Journal of Machine Learning Research 20 (2019) 1-56 Submitted 10/17; Revised 11/18; Published 4/19

Non-Convex Matrix Completion and Related Problems via Strong Duality

Maria-Florina Balcan NINAMF@CS.CMU.EDU
Carnegie Mellon University

Yingyu Liang YLIANG@CS.WISC.EDU
University of Wisconsin-Madison

Zhao Song ZHAOS@G.HARVARD.EDU
UT-Austin & Harvard University

David P. Woodruff DWOODRUF@CS.CMU.EDU
Carnegie Mellon University

Hongyang Zhang* HONGYANZ@CS.CMU.EDU
Carnegie Mellon University & TTIC

Editor: Moritz Hardt

Abstract

This work studies the *strong duality of non-convex matrix factorization problems*: we show that under certain dual conditions, these problems and the 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 prove that under certain dual conditions, the optimal solution of the matrix factorization program is the same as that of 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 is 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. These are examples of efficiently recovering a hidden matrix given limited reliable observations. Our framework shows that exact recoverability and strong duality hold with nearly-optimal sample complexity for the two problems.

Keywords: strong duality, non-convex optimization, matrix factorization, matrix completion, robust principal component analysis, sample complexity

1. Introduction

Non-convex matrix factorization problems have been an emerging object of study in theoretical computer science (Balcan et al., 2019, 2018; Arora et al., 2012; Jain et al., 2013; Hardt, 2014; Sun and Luo, 2015; Moitra, 2016; Razenshteyn et al., 2016; Song et al., 2017, 2019), optimization (Wen et al., 2012; Shen et al., 2014), machine learning (Bhojanapalli et al., 2016b; Ge et al., 2016, 2015; Jain et al., 2010; Li et al., 2016; Wang and Xu, 2012), and many other domains (Bouwmans et al., 2018). In theoretical computer science and optimization, the study of such models has led to significant advances in provable algorithms that converge to local minima in linear time (Jain et al., 2013; Hardt,

*. Corresponding author.

2014; Sun and Luo, 2015; Agarwal et al., 2017; Allen-Zhu, 2017). 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 (Koren et al., 2009). 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 $H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2$ or $H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2$ 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 $\mathbf{X}^*$. 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 and show their strong duality, which leads to exact recoverability for matrix completion and robust PCA via the solutions to convex optimization problems. Strong duality is well understood for convex optimization, but very few non-convex problems were known to have this property. The results in this work thus significantly expand the set of non-convex problems with strong duality. Furthermore, our framework also shows exact recoverability of the two prototypical examples matrix completion and robust PCA with nearly-optimal sample complexity.

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. 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) (Jain et al., 2013; Sun and Luo, 2015; Ge et al., 2016). 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 (Hardt, 2014; Zhang et al., 2016; Balcan and Zhang, 2016). The second area is robust PCA, a fundamental problem of interest in data processing. It aims at recovering both the low-rank and the sparse components exactly from their superposition (Candès et al., 2011; Netrapalli et al., 2014; Gu et al., 2016; Zhang et al., 2015a, 2016; Yi et al., 2016), 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 (Candès et al., 2011; Awasthi et al., 2016). Besides these two areas, we believe that our analytical framework can be potentially applied to other non-convex problems more broadly, e.g., matrix sensing (Tu et al., 2016), dictionary learning (Sun et al., 2017b), weighted low-rank approximation (Razenshteyn et al., 2016; Li et al., 2016), and deep linear neural network (Kawaguchi, 2016), 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 (Hardt et al., 2014; Zhang et al., 2013). 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 (Candès and Recht, 2009; Chen, 2015; Foygel and Srebro, 2011). Some other methods run local search algorithms given an initialization close enough to the global solution in the basin of attraction (Jain et al., 2013; Hardt, 2014; Sun and Luo, 2015; Ge et al., 2015; Jin et al., 2017). However, these methods have sample complexity significantly larger than the information-theoretic lower bound; see Table 1.1 for a comparison. The problem becomes even more challenging when the number of samples is small enough that the sample-based initialization can be 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 (Ge et al., 2016), matrix sensing (Bhojanapalli et al., 2016b), dictionary learning (Sun et al., 2017b), phase retrieval (Sun et al., 2016), linear deep neural networks (Kawaguchi, 2016), 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 (Anandkumar and Ge, 2016), 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 a variety of non-convex matrix factorization problems, and the goal is to provide a unified framework to analyze a large class of matrix factorization problems and to provide efficient algorithms to achieve global optimum. 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.

To state the main theorem of our framework, recall that a function $H(\cdot)$ is closed if for each $\alpha \in \mathbb{R}$, the sub-level set $\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} : H(\mathbf{X}) \leq \alpha\}$ is a closed set. Also, recall the nuclear norm (a.k.a. trace norm) of a matrix $\mathbf{X}$ is $\|\mathbf{X}\|_* = \sum_{i=1}^{r} \sigma_i(\mathbf{X})$. Define the r^*-norm to be $\|\mathbf{X}\|_{r*} = \max_{\mathbf{M}} \langle \mathbf{M}, \mathbf{X} \rangle - \frac{1}{2}\|\mathbf{M}\|_r^2$ where $\|\mathbf{M}\|_r^2 = \sum_{i=1}^{r} \sigma_i^2(\mathbf{M})$ is the sum of the first r largest squared singular values. Note that both $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_{r*}$ are convex functions. Our main results are as follows.

Theorems 3 and 4 (Strong Duality. Informal). *Under certain dual conditions, strong duality holds for the non-convex optimization problem*

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

where $H(\cdot)$ is convex and closed. 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 $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$.

Similarly, under certain dual conditions, strong duality holds for the non-convex optimization problem

$$(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \operatorname*{argmin}_{\mathbf{A} \in \mathbb{R}^{n_1 \times r}, \mathbf{B} \in \mathbb{R}^{r \times n_2}} H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2, \tag{3}$$

where $H(\cdot)$ is convex and closed. In other words, problem (1) *and its bi-dual problem*

$$\bar{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}} H(\mathbf{X}) + \|\mathbf{X}\|_*, \tag{4}$$

have exactly the same optimal solutions in the sense that $\bar{\mathbf{A}}\bar{\mathbf{B}} = \bar{\mathbf{X}}$.

Description of Dual Conditions. Intuitively, the dual conditions in the above-mentioned theorems state that the angle between $\partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$ and the row and column spaces of $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ is small. In other words, there is a matrix in the sub-differential set $\partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$ which has almost the same row and

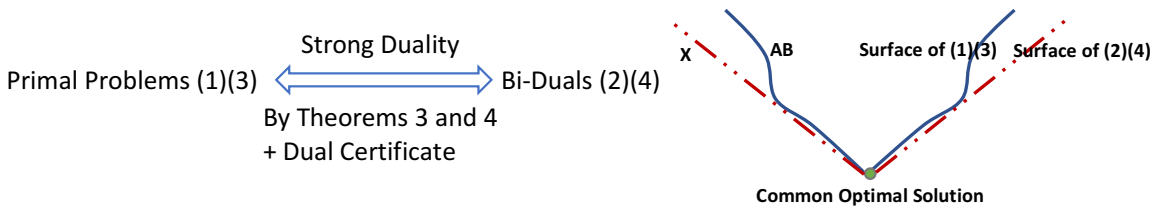

Figure 1: Strong duality of matrix factorizations.

column spaces as matrix $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$. For example, we have $\partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) = \Omega$ for the matrix completion problem, where Ω represents the subspace of matrices supported on the observed indices. Then the dual conditions require that there is a matrix which is supported on the observed indices and shares almost the same row and column spaces as $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$.

Theorem 3 connects the non-convex programs (1) to its convex counterpart (2) via strong duality; see Figure 1. Note that strong duality rarely holds in the non-convex optimization region: low-rank matrix approximation (Overton and Womersley, 1992; Zhang et al., 2019) and quadratic optimization with two quadratic constraints (Beck and Eldar, 2006) 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 theorem connects the regularization $\frac{1}{2}\|\mathbf{AB}\|_F^2$ to the r^* norm $\|\mathbf{X}\|_{r*}$. This regularization is of special interest to many matrix factorization problems. For example, when $H(\mathbf{AB}) = \frac{1}{2}\|\mathbf{X}\|_F^2 - \langle \mathbf{X}, \mathbf{AB}\rangle$, problem (1) reduces to the PCA problem: $\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{X} - \mathbf{AB}\|_F^2$. When $H(\mathbf{AB}) = \frac{1}{2}\|\mathbf{X}\|_F^2 - \langle \mathbf{X}, \mathbf{AB}\rangle + \gamma\|\mathbf{A}\|_F^2 + \gamma\|\mathbf{B}\|_F^2$, problem (1) reduces to the quadratically regularized PCA problem (Udell et al., 2016): $\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{X} - \mathbf{AB}\|_F^2 + \gamma\|\mathbf{A}\|_F^2 + \gamma\|\mathbf{B}\|_F^2$. Our framework of strong duality is then applicable to all these problems.

Furthermore, Theorem 4 also connects the non-convex programs (3) to its convex counterpart (4): the theorem connects $\frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2$ to the nuclear norm $\|\mathbf{X}\|_*$. This gives new insights for the nuclear norm relaxation technique commonly used for optimization problems with low rank constraints from the perspective of strong duality.

The positive results of our framework are complemented by a lower bound to formalize the hardness of the above problem in general. Assuming that the random 4-SAT problem (Razenshteyn et al., 2016) is hard (see Conjecture 32), we give a strong negative result for deterministic algorithms. If also $\mathsf{BPP} = \mathsf{P}$ (see Section 7 for a discussion), then the same conclusion holds for randomized algorithms succeeding with probability at least $2/3$.

Theorem 10 (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* $\epsilon \leq \epsilon_0$ *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* $2/3$*.*

Now we turn to the application of our framework. This only requires the verification of the dual conditions in Theorem 3. 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 $\mathsf{rank}(\mathbf{X}^*) \leq r$ 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 $[n_1] \times [n_2]$ 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 $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ (Recht, 2011; Candès and Tao, 2010):

$$\|\mathbf{U}^T\mathbf{e}_i\|_2 \leq \sqrt{\frac{\mu r}{n_1}} \quad \text{for all } i \in [n_1], \qquad \text{and} \qquad \|\mathbf{V}^T\mathbf{e}_i\|_2 \leq \sqrt{\frac{\mu r}{n_2}} \quad \text{for all } i \in [n_2], \tag{5}$$

where $\mathbf{e}_i$'s are basis 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 standard in the matrix completion literature. Under this standard condition, we have the following results.

Theorems 6, 7, and 8 (Matrix Completion. Informal). *There exist optimization problems for matrix completion in the forms of* (1) *and* (2) *that enjoy strong duality with each other and exactly recovers* $\mathbf{X}^*$ *with high probability, provided that* $m = \mathcal{O}(\kappa^2\mu(n_1+n_2)r\log(n_1+n_2)\log_{2\kappa}(n_1+n_2))$ *(for the formulation in Theorem 8) or* $m = \mathcal{O}(\mu(n_1+n_2)r\log^2(n_1+n_2))$ *(for the formulation in Theorem 7), where* κ *is the condition number of* $\mathbf{X}^*$*. The sample complexity lower bound is* $\Omega(\mu r(n_1+n_2)\log(n_1+n_2))$.

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 1.1 compares our results with prior results. Ours match the best known results but further provide strong duality. Also, our results are achieved by a clean framework for a class of related problems.

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 recover an incoherent low-rank component $\mathbf{X}^*$ and a sparse component $\mathbf{S}^*$ from their sum (Candès et al., 2011; Agarwal et al., 2012). We obtain the following theorem for robust PCA.

Theorems 9 (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* $m = \Theta(n_1n_2)$*, where the support set of* $\mathbf{S}^*$ *is uniformly distributed among all sets of cardinality* m*, and the incoherence parameter* μ *satisfies the incoherence condition* (5) *and* $\|\mathbf{X}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1n_2}}\sigma_r(\mathbf{X}^*)$.

The bounds in Theorem 9 match the best known results in the robust PCA literature when the supports of $\mathbf{S}^*$ are uniformly sampled (Candès et al., 2011), while our assumption is arguably more intuitive; see Section 6. 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. (2017) developed a framework to analyze the loss surface of low-rank problems, and applied the framework to matrix completion and robust PCA. For matrix

1. This lower bound is information-theoretic (Candès and Tao, 2010).

Work	Sample Complexity	Incoherence
Jain et al. (2013)	$\mathcal{O}\left(\kappa^4\mu^2r^{4.5}n_{(1)}\log n_{(1)}\log\left(\frac{r\|\mathbf{X}^*\|_F}{\epsilon}\right)\right)$	(5)
Hardt (2014)	$\mathcal{O}\left(\mu r n_{(1)}(r+\log\left(\frac{n_{(1)}\|\mathbf{X}^*\|_F}{\epsilon}\right)\frac{\|\mathbf{X}^*\|_F^2}{\sigma_r^2}\right)$	(5)
Ge et al. (2016)	$\mathcal{O}(\max\{\mu^6\kappa^{16}r^4,\mu^4\kappa^4r^6\}n_{(1)}\log^2 n_{(1)})$	$\|\mathbf{X}^*_{i:}\|_2\leq\frac{\mu\|\mathbf{X}^*\|_F}{\sqrt{n_{(2)}}}$
Sun and Luo (2015)	$\mathcal{O}(rn_{(1)}\kappa^2\max\left\{\mu\log n_{(2)},\sqrt{\frac{n_{(1)}}{n_{(2)}}}\mu^2r^6\kappa^4\right\}$	(5)
Zheng and Lafferty (2016)	$\mathcal{O}(\mu r^2 n_{(1)}\kappa^2\max(\mu,\log n_{(1)}))$	(5)
Gamarnik et al. (2017)	$\mathcal{O}\Big(\Big(\mu^2r^4\kappa^2+\mu r\log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\Big)n_{(1)}\log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\Big)$	(5)
Zhao et al. (2015)	$\mathcal{O}\left(\mu r^3 n_{(1)}\log n_{(1)}\log\left(\frac{1}{\epsilon}\right)\right)$	(5)
Keshavan et al. (2010a)	$\mathcal{O}\Big(n_{(2)}r\sqrt{\frac{n_{(1)}}{n_{(2)}}}\kappa^2\max\left\{\mu\log n_{(2)},\mu^2r\sqrt{\frac{n_{(1)}}{n_{(2)}}}\kappa^4\right\}\Big)$	(5) and (20)
Gross (2011)	$\mathcal{O}(\mu r n_{(1)}\log^2 n_{(1)})$	(5) and (20)
Chen (2015)	$\mathcal{O}(\mu r n_{(1)}\log^2 n_{(1)})$	(5)
Ours, Theorem 7	$\mathcal{O}(\kappa^2\mu r n_{(1)}\log(n_{(1)})\log_{2\kappa}(n_{(1)}))$	(5)
Ours, Theorem 8	$\mathcal{O}(\mu r n_{(1)}\log^2 n_{(1)})$	(5)
Lower Bound[1]	$\Omega(\mu r n_{(1)}\log n_{(1)})$	(5)

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 $n_{(2)}=\min\{n_1,n_2\}$.

completion, their sample complexity is $\mathcal{O}(\kappa^6\mu^4r^6(n_1+n_2)\log(n_1+n_2))$, significantly larger than our bound. For robust PCA, the number of the outlier entries that their method can tolerate is $\mathcal{O}\left(\frac{n_1n_2}{\mu r\kappa^5}\right)$, but their result is for deterministic outlier entries and thus are not directly comparable to ours. Zhang et al. (2017) also studied the robust PCA problem using non-convex optimization, where the outlier entries are also deterministic and the number of outliers that their algorithm can tolerate is $\mathcal{O}\left(\frac{n_1n_2}{r\kappa}\right)$.

1.2 Our Techniques

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

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r},\mathbf{B}\in\mathbb{R}^{r\times n_2}}\frac{1}{2}\|-\widetilde{\mathbf{\Lambda}}-\mathbf{AB}\|_F^2. \tag{6}$$

We know that all local solutions of (6) are globally optimal (see Lemma 1) and that strong duality holds for any given matrix $-\widetilde{\mathbf{\Lambda}}\in\mathbb{R}^{n_1\times n_2}$ (Grussler et al., 2016). To extend this property to our more general problem (1), our main insight is to reduce problem (1) to the form of (6) 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 (Haeffele et al., 2014; Grussler et al., 2016), or their conclusions relied on strong assumptions on the objective functions, e.g., that the objective functions are twice-differentiable (Haeffele and Vidal, 2015). For example, the conditions that all local solutions are rank-deficient break down even

for the PCA problem, and the assumptions that the objective function is twice-differential preclude $H(\cdot)$ in (1) and (3) from encoding hard constraints. 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 (Razenshteyn et al., 2016) 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 result was shown for deterministic matrix completion (Hardt and Moitra, 2013). Thus we should utilize randomness to analyze the existence of a dual certificate. For specific applications such as 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 $\mathcal{T}$; otherwise, the same construction of Neumann series may diverge as the norm concerning the projection operators on the spaces Ω and $\mathcal{T}$ is larger than 1 in the deterministic worst case. The remainder of the proof is to show that such a construction obeys the dual conditions. To show this, 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 Candès and Tao (2010); Recht (2011); Chandrasekaran et al. (2012) for convex optimization break down for our non-convex problem, since the subgradient of a non-convex objective function may not even exist Boyd and Vandenberghe (2004). Instead, we apply tools from geometric analysis Vershynin (2009) to analyze the geometry of problem (1). Our non-convex geometric analysis is in stark contrast to prior techniques of convex geometric analysis Vershynin (2015) 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.

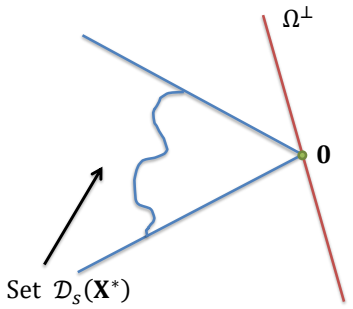

Figure 2: Feasibility.

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

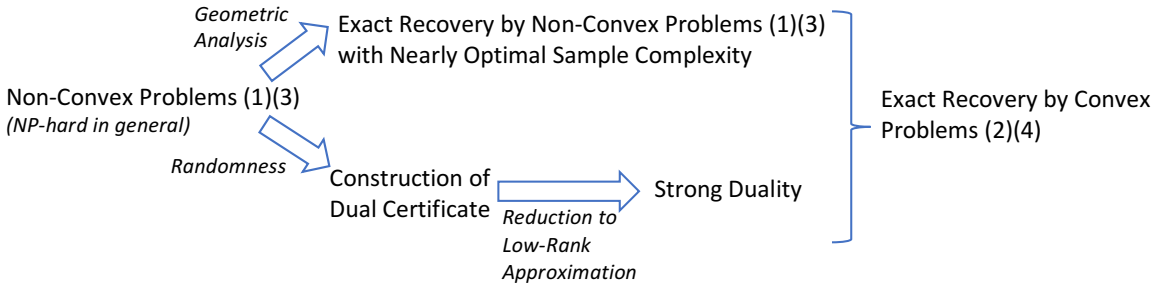

Figure 3: New analytical framework in this paper.

problem (1) from a geometric perspective: $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$ is the unique optimal solution to problem (1) if and only if starting from $\mathbf{X}^*$, either the rank of $\mathbf{X}^* + \mathbf{D}$ or $\|\mathbf{X}^* + \mathbf{D}\|_F$ increases for all directions $\mathbf{D}$'s in the constraint set $\Omega^\perp = \{\mathbf{D} \in \mathbb{R}^{n_1 \times n_2} : \mathcal{P}_\Omega(\mathbf{X}^* + \mathbf{D}) = \mathcal{P}_\Omega(\mathbf{X}^*)\}$. This can be geometrically interpreted as the requirement that the set $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) = \{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \leq r, \|\mathbf{X}\|_F \leq \|\mathbf{X}^*\|_F\}$ and the constraint set $\Omega^\perp$ must intersect uniquely at $\mathbf{0}$ (see Figure 2). This can then be shown by a dual certificate argument.

Putting Things Together. We summarize our new analytical framework with Figure 3.

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 $\|\cdot\|_{r*}$ has similar properties as 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 (Candès et al., 2011).

1.3 Paper Organization

Section 2 reviews related work and Section 3 defines the notations. The main theorems for our framework are presented in Section 4. Their applications to matrix completion are in Section 5, and those to robust PCA are in Section 6. The computational aspects of the problem are discussed in Section 7, and experimental results on synthetic and real data sets are presented in Section 8. Section 9 concludes the paper. For clarity of presentation, some proofs in our analysis are deferred to the appendix.

2. Related Work

Non-convex matrix factorization is a popular topic studied in theoretical computer science (Jain et al., 2013; Hardt, 2014; Sun and Luo, 2015; Razenshteyn et al., 2016), machine learning (Bhojanapalli et al., 2016b; Ge et al., 2016, 2015; Jain et al., 2010; Li et al., 2016), and optimization (Wen et al., 2012; Shen et al., 2014). We review several lines of research on studying the global optimality of such optimization problems.

Global Optimality of Matrix Factorization. While lots of matrix factorization problems have been shown to have no spurious local minima, they either require additional conditions on the local minima, or are based on particular forms of the objective function. Specifically, Bach et al.

(2008) and Journée et al. (2010) proved that $\mathbf{X} = \mathbf{A}\mathbf{A}^T$ is a global minimizer of $F(\mathbf{X})$, if $\mathbf{A}$ is a rank-deficient local minimizer of $F(\mathbf{A}\mathbf{A}^T)$ and $F(\mathbf{X})$ is a twice differentiable convex function. Haeffele and Vidal (2015) further extended this result by allowing a more general form of objective function $F(\mathbf{X}) = G(\mathbf{X}) + H(\mathbf{X})$, where G is a twice differentiable convex function with compact level set and H is a proper convex function such that F is lower semi-continuous. However, a major drawback of this line of research is that these result fails when the local minimizer is of full rank.

Matrix Completion. Matrix completion is a prototypical example of matrix factorization. One line of work on matrix completion builds on convex relaxation (e.g., Srebro and Shraibman (2005); Candès and Recht (2009); Candès and Tao (2010); Recht (2011); Chandrasekaran et al. (2012); Negahban and Wainwright (2012)), which achieve nearly optimal sample complexity. However, the computational complexity is relatively high compared with the non-convex methods. Recently, Ge et al. (2016) showed that matrix completion has no spurious local optimum, when $|\Omega|$ is sufficiently large and the matrix $\mathbf{Y}$ is incoherent. The result is only for positive semi-definite matrices and their sample complexity is not optimal. Independently of our work, Ge et al. (2017) developed a framework to analyze the loss surface of low-rank problems, and applied the framework to matrix completion and robust PCA. For matrix completion, their sample complexity is $\mathcal{O}(\kappa^6\mu^4 r^6(n_1+n_2)\log(n_1+n_2))$, significantly larger than our bound.

Another line of work is built upon good initialization for global convergence. Recent attempts showed that one can first compute some form of initialization (e.g., by singular value decomposition) that is close to the global minimizer and then use non-convex approaches to reach global optimality, such as alternating minimization, block coordinate descent, and gradient descent (Keshavan et al., 2010b,a; Jain et al., 2013; Keshavan, 2012; Hardt, 2014; Bhojanapalli et al., 2016a; Zheng and Lafferty, 2015; Zhao et al., 2015; Tu et al., 2016; Chen and Wainwright, 2015; Sun and Luo, 2015). The advantage of such line of research is that the computational complexity is low compared with convex relaxation based approach, but the sample complexity is relatively high. In our result, we try to bridge these two lines of research, under the same sample complexity as the best-known result of matrix completion (Chen, 2015).

Robust PCA. Robust PCA is also a prototypical example of matrix factorization. The goal is to recover both the low-rank and the sparse components exactly from their superposition (Chandrasekaran et al., 2011; Candès et al., 2011; Netrapalli et al., 2014; Gu et al., 2016; Zhang et al., 2015a, 2016; Yi et al., 2016). It has been widely applied to various tasks, such as video denoising, background modeling, image alignment, photometric stereo, texture representation, subspace clustering, and spectral clustering.

There are typically two settings in the robust PCA literature: (a) the support set of the sparse matrix is uniformly sampled (Chandrasekaran et al., 2011; Candès et al., 2011; Zhang et al., 2016); b) the support set of the sparse matrix is deterministic, but the non-zero entries in each row or column of the matrix cannot be too large (Yi et al., 2016; Ge et al., 2017). In this work, we discuss the first case. Our framework provides results that match the best known work in setting (a) (Candès et al., 2011). For setting b), Ge et al. (2017) studied the robust PCA and the number of the outlier entries that their method can tolerate is $\mathcal{O}\left(\frac{n_1 n_2}{\mu r \kappa^5}\right)$, but their result is for deterministic outlier entries and thus are not directly comparable to ours. Zhang et al. (2017) also studied the robust PCA problem using non-convex optimization, where the outlier entries are also deterministic and the number of outliers that their algorithm can tolerate is $\mathcal{O}\left(\frac{n_1 n_2}{r\kappa}\right)$.

Other Matrix Factorization Problems. Matrix sensing is another typical matrix factorization problem (Chandrasekaran et al., 2012; Jain et al., 2013; Zhao et al., 2015). Bhojanapalli et al. (2016b), Park et al. (2017) and Tu et al. (2016) showed that the matrix recovery model $\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathcal{A}(\mathbf{AB} - \mathbf{Y})\|_F^2$, achieves optimality for every local minimum, if the operator $\mathcal{A}$ satisfies the restricted isometry property. They further gave a lower bound and showed that the unstructured operator $\mathcal{A}$ may easily lead to a local minimum which is not globally optimal.

Some other matrix factorization problems are also shown to have nice geometric properties such as the property that all local minima are global minima. Examples include dictionary learning (Sun et al., 2017b), phase retrieval (Sun et al., 2016), and linear deep neural networks (Kawaguchi, 2016). In multi-layer linear neural networks where the goal is to learn a multi-linear projection $\mathbf{X}^* = \prod_i \mathbf{W}_i$, each $\mathbf{W}_i$ represents the weight matrix that connects the hidden units in the i-th and $(i+1)$-th layers. The study of such linear models can help the theoretical understanding of the loss surface of deep neural networks with non-linear activation functions (Kawaguchi, 2016; Choromanska et al., 2015). In dictionary learning, we aim to recover a dictionary matrix $\mathbf{A}$ from a given signal $\mathbf{X}$ in the form of $\mathbf{X} = \mathbf{AB}$, provided that the representation coefficient $\mathbf{B}$ is sufficiently sparse. This problem centers around solving a non-convex matrix factorization problem with a sparsity constraint on the representation coefficient $\mathbf{B}$ (Bach et al., 2008; Sun et al., 2017b,a; Arora et al., 2014). Other high-impact examples of matrix factorization models range from the classic unsupervised learning problems like PCA, independent component analysis, and clustering, to the more recent problems such as non-negative matrix factorization, weighted low-rank matrix approximation, sparse coding, tensor decomposition (Bhaskara et al., 2014; Anandkumar et al., 2014), subspace clustering (Zhang et al., 2015b, 2014), etc. Applying our framework to these other problems is left for future work.

Atomic Norms. The atomic norm is a recently proposed function for linear inverse problems (Chandrasekaran et al., 2012). Many well-known norms, e.g., the ℓ_1 norm and the nuclear norm, serve as special cases of atomic norms. It has been widely applied to the problems of compressed sensing (Tang et al., 2013), low-rank matrix recovery (Candès and Recht, 2013), blind deconvolution (Ahmed et al., 2014), etc. The norm is defined by the Minkowski functional associated with the convex hull of a set $\mathcal{A}$: $\|\mathbf{X}\|_{\mathcal{A}} = \inf\{t > 0 : \mathbf{X} \in t\mathcal{A}\}$. In particular, if we set $\mathcal{A}$ to be the convex hull of the infinite set of unit-ℓ_2-norm rank-one matrices, then $\|\cdot\|_{\mathcal{A}}$ equals to the nuclear norm. We mention that our objective term $\|\mathbf{AB}\|_F$ in problem (1) is similar to the atomic norm, but with slight differences: unlike the atomic norm, we set $\mathcal{A}$ to be the infinite set of unit-ℓ_2-norm matrices for $\mathsf{rank}(\mathbf{X}) \leq r$. This leads to better sample complexity guarantees than the atomic-norm based methods.

Comparison with Our Work. Despite a large amount of works on matrix factorizations and their convex relaxation, few works study the connection between them. Hereby, we emphasize the difference between exact recoverability and tightness of convex relaxation: exact recoverability studies whether one algorithm can exactly recover the underlying matrix, while the tightness of convex relaxation characterizes the connection between non-convex problem and its convex counterpart. The standard proof technique only focuses on the former problem for both matrix completion and robust PCA. In contrast, this work explores the latter problem and bridges the gap between the non-convex matrix factorization and its convex counterpart.

3. Notations

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 SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ of matrix $\mathbf{X}$, we define $\mathsf{svd}_r(\mathbf{X}) = \mathbf{U}_{:,1:r}\mathbf{\Sigma}_{(1:r):(1:r)}\mathbf{V}^\top_{:,1:r}$. 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 $\mathbf{X} = \mathbf{A}\mathbf{B}$, by restricting the number of columns of $\mathbf{A}$ and rows of $\mathbf{B}$ to be r. For norms, 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 $\|\mathbf{X}\|_\infty = \max_{ij} |\mathbf{X}_{ij}|$. Define the r^*-norm to be $\|\mathbf{X}\|_{r*} = \max_{\mathbf{M}} \langle \mathbf{M}, \mathbf{X} \rangle - \frac{1}{2}\|\mathbf{M}\|_r^2$ where $\|\mathbf{M}\|_r^2 = \sum_{i=1}^r \sigma_i^2(\mathbf{M})$ is the sum of the first r largest squared singular values. 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 to the 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 $\mathcal{N}(0, 1)$ the standard Gaussian distribution, $\mathsf{Uniform}(m)$ the uniform distribution of cardinality m, and $\mathsf{Ber}(p)$ the Bernoulli distribution with success probability p.

4. Strong Duality of Matrix Factorizations: A New Analytical Framework

This section develops a novel framework to analyze non-convex matrix factorization problems. The framework can be applied to different specific problems and leads to nearly optimal sample complexity guarantees.

4.1 $\frac{1}{2}\|\mathbf{A}\mathbf{B}\|_F^2$ Regularization

We first study the $\frac{1}{2}\|\mathbf{A}\mathbf{B}\|_F^2$-regularized matrix factorization problem

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

We will show in this section that under suitable conditions the duality gap between (**P**) and its dual (bi-dual) problem is zero, so problem (**P**) can be converted to an equivalent convex problem.

First, we 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. Its proof is basically to calculate the gradient of $f(\mathbf{A}, \mathbf{B})$ and let it equal to zero (Srebro, 2004); see Appendix A for details.

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.*

Now we turn to the general case. Given Lemma 1, 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:

$$\begin{aligned} F(\mathbf{A}, \mathbf{B}) &= \frac{1}{2}\|\mathbf{AB}\|_F^2 + H(\mathbf{AB}) \\ &= \frac{1}{2}\|\mathbf{AB}\|_F^2 + H^{**}(\mathbf{AB}) \\ &= \max_{\mathbf{\Lambda}} \frac{1}{2}\|\mathbf{AB}\|_F^2 + \langle\mathbf{\Lambda}, \mathbf{AB}\rangle - H^*(\mathbf{\Lambda}) \\ &= \max_{\mathbf{\Lambda}} \frac{1}{2}\| - \mathbf{\Lambda} - \mathbf{AB}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - H^*(\mathbf{\Lambda}) \\ &\triangleq \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}), \end{aligned} \tag{7}$$

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 (**P**),[2] and the second equality in Eqn. (7) holds because H is closed and convex with respect to the argument $\mathbf{AB}$.

By Lemma 1 and the definition of $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$, for any fixed value of $\mathbf{\Lambda}$, 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 $\mathbf{\Lambda}$.

So the remaining part of our analysis is to choose a proper $\widetilde{\mathbf{\Lambda}}$ for a solution $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ of problem (**P**), such that $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$, which then implies strong duality. For this, we introduce the following condition.

Condition 1 *For a solution $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ to problem (**P**), there exists a $\widetilde{\mathbf{\Lambda}} \in \partial_{\mathbf{X}} H(\mathbf{X})|_{\mathbf{X}=\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}}$ such that*

$$-\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T = \widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{B}}^T \quad \textit{and} \quad \widetilde{\mathbf{A}}^T(-\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) = \widetilde{\mathbf{A}}^T\widetilde{\mathbf{\Lambda}}. \tag{8}$$

Explanation of Condition 1. We note that $\nabla_{\mathbf{A}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = \mathbf{ABB}^T + \mathbf{\Lambda}\mathbf{B}^T$ 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 (8), 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 1 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 $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ is indeed a local minimizer, then $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$ and strong duality holds.

2. 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 (**P**) as well.

Lemma 2 (Dual Certificate) *Let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ be a solution to problem* (**P**)*. If there exists a dual certificate $\widetilde{\mathbf{\Lambda}}$ satisfying Condition 1 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 We begin by showing that $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$. 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 $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$.

Now we 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, we have

$$\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = F(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \le \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}) \le \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}).$$

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, $\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$, i.e., strong duality holds.

Finally,

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

as desired. ■

This lemma then leads to the following theorem.

Theorem 3 *Let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ denote an optimal solution of problem* (**P**)*. Define a matrix space*

$$\mathcal{T} \triangleq \{\widetilde{\mathbf{A}}\mathbf{X}^T + \mathbf{Y}\widetilde{\mathbf{B}}, \ \mathbf{X} \in \mathbb{R}^{n_2 \times r}, \ \mathbf{Y} \in \mathbb{R}^{n_1 \times r}\}. \tag{9}$$

Then strong duality holds for problem (**P**)*, provided that*

$$(1)\ \widetilde{\mathbf{\Lambda}} \in \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \triangleq \Psi, \qquad (2)\ \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \qquad (3)\ \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \tag{10}$$

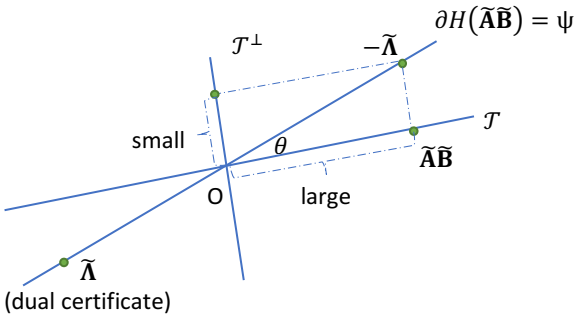

Figure 4: Geometry of dual condition (10) for general matrix factorization problems.

Proof The proof idea is to construct a dual certificate $\widetilde{\mathbf{\Lambda}}$ so that the conditions in Lemma 2 hold. It suffices for $\widetilde{\mathbf{\Lambda}}$ to satisfy the following:

$$\begin{aligned}
&\text{(a)} \quad \widetilde{\mathbf{\Lambda}} \in \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \qquad \text{(by Condition 1)} \\
&\text{(b)} \quad (\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}})\widetilde{\mathbf{B}}^T = \mathbf{0} \quad \text{and} \quad \widetilde{\mathbf{A}}^T(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}}) = \mathbf{0}, \qquad \text{(by Condition 1)} \\
&\text{(c)} \quad \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}}). \qquad \text{(by the local minimizer assumption and Lemma 1)}
\end{aligned} \tag{11}$$

It turns out that for any matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, $\mathcal{P}_{\mathcal{T}^\perp}\mathbf{M} = (\mathbf{I} - \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger)\mathbf{M}(\mathbf{I} - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\dagger)$ and so $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{M}\| \leq \|\mathbf{M}\|$, a fact that we will frequently use in the sequel. 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 $\mathcal{T} = \mathcal{U} + \mathcal{V}$. Therefore, $\mathcal{T}^\perp = (\mathcal{U} + \mathcal{V})^\perp = \mathcal{U}^\perp \cap \mathcal{V}^\perp$.

With this, we will show the conditions in (11) are equivalent to those in (10). Condition (b) implies $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}} \in \mathsf{Null}(\widetilde{\mathbf{A}}^T) = \mathsf{Col}(\widetilde{\mathbf{A}})^\perp$ 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. Condition (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 conditions in (10) are equivalent to the conditions in (11), which imply the conditions in Lemma 2 and give strong duality as desired. ■

To show the dual condition in Theorem 3, intuitively, we need to show that the angle θ between the subspaces $\mathcal{T}$ and Ψ is small (see Figure 4) for a specific function $H(\cdot)$. In Section 5, we will demonstrate applications that, with randomness, obey this dual condition with high probability.

4.2 $\frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2$ Regularization

We now study another class of matrix factorization problems

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

The analysis is similar to the analysis in Section 4.1. To see this, we first note that $\frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2$ has a natural variational form of the nuclear norm:

$$\|\mathbf{AB}\|_* = \min_{\mathbf{A}'\in\mathbb{R}^{n_1\times r}, \mathbf{B}'\in\mathbb{R}^{r\times n_2}, \mathbf{A}'\mathbf{B}'=\mathbf{AB}} \frac{1}{2}\|\mathbf{A}'\|_F^2 + \frac{1}{2}\|\mathbf{B}'\|_F^2.$$

Therefore, problem (**P'**) is equivalent to

$$(\mathbf{P''}) \qquad (\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{R}^{n_1\times r}, \mathbf{B}\in\mathbb{R}^{r\times n_2}}{\operatorname{argmin}} F(\mathbf{A},\mathbf{B}) = H(\mathbf{AB}) + \|\mathbf{AB}\|_*,\ H(\cdot) \text{ is convex and closed.}$$

Thus we can analyze problem **(P'')** as in Section 4.1. More specifically, we have the following result.

Theorem 4 *Let $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ denote an optimal solution of problem* **(P'')**. *Let $\mathbf{U\Sigma V}^\top$ denote the skinny SVD of $\bar{\mathbf{A}}\bar{\mathbf{B}}$. Define a matrix space*

$$\mathcal{T} \triangleq \{\mathbf{UX}^T + \mathbf{YV}^T,\ \mathbf{X}\in\mathbb{R}^{n_2\times r},\ \mathbf{Y}\in\mathbb{R}^{n_1\times r}\}. \tag{12}$$

If there exists a dual certificate $\widetilde{\mathbf{\Lambda}}$ such that

$$\begin{aligned} &(a) \qquad \widetilde{\mathbf{\Lambda}} \in \partial H(\bar{\mathbf{A}}\bar{\mathbf{B}}) \triangleq \Psi, \\ &(b) \qquad \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \mathbf{UV}^\top, \\ &(c) \qquad \|\mathcal{P}_{\mathcal{T}^\perp}(-\widetilde{\mathbf{\Lambda}})\| < \frac{1}{2}, \end{aligned} \tag{13}$$

then strong duality holds for problems **(P'')** *and* **(P')**.

Remark 5 *There is a key difference between Theorem 4 and the standard optimality result by Candès and Recht (2009). On one hand, Theorem 4 shows that under certain conditions, the non-convex matrix completion and its convex counterpart share a common global optimality, i.e., the strong duality. On the other hand, the standard optimality result in Lemma 3.1 of (Candès and Recht, 2009) studies the conditions under which the convex matrix completion exactly recovers the underlying low-rank matrix. Their lemma does not concern about the non-convex matrix completion formulation.*

Proof We note that

$$\begin{aligned} F(\mathbf{A},\mathbf{B}) &= \|\mathbf{AB}\|_* + H(\mathbf{AB}) \\ &= \|\mathbf{AB}\|_* + H^{**}(\mathbf{AB}) \\ &= \|\mathbf{AB}\|_* + \max_{\mathbf{\Lambda}\in\mathbb{R}^{n_1\times n_2}} \langle \mathbf{\Lambda}, \mathbf{AB}\rangle - H^*(\mathbf{\Lambda}) \\ &= \max_{\mathbf{\Lambda}\in\mathbb{R}^{n_1\times n_2}} \|\mathbf{AB}\|_* + \langle \mathbf{\Lambda}, \mathbf{AB}\rangle - H^*(\mathbf{\Lambda}) \\ &= \max_{\mathbf{\Lambda}\in\mathbb{R}^{n_1\times n_2}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}), \end{aligned}$$

where the second equality holds because $H(\cdot)$ is closed and convex with respect to the argument $\mathbf{AB}$ and the third equality holds by the definition of conjugate function. To prove Theorem 4, we need the following two claims.

Claim 1 $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\Lambda} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$.

Proof We only need to show $\mathbf{0} \in \partial_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}}) = \bar{\mathbf{A}}\bar{\mathbf{B}} - \partial_{\mathbf{\Lambda}} H^*(\widetilde{\mathbf{\Lambda}})$, because $L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$ is a concave function of $\mathbf{\Lambda}$ for the fixed $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. Note that $\bar{\mathbf{A}}\bar{\mathbf{B}} \in \partial_{\mathbf{\Lambda}} H^*(\widetilde{\mathbf{\Lambda}})$ is implied by $\widetilde{\mathbf{\Lambda}} \in \partial H(\bar{\mathbf{A}}\bar{\mathbf{B}})$ by the convexity of function $H(\cdot)$. Therefore, condition (a) immediately implies $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$. ■

Claim 2 $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$.

Proof First, we have

$$\begin{aligned}\operatorname*{argmin}_{\mathbf{A}\in\mathbb{R}^{n_1\times r},\mathbf{B}\in\mathbb{R}^{r\times n_2}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}}) &= \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB}\rangle - H^*(\widetilde{\mathbf{\Lambda}}) \\ &= \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB}\rangle \\ &= \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \sum_{i=1}^r \sigma_i(\mathbf{AB}) - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB}\rangle,\end{aligned}$$

where the first step follows by definition of $L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$, the second step follows since $\mathbf{\Lambda}$ is fixed, the third step follows by definition of nuclear norm.

Using conditions (b) and (c) in (13), $-\widetilde{\mathbf{\Lambda}}$ can be rewritten as

$$-\widetilde{\mathbf{\Lambda}} = \underbrace{\mathbf{U}\mathbf{V}^T}_{\text{in space } \mathcal{T} \text{ with eigenvalues } 1} + \underbrace{\mathcal{P}_{\mathcal{T}^\perp}(-\widetilde{\mathbf{\Lambda}})}_{\text{in space } \mathcal{T}^\perp \text{ with eigenvalues } <1/2},$$

which implies $\forall i \in [r], \sigma_i(-\widetilde{\mathbf{\Lambda}}) = 1$.

Using von Neumann's trace inequality (see Lemma 25), we can show

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^r \sigma_i(\mathbf{AB}) - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB}\rangle &\geq \min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^r \sigma_i(\mathbf{AB}) - \sum_{i=1}^r \sigma_i(-\widetilde{\mathbf{\Lambda}})\sigma_i(\mathbf{AB}) \\ &= \min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^r \sigma_i(\mathbf{AB}) - \sum_{i=1}^r \sigma_i(\mathbf{AB}) \qquad \text{by } \sigma_i(-\mathbf{\Lambda}) = 1, \forall i \in [r] \\ &= 0.\end{aligned}$$

Therefore, on the one hand, we already have

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r},\mathbf{B}\in\mathbb{R}^{r\times n_2}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}}) \geq -H^*(\widetilde{\mathbf{\Lambda}}).$$

On the other hand, according to definition of $\bar{\mathbf{A}}, \bar{\mathbf{B}}$, and $-\widetilde{\mathbf{\Lambda}}$,

$$\begin{aligned}L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}}) &= \|\bar{\mathbf{A}}\bar{\mathbf{B}}\|_* - \langle -\widetilde{\mathbf{\Lambda}}, \bar{\mathbf{A}}\bar{\mathbf{B}}\rangle - H^*(\widetilde{\mathbf{\Lambda}}) \\ &= -H^*(\widetilde{\mathbf{\Lambda}}).\end{aligned}$$

Thus, we can conclude that $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$. ■

Claim 1 and Claim 2 together show that $(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of the Lagrangian $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$.

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 $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$, we have

$$F(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}}) \leq L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}), \ \ \forall \mathbf{A}, \mathbf{B},$$

where the inequality holds because $(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\cdot,\cdot,\cdot)$. So on the one hand, we have

$$\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) = F(\bar{\mathbf{A}},\bar{\mathbf{B}}) \le \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}}) \le \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}).$$

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, $\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) = \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$, i.e., strong duality holds. ■

Similarly as for Theorem 3, to show the dual condition in Theorem 4, we need to show that the angle θ between the subspaces $\mathcal{T}$ and Ψ is small for a specific function $H(\cdot)$. We will also demonstrate how randomness implies this dual condition with high probability in concrete applications.

5. Matrix Completion

In matrix completion, there is a hidden matrix $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ with rank at most 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 4 to matrix completion, by setting $H(\cdot) = \mathbf{I}_{\{\mathbf{M}:\mathcal{P}_\Omega(\mathbf{M})=\mathcal{P}_\Omega(\mathbf{X}^*)\}}(\cdot)$. (Recall that $\mathbf{I}_{\mathcal{C}}$ is the indicator function of the set $\mathcal{C}$.)

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. The μ-incoherence condition (5) was introduced for the low-rank matrix $\mathbf{X}$ (Candès et al., 2011; Candès and Recht, 2009; Recht, 2011; Zhang et al., 2016). For this condition, it can be shown that $1 \le \mu \le \frac{n_{(1)}}{r}$. It holds for many random matrices with incoherence parameter μ about $\sqrt{r \log n_{(1)}}$ (Keshavan et al., 2010a).

We have two positive results for matrix completion, which are achieved via efficient algorithms by applying Theorem 3 and Theorem 4, respectively.

Before proceeding, we first cite a lower bound from prior work. It shows that our two positive results are nearly optimal.

Theorem 6 (Information-Theoretic Lower Bound. Candès and Tao (2010), Theorem 1.7) *Denote by $\Omega \sim \mathsf{Uniform}(m)$ the support set uniformly distributed among all sets of cardinality m. Suppose that $m \le c\mu n_{(1)} r \log n_{(1)}$ 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* (5) *such that $\mathcal{P}_\Omega(\mathbf{X}') = \mathcal{P}_\Omega(\mathbf{X}^*)$, with probability at least $1 - n_{(1)}^{-10}$.*

Our first positive result converts a non-convex rank-constrained problem to a convex optimization problem, which can be *efficiently* solved. The proof is simply by applying Theorem 3, so the details are deferred to Appendix D.

Theorem 7 (Efficient Matrix Completion I) *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Denote the condition number of matrix $\mathbf{X}^*$ as $\kappa =$*

$\sigma_1(\mathbf{X}^*)/\sigma_r(\mathbf{X}^*)$. *Let* $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \in \mathbb{R}^{n_1\times r} \times \mathbb{R}^{r\times n_2}$ *denote the output of the non-convex matrix factorization problem*

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r}, \mathbf{B}\in\mathbb{R}^{r\times n_2}} \frac{1}{2}\|\mathbf{AB}\|_F^2, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*). \tag{14}$$

Then there are absolute constants c *and* c_0 *such that with probability at least* $1 - cn^{-10}$*, the strong duality* $\mathbf{X}^* = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$ *holds, 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* (5)*, where* $\widetilde{\mathbf{X}}$ *is the output of the nuclear norm minimization*

$$\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \|\mathbf{X}\|_{r*}, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*). \tag{15}$$

Proof Sketch. We first show that the problem

$$(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{AB}\|_F^2, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*),$$

exactly recovers $\mathbf{X}^*$, i.e., $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$, with the optimal sample complexity (see Appendix D). So if strong duality holds, this non-convex optimization problem can be equivalently converted to the convex program (15), proving Theorem 7.

It now suffices to apply our unified framework in Section 4 to prove the strong duality. We show that the dual condition in Theorem 3 holds with high probability by the following arguments. Let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ be a global solution to problem (15). 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

$$(1)\ \widetilde{\mathbf{\Lambda}} \in \Omega, \qquad (2)\ \mathcal{P}_\mathcal{T}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \qquad (3)\ \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{2}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \tag{16}$$

It is interesting to see that dual condition (16) can be satisfied if the angle θ between subspace Ω and subspace $\mathcal{T}$ is very small; see Figure 4. When the sample size $|\Omega|$ becomes larger, the angle θ becomes smaller (e.g., when $|\Omega| = n_1 n_2$, 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 (16) to hold. □

Our second result further improves the sample complexity of Theorem 7 by applying Theorem 4. The reduced sample complexity matches the best known result, which was achieved by nuclear norm minimization. However, the strong duality is new here, which illustrates the reason why nuclear norm works for matrix completion from another viewpoint of non-convex optimization. The proof is deferred to Appendix E.

Theorem 8 (Efficient Matrix Completion II) *Let* $\Omega \sim \mathsf{Uniform}(m)$ *be the support set uniformly distributed among all sets of cardinality* m*. Let* $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \in \mathbb{R}^{n_1\times r} \times \mathbb{R}^{r\times n_2}$ *denote the output of the non-convex matrix factorization problem*

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r}, \mathbf{B}\in\mathbb{R}^{r\times n_2}} \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*), \tag{17}$$

Then there are absolute constants c and c_0 such that with probability at least $1 - cn^{-10}$, the strong duality $\mathbf{X}^* = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$ holds, provided that $m \geq c_0 \mu r n_{(1)} \log^2 n_{(1)}$ and $\mathbf{X}^*$ obeys μ-incoherence (5)*, where $\widetilde{\mathbf{X}}$ is the output of the nuclear norm minimization*

$$\min_{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{X}\|_*, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*). \tag{18}$$

6. Robust Principal Component Analysis

This section develops 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 which is incoherent and low rank, 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 spreads evenly throughout the matrix, the matrix cannot have one entry whose absolute value is significantly larger than other entries. Candès et al. (2011) introduced an extra incoherence condition (Recall that $\mathbf{X}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the skinny SVD of $\mathbf{X}^*$)

$$\|\mathbf{U}\mathbf{V}^T\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}. \tag{19}$$

In this work, we make the following incoherence assumption for robust PCA instead of (19):

$$\|\mathbf{X}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}} \sigma_r(\mathbf{X}^*). \tag{20}$$

Note that condition (20) is very similar to the incoherence condition (19) for the robust PCA problem, but the two notions are not directly comparable. On the other hand, the condition (20) 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 9 (Robust PCA) *Suppose $\mathbf{X}^*$ is an $n_1 \times n_2$ matrix of rank r, and obeys incoherence* (5) *and* (20)*. 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*

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

with $\lambda = \frac{\sigma_r(\mathbf{X}^*)}{\sqrt{n_{(1)}}}$ is exact, namely, $\widetilde{\mathbf{X}} = \mathbf{X}^*$ and $\widetilde{\mathbf{S}} = \mathbf{S}^*$, provided that $\mathsf{rank}(\mathbf{X}^*) \leq \rho_r \frac{n_{(2)}}{\mu \log^2 n_{(1)}}$ and $m \leq \rho_s n_1 n_2$, where c, ρ_r, and ρ_s are all positive absolute constants, and function $\|\cdot\|_{r*}$ is given by (21)*.*

The bounds on the rank of $\mathbf{X}^*$ and the sparsity of $\mathbf{S}^*$ in Theorem 9 match the best known results for robust PCA in prior work when we assume the support set of $\mathbf{S}^*$ is sampled uniformly (Candès et al., 2011).

7. 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 (see Appendix H.1)

$$\begin{aligned} &(\textbf{Dual, D1}) \qquad \max_{\mathbf{\Lambda}\in\mathbb{R}^{n_1\times n_2}} -H^*(\mathbf{\Lambda}) - \frac{1}{2}\|\mathbf{\Lambda}\|_r^2, \quad \text{where } \|\mathbf{\Lambda}\|_r^2 = \sum_{i=1}^{r}\sigma_i^2(\mathbf{\Lambda}), \\ &(\textbf{Bi-Dual, D2}) \qquad \min_{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}} H(\mathbf{M}) + \|\mathbf{M}\|_{r*}, \quad \text{where } \|\mathbf{M}\|_{r*} = \max_{\mathbf{X}}\langle \mathbf{M}, \mathbf{X}\rangle - \frac{1}{2}\|\mathbf{X}\|_r^2. \end{aligned} \tag{21}$$

and the dual and bi-dual of primal problem (**P'**) are given by (see Appendix H.2)

$$\begin{aligned} &(\textbf{Dual, D1'}) \qquad \max_{\|\mathbf{\Lambda}\|\le 1} -H^*(\mathbf{\Lambda}), \\ &(\textbf{Bi-Dual, D2'}) \qquad \min_{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}} H(\mathbf{M}) + \|\mathbf{M}\|_*. \end{aligned}$$

Problems (**D1**), (**D1'**) and (**D2**), (**D2'**) can be solved efficiently due to their convexity. In particular, Grussler et al. (2016) provided a computationally efficient algorithm to compute the proximal operators of functions $\frac{1}{2}\|\cdot\|_r^2$ and $\|\cdot\|_{r*}$. Hence, there exist algorithms that can find global minimum up to an ϵ error in function value in time $\mathsf{poly}(1/\epsilon)$, e.g., the Douglas-Rachford algorithm (He and Yuan, 2012).

Computational Lower Bounds. Unfortunately, strong duality does not always hold for general non-convex problems (**P**) (and (**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 (Feige, 2002) and gives us constant factor inapproximability of problem (**P**) (and (**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 pseudo-random generators or if any problem in deterministic exponential time has exponential size circuits (Impagliazzo and Wigderson, 1997). Therefore, any sub-exponential time algorithm achieving a sufficiently small constant factor approximation to problem (**P**) (and (**P'**)) in general would imply a major breakthrough in complexity theory.

The lower bound is proved by a reduction from the Maximum Edge Bi-clique problem (Ambühl et al., 2011). The details are presented in Appendix G.

Theorem 10 (Computational Lower Bound) *Assume the hardness of Random 4-SAT (See Conjecture 32 in Appendix G). 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**) *(and* (**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 $2/3$.*

It is clear that the bi-dual is typically solvable in polynomial time. Thus, the hardness result implies that (a) under certain circumstances, strong duality does not hold, and (b) the original problem is hard, while the bi-dual is easy.

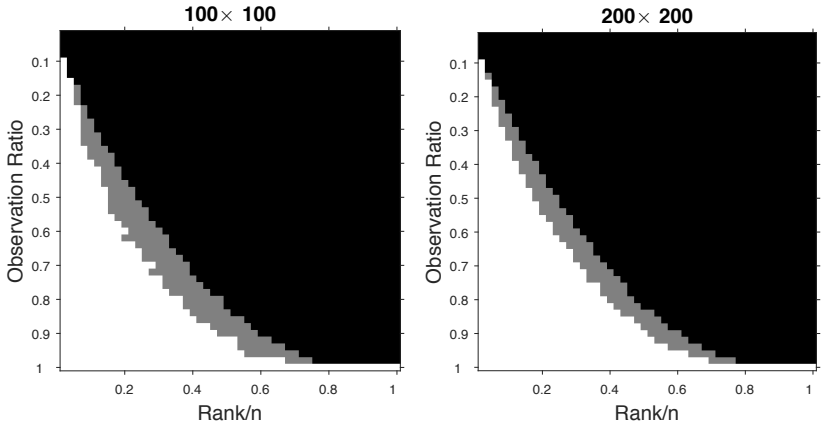

Figure 5: Exact recoverability of matrix completion with varying ranks and sample sizes. **White Region:** nuclear norm minimization succeeds. **White and Gray Regions:** r^* minimization succeeds. **Black Region:** both algorithms fail. It shows that the success region of r^* minimization slightly contains that of the nuclear minimization method.

8. Experiments

In this section, we will present our experimental results.

8.1 Experiments on Synthetic Data

We verify the exact recoverability of the r^* minimization (15) and the nuclear norm minimization (18) on the matrix completion problem by experiments on the synthetic data. The synthetic data are generated as follows. We construct the ground-truth matrix $\mathbf{X}^* = \mathbf{AB}$ as a product of matrices $\mathbf{A}$ of size $n \times r$ and $\mathbf{B}$ of size $r \times n$, whose entries are i.i.d. $\mathcal{N}(0, 1)$. We then uniformly sample m entries from $\mathbf{X}^*$ as the observations. For each size of the problem ($\mathbf{X}^*$ is 100×100 or 200×200), we test with different rank ratios r/n and observation ratios m/n^2. Each set of parameters is run 5 times, and the algorithm is said to succeed if $\|\widetilde{\mathbf{X}} - \mathbf{X}^*\|_F / \|\mathbf{X}^*\|_F \leq 10^{-3}$ for all five experiments, where $\widetilde{\mathbf{X}}$ is the output of the algorithms. We set the parameter r in r^* minimization (15) as the true rank, and use the Augmented Lagrange Multiplier Method (Chen et al., 2009) for optimization, where the proximal map of r^* norm is computed as in (Grussler et al., 2016).

The two figures in Figure 5 plots the fraction of exact recoveries: the white region represents the exact recovery by nuclear norm minimization, the white+gray region represents the exact recovery by r^* minimization (15), and the black region indicates the failure for both algorithms. It is clear that both algorithms succeed for a wide range of parameters. The success region of r^* minimization is slightly larger and contains the success region of the nuclear norm minimization for both 100×100 and 200×200 matrix completion problems.

8.2 Experiments on Real Data

To verify the performance of the algorithms on real data, we conduct experiments on the Hopkins 155 data set. This data set consists of 155 tasks/matrices, each of which consists of multiple data points

#Task	Size	$m = 0.05n_1n_2$ Nuclear	$m = 0.05n_1n_2$ r^*	$m = 0.1n_1n_2$ Nuclear	$m = 0.1n_1n_2$ r^*
Average over all 155 tasks	–	0.8249	0.8114	0.5689	0.5409
#1	59×459	0.7438	0.5948	0.5115	0.5117
#2	49×482	0.8235	0.6564	0.6371	0.5919
#3	49×153	0.7803	0.9174	0.5386	0.5386
#4	49×379	0.8500	0.9583	0.7287	0.7691
#5	49×432	0.8174	0.6353	0.4476	0.4477

Table 2: Relative error by matrix completion algorithms on the Hopkins 155 data set.

drawn from 2 or 3 moving objects. The trajectory of each object lies in a low-dimensional subspace, so the matrix for each task is supposed to be approximately low rank. We uniformly sample m entries from the matrix as our observations and run the matrix completion algorithms. The parameter r in the r^* minimization is set as the number of moving objects which is known to us in the data set.

Table 8.2 shows the the relative errors $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F / \|\mathbf{X}^*\|_F$ of the nuclear norm minimization (18) and r^* minimization (15). On average, r^* minimization slightly outperforms the competitor, while sometimes the nuclear norm minimization is better. Table 8.2 also shows the errors on the first five tasks in the data set. It shows that when the number of observations is relatively large (10% observations or higher), the performance of the two algorithms are competitive to each other. When the number of observations is small (5% observed entries), there is a larger variance, but on average r^* minimization has an slight advantage.

9. Conclusions

This paper studied the strong duality of non-convex matrix factorization problems. It was shown that under certain dual conditions, a wide class of non-convex matrix factorization problems and their dual have the same optimum. This strong duality phenomenon is rarely discovered in the previous work. Hardness results were provided to show that strong duality is impossible without the randomness of sampling. The analytical framework may be of independent interest to non-convex optimization more broadly.

The proposed framework was applied to two prototypical matrix factorization problems in the machine learning community: matrix completion and robust PCA. This gave several efficient algorithms with nearly optimal sample complexity bounds which match the best-known previous results and also illustrated why the nuclear norm technique works from the new viewpoint of non-convex optimization.

Acknowledgments

We thank the anonymous reviewers for valuable comments, and Rina Foygel Barber, Rong Ge, Jason D. Lee, Zhouchen Lin, Guangcan Liu, Tengyu Ma, Benjamin Recht and Tuo Zhao for useful discussions. 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. We would like to also thank Christian Grussler, Anders Rantzer, and Pontus Giselsson who kindly provided the code for computing the proximal map of the r^* function.

A. Proof of Lemma 1

Lemma 1 (Restated)**.** *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.*

Proof We have that (**A**,**B**) is a critical point of $f(\mathbf{A},\mathbf{B})$ if and only if $\nabla_{\mathbf{A}} f(\mathbf{A},\mathbf{B}) = \mathbf{0}$ and $\nabla_{\mathbf{B}} f(\mathbf{A},\mathbf{B}) = \mathbf{0}$, or equivalently,

$$\mathbf{ABB}^T = \widehat{\mathbf{Y}}\mathbf{B}^T \quad \text{and} \quad \mathbf{A}^T\mathbf{AB} = \mathbf{A}^T\widehat{\mathbf{Y}}. \tag{22}$$

Note that for any fixed matrix $\mathbf{A}$ (respectively $\mathbf{B}$), the function $f(\mathbf{A},\mathbf{B})$ is convex in the coefficients of $\mathbf{B}$ (respectively $\mathbf{A}$).

To prove the desired lemma, we need the following claim.

Claim 3 *If two matrices $\mathbf{A}$ and $\mathbf{B}$ define a critical point of $f(\mathbf{A},\mathbf{B})$, then the global mapping $\mathbf{M} = \mathbf{AB}$ is of the form*

$$\mathbf{M} = \mathcal{P}_{\mathbf{A}}\widehat{\mathbf{Y}},$$

with $\mathbf{A}$ satisfying

$$\mathbf{AA}^\dagger\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T = \mathbf{AA}^\dagger\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T\mathbf{AA}^\dagger = \widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T\mathbf{AA}^\dagger. \tag{23}$$

Proof If $\mathbf{A}$ and $\mathbf{B}$ define a critical point of $f(\mathbf{A},\mathbf{B})$, then (22) holds and the general solution to (22) satisfies

$$\mathbf{B} = (\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T\widehat{\mathbf{Y}} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{L}, \tag{24}$$

for some matrix $\mathbf{L}$. So $\mathbf{M} = \mathbf{AB} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T\widehat{\mathbf{Y}} = \mathbf{AA}^\dagger\widehat{\mathbf{Y}} = \mathcal{P}_{\mathbf{A}}\widehat{\mathbf{Y}}$ by the property of the Moore-Penrose pseudo-inverse $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T$.

By (22), we also have

$$\mathbf{ABB}^T\mathbf{A}^T = \widehat{\mathbf{Y}}\mathbf{B}^T\mathbf{A}^T \quad \text{or equivalently} \quad \mathbf{MM}^T = \widehat{\mathbf{Y}}\mathbf{M}^T.$$

Plugging in the relation $\mathbf{M} = \mathbf{AA}^\dagger\widehat{\mathbf{Y}}$, (A) can be rewritten as

$$\mathbf{AA}^\dagger\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T\mathbf{AA}^\dagger = \widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T\mathbf{AA}^\dagger.$$

Note that the matrix $\mathbf{AA}^\dagger\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T\mathbf{AA}^\dagger$ is symmetric. Thus

$$\mathbf{AA}^\dagger\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T = \mathbf{AA}^\dagger\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T\mathbf{AA}^\dagger,$$

as desired. ■

To prove Lemma 1, we also need the following claim.

Claim 4 *Denote by $\mathcal{I} = \{i_1, i_2, ..., i_r\}$ any ordered r-index set (ordered by $\lambda_{i_j}, j \in [r]$ from the largest to the smallest) and λ_i, $i \in [n_1]$, the eigenvalues of $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T \in \mathbb{R}^{n_1 \times n_1}$ with p distinct values. Let $\mathbf{U}_{\mathcal{I}} = [\mathbf{u}_{i_1}, \mathbf{u}_{i_2}, ..., \mathbf{u}_{i_r}]$ denote the matrix formed by the orthonormal eigenvectors $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{n_1}]$ of $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T \in \mathbb{R}^{n_1 \times n_1}$ associated with the ordered p eigenvalues, whose multiplicities are $m_1, m_2, ..., m_p$ ($m_1 + m_2 + ... + m_p = n_1$). Then two matrices $\mathbf{A}$ and $\mathbf{B}$ define a critical point of $f(\mathbf{A}, \mathbf{B})$ if and only if there exists an ordered r-index set $\mathcal{I}$, an invertible matrix $\mathbf{C}$, and an $r \times n$ matrix $\mathbf{L}$ such that*

$$\mathbf{A} = (\mathbf{UD})_{:\mathcal{I}}\mathbf{C} \quad \textit{and} \quad \mathbf{B} = \mathbf{A}^\dagger\widehat{\mathbf{Y}} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{L}, \tag{25}$$

where $\mathbf{D}$ is a p-block-diagonal matrix with each block equal to the orthogonal projector of dimension m_i. For such a critical point, we have

$$\mathbf{AB} = \mathcal{P}_{\mathbf{A}}\widehat{\mathbf{Y}},$$

$$f(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\left(\mathsf{tr}(\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T) - \sum_{i \in \mathcal{I}} \lambda_i\right) = \frac{1}{2}\sum_{i \notin \mathcal{I}} \lambda_i. \tag{26}$$

Proof Note that $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T$ is a real symmetric covariance matrix. So it can always be represented as $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U} \in \mathbb{R}^{n_1 \times n_1}$ is an orthonormal matrix consisting of eigenvectors of $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T$ and $\mathbf{\Lambda} \in \mathbb{R}^{n_1 \times n_1}$ is a diagonal matrix with non-increasing eigenvalues of $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T$.

If $\mathbf{A}$ and $\mathbf{B}$ satisfy (25) for some $\mathbf{C}$, $\mathbf{L}$, and $\mathcal{I}$, then

$$\mathbf{ABB}^T = \widehat{\mathbf{Y}}\mathbf{B}^T \quad \text{and} \quad \mathbf{A}^T\mathbf{AB} = \mathbf{A}^T\widehat{\mathbf{Y}},$$

which is (22). So $\mathbf{A}$ and $\mathbf{B}$ define a critical point of $f(\mathbf{A}, \mathbf{B})$.

For the converse, notice that

$$\mathcal{P}_{\mathbf{U}^T\mathbf{A}} = \mathbf{U}^T\mathbf{A}(\mathbf{U}^T\mathbf{A})^\dagger = \mathbf{U}^T\mathbf{A}\mathbf{A}^\dagger\mathbf{U} = \mathbf{U}^T\mathcal{P}_{\mathbf{A}}\mathbf{U},$$

or equivalently, $\mathcal{P}_{\mathbf{A}} = \mathbf{U}\mathcal{P}_{\mathbf{U}^T\mathbf{A}}\mathbf{U}^T$. Thus (23) yields

$$\mathbf{U}\mathcal{P}_{\mathbf{U}^T\mathbf{A}}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U}\mathcal{P}_{\mathbf{U}^T\mathbf{A}}\mathbf{U}^T,$$

or equivalently, $\mathcal{P}_{\mathbf{U}^T\mathbf{A}}\mathbf{\Lambda} = \mathbf{\Lambda}\mathcal{P}_{\mathbf{U}^T\mathbf{A}}$. Notice that $\mathbf{\Lambda} \in \mathbb{R}^{n_1 \times n_1}$ is a diagonal matrix with p distinct eigenvalues of $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T$. So $\mathcal{P}_{\mathbf{U}^T\mathbf{A}}$ is a block-diagonal matrix with p blocks, each of which is an orthogonal projector of dimension m_i, corresponding to the eigenvalues λ_i, $i \in [p]$. Therefore, there exists an index set $\mathcal{I}$ such that $\mathcal{P}_{\mathbf{U}^T\mathbf{A}} = \mathbf{D}_{:\mathcal{I}}\mathbf{D}_{:\mathcal{I}}^T$, where $\mathbf{D}$ is a block-diagonal matrix. It follows that

$$\mathcal{P}_{\mathbf{A}} = \mathbf{U}\mathcal{P}_{\mathbf{U}^T\mathbf{A}}\mathbf{U}^T = \mathbf{UD}_{:\mathcal{I}}\mathbf{D}_{:\mathcal{I}}^T\mathbf{U}^T = (\mathbf{UD})_{:\mathcal{I}}(\mathbf{UD})_{:\mathcal{I}}^T.$$

Since the column space of $\mathbf{A}$ coincides with the column space of $(\mathbf{UD})_{:\mathcal{I}}$, $\mathbf{A}$ is of the form $\mathbf{A} = (\mathbf{UD})_{:\mathcal{I}}\mathbf{C}$, and $\mathbf{B}$ is given by (24). Thus $\mathbf{AB} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T\widehat{\mathbf{Y}} + \mathbf{A}(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{L} = \mathcal{P}_{\mathbf{A}}\widehat{\mathbf{Y}}$

and

$$\begin{aligned} f(\mathbf{A},\mathbf{B}) &= \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{A}\mathbf{B}\|_F^2 \\ &= \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathcal{P}_{\mathbf{A}}\widehat{\mathbf{Y}}\|_F^2 \\ &= \frac{1}{2}\|\mathcal{P}_{\mathbf{A}^\perp}\widehat{\mathbf{Y}}\|_F^2 \\ &= \frac{1}{2}\sum_{i\notin\mathcal{I}}\lambda_i. \end{aligned}$$

■

So the local minimizer of $f(\mathbf{A},\mathbf{B})$ is given by (25) with $\mathcal{I} = \Phi$, which is globally optimal according to (26), where Φ is the index set corresponding to the r largest eigenvalues of $\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T$. We then show that when $\mathcal{I}$ consists of other combinations of indices of eigenvalues, i.e., $\mathcal{I} \neq \Phi$, the corresponding pair $(\mathbf{A},\mathbf{B})$ given by (25) is a strict saddle point.

Claim 5 *If $\mathcal{I} \neq \Phi$, then the pair $(\mathbf{A},\mathbf{B})$ given by* (25) *is a strict saddle point.*

Proof Let $i \in \Phi$ but $i \notin \mathcal{I}$, and denote by $\mathbf{UD} = \mathbf{R}$. It is enough to slightly perturb the column space of $\mathbf{A}$ towards the direction of an eigenvector of λ_i. More precisely, fix two indices i and j such that $i \in \Phi$, $i \notin \mathcal{I}$, and j is the largest index in $\mathcal{I}$. For any ϵ, let $\widetilde{\mathbf{R}}_{:j} = (1+\epsilon^2)^{-1/2}(\mathbf{R}_{:j} + \epsilon\mathbf{R}_{:i})$. Notice that $i \notin \mathcal{I}$. Thus $\widetilde{\mathbf{R}}_{:j}^T\widetilde{\mathbf{R}}_{:j} = \mathbf{I}$. Let $\widetilde{\mathbf{A}} = \widetilde{\mathbf{R}}_{\mathcal{I}}\mathbf{C}$ and $\widetilde{\mathbf{B}} = \widetilde{\mathbf{A}}^\dagger\mathbf{Y} + (\mathbf{I} - \widetilde{\mathbf{A}}^\dagger\widetilde{\mathbf{A}})\mathbf{L}$. A direct calculation shows that

$$f(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}}) = f(\mathbf{A},\mathbf{B}) - \epsilon^2(\lambda_i - \lambda_j)/(2+2\epsilon^2).$$

Hence,

$$\lim_{\epsilon\to 0}\frac{f(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}}) - f(\mathbf{A},\mathbf{B})}{\epsilon^2} = -\frac{1}{2}(\lambda_i - \lambda_j) < 0.$$

■

Note that all critical points of $f(\mathbf{A},\mathbf{B})$ are in the form of (25), and if $\mathcal{I} \neq \Phi$, the pair $(\mathbf{A},\mathbf{B})$ given by (25) is a strict saddle point, while if $\mathcal{I} = \Phi$, then the pair $(\mathbf{A},\mathbf{B})$ given by (25) is a local minimum. We conclude that $f(\mathbf{A},\mathbf{B})$ has no local maximum. The proof is completed. ■

B. Existence of Dual Certificate for Matrix Completion I

Let $\widetilde{\mathbf{A}} \in \mathbb{R}^{n_1\times r}$ and $\widetilde{\mathbf{B}} \in \mathbb{R}^{r\times n_2}$ such that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$. Then we have the following lemma.

Lemma 11 *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Suppose that $m \geq c\kappa^2\mu n_{(1)} r\log n_{(1)}\log_{2\kappa} n_{(1)}$ for an absolute constant c and $\mathbf{X}^*$ obeys*

μ-incoherence (5). *Then there exists $\widetilde{\mathbf{\Lambda}}$ such that*

$$\begin{aligned} &(1) \quad \widetilde{\mathbf{\Lambda}} \in \Omega, \\ &(2) \quad \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \\ &(3) \quad \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{2}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \end{aligned} \tag{27}$$

with probability at least $1 - n_{(1)}^{-10}$.

The rest of the section is devoted to the proof of Lemma 11. We begin with the following lemma.

Lemma 12 *If we can construct an $\mathbf{\Lambda}$ such that*

$$\begin{aligned} &\text{(a)} \quad \mathbf{\Lambda} \in \Omega, \\ &\text{(b)} \quad \|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_F \leq \sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \\ &\text{(c)} \quad \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| < \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \end{aligned} \tag{28}$$

then we can construct an $\widetilde{\mathbf{\Lambda}}$ such that Eqn. (27) *holds with probability at least* $1 - n_{(1)}^{-10}$.

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

Theorem 13 (Candès and Recht (2009), 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* (5) *holds. Then there is an absolute constant C_R such that for $\beta > 1$, we have*

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

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 (28) 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 $\|\mathcal{P}_{\mathcal{T}}\mathbf{Y}\|_F \leq \sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. We now prove Condition (3) in Eqn. (27). Observe that

$$\begin{aligned} \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| &\leq \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| \\ &\leq \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\|. \end{aligned} \tag{29}$$

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

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

$$\mathcal{P}^{-1} = \sum_{k\geq 0}(\mathcal{P}_{\mathcal{T}} - \mathcal{P})^k \text{ whenever } \|\mathcal{P}_{\mathcal{T}} - \mathcal{P}\| < 1.$$

So according to Theorem 13, the operator $p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1}$ can be represented as a *convergent* Neumann series

$$p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1} = \sum_{k\geq 0}(-1)^k(\mathcal{P}_{\mathcal{T}}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}})^k,$$

because $\|\mathcal{P}_{\mathcal{T}}\mathcal{Q}_{\Omega}\mathcal{P}_{\mathcal{T}}\| \leq \epsilon < \frac{1}{2}$ once $m \geq C\mu n_{(1)} r \log n_{(1)}$ 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

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

with high probability. The proof is completed. ■

It thus suffices to construct a dual certificate $\mathbf{\Lambda}$ such that all conditions in (28) 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 13,

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

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

The following lemma together implies the strong duality of (15) straightforwardly.

Lemma 14 *Under the assumptions of Theorem 7, the dual certification $\mathbf{\Lambda}_b$ obeys the dual condition* (28) *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; see Section I for a brief justification. By the equivalence, we can suppose $\Omega \sim \mathsf{Ber}(p)$.

To prove Lemma 14, as a preliminary, we need the following lemmas.

Lemma 15 (Chen (2015), 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 16 (Candès et al. (2011), 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 17 (Chen (2015), 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 with 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 16,

$$\|\mathbf{W}_j\|_\infty \leq \left(\frac{1}{2}\right)^j \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty,$$

and by Lemma 17,

$$\|\mathbf{W}_j\|_{\infty,2} \leq \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{W}_{j-1}\|_\infty + \frac{1}{2}\|\mathbf{W}_{j-1}\|_{\infty,2}.$$

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,

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \\
&\leq \sum_{j=1}^{b} \|\frac{n_1 n_2}{q}\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1}\| \\
&= \sum_{j=1}^{b} \|\mathcal{P}_{\mathcal{T}^\perp}(\frac{n_1 n_2}{q}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1} - \mathbf{W}_{j-1})\| \\
&\leq \sum_{j=1}^{b} \|(\frac{n_1 n_2}{q}\mathcal{P}_{\Omega_j} - \mathcal{I})(\mathbf{W}_{j-1})\|.
\end{aligned}$$

Let p denote $\Theta\left(\frac{q}{n_1 n_2}\right)$. By Lemma 15,

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \\
&\leq C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_{\infty,2} \\
&\leq C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\left(\frac{1}{2}\right)^j\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty \\
&\quad + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\left[j\left(\frac{1}{2}\right)^j\sqrt{\frac{n_{(1)}}{\mu r}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + \left(\frac{1}{2}\right)^j\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}\right] \\
&\leq C_0'\frac{\log n_{(1)}}{p}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + 2C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sqrt{\frac{n_{(1)}}{\mu r}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}.
\end{aligned}$$

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

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

and that

$$\begin{aligned}
\|\mathbf{X}^*\|_\infty &= \max_{ij}\langle\mathbf{X}^*, \mathbf{e}_i\mathbf{e}_j^T\rangle = \max_{ij}\langle\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \mathbf{e}_i\mathbf{e}_j^T\rangle = \max_{ij}\langle\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}, \mathbf{e}_j^T\mathbf{V}\rangle \\
&\leq \max_{ij}\|\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\|_2\|\mathbf{e}_j^T\mathbf{V}\|_2 \leq \max_j\|\mathbf{X}^*\|_{\infty,2}\|\mathbf{e}_j^T\mathbf{V}\|_2 \leq \frac{\mu r\kappa}{\sqrt{n_1 n_2}}\sigma_r(\mathbf{X}^*).
\end{aligned}$$

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 $\|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}_b\| < \frac{1}{3}\sigma_r(\mathbf{X}^*)$. The proof is completed. ■

C. Subgradient of the $r*$ Function

Lemma 18 *Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the skinny SVD of matrix $\mathbf{X}^*$ of rank r. The subdifferential of $\|\cdot\|_{r*}$ evaluated at $\mathbf{X}^*$ is given by*

$$\partial\|\mathbf{X}^*\|_{r*} = \{\mathbf{X}^* + \mathbf{W} : \mathbf{U}^T\mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\| \leq \sigma_r(\mathbf{X}^*)\}. \tag{30}$$

Proof Note that for any fixed function $f(\cdot)$, the set of all optimal solutions of the problem

$$f^*(\mathbf{X}^*) = \max_{\mathbf{Y}} \langle \mathbf{X}^*, \mathbf{Y}\rangle - f(\mathbf{Y}) \tag{31}$$

form the subdifferential of the conjugate function $f^*(\cdot)$ evaluated at $\mathbf{X}^*$. Set $f(\cdot)$ to be $\frac{1}{2}\|\cdot\|_r^2$ and notice that the function $\frac{1}{2}\|\cdot\|_r^2$ is unitarily invariant. By Von Neumann's trace inequality, the optimal solutions to problem (31) are given by

$$[\mathbf{U}, \mathbf{U}^\perp]\mathsf{Diag}([\sigma_1(\mathbf{Y}), ..., \sigma_r(\mathbf{Y}), \sigma_{r+1}(\mathbf{Y}), ..., \sigma_{n_{(2)}}(\mathbf{Y})])[\mathbf{V}, \mathbf{V}^\perp]^T,$$

where $\{\sigma_i(\mathbf{Y})\}_{i=r+1}^{n_{(2)}}$ can be any value no larger than $\sigma_r(\mathbf{Y})$ and $\{\sigma_i(\mathbf{Y})\}_{i=1}^r$ are given by the optimal solution to the problem

$$\max_{\{\sigma_i(\mathbf{Y})\}_{i=1}^r} \sum_{i=1}^r \sigma_i(\mathbf{X}^*)\sigma_i(\mathbf{Y}) - \frac{1}{2}\sum_{i=1}^r \sigma_i^2(\mathbf{Y}).$$

The solution is unique such that $\sigma_i(\mathbf{Y}) = \sigma_i(\mathbf{X}^*)$, $i = 1, 2, ..., r$. The proof is complete. ■

D. Proof of Theorem 7

We will prove Theorem 7 in this section.

D.1 Exact Recoverability of Non-Convex Formulation

Theorem 19 (Uniqueness of Solution) *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Suppose that $m \geq c\kappa^2 \mu n_{(1)} r \log n_{(1)} \log_{2\kappa} n_{(1)}$ for an absolute constant c and $\mathbf{X}^*$ obeys μ-incoherence* (5)*. Then $\mathbf{X}^*$ is the unique solution of non-convex optimization*

$$\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{A}\mathbf{B}\|_F^2, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{A}\mathbf{B}) = \mathcal{P}_\Omega(\mathbf{X}^*),$$

with probability at least $1 - n_{(1)}^{-10}$.

Proof We note that a recovery result under the Bernoulli model automatically implies a corresponding result for the uniform model Candès et al. (2011); see Section I for the details. So in the following, we assume the Bernoulli model.

Consider the feasibility of the matrix completion problem:

$$\text{Find a matrix } \mathbf{X} \in \mathbb{R}^{n_1 \times n_2} \text{ such that } \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*), \quad \|\mathbf{X}\|_F \le \|\mathbf{X}^*\|_F, \quad \mathsf{rank}(\mathbf{X}) \le r. \tag{32}$$

Note that if $\mathbf{X}^*$ is the unique solution of (32), then $\mathbf{X}^*$ is the unique solution of (14). We now show the former. Our proof first identifies a feasibility condition for problem (32), 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

$$\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) = \{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \le r,\ \|\mathbf{X}\|_F \le \|\mathbf{X}^*\|_F\},$$

and

$$\mathcal{T} = \{\mathbf{U}\mathbf{X}^T + \mathbf{Y}\mathbf{V}^T,\ \mathbf{X} \in \mathbb{R}^{n_2 \times r},\ \mathbf{Y} \in \mathbb{R}^{n_1 \times r}\},$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the skinny SVD of $\mathbf{X}^*$.

We have the following proposition for the feasibility of problem (32).

Proposition 20 (Feasibility Condition) *$\mathbf{X}^*$ is the unique feasible solution to problem* (32) *if* $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$.

Proof Notice that problem (32) is equivalent to another feasibility problem

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

Suppose that $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$. Since $\mathsf{rank}(\mathbf{X}^* + \mathbf{D}) \le r$ and $\|\mathbf{X}^* + \mathbf{D}\|_F \le \|\mathbf{X}^*\|_F$ are equivalent to $\mathbf{D} \in \mathcal{D}_{\mathcal{S}}(\mathbf{X}^*)$, 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 (32). ■

The remainder of the proof is to show $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$. To proceed, we note that

$$\begin{aligned}
\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) &= \left\{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \le r,\ \frac{1}{2}\|\mathbf{X}\|_F^2 \le \frac{1}{2}\|\mathbf{X}^*\|_F^2\right\} \\
&\subseteq \{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \|\mathbf{X}\|_{r*} \le \|\mathbf{X}^*\|_{r*}\} \quad \left(\text{since } \frac{1}{2}\|\mathbf{Y}\|_F^2 = \|\mathbf{Y}\|_{r*} \text{ for any rank-}r \text{ matrix}\right) \\
&\triangleq \mathcal{D}_{\mathcal{S}_*}(\mathbf{X}^*).
\end{aligned}$$

We now show that

$$\mathcal{D}_{\mathcal{S}_*}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}, \tag{33}$$

when $m \ge c\kappa^2 \mu r n_{(1)} \log_{2\kappa}(n_{(1)}) \log(n_{(1)})$, which will prove $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$ as desired.

By Lemma 11, there exists a $\mathbf{\Lambda}$ such that

$$\begin{aligned}
&(1) \qquad \mathbf{\Lambda} \in \Omega, \\
&(2) \qquad \mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) = \mathbf{X}^*, \\
&(3) \qquad \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| < \frac{2}{3}\sigma_r(\mathbf{X}^*).
\end{aligned}$$

Consider any $\mathbf{D} \in \Omega^\perp$ such that $\mathbf{D} \neq \mathbf{0}$. By Lemma 18, for any $\mathbf{W} \in \mathcal{T}^\perp$ and $\|\mathbf{W}\| \leq \sigma_r(\mathbf{X}^*)$,

$$\|\mathbf{X}^* + \mathbf{D}\|_{r*} \geq \|\mathbf{X}^*\|_{r*} + \langle \mathbf{X}^* + \mathbf{W}, \mathbf{D} \rangle.$$

Since $\langle \mathbf{W}, \mathbf{D} \rangle = \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{W}, \mathbf{D} \rangle = \langle \mathbf{W}, \mathcal{P}_{\mathcal{T}^\perp} \mathbf{D} \rangle$, we can choose $\mathbf{W}$ such that

$$\langle \mathbf{W}, \mathbf{D} \rangle = \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_*.$$

Then

$$\begin{aligned}
\|\mathbf{X}^* + \mathbf{D}\|_{r*} &\geq \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathbf{X}^*, \mathbf{D} \rangle \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathbf{X}^* + \mathbf{\Lambda}, \mathbf{D} \rangle \quad \text{(since } \mathbf{\Lambda} \in \Omega \text{ and } \mathbf{D} \in \Omega^\perp\text{)} \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathbf{X}^* + \mathcal{P}_{\mathcal{T}} \mathbf{\Lambda}, \mathbf{D} \rangle + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathbf{D} \rangle \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathbf{D} \rangle \quad \text{(by condition (2))} \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathbf{D} \rangle \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathcal{P}_{\mathcal{T}^\perp} \mathbf{D} \rangle \\
&\geq \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* - \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}\| \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* \quad \text{(by Hölder's inequality)} \\
&\geq \|\mathbf{X}^*\|_{r*} + \frac{1}{3} \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* \quad \text{(by condition (3))}.
\end{aligned}$$

So if $\mathcal{T} \cap \Omega^\perp = \{\mathbf{0}\}$, since $\mathbf{D} \in \Omega^\perp$ and $\mathbf{D} \neq \mathbf{0}$, we have $\mathbf{D} \notin \mathcal{T}$. Therefore,

$$\|\mathbf{X}^* + \mathbf{D}\|_{r*} > \|\mathbf{X}^*\|_{r*}$$

which then leads to $\mathcal{D}_{\mathcal{S}_*}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$.

The rest of proof is to show that $\mathcal{T} \cap \Omega^\perp = \{\mathbf{0}\}$. We have the following lemma.

Lemma 21 *Assume that $\Omega \sim \mathsf{Ber}(p)$ and the incoherence condition* (5) *holds. Then with probability at least $1 - n_{(1)}^{-10}$, we have $\|\mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\| \leq \sqrt{1 - p + \epsilon p}$, provided that $p \geq C_0 \epsilon^{-2} (\mu r \log n_{(1)})/n_{(2)}$, where C_0 is an absolute constant.*

Proof If $\Omega \sim \mathsf{Ber}(p)$, we have, by Theorem 13, that with high probability

$$\|\mathcal{P}_{\mathcal{T}} - p^{-1} \mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega} \mathcal{P}_{\mathcal{T}}\| \leq \epsilon,$$

provided that $p \geq C_0 \epsilon^{-2} \frac{\mu r \log n_{(1)}}{n_{(2)}}$. 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

$$\|\mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\| \leq \epsilon p + (1 - p).$$

Since $\|\mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\|^2 \leq \|\mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\|$, 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 of latter part of the theorem is completed. ■

D.2 Strong Duality

We have shown in Theorem 19 that the problem $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{AB}\|_F^2$, 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 nearly optimal sample complexity. So if strong duality holds, this non-convex optimization problem can be equivalently converted to the convex program (15). Then Theorem 7 is straightforward from strong duality.

It now suffices to apply our unified framework in Section 4 to prove the strong duality. Let

$$H(\mathbf{X}) = \mathbf{I}_{\{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}:\ \mathcal{P}_\Omega\mathbf{M}=\mathcal{P}_\Omega\mathbf{X}^*\}}(\mathbf{X})$$

in Problem (**P**), and let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ be a global solution to the problem. Then by Theorem 19, $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$. For Problem (**P**) with this special $H(\mathbf{X})$, 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 $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$. Combining with Lemma 11 shows that the dual condition in Theorem 3 holds with high probability, which leads to strong duality and thus proving Theorem 7.

E. Proof of Theorem 8

Given support set $\Omega \subset [n_1]\times[n_2]$ and rank-r matrix $\mathbf{X}^* \in \mathbb{R}^{n_1\times n_2}$, define the following optimization problem

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r},\mathbf{B}\in\mathbb{R}^{r\times n_2}} \|\mathbf{AB}\|_*, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*). \tag{34}$$

We have the following lemma.

Lemma 22 *For any fixed support set Ω and rank-r matrix $\mathbf{X}^* \in \mathbb{R}^{n_1\times n_2}$, problem* (17) *and the non-convex problem* (34) *have the same solution, i.e.,*

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \bar{\mathbf{A}}\bar{\mathbf{B}},$$

where $\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}$ is the optimal solution of (17) *and $\bar{\mathbf{A}}\bar{\mathbf{B}}$ is the optimal solution of* (34)*.*

Proof First, introduce another optimization problem as a connection between the two in the lemma. For any matrix $\mathbf{Z}\in\mathbb{R}^{n_1\times n_2}$ of rank at most r, we consider the following optimization problem

$$\begin{aligned}\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r},\mathbf{B}\in\mathbb{R}^{r\times n_2}} \ & \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2,\\ \text{s.t.} \ & \mathbf{AB} = \mathbf{Z},\\ & \mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*).\end{aligned} \tag{35}$$

We want to show that the optimal cost of objective function (35) is $\|\mathbf{Z}\|_*$ and the minimum is achieved when $\|\mathbf{A}\|_F = \|\mathbf{B}\|_F$. Denote by $\mathbf{U\Sigma V}^\top$ the skinny SVD of matrix $\mathbf{Z} = \mathbf{AB}$.

On one hand, we have

$$\begin{aligned}
\|\mathbf{Z}\|_* &= \mathsf{tr}(\mathbf{\Sigma}) \\
&= \mathsf{tr}(\mathbf{U}^\top \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top \mathbf{V}) && (\text{ by } \mathbf{U}^\top \mathbf{U} = \mathbf{I}, \mathbf{V}^\top \mathbf{V} = \mathbf{I} \in \mathbb{R}^{r \times r}) \\
&= \mathsf{tr}(\mathbf{U}^\top \mathbf{A} \mathbf{B} \mathbf{V}) && (\text{ by } \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{Z} = \mathbf{A}\mathbf{B}) \\
&\leq \sum_{i=1}^r \sigma_i(\mathbf{U}^\top \mathbf{A}) \cdot \sigma_i(\mathbf{B}\mathbf{V}) && (\text{ by trace inequality}) \\
&\leq \frac{1}{2}\|\mathbf{U}^\top \mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\mathbf{V}\|_F^2 && (\forall a, b : ab \leq \frac{1}{2}(a^2 + b^2)) \\
&= \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2.
\end{aligned}$$

Therefore, $\|\mathbf{Z}\|_* \leq \min_{\mathbf{A},\mathbf{B}, \text{ s.t. } \mathbf{AB}=\mathbf{Z}, \mathcal{P}_\Omega(\mathbf{AB})=\mathcal{P}_\Omega(\mathbf{X}^*)} \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2$ for any matrix $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ satisfies that $\mathsf{rank}(\mathbf{Z}) \leq r$ and $\mathcal{P}_\Omega(\mathbf{Z}) = \mathcal{P}_\Omega(\mathbf{X}^*)$.

On the other hand, taking $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}^{1/2}$ and $\mathbf{B} = \mathbf{\Sigma}^{1/2}\mathbf{V}^\top$, we can verify these properties

$$\begin{aligned}
&\mathbf{AB} = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{V}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{Z}, \\
&\mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*), \\
&\|\mathbf{A}\|_F = \|\mathbf{\Sigma}^{1/2}\|_F = \|\mathbf{B}\|_F, \\
&\frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2 = \|\mathbf{\Sigma}^{1/2}\|_F^2 = \|\mathbf{Z}\|_*.
\end{aligned}$$

Thus (35) holds and the minimum is achieved when $\|\mathbf{A}\|_F = \|\mathbf{B}\|_F$.

We now prove Lemma 22 by contradiction. Assume for contradiction that $(\mathbf{A}^*, \mathbf{B}^*)$ is a solution to (34) such that $\|\mathbf{A}^*\|_F = \|\mathbf{B}^*\|_F$ (for any solution $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ to problem (34), we can always find the pair $(\mathbf{A}^*, \mathbf{B}^*)$ such that $\mathbf{A}^*\mathbf{B}^* = \bar{\mathbf{A}}\bar{\mathbf{B}}$ and $\|\mathbf{A}^*\|_F = \|\mathbf{B}^*\|_F$), while it is not a solution to (17). So there exists $(\mathbf{A}', \mathbf{B}')$ such that $\mathcal{P}_\Omega(\mathbf{A}'\mathbf{B}') = \mathcal{P}_\Omega(\mathbf{X}^*)$ and

$$\begin{aligned}
\|\mathbf{A}'\mathbf{B}'\|_* &= \frac{1}{2}\|\mathbf{A}'\|_F^2 + \frac{1}{2}\|\mathbf{B}'\|_F^2 && (\text{by Problem 35}) \\
&< \frac{1}{2}\|\mathbf{A}^*\|_F^2 + \frac{1}{2}\|\mathbf{B}^*\|_F^2 && (\text{by } (\mathbf{A}', \mathbf{B}') \text{ is not a solution to Problem 17}) \\
&= \|\mathbf{A}^*\mathbf{B}^*\|_*, && (\text{by Problem 35})
\end{aligned}$$

which is contradictory with the optimality of $(\mathbf{A}^*, \mathbf{B}^*)$ to problem (34).

Similarly, assume for contradiction that $(\mathbf{A}^*, \mathbf{B}^*)$ is a solution to (17), while it is not a solution to (34). So there exists $(\mathbf{A}', \mathbf{B}')$ such that $\mathcal{P}_\Omega(\mathbf{A}'\mathbf{B}') = \mathcal{P}_\Omega(\mathbf{X}^*)$ and

$$\begin{aligned}
\frac{1}{2}\|\mathbf{A}'\|_F^2 + \frac{1}{2}\|\mathbf{B}'\|_F^2 &= \|\mathbf{A}'\mathbf{B}'\|_* \\
&< \|\mathbf{A}^*\mathbf{B}^*\|_* \\
&= \frac{1}{2}\|\mathbf{A}^*\|_F^2 + \frac{1}{2}\|\mathbf{B}^*\|_F^2,
\end{aligned}$$

which is contradictory with the optimality of $(\mathbf{A}^*, \mathbf{B}^*)$ to problem (17). The proof is completed. ■

Before proceeding, we provide the definition of a matrix space.

Definition 23 *Given matrices* $\bar{\mathbf{A}} \in \mathbb{R}^{n_1 \times r}$ *and* $\bar{\mathbf{B}} \in \mathbb{R}^{r \times n_2}$*, we denote by* $\mathcal{T} \triangleq \{\bar{\mathbf{A}}\mathbf{X}^\top + \mathbf{Y}\bar{\mathbf{B}} : \mathbf{X} \in \mathbb{R}^{n_2 \times r},\ \mathbf{Y} \in \mathbb{R}^{n_1 \times r}\}$ *a matrix space. For any matrix* $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$*, we have*

$$\mathcal{P}_{\mathcal{T}}\mathbf{M} = \bar{\mathbf{A}}\bar{\mathbf{A}}^\dagger\mathbf{M} + \mathbf{M}\bar{\mathbf{B}}\bar{\mathbf{B}}^\dagger - \bar{\mathbf{A}}\bar{\mathbf{A}}^\dagger\mathbf{M}\bar{\mathbf{B}}\bar{\mathbf{B}}^\dagger,$$

and

$$\mathcal{P}_{\mathcal{T}^\perp}\mathbf{M} = (\mathbf{I} - \bar{\mathbf{A}}\bar{\mathbf{A}}^\dagger)\mathbf{M}(\mathbf{I} - \bar{\mathbf{B}}\bar{\mathbf{B}}^\dagger).$$

Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the skinny SVD of matrix $\bar{\mathbf{A}}\bar{\mathbf{B}}$. Denote by $\mathcal{U} = \{\mathbf{U}\mathbf{X}^\top : \mathbf{X} \in \mathbb{R}^{n_2 \times r}\}$ and $\mathcal{V} = \{\mathbf{Y}\mathbf{V}^\top : \mathbf{Y} \in \mathbb{R}^{n_1 \times r}\}$. Then the linear space $\mathcal{T}$ can be equivalently represented as $\mathcal{T} = \mathcal{U} + \mathcal{V}$. Therefore, $\mathcal{T}^\perp = (\mathcal{U} + \mathcal{V})^\perp = \mathcal{U}^\perp \cap \mathcal{V}^\perp$.

The following lemma is a restated result of Theorem 4 for the matrix completion problem. For completeness, we include its proof here.

Lemma 24 (Dual Certificate) *Let* $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ *be a solution of problem* (34)*. Let* $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ *denote the skinny SVD of* $\bar{\mathbf{A}}\bar{\mathbf{B}}$*. If there exists a dual certificate* $\widetilde{\mathbf{\Lambda}}$ *such that*

$$\begin{aligned} &(a) \quad \widetilde{\mathbf{\Lambda}} \in \Omega, \\ &(b) \quad \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \mathbf{U}\mathbf{V}^\top, \\ &(c) \quad \|\mathcal{P}_{\mathcal{T}^\perp}(-\widetilde{\mathbf{\Lambda}})\| < \frac{1}{2}, \end{aligned} \tag{36}$$

then strong duality holds, i.e., problem (18) *and problem* (17) *have the same solution.*

Proof For any convex set $\mathcal{C}$, we define indicator function $\mathbf{I}_{\mathcal{C}} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ such that

$$\mathbf{I}_{\mathcal{C}}(\mathbf{X}) = \begin{cases} 0, & \text{if } \mathbf{X} \in \mathcal{C}; \\ +\infty, & \text{otherwise.} \end{cases}$$

We set $\mathcal{C}$ to be $\{\mathbf{X} : \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*), \mathbf{X} \in \mathbb{R}^{n_1 \times n_2}\}$, and define function $H : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ such that for any $\mathbf{X}$, $H(\mathbf{X}) = \mathbf{I}_{\mathcal{C}}(\mathbf{X})$.

Then problem (34) can be equivalently represented by

$$(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \underset{\mathbf{A} \in \mathbb{R}^{n_1 \times r}, \mathbf{B} \in \mathbb{R}^{r \times n_2}}{\operatorname{argmin}} F(\mathbf{A}, \mathbf{B}) = \|\mathbf{A}\mathbf{B}\|_* + H(\mathbf{A}\mathbf{B}).$$

Define $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$ to be $\|\mathbf{A}\mathbf{B}\|_* + \langle \mathbf{\Lambda}, \mathbf{A}\mathbf{B} \rangle - H^*(\mathbf{\Lambda})$. Then we have

$$\begin{aligned} F(\mathbf{A}, \mathbf{B}) &= \|\mathbf{A}\mathbf{B}\|_* + H(\mathbf{A}\mathbf{B}) \\ &= \|\mathbf{A}\mathbf{B}\|_* + H^{**}(\mathbf{A}\mathbf{B}) \\ &= \|\mathbf{A}\mathbf{B}\|_* + \max_{\mathbf{\Lambda} \in \mathbb{R}^{n_1 \times n_2}} \langle \mathbf{\Lambda}, \mathbf{A}\mathbf{B} \rangle - H^*(\mathbf{\Lambda}) \\ &= \max_{\mathbf{\Lambda} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{A}\mathbf{B}\|_* + \langle \mathbf{\Lambda}, \mathbf{A}\mathbf{B} \rangle - H^*(\mathbf{\Lambda}) \\ &= \max_{\mathbf{\Lambda} \in \mathbb{R}^{n_1 \times n_2}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}), \end{aligned}$$

where the second equality holds because $H(\cdot)$ is closed and convex with respect to the argument $\mathbf{AB}$ and the third equality holds by the definition of the conjugate function.

We first show that $(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of the Lagrangian $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$, namely, $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$ and $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$. Using Claim 6, we have $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$. Using Claim 7, we have $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$. Therefore, $(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of the Lagrangian $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$.

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 $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$, we have

$$F(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}}) \leq L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}), \ \ \forall \mathbf{A}, \mathbf{B}.$$

where the inequality holds because $(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\cdot, \cdot, \cdot)$. So on the one hand, we have

$$\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = F(\bar{\mathbf{A}}, \bar{\mathbf{B}}) \leq \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}) \leq \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}).$$

On the other hand, by weak duality,

$$\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) \geq \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}).$$

Therefore, $\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$, i.e., strong duality holds. ■

Claim 6 $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\Lambda} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$.

Proof We only need to show $\mathbf{0} \in \partial_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}}) = \bar{\mathbf{A}}\bar{\mathbf{B}} - \partial_{\mathbf{\Lambda}} H^*(\widetilde{\mathbf{\Lambda}})$, because $L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$ is a concave function of $\mathbf{\Lambda}$ for the fixed $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. Note that $\bar{\mathbf{A}}\bar{\mathbf{B}} \in \partial_{\mathbf{\Lambda}} H^*(\widetilde{\mathbf{\Lambda}})$ is implied by $\widetilde{\mathbf{\Lambda}} \in \partial H(\bar{\mathbf{A}}\bar{\mathbf{B}})$ by the convexity of function $H(\cdot)$, and

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

Therefore, condition (a) immediately implies $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{\Lambda})$. ■

Claim 7 $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$.

Proof To see $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$ for the fixed $\widetilde{\mathbf{\Lambda}}$ with conditions (b) and (c), we have

$$\begin{aligned}
\operatorname*{argmin}_{\mathbf{A} \in \mathbb{R}^{n_1 \times r}, \mathbf{B} \in \mathbb{R}^{r \times n_2}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}) &= \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB} \rangle - H^*(\widetilde{\mathbf{\Lambda}}) \\
&= \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB} \rangle \\
&= \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{r} \sigma_i(\mathbf{AB}) - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB} \rangle
\end{aligned}$$

where the first step follows by definition of $L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$, the second step follows by $\mathbf{\Lambda}$ is fixed, the third step follows by definition of nuclear norm.

Using conditions (b) and (c) in Eqn. (36), we can rewrite $-\widetilde{\mathbf{\Lambda}}$ as follows,

$$-\widetilde{\mathbf{\Lambda}} = \underbrace{\mathbf{U}\mathbf{V}^\top}_{\text{in space } \mathcal{T} \text{ with eigenvalues } 1} + \underbrace{\mathcal{P}_{\mathcal{T}^\perp}(-\widetilde{\mathbf{\Lambda}})}_{\text{in space } \mathcal{T}^\perp \text{ with eigenvalues } <1/2},$$

which implies $\forall i \in [r], \sigma_i(-\widetilde{\mathbf{\Lambda}}) = 1$.

We also need the following von Neumann's trace inequality.

Lemma 25 (von Neumann's trace inequality, Lin and Zhang (2017)) *For any matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ $(m \le n)$, $\mathsf{tr}(\mathbf{A}^\top\mathbf{B}) \le \sum_{i=1}^m \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$. The equality holds if and only if there exists column orthonormal matrices $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{A} = \mathbf{U}\mathit{Diag}(\sigma(\mathbf{A}))\mathbf{V}^\top$ and $\mathbf{B} = \mathbf{U}\mathit{Diag}(\sigma(\mathbf{B}))\mathbf{V}^\top$ are the SVDs of $\mathbf{A}$ and $\mathbf{B}$, simultaneously.*

Using von Neumann's trace inequality, we can show

$$\begin{aligned}
\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^r \sigma_i(\mathbf{AB}) - \langle -\widetilde{\mathbf{\Lambda}}, \mathbf{AB}\rangle &\ge \min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^r \sigma_i(\mathbf{AB}) - \sum_{i=1}^r \sigma_i(-\widetilde{\mathbf{\Lambda}})\sigma_i(\mathbf{AB}) \\
&= \min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^r \sigma_i(\mathbf{AB}) - \sum_{i=1}^r \sigma_i(\mathbf{AB}) \qquad (\text{by } \sigma_i(-\mathbf{\Lambda}) = 1, \forall i \in [r]) \\
&= 0,
\end{aligned}$$

Therefore, on one hand, we already have

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r}, \mathbf{B}\in\mathbb{R}^{r\times n_2}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}) \ge -H^*(\widetilde{\mathbf{\Lambda}}).$$

On the other hand, according to definition of $\mathbf{A}, \mathbf{B}$ and $-\widetilde{\mathbf{\Lambda}}$, we can show

$$\begin{aligned}
L(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \widetilde{\mathbf{\Lambda}}) &= \|\bar{\mathbf{A}}\bar{\mathbf{B}}\|_* - \langle -\widetilde{\mathbf{\Lambda}}, \bar{\mathbf{A}}\bar{\mathbf{B}}\rangle - H^*(\widetilde{\mathbf{\Lambda}}) \\
&= -H^*(\widetilde{\mathbf{\Lambda}}).
\end{aligned}$$

Thus, we can conclude $(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$. ■

To show the dual condition in Lemma 24, intuitively, we need to show that the angle θ between subspace $\mathcal{T}$ and Ψ is small (see Figure 4) for a specific function $H(\cdot)$.

The proof of the dual conditions is similar to that in Appendix D, with some slight differences. We include the proof for completeness.

Lemma 26 *If we can construct an $\mathbf{\Lambda}$ such that*

$$\begin{aligned}
&\text{(a)} \quad \mathbf{\Lambda} \in \Omega, \\
&\text{(b)} \quad \|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top\|_F \le \sqrt{\frac{r}{3n_{(1)}^2}}, \\
&\text{(c)} \quad \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| < \frac{1}{3},
\end{aligned} \tag{37}$$

then we can construct an $\widetilde{\mathbf{\Lambda}}$ such that (36) *holds with probability at least $1 - n_{(1)}^{-10}$.*

Proof Suppose that Condition (37) 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 $\mathbf{Y} = \mathcal{P}_\Omega \mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}})^{-1}(\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top)$. So $\|\mathcal{P}_{\mathcal{T}}\mathbf{Y}\|_F \leq \sqrt{\frac{r}{3n_{(1)}^2}}$. We now prove Condition (3) in (36). Observe that

$$\begin{aligned} \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| &\leq \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| \\ &\leq \frac{1}{3}\sigma_r(\mathbf{U}\mathbf{V}^\top) + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\|. \end{aligned} \tag{38}$$

So we only need to show $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| \leq \frac{1}{3}$.

Before proceeding, we begin by introducing $\mathcal{Q}_\Omega : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$, which is a normalized version of $\mathcal{P}_\Omega$, defined as

$$\mathcal{Q}_\Omega = p^{-1}\mathcal{P}_\Omega - \mathcal{I}.$$

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

$$\mathcal{P}^{-1} = \sum_{k\geq 0}(\mathcal{P}_{\mathcal{T}} - \mathcal{P})^k \text{ whenever } \|\mathcal{P}_{\mathcal{T}} - \mathcal{P}\| < 1.$$

So according to Theorem 13, the operator $p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}})^{-1}$ can be represented as a *convergent* Neumann series

$$p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}})^{-1} = \sum_{k\geq 0}(-1)^k(\mathcal{P}_{\mathcal{T}}\mathcal{Q}_\Omega\mathcal{P}_{\mathcal{T}})^k,$$

because $\|\mathcal{P}_{\mathcal{T}}\mathcal{Q}_\Omega\mathcal{P}_{\mathcal{T}}\| \leq \epsilon < \frac{1}{2}$ once $m \geq C\mu n_{(1)} r \log n_{(1)}$ 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

$$\begin{aligned} \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| &= \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}})^{-1}(\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top))\| \\ &= \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_\Omega\mathcal{P}_{\mathcal{T}}p(\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}})^{-1}((\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top))\| \\ &= \|\sum_{k\geq 0}(-1)^k\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_\Omega(\mathcal{P}_{\mathcal{T}}\mathcal{Q}_\Omega\mathcal{P}_{\mathcal{T}})^k((\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top))\| \\ &\leq \sum_{k\geq 0}\|(-1)^k\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_\Omega(\mathcal{P}_{\mathcal{T}}\mathcal{Q}_\Omega\mathcal{P}_{\mathcal{T}})^k((\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top))\|_F \\ &\leq \|\mathcal{Q}_\Omega\|\sum_{k\geq 0}\|\mathcal{P}_{\mathcal{T}}\mathcal{Q}_\Omega\mathcal{P}_{\mathcal{T}}\|^k\|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top))\|_F \\ &\leq \frac{4}{p}\|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \mathbf{U}\mathbf{V}^\top)\|_F \\ &\leq \Theta\left(\frac{n_1 n_2}{m}\right)\sqrt{\frac{r}{3n_{(1)}^2}} \\ &\leq \frac{1}{3} \end{aligned}$$

with high probability. The proof is completed. ■

It thus suffices to construct a dual certificate $\mathbf{\Lambda}$ such that all conditions in (37) hold. The proof follows from (Recht, 2011). Partition $\Omega = \Omega_1 \cup \Omega_2 \cup ... \cup \Omega_b$ into b partitions of size q. By assumption, we may choose

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

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 $\mathbf{W}_0 = \mathbf{U}\mathbf{V}^\top$ and set $\mathbf{\Lambda}_k = \frac{n_1 n_2}{q} \sum_{j=1}^{k} \mathcal{P}_{\Omega_j}(\mathbf{W}_{j-1})$, $\mathbf{W}_k = \mathbf{U}\mathbf{V}^\top - \mathcal{P}_{\mathcal{T}}(\mathbf{\Lambda}_k)$ for $k = 1, 2, ..., b$. Then by Theorem 13,

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

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

The following lemma together implies the strong duality of (18) straightforwardly.

Lemma 27 *Under the assumptions of Theorem 8, the dual certification* $\mathbf{W}_b$ *obeys the dual condition* (37) *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; see Appendix I for a brief justification. By the equivalence, we can suppose $\Omega \sim \mathsf{Ber}(p)$.

Observe that by Lemma 16,

$$\|\mathbf{W}_j\|_\infty \leq \left(\frac{1}{2}\right)^j \|\mathbf{U}\mathbf{V}^\top\|_\infty,$$

and by Lemma 17,

$$\|\mathbf{W}_j\|_{\infty,2} \leq \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{W}_{j-1}\|_\infty + \frac{1}{2}\|\mathbf{W}_{j-1}\|_{\infty,2}.$$

Combining the two leads to

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

Therefore,

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \\
\leq &\sum_{j=1}^{b}\|\frac{n_1n_2}{q}\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1}\| \\
= &\sum_{j=1}^{b}\|\mathcal{P}_{\mathcal{T}^\perp}(\frac{n_1n_2}{q}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1}-\mathbf{W}_{j-1})\| \\
\leq &\sum_{j=1}^{b}\|(\frac{n_1n_2}{q}\mathcal{P}_{\Omega_j}-\mathcal{I})(\mathbf{W}_{j-1})\|.
\end{aligned}$$

Let p denote $\Theta\left(\frac{q}{n_1n_2}\right)$. By Lemma 15,

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \\
\leq &\, C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_{\infty,2} \\
\leq &\, C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\left(\frac{1}{2}\right)^j\|\mathbf{U}\mathbf{V}^\top\|_\infty \\
&+ C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\left[j\left(\frac{1}{2}\right)^j\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{U}\mathbf{V}^\top\|_\infty + \left(\frac{1}{2}\right)^j\|\mathbf{U}\mathbf{V}^\top\|_{\infty,2}\right] \\
\leq &\, C_0'\frac{\log n_{(1)}}{p}\|\mathbf{U}\mathbf{V}^\top\|_\infty + 2C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{U}\mathbf{V}^\top\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\|\mathbf{U}\mathbf{V}^\top\|_{\infty,2}.
\end{aligned}$$

We note the facts that (assume without loss of generality that $n_2 \geq n_1$)

$$\|\mathbf{U}\mathbf{V}^\top\|_{\infty,2} = \max_i\|\mathbf{e}_i^T\mathbf{U}\mathbf{V}^T\|_2 = \max_i\|\mathbf{e}_i^T\mathbf{U}\|_2 \leq \sqrt{\frac{\mu r}{n_1}},$$

and that

$$\begin{aligned}
\|\mathbf{U}\mathbf{V}^\top\|_\infty &= \max_{ij}\langle\mathbf{U}\mathbf{V}^\top, \mathbf{e}_i\mathbf{e}_j^T\rangle = \max_{ij}\langle\mathbf{e}_i^T\mathbf{U}, \mathbf{e}_j^T\mathbf{V}\rangle \\
&\leq \max_{ij}\|\mathbf{e}_i^T\mathbf{U}\|_2\|\mathbf{e}_j^T\mathbf{V}\|_2 \leq \frac{\mu r}{\sqrt{n_1n_2}}.
\end{aligned}$$

Substituting $p = \Theta\left(\frac{\mu r n_{(1)}\log^2(n_{(1)})}{n_1n_2}\right)$, we obtain $\|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{1}{3}$. The proof is completed. ■

We note that the bi-dual problem of problem (34) is problem (18); see Appendix H.2 for the details. By Lemmas 24, 26 and 27, we have $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$. We also notice from Theorem 1 of Chen (2015) that $\widetilde{\mathbf{X}} = \mathbf{X}^*$. The proof of Theorem 8 is then completed.

F. Proof of Theorem 9

Theorem 9 (Robust PCA. Restated). *Suppose $\mathbf{X}^*$ is $n_1 \times n_2$, obeys incoherence* (5) *and* (20). *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*

$$(\widetilde{\mathbf{X}}, \widetilde{\mathbf{S}}) = \underset{\mathbf{X},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{X}\|_{r*} + \lambda\|\mathbf{S}\|_1, \quad \text{s.t.} \quad \mathbf{D} = \mathbf{X} + \mathbf{S}, \tag{39}$$

with $\lambda = \frac{\sigma_r(\mathbf{X}^*)}{\sqrt{n_{(1)}}}$ is exact, i.e., $\widetilde{\mathbf{X}} = \mathbf{X}^*$ and $\widetilde{\mathbf{S}} = \mathbf{S}^*$, provided that $\mathsf{rank}(\mathbf{X}^*) \leq \frac{\rho_r n_{(2)}}{\mu \log^2 n_{(1)}}$ and $m \leq \rho_s n_1 n_2$, where c, ρ_r, and ρ_s are all positive absolute constants, and function $\|\cdot\|_{r*}$ is given by (21).

F.1 Dual Certificates

Lemma 28 *Assume that $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \leq 1/2$ and $\lambda < \sigma_r(\mathbf{X}^*)$. Then $(\mathbf{X}^*, \mathbf{S}^*)$ is the unique solution to problem* (9) *if there exists a pair $(\mathbf{W}, \mathbf{F})$ for which*

$$\mathbf{X}^* + \mathbf{W} = \lambda(\mathsf{sign}(\mathbf{S}^*) + \mathbf{F} + \mathcal{P}_\Omega \mathbf{K}),$$

where $\mathbf{W} \in \mathcal{T}^\perp$, $\|\mathbf{W}\| \leq \frac{\sigma_r(\mathbf{X}^*)}{2}$, $\mathbf{F} \in \Omega^\perp$, $\|\mathbf{F}\|_\infty \leq \frac{1}{2}$, and $\|\mathcal{P}_\Omega \mathbf{K}\|_F \leq \frac{1}{4}$.

Proof Let $(\mathbf{X}^* + \mathbf{H}, \mathbf{S}^* - \mathbf{H})$ be any optimal solution to problem (39). By the definition of the subgradient, the inequality follows

$$\begin{aligned}
&\|\mathbf{X}^* + \mathbf{H}\|_{r*} + \lambda\|\mathbf{S}^* - \mathbf{H}\|_1 \\
&\geq \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \mathbf{X}^* + \mathbf{W}^*, \mathbf{H}\rangle - \lambda\langle \mathsf{sign}(\mathbf{S}^*) + \mathbf{F}^*, \mathbf{H}\rangle \\
&= \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \mathbf{X}^* - \lambda\mathsf{sign}(\mathbf{S}^*), \mathbf{H}\rangle + \langle \mathbf{W}^*, \mathbf{H}\rangle - \lambda\langle \mathbf{F}^*, \mathbf{H}\rangle \\
&= \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \mathbf{X}^* - \lambda\mathsf{sign}(\mathbf{S}^*), \mathbf{H}\rangle + \sigma_r(\mathbf{X}^*)\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 \\
&= \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \lambda\mathbf{F} + \lambda\mathcal{P}_\Omega\mathbf{K} - \mathbf{W}, \mathbf{H}\rangle + \sigma_r(\mathbf{X}^*)\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 \\
&= \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \frac{\sigma_r(\mathbf{X}^*)}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \frac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 - \frac{\lambda}{4}\|\mathcal{P}_\Omega\mathbf{H}\|_F.
\end{aligned}$$

We note that

$$\begin{aligned}
\|\mathcal{P}_\Omega\mathbf{H}\|_F &\leq \|\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathbf{H}\|_F + \|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F \\
&\leq \frac{1}{2}\|\mathbf{H}\|_F + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F \\
&\leq \frac{1}{2}\|\mathcal{P}_\Omega\mathbf{H}\|_F + \frac{1}{2}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_F + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F,
\end{aligned}$$

which implies that $\frac{\lambda}{4}\|\mathcal{P}_\Omega\mathbf{H}\|_F \leq \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_F + \frac{\lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F \leq \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 + \frac{\lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_*$. Therefore,

$$\begin{aligned}
\|\mathbf{X}^* + \mathbf{H}\|_{r*} + \lambda\|\mathbf{S}^* - \mathbf{H}\|_1 &\geq \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \frac{\sigma_r(\mathbf{X}^*) - \lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 \\
&\geq \|\mathbf{X}^* + \mathbf{H}\|_{r*} + \lambda\|\mathbf{S}^* - \mathbf{H}\|_1 + \frac{\sigma_r(\mathbf{X}^*) - \lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1,
\end{aligned}$$

where the second inequality holds because $(\mathbf{X}^* + \mathbf{H}, \mathbf{S}^* - \mathbf{H})$ is optimal. Thus $\mathbf{H} \in \mathcal{T} \cap \Omega$. Note that $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| < 1$ implies $\mathcal{T} \cap \Omega = \{0\}$. This completes the proof. ■

According to Lemma 28, to show the exact recoverability of problem (39), it is sufficient to find an appropriate $\mathbf{W}$ for which

$$\begin{cases} \mathbf{W} \in \mathcal{T}^\perp, \\ \|\mathbf{W}\| \leq \frac{\sigma_r(\mathbf{X}^*)}{2}, \\ \|\mathcal{P}_\Omega(\mathbf{X}^* + \mathbf{W} - \lambda \mathsf{sign}(S^*))\|_F \leq \frac{\lambda}{4}, \\ \|\mathcal{P}_{\Omega^\perp}(\mathbf{X}^* + \mathbf{W})\|_\infty \leq \frac{\lambda}{2}. \end{cases} \tag{40}$$

F.2 Dual Certification by Least Squares and the Golfing Scheme

The remainder of the proof is to construct $\mathbf{W}$ such that the dual condition (40) holds true. Before introducing our construction, we assume $\Omega \sim \mathsf{Ber}(p)$, or equivalently $\Omega^\perp \sim \mathsf{Ber}(1-p)$, where p is allowed be as large as an absolute constant. Note that $\Omega^\perp$ has the same distribution as that of $\Omega_1 \cup \Omega_2 \cup ... \cup \Omega_{j_0}$, where the Ω_j's are drawn independently with replacement from $\mathsf{Ber}(q)$, $j_0 = \lceil \log n_{(1)} \rceil$, and q obeys $p = (1-q)^{j_0}$ ($q = \Omega(1/\log n_{(1)})$ implies $p = \mathcal{O}(1)$). We construct $\mathbf{W}$ based on such a distribution.

Our construction separates $\mathbf{W}$ into two terms: $\mathbf{W} = \mathbf{W}^L + \mathbf{W}^S$. To construct $\mathbf{W}^L$, we apply the golfing scheme introduced in (Gross, 2011; Recht, 2011). Specifically, $\mathbf{W}^L$ is constructed by an inductive procedure:

$$\begin{aligned} \mathbf{Y}_j = \mathbf{Y}_{j-1} + q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_\mathcal{T}(\mathbf{X}^* - \mathbf{Y}_{j-1}),\ \mathbf{Y}_0 = \mathbf{0}, \\ \mathbf{W}^L = \mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}_{j_0}. \end{aligned} \tag{41}$$

To construct $\mathbf{W}^S$, we apply the method of least squares by Candès et al. (2011), which is

$$\mathbf{W}^S = \lambda \mathcal{P}_{\mathcal{T}^\perp} \sum_{k \geq 0} (\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^k \mathsf{sign}(\mathbf{S}^*). \tag{42}$$

Note that $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \leq 1/2$. Thus $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega\| \leq 1/4$ and the Neumann series in (42) is well-defined. Observe that $\mathcal{P}_\Omega \mathbf{W}^S = \lambda(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^{-1} \mathsf{sign}(\mathbf{S}^*) = \lambda \mathsf{sign}(\mathbf{S}^*)$. So to prove the dual condition (40), it suffices to show that

$$\begin{aligned} &\text{(a)} \quad \|\mathbf{W}^L\| \leq \frac{\sigma_r(\mathbf{X}^*)}{4}, \\ &\text{(b)} \quad \|\mathcal{P}_\Omega(\mathbf{X}^* + \mathbf{W}^L)\|_F \leq \frac{\lambda}{4}, \\ &\text{(c)} \quad \|\mathcal{P}_{\Omega^\perp}(\mathbf{X}^* + \mathbf{W}^L)\|_\infty \leq \frac{\lambda}{4}, \end{aligned} \tag{43}$$

and

$$\begin{aligned} &\text{(d)} \quad \|\mathbf{W}^S\| \leq \frac{\sigma_r(\mathbf{X}^*)}{4}, \\ &\text{(e)} \quad \|\mathcal{P}_{\Omega^\perp}\mathbf{W}^S\|_\infty \leq \frac{\lambda}{4}. \end{aligned} \tag{44}$$

F.3 Proof of Dual Conditions

Since we have constructed the dual certificate $\mathbf{W}$, the remainder is to show that $\mathbf{W}$ obeys dual conditions (43) and (44) with high probability. We have the following.

Lemma 29 *Assume* $\Omega_j \sim \mathsf{Ber}(q)$, $j = 1, 2, ..., j_0$, *and* $j_0 = 2\lceil \log n_{(1)} \rceil$. *Then under the other assumptions of Theorem 9,* $\mathbf{W}^L$ *given by* (41) *obeys dual condition* (43).

Proof Let $\mathbf{Z}_j = \mathcal{P}_{\mathcal{T}}(\mathbf{X}^* - \mathbf{Y}_j) \in \mathcal{T}$. Then we have

$$\mathbf{Z}_j = \mathcal{P}_{\mathcal{T}}\mathbf{Z}_{j-1} - q^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_j}\mathcal{P}_{\mathcal{T}}\mathbf{Z}_{j-1} = (\mathcal{P}_{\mathcal{T}} - q^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega_j}\mathcal{P}_{\mathcal{T}})\mathbf{Z}_{j-1},$$

and $\mathbf{Y}_j = \sum_{k=1}^{j} q^{-1}\mathcal{P}_{\Omega_k}\mathbf{Z}_{k-1} \in \Omega^{\perp}$. We set $q = \Omega(\epsilon^{-2}\mu r \log n_{(1)}/n_{(2)})$.

Proof of (a). It holds that

$$\begin{aligned}
\|\mathbf{W}^L\| = \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}_{j_0}\| &\le \sum_{k=1}^{j_0} \|q^{-1}\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\Omega_k}\mathbf{Z}_{k-1}\| \\
&= \sum_{k=1}^{j_0} \|\mathcal{P}_{\mathcal{T}^\perp}(q^{-1}\mathcal{P}_{\Omega_k}\mathbf{Z}_{k-1} - \mathbf{Z}_{k-1})\| \\
&\le \sum_{k=1}^{j_0} \|q^{-1}\mathcal{P}_{\Omega_k}\mathbf{Z}_{k-1} - \mathbf{Z}_{k-1}\| \\
&\le C_0' \left(\frac{\log n_{(1)}}{q} \sum_{k=1}^{j_0} \|\mathbf{Z}_{k-1}\|_\infty + \sqrt{\frac{\log n_{(1)}}{q}} \sum_{k=1}^{j_0} \|\mathbf{Z}_{k-1}\|_{\infty,2} \right). \quad \text{(by Lemma 15)}
\end{aligned}$$

We note that by Lemma 16 and Lemma 17, respectively,

$$\begin{aligned}
\|\mathbf{Z}_{k-1}\|_\infty &\le \left(\frac{1}{2}\right)^{k-1} \|\mathbf{Z}_0\|_\infty, \\
\|\mathbf{Z}_{k-1}\|_{\infty,2} &\le \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_{k-2}\|_\infty + \frac{1}{2}\|\mathbf{Z}_{k-2}\|_{\infty,2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathbf{Z}_{k-1}\|_{\infty,2} &\le \left(\frac{1}{2}\right)^{k-1} \sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_0\|_\infty + \frac{1}{2}\|\mathbf{Z}_{k-2}\|_{\infty,2} \\
&\le (k-1)\left(\frac{1}{2}\right)^{k-1} \sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_0\|_\infty + \left(\frac{1}{2}\right)^{k-1} \|\mathbf{Z}_0\|_{\infty,2},
\end{aligned}$$

and so we have

$$\begin{aligned}
\|\mathbf{W}^L\| &\le C_0'\Bigg[\frac{\log n_{(1)}}{q}\sum_{k=1}^{j_0}\left(\frac{1}{2}\right)^{k-1}\|\mathbf{Z}_0\|_\infty \\
&\quad + \sqrt{\frac{\log n_{(1)}}{q}}\sum_{k=1}^{j_0}\left((k-1)\left(\frac{1}{2}\right)^{k-1}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_0\|_\infty + \left(\frac{1}{2}\right)^{k-1}\|\mathbf{Z}_0\|_{\infty,2}\right)\Bigg] \\
&\le 2C_0'\left[\frac{\log n_{(1)}}{q}\|\mathbf{X}^*\|_\infty + \sqrt{\frac{n_{(1)}\log n_{(1)}}{q\mu r}}\|\mathbf{X}^*\|_\infty + \sqrt{\frac{\log n_{(1)}}{q}}\|\mathbf{X}^*\|_{\infty,2}\right] \\
&\le 2C_0\left[\frac{n_{(2)}}{\mu r}\|\mathbf{X}^*\|_\infty + \frac{\sqrt{n_1 n_2}}{\mu r}\|\mathbf{X}^*\|_\infty + \sqrt{\frac{n_{(2)}}{\mu r}}\|\mathbf{X}^*\|_{\infty,2}\right] \quad \text{(since } q = \Omega(\mu r \log n_{(1)})/n_{(2)}\text{)} \\
&\le \frac{\sigma_r(\mathbf{X}^*)}{4}, \quad \text{(by incoherence (20))}
\end{aligned}$$

where we have used the fact that

$$\|\mathbf{X}^*\|_{\infty,2} \le \sqrt{n_{(1)}}\|\mathbf{X}^*\|_\infty \le \sqrt{\frac{\mu r}{n_{(2)}}}\sigma_r(\mathbf{X}^*).$$

Proof of (b). Because $\mathbf{Y}_{j_0} \in \Omega^\perp$, we have $\mathcal{P}_\Omega(\mathbf{X}^* + \mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}_{j_0}) = \mathcal{P}_\Omega(\mathbf{X}^* - \mathcal{P}_{\mathcal{T}}\mathbf{Y}_{j_0}) = \mathcal{P}_\Omega\mathbf{Z}_{j_0}$. It then follows from Theorem 13 that

$$\begin{aligned}
\|\mathbf{Z}_{j_0}\|_F &\le t^{j_0}\|\mathbf{X}^*\|_F \\
&\le t^{j_0}\sqrt{n_1 n_2}\|\mathbf{X}^*\|_\infty \\
&\le t^{j_0}\sqrt{n_1 n_2}\sqrt{\frac{\mu r}{n_1 n_2}}\sigma_r(\mathbf{X}^*) \\
&\le \frac{\lambda}{8}. \quad (t^{j_0} \le e^{-2\log n_{(1)}} \le n_{(1)}^{-2})
\end{aligned}$$

Proof of (c). By definition, $\mathbf{X}^* + \mathbf{W}^L = \mathbf{Z}_{j_0} + \mathbf{Y}_{j_0}$. Since we have shown $\|\mathbf{Z}_{j_0}\|_F \le \lambda/8$, it suffices to prove $\|\mathbf{Y}_{j_0}\|_\infty \le \lambda/8$. We have

$$\begin{aligned}
\|\mathbf{Y}_{j_0}\|_\infty &\le q^{-1}\sum_{k=1}^{j_0}\|\mathcal{P}_{\Omega_k}\mathbf{Z}_{k-1}\|_\infty \\
&\le q^{-1}\sum_{k=1}^{j_0}\epsilon^{k-1}\|\mathbf{X}^*\|_\infty \quad \text{(by Lemma 16)} \\
&\le \frac{n_{(2)}\epsilon^2}{C_0\mu r\log n_{(1)}}\sqrt{\frac{\mu r}{n_1 n_2}}\sigma_r(\mathbf{X}^*) \quad \text{(by incoherence (20))} \\
&\le \frac{\lambda}{8},
\end{aligned}$$

if we choose $\epsilon = C\left(\frac{\mu r(\log n_{(1)})^2}{n_{(2)}}\right)^{1/4}$ for an absolute constant C. This can be true once the constant ρ_r is sufficiently small. ■

We now prove that $\mathbf{W}^S$ given by (42) obeys dual condition (44).

Lemma 30 *Assume $\Omega \sim \mathsf{Ber}(p)$. Then under the other assumptions of Theorem 9, $\mathbf{W}^S$ given by* (42) *obeys dual condition* (44)*.*

Proof According to the standard de-randomization argument (Candès et al., 2011), it is equivalent to studying the case when the signs δ_{ij} of $\mathbf{S}^*_{ij}$ are independently distributed as

$$\delta_{ij} = \begin{cases} 1, & \text{w.p. } p/2, \\ 0, & \text{w.p. } 1-p, \\ -1, & \text{w.p. } p/2. \end{cases}$$

Proof of (d). Recall that

$$\begin{aligned} \mathbf{W}^S &= \lambda \mathcal{P}_{\mathcal{T}^\perp} \sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_{\mathcal{T}} \mathcal{P}_\Omega)^k \mathsf{sign}(\mathbf{S}^*) \\ &= \lambda \mathcal{P}_{\mathcal{T}^\perp} \mathsf{sign}(\mathbf{S}^*) + \lambda \mathcal{P}_{\mathcal{T}^\perp} \sum_{k\geq 1} (\mathcal{P}_\Omega \mathcal{P}_{\mathcal{T}} \mathcal{P}_\Omega)^k \mathsf{sign}(\mathbf{S}^*). \end{aligned}$$

To bound the first term, we have $\|\mathsf{sign}(\mathbf{S}^*)\| \leq 4\sqrt{n_{(1)}p}$ as shown in (Vershynin, 2010). So $\|\lambda \mathcal{P}_{\mathcal{T}^\perp} \mathsf{sign}(\mathbf{S}^*)\| \leq \lambda \|\mathsf{sign}(\mathbf{S}^*)\| \leq 4\sqrt{p}\sigma_r(\mathbf{X}^*) \leq \sigma_r(\mathbf{X}^*)/8$.

We now bound the second term. Let $\mathcal{G} = \sum_{k\geq 1} (\mathcal{P}_\Omega \mathcal{P}_{\mathcal{T}} \mathcal{P}_\Omega)^k$, which is self-adjoint, and denote by N_{n_1} and N_{n_2} the $\frac{1}{2}$-nets of $\mathbb{S}^{n_1-1}$ and $\mathbb{S}^{n_1-1}$ of sizes at most 6^{n_1} and 6^{n_2}, respectively (Ledoux, 2005). Vershynin (2010) has shown that

$$\begin{aligned} \|\mathcal{G}(\mathsf{sign}(\mathbf{S}^*))\| &= \sup_{\mathbf{x}\in\mathbb{S}^{n_2-1}, \mathbf{y}\in\mathbb{S}^{n_1-1}} \langle \mathcal{G}(\mathbf{y}\mathbf{x}^T), \mathsf{sign}(\mathbf{S}^*)\rangle \\ &\leq 4 \sup_{\mathbf{x}\in N_{n_2}, \mathbf{y}\in N_{n_1}} \langle \mathcal{G}(\mathbf{y}\mathbf{x}^T), \mathsf{sign}(\mathbf{S}^*)\rangle. \end{aligned}$$

Consider the random variable $X(\mathbf{x},\mathbf{y}) = \langle \mathcal{G}(\mathbf{y}\mathbf{x}^T), \mathsf{sign}(\mathbf{S}^*)\rangle$ which has zero expectation. By Hoeffding's inequality, we have

$$\Pr(|X(\mathbf{x},\mathbf{y})| > t|\Omega) \leq 2\exp\left(-\frac{2t^2}{\|\mathcal{G}(\mathbf{x}\mathbf{y}^T)\|_F^2}\right) \leq 2\exp\left(-\frac{2t^2}{\|\mathcal{G}\|^2}\right).$$

Therefore, by a union bound,

$$\Pr(\|\mathcal{G}(\mathsf{sign}(\mathbf{S}^*))\| > t|\Omega) \leq 2\times 6^{n_1+n_2}\exp\left(-\frac{t^2}{8\|\mathcal{G}\|^2}\right).$$

Note that conditioned on the event $\{\|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\| \leq \sigma\}$, we have $\|\mathcal{G}\| = \left\|\sum_{k\geq 1}(\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega)^k\right\| \leq \frac{\sigma^2}{1-\sigma^2}$. So

$$\begin{aligned} &\Pr(\lambda\|\mathcal{G}(\mathsf{sign}(\mathbf{S}^*))\| > t) \\ &\leq 2\times 6^{n_1+n_2}\exp\left(-\frac{t^2}{8\lambda^2}\left(\frac{1-\sigma^2}{\sigma^2}\right)^2\right)\Pr(\|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\| \leq \sigma) + \Pr(\|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\| > \sigma). \end{aligned}$$

The following lemma guarantees that event $\{\|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\| \leq \sigma\}$ holds with high probability for a very small absolute constant σ.

Lemma 31 (Candès et al. (2011), Corollary 2.7) *Suppose that* $\Omega \sim \mathsf{Ber}(p)$ *and incoherence* (5) *holds. Then with probability at least* $1 - n_{(1)}^{-10}$, $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\|^2 \leq p + \epsilon$, *provided that* $1 - p \geq C_0 \epsilon^{-2} \mu r \log n_{(1)}/n_{(2)}$ *for an absolute constant* C_0.

Setting $t = \frac{\sigma_r(\mathbf{X}^*)}{8}$, this completes the proof of (d). ■

Proof of (e). Recall that $\mathbf{W}^S = \lambda \mathcal{P}_{\mathcal{T}^\perp} \sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^k \mathsf{sign}(\mathbf{S}^*)$ and so

$$\begin{aligned}\mathcal{P}_{\Omega^\perp} \mathbf{W}^S &= \lambda \mathcal{P}_{\Omega^\perp} (\mathcal{I} - \mathcal{P}_\mathcal{T}) \sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^k \mathsf{sign}(\mathbf{S}^*) \\ &= -\lambda \mathcal{P}_{\Omega^\perp} \mathcal{P}_\mathcal{T} \sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^k \mathsf{sign}(\mathbf{S}^*).\end{aligned}$$

Then for any $(i,j) \in \Omega^\perp$, we have

$$\mathbf{W}^S_{ij} = \langle \mathbf{W}^S, \mathbf{e}_i \mathbf{e}_j^T \rangle = \left\langle \lambda \mathsf{sign}(\mathbf{S}^*), -\sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^k \mathcal{P}_\Omega \mathcal{P}_\mathcal{T} (\mathbf{e}_i \mathbf{e}_j^T) \right\rangle.$$

Let $X(i,j) = -\sum_{k\geq 0} (\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^k \mathcal{P}_\Omega \mathcal{P}_\mathcal{T} (\mathbf{e}_i \mathbf{e}_j^T)$. By Hoeffding's inequality and a union bound,

$$\Pr\left(\sup_{ij} |\mathbf{W}^S_{ij}| > t | \Omega \right) \leq 2 \sum_{ij} \exp\left(-\frac{2t^2}{\lambda^2 \|X(i,j)\|_F^2} \right).$$

We note that conditioned on the event $\{\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \leq \sigma\}$, for any $(i,j) \in \Omega^\perp$,

$$\begin{aligned}\|X(i,j)\|_F &\leq \frac{1}{1-\sigma^2} \sigma \|\mathcal{P}_\mathcal{T}(\mathbf{e}_i \mathbf{e}_j^T)\|_F \\ &\leq \frac{1}{1-\sigma^2} \sigma \sqrt{1 - \|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{e}_i \mathbf{e}_j^T)\|_F^2} \\ &= \frac{1}{1-\sigma^2} \sigma \sqrt{1 - \|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{e}_i\|_2^2 \|(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{e}_j\|_2^2} \\ &\leq \frac{1}{1-\sigma^2} \sigma \sqrt{1 - \left(1 - \frac{\mu r}{n_1}\right)\left(1 - \frac{\mu r}{n_2}\right)} \\ &\leq \frac{1}{1-\sigma^2} \sigma \sqrt{\frac{\mu r}{n_1} + \frac{\mu r}{n_2}}.\end{aligned}$$

Then unconditionally,

$$\begin{aligned}\Pr\left(\sup_{ij} |\mathbf{W}^S_{ij}| > t \right) \leq & 2 n_1 n_2 \exp\left(-\frac{2t^2}{\lambda^2} \frac{(1-\sigma^2)^2 n_1 n_2}{\sigma^2 \mu r (n_1 + n_2)} \right) \Pr(\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \leq \sigma) \\ & + \Pr(\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| > \sigma).\end{aligned}$$

By Lemma 31 and setting $t = \lambda/4$, the proof of (e) is completed.

G. Proof of Theorem 10

Our computational lower bound for problem (**P**) (and (**P'**)) assumes the average-case hardness of SAT, which has been used in (Razenshteyn et al., 2016; Song et al., 2019).

Conjecture 32 (Random 4-SAT, Razenshteyn et al. (2016)) *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 32, we have the following computational lower bound for problem (**P**) (and (**P'**)). We show that problem (**P**) (and (**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 10 (Computational Lower Bound. Restated)**.** *Assume Conjecture 32. Then there exists an absolute constant $\epsilon_0 > 0$ for which any algorithm that achieves $(1+\epsilon)$OPT in objective function value for problem* (**P**) *(and* (**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 10 is proved by using the hypothesis that random 4-SAT is hard to show hardness of the Maximum Edge Biclique problem for deterministic algorithms. The same construction is used for both problem (**P**) and problem (**P'**), so we only present the proof for the former in the following.

Definition 33 (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.

Goerdt and Lanka (2004) showed that under the random 4-SAT assumption there exist two constants $1 > \epsilon_1 > \epsilon_2 > 0$ 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 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 $(n/16)^2(1+\epsilon_1)$ 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 (Goerdt and Lanka, 2004) 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 (**P**) except the input description of H is padded with a string of 1's of length $2^{(\alpha/2)(n_1+n_2)}$. This string is irrelevant for solving problem (**P**) but changes the input size to $N = \mathrm{poly}(n_1, n_2) + 2^{(\alpha/2)(n_1+n_2)}$. 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 $\mathrm{poly}(N)$ time. This contradicts the assumption that BPP = P. This completes the proof. ■

H. Dual and Bi-Dual Problems

In this section, we will present the dual and bi-dual problems of r^* minimization and nuclear norm minimization, respectively.

H.1 r^* Minimization

We derive the dual and bi-dual problems of non-convex program (**P**). According to (7), the primal problem (**P**) is equivalent to

$$\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} \frac{1}{2}\| - \mathbf{\Lambda} - \mathbf{AB}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - H^*(\mathbf{\Lambda}).$$

Therefore, the dual problem is given by

$$\begin{aligned}
&\max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\| - \mathbf{\Lambda} - \mathbf{AB}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - H^*(\mathbf{\Lambda}) \\
&= \max_{\mathbf{\Lambda}} \frac{1}{2}\sum_{i=r+1}^{n_{(2)}} \sigma_i^2(-\mathbf{\Lambda}) - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - H^*(\mathbf{\Lambda}) \\
&= \max_{\mathbf{\Lambda}} -\frac{1}{2}\|\mathbf{\Lambda}\|_r^2 - H^*(\mathbf{\Lambda}), \qquad \textbf{(D1)}
\end{aligned}$$

where $\|\mathbf{\Lambda}\|_r^2 = \sum_{i=1}^r \sigma_i^2(\mathbf{\Lambda})$. The bi-dual problem is derived by

$$\begin{aligned}
&\min_{\mathbf{M}} \max_{\mathbf{\Lambda},\mathbf{\Lambda}'} -\frac{1}{2}\|\mathbf{\Lambda}\|_r^2 - H^*(\mathbf{\Lambda}') + \langle \mathbf{M}, \mathbf{\Lambda}' - \mathbf{\Lambda}\rangle \\
&= \min_{\mathbf{M}} \max_{-\mathbf{\Lambda}} \left[\langle \mathbf{M}, -\mathbf{\Lambda}\rangle - \frac{1}{2}\| - \mathbf{\Lambda}\|_r^2\right] + \max_{\mathbf{\Lambda}'} \left[\langle \mathbf{M}, \mathbf{\Lambda}'\rangle - H^*(\mathbf{\Lambda}')\right] \\
&= \min_{\mathbf{M}} \|\mathbf{M}\|_{r*} + H(\mathbf{M}), \qquad \textbf{(D2)}
\end{aligned}$$

where $\|\mathbf{M}\|_{r*} = \max_{\mathbf{X}} \langle \mathbf{M}, \mathbf{X}\rangle - \frac{1}{2}\|\mathbf{X}\|_r^2$ is a convex function, and the argument $H(\mathbf{M}) = \max_{\mathbf{\Lambda}'} \left[\langle \mathbf{M}, \mathbf{\Lambda}'\rangle - H^*(\mathbf{\Lambda}')\right]$ holds by the definition of conjugate function.

H.2 Nuclear Norm Minimization

In this section, we derive the dual and bi-dual problems of non-convex program (34). First note that

$$\begin{aligned}
&\min_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* + H(\mathbf{AB}) \\
=&\min_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* + H^{**}(\mathbf{AB}) \\
=&\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} \|\mathbf{AB}\|_* + \langle \mathbf{\Lambda}, \mathbf{AB}\rangle - H^*(\mathbf{\Lambda}).
\end{aligned}$$

Therefore, the dual problem is given by

$$\begin{aligned}
&\max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* + \langle \mathbf{\Lambda}, \mathbf{AB}\rangle - H^*(\mathbf{\Lambda}) \\
=&\max_{\|\mathbf{\Lambda}\|\le 1} -H^*(\mathbf{\Lambda}), \qquad \textbf{(D1'')}
\end{aligned}$$

where $\|\mathbf{\Lambda}\| \le 1$ because otherwise, $\min_{\mathbf{A},\mathbf{B}} \|\mathbf{AB}\|_* + \langle \mathbf{\Lambda}, \mathbf{AB}\rangle - H^*(\mathbf{\Lambda}) = -\infty$ while we would like to maximize it over all $\mathbf{\Lambda}$.

The bi-dual problem is derived by

$$\begin{aligned}
&\min_{\mathbf{M}} \max_{\|\mathbf{\Lambda}\|\le 1, \mathbf{\Lambda}'} -H^*(\mathbf{\Lambda}') + \langle \mathbf{M}, \mathbf{\Lambda}' - \mathbf{\Lambda}\rangle \\
=&\min_{\mathbf{M}} \max_{\|-\mathbf{\Lambda}\|\le 1} \langle \mathbf{M}, -\mathbf{\Lambda}\rangle + \max_{\mathbf{\Lambda}'} \left[\langle \mathbf{M}, \mathbf{\Lambda}'\rangle - H^*(\mathbf{\Lambda}')\right] \\
=&\min_{\mathbf{M}} \|\mathbf{M}\|_* + H(\mathbf{M}), \qquad \textbf{(D2'')}
\end{aligned}$$

where $H(\mathbf{M}) = \max_{\mathbf{\Lambda}'} \left[\langle \mathbf{M}, \mathbf{\Lambda}'\rangle - H^*(\mathbf{\Lambda}')\right]$ holds by the definition of conjugate function, and $\|\mathbf{M}\|_* = \max_{\|-\mathbf{\Lambda}\|\le 1} \langle \mathbf{M}, -\mathbf{\Lambda}\rangle$ because $\|\cdot\|_*$ and $\|\cdot\|$ are a pair of dual norms.

I. Equivalence of Bernoulli and Uniform Models

We begin by arguing that a recovery result under the Bernoulli model with some probability automatically implies a corresponding result for the uniform model with at least the same probability. The argument follows Section 7.1 of Candès et al. (2011). For completeness, we provide the proof here.

Denote by $\Pr_{\mathsf{Unif}(m)}$ and $\Pr_{\mathsf{Ber}(p)}$ probabilities calculated under the uniform and Bernoulli models and let "Success" be the event that the algorithm succeeds. We have

$$\begin{aligned}
\Pr_{\mathsf{Ber}(p)}(\text{Success}) &= \sum_{k=0}^{n_1 n_2} \Pr_{\mathsf{Ber}(p)}(\text{Success} \mid |\Omega| = k) \Pr_{\mathsf{Ber}(p)}(|\Omega| = k) \\
&\le \sum_{k=0}^{m} \Pr_{\mathsf{Unif}(k)}(\text{Success} \mid |\Omega| = k) \Pr_{\mathsf{Ber}(p)}(|\Omega| = k) + \sum_{k=m+1}^{n_1 n_2} \Pr_{\mathsf{Ber}(p)}(|\Omega| = k) \\
&\le \Pr_{\mathsf{Unif}(m)}(\text{Success}) + \Pr_{\mathsf{Ber}(p)}(|\Omega| > m),
\end{aligned}$$

where we have used the fact that for $k \le m$, $\Pr_{\mathsf{Unif}(k)}(\text{Success}) \le \Pr_{\mathsf{Unif}(m)}(\text{Success})$, and that the conditional distribution of $|\Omega|$ is uniform. Thus

$$\Pr_{\mathsf{Unif}(m)}(\text{Success}) \ge \Pr_{\mathsf{Ber}(p)}(\text{Success}) - \Pr_{\mathsf{Ber}(p)}(|\Omega| > m).$$

Take $p = m/(n_1 n_2) - \epsilon$, where $\epsilon > 0$. The conclusion follows from $\Pr_{\mathsf{Ber}(p)}(|\Omega| > m) \le e^{-\frac{\epsilon^2 n_1 n_2}{2p}}$.

References

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.

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 the Theory of Computing*, pages 1195–1199, 2017.

Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

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.

Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *Annual Conference on Learning Theory*, 2016.

Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1): 2773–2832, 2014.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *ACM Symposium on Theory of Computing*, pages 145–162, 2012.

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, 2014.

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.

Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

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.

Maria-Florina Balcan, Yingyu Liang, David P. Woodruff, and Hongyang Zhang. Matrix completion and related problems via strong duality. In *Innovations in Theoretical Computer Science*, volume 94, 2018.

Maria-Florina Balcan, Yi Li, David P Woodruff, and Hongyang Zhang. Testing matrix rank, optimally. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 727–746, 2019.

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.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *ACM Symposium on Theory of Computing*, pages 594–603, 2014.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Annual Conference on Learning Theory*, pages 530–582, 2016a.

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, 2016b.

Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the applications of robust pca in image and video processing. *Proceedings of the IEEE*, 106(8): 1427–1457, 2018.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J. Candès and Benjamin Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, pages 1–13, 2013.

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.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

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.

Minming Chen, Zhouchen Lin, Yi Ma, and Leqin Wu. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Coordinated Science Laboratory Report no. UILU-ENG-09-2215*, 2009.

Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

Yudong Chen and Martin J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *International Conference on Artificial Intelligence and Statistics*, 2015.

Uriel Feige. Relations between average case complexity and approximation complexity. In *IEEE Conference on Computational Complexity*, page 5, 2002.

Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *Annual Conference on Learning Theory*, pages 315–340, 2011.

David Gamarnik, Quan Li, and Hongyi Zhang. Matrix completion from $o(n)$ samples in linear time. In *Annual Conference on Learning Theory*, volume 65, pages 940–947, 2017.

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.

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.

Rong Ge, Chi Jin, and Zheng Yi. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *ICML*, 2017.

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

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

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

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.

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.

Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

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

Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. *COLT*, 2013.

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

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.

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.

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.

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.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *ICML*, 2017.

Michel Journée, Francis Bach, P-A Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

Kenji Kawaguchi. Deep learning without poor local minima. *arXiv preprint arXiv:1605.07110*, 2016.

Raghunandan H Keshavan. *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University, 2012.

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

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010b.

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

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2005.

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.

Zhouchen Lin and Hongyang Zhang. *Low Rank Models for Visual Analysis: Theories, Algorithms and Applications*. Elsevier, 2017.

Ankur Moitra. An almost optimal algorithm for computing nonnegative rank. *SIAM Journal on Computing*, 45(1):156–173, 2016.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.

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.

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.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *International Conference on Artificial Intelligence and Statistics*, volume 54, pages 65–74, 2017.

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.

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

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.

Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise ℓ_1-norm error. In *Annual Symposium on the Theory of Computing*, 2017.

Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 2772–2789, 2019.

Nathan Srebro. *Learning with matrix factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560, 2005.

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

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017a.

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, 2017b.

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

Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE Transactions on Information Theory*, 59(11):7465–7490, 2013.

Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *ICML*, 2016.

Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

Roman Vershynin. Lectures in geometric functional analysis. pages 1–76, 2009.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint: 1011.3027*, 2010.

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

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

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.

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.

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.

Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Junbin Gao. Robust latent low rank representation for subspace clustering. *Neurocomputing*, 145:369–373, 2014.

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, 2015a.

Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Junbin Gao. Relations among some low rank subspace recovery models. *Neural Computation*, 27:1915–1950, 2015b.

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.

Hongyang Zhang, Junru Shao, and Ruslan Salakhutdinov. Deep neural networks with multi-branch architectures are intrinsically less non-convex. In *International Conference on Artificial Intelligence and Statistics*, pages 1099–1109, 2019.

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

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.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

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