|\mathbf{W}_j\|_\infty \leq \left(\frac{1}{2}\right)^j \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty$, and by Lemma 19, we have $\|\mathbf{W}_j\|_{\infty,2} \leq \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}} \|\mathbf{W}_{j-1}\|_\infty + \frac{1}{2}\|\mathbf{W}_{j-1}\|_{\infty,2}$. So

$$\begin{aligned} \|\mathbf{W}_j\|_{\infty,2} &\leq \left(\frac{1}{2}\right)^j \sqrt{\frac{n_{(1)}}{\mu r}} \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + \frac{1}{2}\|\mathbf{W}_{j-1}\|_{\infty,2} \\ &\leq j \left(\frac{1}{2}\right)^j \sqrt{\frac{n_{(1)}}{\mu r}} \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + \left(\frac{1}{2}\right)^j \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}. \end{aligned}$$

Therefore, we have $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \leq \sum_{j=1}^{b} \|\frac{n_1 n_2}{q} \mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1}\| = \sum_{j=1}^{b} \|\mathcal{P}_{\mathcal{T}^\perp}(\frac{n_1 n_2}{q}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1} - \mathbf{W}_{j-1})\| \leq \sum_{j=1}^{b} \|(\frac{n_1 n_2}{q}\mathcal{P}_{\Omega_j} - \mathcal{I})(\mathbf{W}_{j-1})\|$. Let p denote $\Theta\left(\frac{q}{n_1 n_2}\right)$. By Lemma 17,

$$\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \le C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_{\infty,2}$$

$$\le C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\left(\frac{1}{2}\right)^j\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\left[j\left(\frac{1}{2}\right)^j\sqrt{\frac{n_{(1)}}{\mu r}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + \left(\frac{1}{2}\right)^j\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}\right]$$

$$\le C_0'\frac{\log n_{(1)}}{p}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + 2C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sqrt{\frac{n_{(1)}}{\mu r}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}.$$

Setting $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$, we note the facts that (we assume WLOG $n_2 \ge n_1$)

$$\|\mathbf{X}^*\|_{\infty,2} = \max_i \|\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\|_2 \le \max_i \|\mathbf{e}_i^T\mathbf{U}\|\sigma_1(\mathbf{X}^*) \le \sqrt{\frac{\mu r}{n_1}}\sigma_1(\mathbf{X}^*) \le \sqrt{\frac{\mu r}{n_1}}\kappa\sigma_r(\mathbf{X}^*),$$

$$\|\mathbf{X}^*\|_\infty = \max_{ij}\langle\mathbf{X}^*, \mathbf{e}_i\mathbf{e}_j^T\rangle = \max_{ij}\langle\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \mathbf{e}_i\mathbf{e}_j^T\rangle = \max_{ij}\langle\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}, \mathbf{e}_j^T\mathbf{V}\rangle$$

$$\le \max_{ij}\|\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\|_2\|\mathbf{e}_j^T\mathbf{V}\|_2 \le \max_j\|\mathbf{X}^*\|_{\infty,2}\|\mathbf{e}_j^T\mathbf{V}\|_2 \le \frac{\mu r\kappa}{\sqrt{n_1 n_2}}\sigma_r(\mathbf{X}^*).$$

Substituting $p = \Theta\left(\frac{\kappa^2\mu r n_{(1)}\log(n_{(1)})\log_{2\kappa}(n_{(1)})}{n_1 n_2}\right)$, we obtain $\|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{1}{3}\sigma_r(\mathbf{X}^*)$. The proof is completed. ◀

C Matrix Completion by Information-Theoretic Upper Bound

Theorem 5 formulates matrix completion as a feasibility problem. However, it is a priori unclear if there is an algorithm for finding $\mathbf{X}^*$ with $\mathcal{O}(\mu n_{(1)} r\log n_{(1)})$ sample complexity and incoherence (3) via solving the feasibility problem. To answer this question, we mention that matrix completion can be solved in finite time under these minimum assumptions, namely, we note that the feasibility problem is equivalent to finding a zero of the polynomial $\sum_{(i,j)\in\Omega}(\mathbf{e}_i^T\mathbf{A}\mathbf{B}\mathbf{e}_j - \mathbf{X}^*_{ij})^2 = 0$ w.r.t. the $(n_1+n_2)r$ unknowns $\mathbf{A}$ and $\mathbf{B}$. Since $\mathbf{A}$ can be assumed to be orthogonal, if the entries of $\mathbf{X}^*$ can be written down with $\mathsf{poly}(n)$ bits, then $\|\mathbf{B}\|_F \le \exp(\mathsf{poly}(n))$, which means if one rounds each of the entries of $\mathbf{B}$ to the nearest additive grid multiple of $1/\exp(\mathsf{poly}(n))$, then we will get a rank-k matrix $\mathbf{B}$ where each entry represents the true entry of the optimal $\mathbf{B}$ up to additive $1/\exp(\mathsf{poly}(n))$ error (of course one cannot write down $\mathbf{B}$ in some cases if the entries are irrational). Such an $\mathbf{A}$ and $\mathbf{B}$ can be found in $\exp((n_1+n_2)r)$ time [47, 48, 9]. This gives an exponential time algorithm to solve the feasibility problem in Theorem 5 for matrix completion.

D Proof of Theorem 9

Our computational lower bound for problem (**P**) assumes the hardness of random 4-SAT.

▶ **Conjecture 20** (Random 4-SAT). *Let $c > \ln 2$ be a constant. Consider a random 4-SAT formula on n variables in which each clause has 4 literals, and in which each of the $16n^4$ clauses is picked independently with probability c/n^3. Then any algorithm which always outputs 1 when the random formula is satisfiable, and outputs 0 with probability at least $1/2$ when the random formula is unsatisfiable, must run in $2^{c'n}$ time on some input, where $c' > 0$ is an absolute constant.*

Based on Conjecture 20, we have the following computational lower bound for problem (**P**). We show that problem (**P**) is in general hard for deterministic algorithms. If we additionally assume BPP = P, then the same conclusion holds for randomized algorithms with high probability.

Theorem 9. (Computational Lower Bound. Restated.) *Assume Conjecture 20. Then there exists an absolute constant $\epsilon_0 > 0$ for which any algorithm that achieves $(1+\epsilon)$OPT in objective function value for problem $(\mathbf{P})$ with $\epsilon \leq \epsilon_0$, and with constant probability, requires $2^{\Omega(n_1+n_2)}$ time, where OPT is the optimum. If in addition, BPP = P, then the same conclusion holds for randomized algorithms succeeding with probability at least $2/3$.*

Proof. Theorem 9 is proved by using the hypothesis that random 4-SAT is hard to show hardness of the Maximum Edge Biclique problem for deterministic algorithms.

▶ **Definition 21** (Maximum Edge Biclique)**.** The problem is
Input: An n-by-n bipartite graph G.
Output: A k_1-by-k_2 complete bipartite subgraph of G, such that $k_1 \cdot k_2$ is maximized.

[24] showed that under the random 4-SAT assumption there exist two constants $\epsilon_1 > \epsilon_2 > 0$ such that no efficient deterministic algorithm is able to distinguish between bipartite graphs $G(U, V, E)$ with $|U| = |V| = n$ which have a clique of size $\geq (n/16)^2(1+\epsilon_1)$ and those in which all bipartite cliques are of size $\leq (n/16)^2(1+\epsilon_2)$. The reduction uses a bipartite graph G with at least tn^2 edges with large probability, for a constant t.

Given a given bipartite graph $G(U, V, E)$, define $H(\cdot)$ as follows. Define the matrix $\mathbf{Y}$ and $\mathbf{W}$: $\mathbf{Y}_{ij} = 1$ if edge $(U_i, V_j) \in E$, $\mathbf{Y}_{ij} = 0$ if edge $(U_i, V_j) \notin E$; $\mathbf{W}_{ij} = 1$ if edge $(U_i, V_j) \in E$, and $\mathbf{W}_{ij} = \text{poly}(n)$ if edge $(U_i, V_j) \notin E$. Choose a large enough constant $\beta > 0$ and let $H(\mathbf{AB}) = \beta \sum_{ij} \mathbf{W}_{ij}^2 (\mathbf{Y}_{ij} - (\mathbf{AB})_{ij})^2$. Now, if there exists a biclique in G with at least $(n/16)^2(1+\epsilon_2)$ edges, then the number of remaining edges is at most $tn^2 - (n/16)^2(1+\epsilon_1)$, and so the solution to $\min H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2$ has cost at most $\beta[tn^2 - (n/16)^2(1+\epsilon_1)] + n^2$. On the other hand, if there does not exist a biclique that has more than $(n/16)^2(1+\epsilon_2)$ edges, then the number of remaining edges is at least $(n/16)^2(1+\epsilon_2)$, and so any solution to $\min H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2$ has cost at least $\beta[tn^2 - (n/16)^2(1+\epsilon_2)]$. Choose β large enough so that $\beta[tn^2 - (n/16)^2(1+\epsilon_2)] > \beta[tn^2 - (n/16)^2(1+\epsilon_1)] + n^2$. This combined with the result in [24] completes the proof for deterministic algorithms.

To rule out randomized algorithms running in time $2^{\alpha(n_1+n_2)}$ for some function α of n_1, n_2 for which $\alpha = o(1)$, observe that we can define a new problem which is the same as problem $(\mathbf{P})$ except the input description of H is padded with a string of 1s of length $2^{(\alpha/2)(n_1+n_2)}$. This string is irrelevant for solving problem $(\mathbf{P})$ but changes the input size to $N = \text{poly}(n_1, n_2) + 2^{(\alpha/2)(n_1+n_2)}$. By the argument in the previous paragraph, any deterministic algorithm still requires $2^{\Omega(n)} = N^{\omega(1)}$ time to solve this problem, which is super-polynomial in the new input size N. However, if a randomized algorithm can solve it in $2^{\alpha(n_1+n_2)}$ time, then it runs in $\text{poly}(N)$ time. This contradicts the assumption that BPP = P. This completes the proof. ◀