One-sided Matrix Completion from Two Observations Per Row

Steven Cao [1] **Percy Liang** [1] **Gregory Valiant** [1]

Abstract

Given only a few observed entries from a low-rank matrix X, matrix completion is the problem of imputing the missing entries, and it formalizes a wide range of real-world settings that involve estimating missing data. However, when there are too few observed entries to complete the matrix, what other aspects of the underlying matrix can be reliably recovered? We study one such problem setting, that of "one-sided" matrix completion, where our goal is to recover the right singular vectors of X, even in the regime where recovering the left singular vectors is impossible, which arises when there are more rows than columns and very few observations. We propose a natural algorithm that involves imputing the missing values of the matrix X^TX and show that even with only two observations per row in X, we can provably recover X^TX as long as we have at least $\Omega(r^2 d \log d)$ rows, where r is the rank and d is the number of columns. We evaluate our algorithm on one-sided recovery of synthetic data and low-coverage genome sequencing. In these settings, our algorithm substantially outperforms standard matrix completion and a variety of direct factorization methods.

1. Introduction

Matrix completion, the problem of recovering a low-rank matrix after observing only a subset of its entries, formalizes a wide range of real-world settings that involve estimating missing data, including recommending movies to users (Koren et al., 2009), reducing MRI scan time via parallel imaging (Shin et al., 2014), and quantifying annotator disagreement in dataset crowdsourcing (Gordon et al., 2021). Over the years, a flurry of research has produced a robust understanding of the problem (Candès & Recht, 2009; Keshavan et al., 2009, *inter alia*). However, most of our understanding is restricted to settings where each row and each column have more observations than the rank of the underlying matrix. It is natural that past work operated under this assumption because full matrix completion is impossible without it: for a rank-r matrix X with shape $m \times d$, one can show that estimating the matrix is impossible with $o(r(m+d))$ observations. Nonetheless, many important applications do not satisfy this assumption: for example, in low-coverage genotype imputation (Li et al., 2009), we might sequence $d = 2{,}000$ people for $10{,}000$ genetic variants each, out of the $m = 10{,}000{,}000$ genetic variants in humans. Represented as a matrix, we have a $10{,}000{,}000 \times 2{,}000$ matrix with $2{,}000 * 10{,}000 = 20{,}000{,}000$ total observations, or about two observations per row on average, which is certainly much less than the rank of the matrix. Given too few observed entries to fully complete the matrix, what other aspects of the underlying matrix can be reliably recovered?

In this paper, we study settings like these and show that even with just *two* entries per row, as long as there are sufficiently many rows, we can perform "one-sided" recovery and estimate the right singular vectors $Q \in \mathbb{R}^{d\times r}$. In the aforementioned genomics example, this means recovering the r underlying genotype variation factors for each of the d people (e.g. ethnicity, sex, and so on; see Figure 1). This result is despite the fact that we have close to no information about the left singular vectors $P \in \mathbb{R}^{m\times r}$.

This result might seem counterintuitive due to how ill-posed the recovery problem is. In particular, let $X = UV^T$ denote the ground truth rank-r matrix with factors $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{d\times r}$. Then, given fewer than r observations per row, for *any* $\tilde{V} \in \mathbb{R}^{d\times r}$ for which every subset of r rows is linearly independent, we can find a $\tilde{U} \in \mathbb{R}^{m\times r}$ using linear inversion such that $\tilde{X} = \tilde{U}\tilde{V}^T$ agrees with the observations.[1] In other words, regardless of how many rows we have, observing two entries per row is not enough

[1]Stanford University. Correspondence to: Steven Cao <shcao@stanford.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]For example, suppose we are given exactly two observations per row $X_{i,a(i)}$ and $X_{i,b(i)}$, where i is the row index and $a(i), b(i)$ are the two observed locations. Then, given any pairwise linearly independent set of rank-2 vectors $v_1, ..., v_d \in \mathbb{R}^2$, we can choose each $u_1, ..., u_m \in \mathbb{R}^2$ by inverting the constraints $\langle u_i, v_{a(i)}\rangle = X_{i,a(i)}$ and $\langle u_i, v_{b(i)}\rangle = X_{i,b(i)}$. Then, stacking the u_i's and v_i's, we have a matrix $\tilde{X} = \tilde{U}\tilde{V}^T$ that agrees with the data.

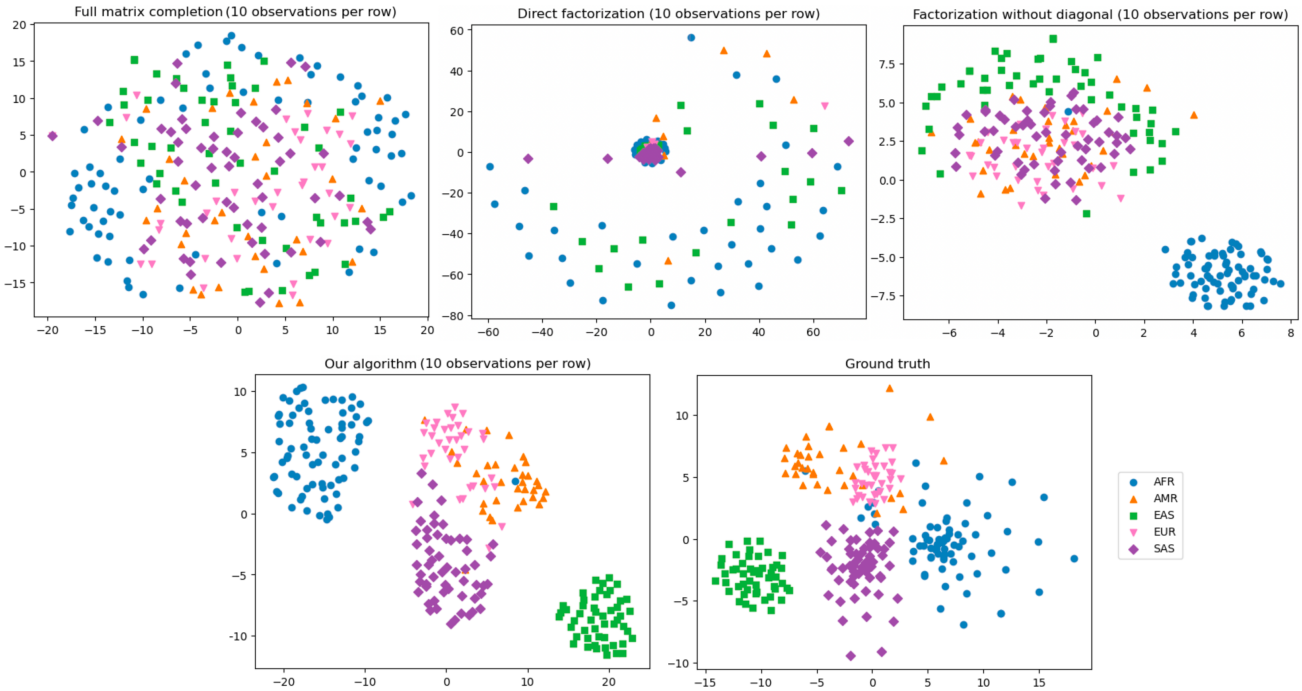

Figure 1. TSNE (van der Maaten & Hinton, 2008) visualization of column factor (i.e. eigenvalue-weighted right singular vector) recovery on the 1000genomes dataset (Fairley et al., 2019) with $m = 1\,500\,000$ bialleles, $d = 250$ people, and $k = 10$ observations per row. Each of the d column factors can be thought of as a vector representation of each person and their underlying genotype variation factors, with ethnicity being the strongest contributor. Colors represent ethnicity (AFR: blue, AMR: orange, EAS: green, EUR: pink, SAS: purple). Methods (left to right, top to bottom): (1) Full matrix completion: factor the output of matrix completion on $P_E(X)$ (where given the observation mask E, P_E sets unobserved entries to zero), (2) Direct factorization: factor $P_E(X)^T P_E(X)$, (3) Factorization without diagonal: factor $P_E(X)^T P_E(X)$ with the diagonal set to zero (Cai et al., 2021), (4) Our algorithm: perform matrix completion to estimate $X^T X$ (see Equation 1) and then factor it, and (5) Ground truth: factor the original fully observed matrix. Our method produces column factors closest to the ground truth, while full matrix completion fails because there are too few observations per row.

to distinguish between a wide range of low-rank matrices, each with drastically different right singular vectors.

So then how is recovery possible? Recall that when constructing the distractors $\tilde{X}$, the left factors $\tilde{U}$ were chosen via linear inversion, making the observation locations and $\tilde{X}$ dependent on one another. Then, to rule out these distractors, we can make the assumption that the observation locations are chosen randomly, independently of the underlying matrix. Intuitively, this assumption lets us rule out matrices that are too "correlated" with the observation locations: for example, we can rule out candidate matrices whose entries consistently are smaller at the observation locations than outside, which is often the case for distractor matrices constructed by linear inversion. While our algorithm does not explicitly rule out matrices in this way, we provide this discussion to emphasize the importance of random sampling in this setting and provide intuition for why the apparent ill-posedness is not fatal.

Algorithmically, our main idea is that each pair of observations in a row, denoted $X_{i,a(i)}$ and $X_{i,b(i)}$, produces a noisy estimate of entry $(a(i), b(i))$ of the matrix $\frac{1}{m}X^T X$. Then, given enough rows, we have sufficiently many observations to impute the missing entries of $\frac{1}{m}X^T X$, which we can then factor to obtain the right-side singular vectors (or equivalently, recover the rowspace) of X. Specifically, we find that $m = \Omega(\alpha^2 r d \log d)$ rows is sufficient to complete $\frac{1}{m}X^T X$ in additive Frobenius error, where α denotes the maximum squared entry $\alpha = \max_{ij} X_{ij}^2$. From this bound, we show that rowspace recovery is possible with $m = \Omega(r^2 d \log d)$ rows, with synthetic experiments suggesting that the r^2 dependence is fundamental.

The idea of operating on $X^T X$ when X is unbalanced has been explored in a variety of papers on noisy matrix factorization and subspace estimation, problems closely related to ours (Gonen et al., 2016; Donoho & Feldman, 2022; Montanari & Wu, 2022, *inter alia*). Broadly speaking, these papers analyze the error of directly factoring a noisy or incomplete version of $X^T X$, while we focus on showing how the missing values of $X^T X$ can be imputed. The papers most directly related to ours are Montanari & Sun (2018) and Cai et al. (2021), who study direct factoriza-

tion for subspace estimation from incomplete observations. The main difference with our work is that they assume that the observations are uniform over the matrix, making it unclear how reliant the algorithm is on the heavy rows which happen to have many observations. In contrast, we study one-sided recovery in the setting where *every* row has just two observations. Nonetheless, their direct factorization algorithms are still applicable to our setting. In our experiments, we find that compared to direct factorization, imputing the missing values of X^TX before factoring produces substantially better subspace estimates.

Empirically, we find that our algorithm indeed recovers the right-side singular vectors even when full matrix completion is impossible. We evaluate on synthetic data and the 1000genomes dataset (Fairley et al., 2019), a real-world example where the matrix is highly incomplete with many more rows than columns. In these settings, our algorithm substantially outperforms standard matrix completion and a variety of direct factorization methods.

2. Main Result

Notation: for a matrices A and B, we use $\|A\|_{\max}$ to denote $\max_{i,j}|A_{ij}|$, $\|A\|_{\text{op}}$ to denote the operator norm, $\|A\|_{\text{nuc}}$ to denote the nuclear norm, $\|A\|_F$ to denote the Frobenius norm, and $\langle A, B\rangle = \text{tr}(A^TB)$ to denote the matrix inner product. We use $E_{i,j} \in \mathbb{R}^{d\times d}$ to denote the matrix with 1 in entry (i,j) and 0 elsewhere, $f \lesssim g$ to denote that there exists a universal constant c such that $f \le cg$, and $[d]$ to denote the set $\{1, 2, ..., d\}$. Given an observation mask $E \in \{0,1\}^{m\times d}$ and a matrix $A \in \mathbb{R}^{m\times d}$, $P_E(A)$ will denote the elementwise multiplication of E and A.

The problem setup is as follows: from a rank-r matrix $X \in \mathbb{R}^{m\times d}$, we randomly observe two entries per row, which we can represent as indices $(a(1), b(1)), ..., (a(m), b(m))$ drawn i.i.d. uniformly from $\{(j_1, j_2) : j_1, j_2 \in [d],\ j_1 \neq j_2\}$. We wish to estimate the matrix $\Theta^* = \frac{1}{m}X^TX$.

As an example to provide intuition, let entry $X_{i,j}$ represent the ith user's rating for the jth item. Writing X as UV^T for $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{d\times r}$, each entry $X_{i,j}$ can be written $\langle u_i, v_j\rangle$ where $u_i, v_j \in \mathbb{R}^r$ are the ith and jth rows of U and V. Then, we can think of u_i as representing the ith user's preferences along r attributes, v_i as representing the jth item's r attributes, and the rating X_{ij} as their inner product. We'll informally refer to U as row (i.e. user) factors and V as column (i.e. item) factors (note that this decomposition of X into U and V is not unique).

Then, in this example, we can write

$$\Theta^* = \frac{1}{m}X^TX = V\bar{S}V^T,$$

where $\bar{S}$ is given by $\bar{S} = \frac{1}{m}\sum_{i=1}^m u_iu_i^T \in \mathbb{R}^{r\times r}$. Written this way, the $(j_1, j_2)$the entry of Θ^* is given by $v_{j_1}^T\bar{S}v_{j_2}$, which is the inner product (induced by $\bar{S}$) between the j_1th and j_2th column factors. Therefore, we can interpret our goal of recovering Θ^* as recovering a matrix of pairwise "column factor similarities."

Why might recovering pairwise similarities be possible? For each pair of observations $X_{i,a(i)} = \langle u_i, v_{a(i)}\rangle$ and $X_{i,b(i)} = \langle u_i, v_{b(i)}\rangle$, while we don't know u_i, the two observations should on average be similar if the inner product $\langle v_{a(i)}, v_{b(i)}\rangle_{\bar{S}}$ is positive, and dissimilar if the inner product is negative. Then, using the product $X_{i,a(i)}X_{i,b(i)}$ as our empirical observation for the similarity between $a(i)$ and $b(i)$, our estimator involves optimizing the following squared loss with a nuclear norm regularizer:[2]

$$\hat{\Theta} \in \underset{\|\Theta\|_{\max}\le\alpha}{\text{argmin}}\ \mathcal{L}(\Theta) + \lambda\|\Theta\|_{\text{nuc}}, \tag{1}$$

$$\begin{aligned}\mathcal{L}(\Theta) = \frac{1}{4m}\sum_{i=1}^m \Big[&(\Theta_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})^2 \\ &+ (\Theta_{b(i),a(i)} - X_{i,b(i)}X_{i,a(i)})^2 \\ &+ (\Theta_{a(i),a(i)} - X_{i,a(i)}^2)^2 \\ &+ (\Theta_{b(i),b(i)} - X_{i,b(i)}^2)^2\Big].\end{aligned} \tag{2}$$

With this estimator, we have the following error bound:

Theorem 2.1 (Main result)**.** *Let $\hat{\Theta}$ be the solution of the optimization problem defined in Equation 1, where λ is set to $16\alpha\sqrt{\frac{\log d+\delta}{dm}}$. Also, suppose that X is rank r with $\|X\|_{max}^2 \le \alpha$, and $m \ge d(\log d+\delta)$. Then, with probability $\ge 1 - 3e^{-\delta}$, we have that*

$$\frac{1}{d^2}\|\hat{\Theta} - \Theta^*\|_F^2 \lesssim \alpha^2\frac{rd(\log d + \delta)}{m}.$$

From this theorem, we can derive the following two corollaries: first, because X and $\Theta^* = \frac{1}{m}X^TX$ have the same rowspace, we can use recovery of Θ^* to estimate the rowspace of X. As is standard, we can measure rowspace recovery error as the error in estimating the right-side singular vectors up to rotation, producing the following:

Corollary 2.2 (Right-side singular vector recovery)**.** *Under the same conditions as Theorem 2.1, let $Q \in \mathbb{R}^{d\times r}$ denote the right-side singular vectors of X, and let $\hat{Q} \in \mathbb{R}^{d\times r}$ be the top r singular vectors of $\hat{\Theta}$. Then, letting σ_r be the rth singular value of $\Theta^* = \frac{1}{m}X^TX$, we have*

$$\min_{R\in\mathbb{R}^{r\times r}:R^TR=I_r}\|\hat{Q}R - Q\|_F^2 \lesssim \left(\frac{d\alpha}{\sigma_r}\right)^2\frac{rd(\log d + \delta)}{m}.$$

[2]Note that this program is convex and can be solved via semidefinite programming as is standard in matrix completion. In our experiments, we instead do non-convex gradient descent for computational efficiency, as discussed in Section 5.

In this corollary, a σ_r factor appears in the denominator because the algorithm is tasked with recovering all r singular directions, even if some have low weight (i.e. σ_r is small). However, in many applications, we often care only about recovering the singular directions with high weight, as the low weight singular directions have little effect on the data. Therefore, from Theorem 2.1 we can also derive a column (i.e. right-side) factor recovery result, which can be thought of as a weighted version of the rowspace recovery result:

Corollary 2.3 (Column factor recovery)**.** *Under the same conditions as Theorem 2.1, let $\hat{Q} \in \mathbb{R}^{d\times r}$ and $\hat{\Lambda} \in \mathbb{R}^{r\times r}$ be the top r singular vectors and singular values of $\hat{\Theta}$. Also, let $\Theta^* = Q\Lambda Q^T$ be the SVD of $\Theta^* = \frac{1}{m}X^TX$, where $Q \in \mathbb{R}^{d\times r}$ and $\Lambda \in \mathbb{R}^{r\times r}$. Then, we have*

$$\min_{\substack{R\in\mathbb{R}^{r\times r}:\\ R^TR=I_r}} \frac{1}{d}\|\hat{Q}\hat{\Lambda}^{1/2}R - Q\Lambda^{1/2}\|_F^2 \lesssim \alpha\sqrt{\frac{r^2d(\log d+\delta)}{m}}.$$

We refer to this corollary as "column factor recovery" because in the users and items example, we can think of the corollary as using our estimate of $\Theta^* = V\bar{S}V^T$ to recover $V\bar{S}^{1/2} \in \mathbb{R}^{d\times r}$, or the column factors "skewed" by $\bar{S}$.

2.1. Interpreting the bounds

While the error metrics in Theorem 2.1 and Corollary 2.3 scale with the magnitude of X, the rowspace recovery error in Corollary 2.2 is scale-invariant. Therefore, while Theorem 2.1 and Corollary 2.3 scale with the maximum entry α, the bound in Corollary 2.2 is in terms of the scale-invariant quantity $\frac{d\alpha}{\sigma_r}$. To make the bounds more comparable, we can convert these additive bounds into multiplicative ones by dividing both sides by the norm of the quantity being estimated ($\|\Theta^*\|_F^2$ in Theorem 2.1, $\|Q\|_F^2 = r$ in Corollary 2.2, and $\|Q\Lambda^{1/2}\|_F^2 = \|\Theta^*\|_{\text{nuc}}$ in Corollary 2.3). Then, we have the following sample complexities:

$$\text{Theorem 2.1: } m \gtrsim \left(\frac{d\alpha}{\|\Theta^*\|_F}\right)^2 rd\log d$$

$$\text{Corollary 2.2: } m \gtrsim \left(\frac{d\alpha}{\sigma_r(\Theta^*)\sqrt{r}}\right)^2 rd\log d$$

$$\text{Corollary 2.3: } m \gtrsim \left(\frac{d\alpha\sqrt{r}}{\|\Theta^*\|_{\text{nuc}}}\right)^2 rd\log d.$$

Each of these scale-invariant constants can be thought of as incoherence constants which capture the "spikiness" of the matrix Θ^*, with similar constants appearing in Negahban & Wainwright (2012), Montanari & Sun (2018), and Cai et al. (2021) (among others). Letting $\mu_1 = \frac{d\alpha}{\|\Theta^*\|_F}$, $\mu_2 = \frac{d\alpha}{\sigma_r(\Theta^*)\sqrt{r}}$, and $\mu_3 = \frac{d\alpha\sqrt{r}}{\|\Theta^*\|_{\text{nuc}}}$, we can interpret these constants by looking at a few examples:

- **All ones matrix**: if X is the all ones matrix, then we have $\mu_1 = \mu_2 = \mu_3 = 1$.
- **Single zero matrix**: if X is 1 everywhere, except with a 0 in entry $(1,1)$, then we have μ_1 and μ_3 constant, while μ_2 is order md. In other words, for approximate recovery of Θ^* and the column factors, predicting all ones is sufficient. On the other hand, to recover the rowspace, the algorithm must know where the 0 is, which requires sampling entry $(1,1)$ and is impossible to do with high probability.
- **Gaussian factors** ($X = UV^T$ with $U, V \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{d\times r}$): In this example, $\Theta^* = V\bar{S}V^T \approx VV^T$ because the empirical covariance $\bar{S} = \frac{1}{m}\sum_{i=1}^m u_iu_i^T$ is close to the identity I_r. Then, up to log and constant factors, we have[3] $\mu_1 \approx \mu_2 \approx \mu_3 \approx \sqrt{r}$, producing a sample complexity of $m \gtrsim r^2d\log d$. We observe this r^2 dependence in synthetic experiments and suspect that it is fundamental.
- **Correlated Gaussian factors**: while the rows of U and V in the previous example were uncorrelated Gaussians, we can also consider the case where they are drawn according to $\mathcal{N}(0, C)$ for some covariance matrix $C \in \mathbb{R}^{r\times r}$. In this case, μ_1 scales roughly as $\text{tr}(C^2)/\sqrt{\text{tr}(C^4)}$, which is $\sqrt{r}$ if C is the identity but potentially smaller if the eigenvalues of C are non-uniform.[4] For example, letting $s = (s_1, ..., s_r)$ denote the eigenvalues of C^2, if the s_i decay according to a power law $s_i = c_0 i^\alpha$, then μ_1 scales as $\log r$ for $\alpha = 1$ and is constant for $\alpha > 1$. In other words, the sample complexity to recover Θ^* is $m \gtrsim rd\log d$ if the r factors are sufficiently correlated.

3. Warmup: Gaussian row factors

In this section, we warm up with a simpler setting to provide intuition. In particular, suppose that the row factors $u_1, ..., u_m \in \mathbb{R}^r$ are drawn i.i.d. from a standard Gaussian $\mathcal{N}(0, I_r)$. Then, for a pair of observations $X_{i,a(i)}$ and $X_{i,b(i)}$, we have the following expectations:

$$\begin{aligned}\mathbb{E}[X_{i,a(i)}X_{i,b(i)}] &= \mathbb{E}[u_i^Tv_{a(i)}u_i^Tv_{b(i)}]\\ &= v_{a(i)}^T\mathbb{E}[u_iu_i^T]v_{b(i)}\\ &= v_{a(i)}^Tv_{b(i)},\\ \mathbb{E}[X_{i,a(i)}^2] &= \|v_{a(i)}\|_2^2.\end{aligned}$$

[3] For $r \ll d$, we have (up to constant and log factors) that $\alpha \approx r$, $\|\Theta^*\|_F \approx d\sqrt{r}$, $\sigma_r(\Theta^*) \approx d$, and $\|\Theta^*\|_{\text{nuc}} \approx rd$.

[4] Let $Z_1 \in \mathbb{R}^{m\times r}$ and $Z_2 \in \mathbb{R}^{d\times r}$ have entries drawn i.i.d. $\mathcal{N}(0,1)$. Then, for $m \gg r$, we have that $X = Z_1CZ_2^T$ and $\Theta^* \approx Z_2C^2Z_2^T$. The entries of X have mean magnitude $\sqrt{\text{tr}(C^2)}$, and the off-diagonal entries of Θ^* have mean magnitude $\sqrt{\text{tr}(C^4)}$, with both having sub-Exponential concentration. Then, up to log factors, $\mu_1 = d\alpha/\|\Theta^*\|_F \approx \text{tr}(C^2)/\sqrt{\text{tr}(C^4)}$.

Then, each pair of observations gives us unbiased estimates of entries $(a(i), b(i))$, $(b(i), a(i))$, $(a(i), a(i))$, and $(b(i), b(i))$ in the pairwise similarity matrix $VV^T \in \mathbb{R}^{d\times d}$. Given a very large number of rows, we can then estimate each entry of VV^T as its empirical average:

Off-diagonal terms ($j_1 \neq j_2$):

$$\hat{\Theta}^{(\text{emp})}_{j_1,j_2} = \frac{1}{|S_{j_1,j_2}|}\sum_{i\in S_{j_1,j_2}} X_{i,j_1}X_{i,j_2},$$

$$S_{j_1,j_2} = \{i : (a(i), b(i)) = (j_1, j_2) \text{ or } (j_2, j_1)\}.$$

Diagonal terms ($j_1 = j_2$):

$$\hat{\Theta}^{(\text{emp})}_{j,j} = \frac{1}{|S_j|}\sum_{i\in S_j} X^2_{i,j},$$

$$S_j = \{i : a(i) = j \text{ or } b(i) = j\}.$$

This empirical average can also be written

$$\hat{\Theta}^{(\text{emp})} = [P_E(X)^T P_E(X)] \oslash [E^T E],$$

where $E \in \mathbb{R}^{m\times d}$ is the observation mask (1 if that entry is observed, 0 otherwise), P_E sets unobserved entries to zero, and $\oslash$ represents element-wise division for entries where the divisor is non-zero, and a no-op otherwise. In other words, the empirical average $\hat{\Theta}$ can be written as a renormalized version of $P_E(X)^T P_E(X)$, where the (j_1, j_2)th entry of the renormalization matrix $E^T E$ is the number of rows where j_1 and j_2 are both observed.

If we don't have a very large number of rows, then $\hat{\Theta}^{(\text{emp})}$ can be thought of as a noisy, sparsely populated version of the column factor similarity matrix VV^T, which is rank r. Then a natural algorithm to estimate VV^T is to minimize the squared loss between the $\hat{\Theta}$ and $\hat{\Theta}^{(\text{emp})}$, plus a nuclear norm regularizer $\|\hat{\Theta}\|_{\text{nuc}}$. Furthermore, we might weight each entry in the squared loss by the number times it was observed, producing the objective

$$\min_{\Theta} \|\Theta - \Theta^{(\text{emp})}\|^2_{E^T E} + \lambda\|\Theta\|_{\text{nuc}},$$

where the weighted loss $\|\Theta - \Theta^{(\text{emp})}\|^2_{E^T E}$ is given by

$$\begin{aligned}\|\Theta - \Theta^{(\text{emp})}\|^2_{E^T E} &= \sum_{j_1=1}^{d}\sum_{j_2=1}^{d}(E^T E)_{j_1,j_2}(\Theta_{j_1,j_2} - \Theta^{(\text{emp})}_{j_1,j_2})^2 \\ &= \text{tr}(\Theta^T E^T E \Theta^{(\text{emp})}). \end{aligned} \tag{3}$$

In fact, some calculation shows that this weighted loss is exactly the loss we defined originally in Equation 2, up to rescaling and removal of terms that don't depend on Θ. Therefore, we can interpret our algorithm as performing weighted matrix completion with respect to $\hat{\Theta}^{(\text{emp})}$, which is a properly renormalized version of $P_E(X)^T P_E(X)$.

From this warmup, one can imagine ways to prove noisy matrix completion error bounds for recovery of VV^T given i.i.d. Gaussian u_i's, as well as extensions to more general distributions (e.g. sub-Gaussian or sub-Exponential), where we instead recover $V\text{Cov}(u)V^T$ where $\text{Cov}(u)$ is the covariance matrix of u. However, it is not always reasonable to assume independently drawn row factors: for example, in the genomics case, the chromosome base pairs are certainly not random or independent of one another. Therefore, we would like to prove recovery results without making distributional assumptions on the row factors.

However, recall from the introduction that the problem is severely underconstrained if we allow arbitrary row factors and masking. Somewhat surprisingly, we find that even with arbitrary row factors, assuming random masking is enough to enable recovery: while the intuition about noisy empirical averages no longer holds, the randomness in the masking is enough to enable a key technical step in the proof involving Radamacher symmetrization (Section A.6). In the following section, we provide a sketch of the proof in this more general setting.

4. Proof sketch

4.1. Outline

In this section, we outline of the proof, which uses restricted strong convexity arguments (Negahban & Wainwright, 2012; Negahban et al., 2012). The proof proceeds as follows: letting Δ denote the error $\hat{\Theta} - \Theta^*$, by the optimality of $\hat{\Theta}$ (along with reverse triangle) we have

$$\begin{aligned} 0 &\geq \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) + \lambda(\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}}) \\ &\geq \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \lambda\|\Delta\|_{\text{nuc}}. \end{aligned}$$

If $\mathcal{L}$ were τ-strongly-convex, then we could write

$$\begin{aligned} &\geq \langle\nabla\mathcal{L}(\Theta^*), \Delta\rangle + \tau\|\Delta\|^2_F - \lambda\|\Delta\|_{\text{nuc}} \\ &\geq -\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}\|\Delta\|_{\text{nuc}} + \tau\|\Delta\|^2_F - \lambda\|\Delta\|_{\text{nuc}}, \end{aligned}$$

where the second inequality follows from Holder's inequality. Then, bounding $\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}$ and $\|\Delta\|_{\text{nuc}}$ and rearranging would result in a bound for the error $\|\Delta\|^2_F$. However, $\mathcal{L}$ is not strongly convex: if we do not restrict to low-rank matrices, then there are multiple matrices that agree with the incomplete observations. Therefore, we'll instead show a form of restricted strong convexity: in particular, we'll show (in Lemma A.7) that the quantity $\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle\nabla\mathcal{L}(\Theta^*), \Delta\rangle$ concentrates around $\frac{1}{d^2}\|\Delta\|^2_F$, but with deviation terms depending on $\|\Delta\|_{\text{nuc}}$ and $\|\Delta\|_{\max}$. Therefore, recovery will depend on $\|\Delta\|_{\text{nuc}}$ being small, which follows from the condition that the regularization strength λ is large enough (Lemma 1 of Negahban & Wainwright (2012)), and $\|\Delta\|_{\max}$ being small, which

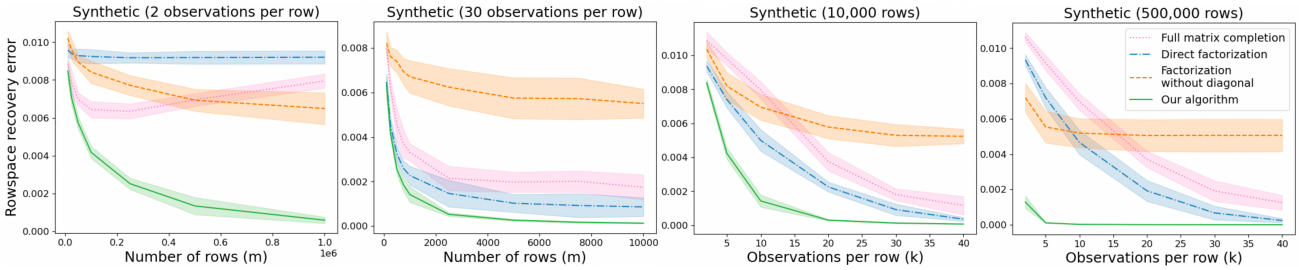

Figure 2. Rowspace recovery for rank-25 i.i.d. Gaussian column and row factors, with $d = 100$ columns, where from left to right, the experiments are as follows: (a) we sample $k = 2$ observations per row and vary the number of rows m from 10,000 to 1,000,000, (b) $k = 30$ with m between 100 and 10,000 (where fewer rows are necessary because k is large), (c) $m = 10{,}000$ with k between 2 and 40, and (d) $m = 500{,}000$ with k between 2 and 40. Our algorithm (in solid green) performs produces the most accurate rowspace estimates for all parameter ranges evaluated on, with the gap being largest for large m and small k.

follows by our assumption that $\|X\|_{\max}^2 \leq \alpha$ and the optimization constraint $\|\hat{\Theta}\|_{\max} \leq \alpha$. With this approach in mind, the following are caricatures of each of the lemmas:

(a) Section A.6 (operator norm bound):

$$\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}} \lesssim \alpha \sqrt{\frac{\log d}{dm}}.$$

(b) Section A.7 (restricted strong convexity):

$$\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle \geq \frac{1}{d^2}\|\Delta\|_F^2 - c\alpha \|\Delta\|_{\text{nuc}} \sqrt{\frac{\log d}{dm}}.$$

(c) Section A.8 (decomposability):

$$\text{If } \lambda \geq 2\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}}, \text{ then } \|\Delta\|_{\text{nuc}} \lesssim \sqrt{r}\|\Delta\|_F.$$

With these lemmas, we can first apply restricted strong convexity, reverse triangle, and Holder's inequality as before:

$$\begin{aligned} 0 &\geq \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) + \lambda(\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}}) \\ &\geq \frac{1}{d^2}\|\Delta\|_F^2 - c\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - \|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}}\|\Delta\|_{\text{nuc}} \\ &\quad - \lambda\|\Delta\|_{\text{nuc}}. \end{aligned}$$

Next, by our operator norm bound (Lemma A.6) and our setting of $\lambda = 16\alpha\sqrt{\frac{\log d}{dm}}$, we can replace the latter two terms with $c\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}}$ and combine, producing

$$\geq \frac{1}{d^2}\|\Delta\|_F^2 - c'\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}}.$$

Finally, applying the bound $\|\Delta\|_{\text{nuc}} \lesssim \sqrt{r}\|\Delta\|_F$ (Lemma 1 of Negahban & Wainwright (2012)) and rearranging produces $\frac{1}{d}\|\Delta\|_F \lesssim \alpha\sqrt{\frac{rd\log d}{m}}$, as desired.

4.2. Operator norm bound

It remains to discuss each of the lemmas, which we do briefly here and at length in the appendix. First, using our definition of the loss in Equation 2, we can compute the gradient of $\mathcal{L}$ at Θ^* as follows (where recall that $E_{i,j}$ is the mask matrix with 1 at (i, j) and 0 elsewhere):

$$\begin{aligned} \nabla \mathcal{L}(\Theta^*) = &\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})E_{a(i),b(i)} \\ &+ \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})E_{b(i),a(i)} \\ &+ \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\Theta^*_{a(i),a(i)} - X^2_{i,a(i)})E_{a(i),a(i)} \\ &+ \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\Theta^*_{b(i),b(i)} - X^2_{i,b(i)})E_{b(i),b(i)}. \end{aligned}$$

We wish to prove that the operator norm of this quantity is small, with high probability with respect to the randomly sampled indices $(a(i), b(i))$. In the example where the row factors u_i were Gaussian, we had that $\mathbb{E}_{u_i \sim Z}[X_{i,j_1}X_{i,j_2}] = \Theta^*_{j_1,j_2}$, turning each summand into a mask matrix multiplied by mean-zero noise. While that approach no longer holds here, we can write

$$X_{i,a(i)}X_{i,b(i)} - \Theta^*_{a(i),b(i)} = v_{a(i)}^T \left(u_i u_i^T - \frac{1}{m}\sum_{i=1}^{m} u_i u_i^T \right) v_{b(i)}.$$

At this point, we can apply our assumption that the mask is chosen independently of the underlying matrix, as it means that the expectation $\mathbb{E}_{a(i),b(i)}\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}}$ is invariant to permutations of the rows. Considering randomly permuted row factors brings us closer to the distributional case, which allows us to apply Radamacher symmetrization arguments (Section A.6). The rest of the bound then proceeds via standard concentration arguments.

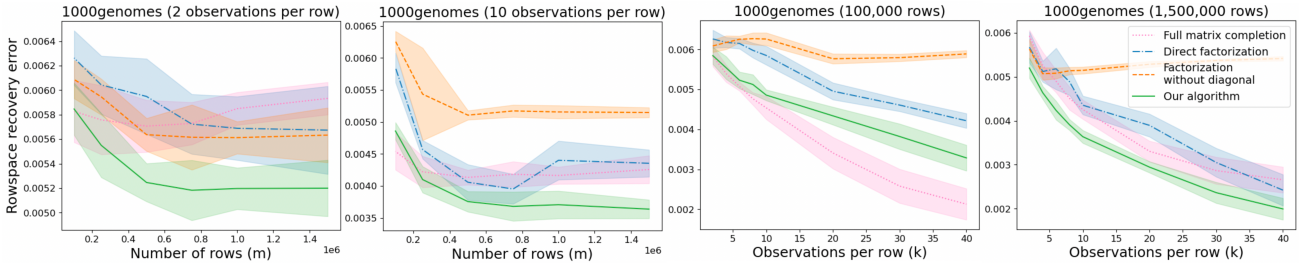

Figure 3. Rowspace recovery on the 1000genomes dataset, where from left to right the experiments are as follows: (a) we observe $k = 2$ entries per row with the number of rows m between 100,000 and 1,500,000, (b) $k = 10$ with m between 100,000 and 1,500,000, (c) $m = 100{,}000$ with k between 2 and 40, and (d) $m = 1{,}500{,}000$ with k between 2 and 40. The number of columns is fixed with $d = 250$ for all experiments. Our algorithm (in solid green) is most accurate for most parameter settings, but is outperformed by full matrix completion (in dotted pink) when m is small and k is large (plot (c)).

4.3. Restricted strong convexity

The other key lemma is restricted strong convexity; by the definition of the loss (Equation 2), we can first compute

$$\begin{aligned}\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle \\ = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}[\Delta^2_{a(i),b(i)} + \Delta^2_{b(i),a(i)}] \\ + \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}[\Delta^2_{a(i),a(i)} + \Delta^2_{b(i),b(i)}].\end{aligned}$$

Recall that we wanted to show that this quantity is lower-bounded by $\frac{1}{d^2}\|\Delta\|_F^2$, minus some concentration terms. The first point to note is that each term has expectation

$$\mathbb{E}[\Delta^2_{a(i),b(i)}] = \frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2,$$
$$\mathbb{E}[\Delta^2_{a(i),a(i)}] = \frac{1}{d}\|P_{\text{diag}}(\Delta)\|_F^2,$$

where $P_{\text{diag}}(\Delta)$ sets the off-diagonal terms of Δ to zero and $P_{\text{off-diag}}(\Delta) = \Delta - P_{\text{diag}}(\Delta)$ sets the diagonal terms to zero. Then, analyzing the off-diagonal terms and diagonal terms separately, we can first show concentration of the given sums around their expectations. The lemma then follows from proper treatment and recombination of the diagonal and off-diagonal terms.

5. Experiments

5.1. Setup

Finally, we evaluate our method on synthetic data and the 1000genomes dataset (Fairley et al., 2019). The 1000genomes dataset contains fully sequenced chromosomes of 2354 subjects. Due to computational limitations, we first subsample to $m = 1\,500\,000$ rows (chosen from the first chromosome in order of mutation frequency) and the subjects to $d = 250$ columns (chosen randomly). We also reduce the number of rows further in some experiments to evaluate how error scales as a function of m. We then randomly sample between $k = 2$ and 40 observations per row and compare the estimated right-side singular vectors to the ground truth, produced by factoring the original fully sampled matrix. Specifically, let the ground truth SVD be given by $X = P\Sigma Q^T$. Then, we compute the error as[5]

$$\min_{R \in \mathbb{R}^{r\times r}: R^T R = I_r} \|\hat{Q}R - Q\|_F^2. \tag{4}$$

The estimate $\hat{Q}$ is produced by the following algorithms:

(a) Full matrix completion: perform rank r matrix completion on $P_E(X)$ and compute the SVD of the result.

(b) Direct factorization: compute the rank r SVD of the matrix $P_E(X)^T P_E(X)$.

(c) Factorization without diagonal (Cai et al., 2021): compute the rank r SVD of $P_{\text{off-diag}}(P_E(X)^T P_E(X))$.

(d) Our algorithm: perform rank r matrix completion with respect to $\mathcal{L}$ (Equation 2) and compute the SVD of the result. The loss function naturally generalizes to $k > 2$ via the formulation in Equation 3.

We implement vanilla matrix completion via the non-convex optimization[6]

$$\min_{\substack{U\in\mathbb{R}^{m\times r}\\ V\in\mathbb{R}^{d\times r}}} \frac{1}{|E|}\|P_E(UV^T) - P_E(X)\|_F^2 + \lambda\left(\frac{1}{m}\sum_{i=1}^{m}\|u_i\|_2^2 + \frac{1}{d}\sum_{i=1}^{d}\|v_i\|_2^2\right).$$

[5]This minimization is known as the Procrustes problem (Schonemann, 1966) and can be solved in closed form.

[6]Note that the L2 regularizer in this objective corresponds to the likelihood under the Gaussian prior, which is the factor distribution that we use in our synthetic experiments. We choose $\lambda = 0.1$ using grid search over the set $(0, 0.001, 0.01, 0.1, 0.5, 1, 10)$.

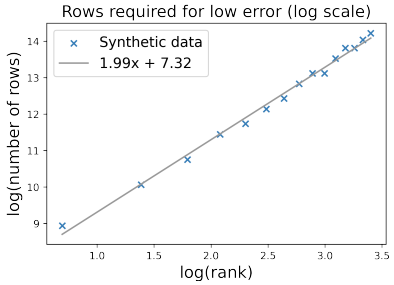

Figure 4. A log plot of the row-rank dependence for our algorithm, where we plot the number of rows m required to achieve low rowspace recovery error for each rank r for i.i.d. Gaussian row and column factors, $k = 2$ observations per row, and $d = 200$ columns. The linear fit has slope almost exactly 2, suggesting a dependence of $m \propto r^2$ and confirming our derived bounds.

We also implement our method via the non-convex optimization $\min_{V \in \mathbb{R}^{d \times r}} \mathcal{L}(VV^T)$, where we fit a symmetric matrix and omit the L2 regularizer because the diagonal terms control the factor norms.

For the synthetic experiments, we sample $X = UV^T$ with $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{d \times r}$ i.i.d. Gaussian $\mathcal{N}(0, r^{-1/2})$, rescaled to ensure that the expected norm of each entry is 1. We set the rank to $r = 25$ with $d = 100$ columns, while varying the number of rows m and the observations per row k depending on the experiment.

We produce each of the plots by resampling the mask ten times for each parameter setting and plotting the mean plus or minus two standard deviations; see Section B for other implementation details and hyperparameters.

5.2. Synthetic experiments

We first discuss the synthetic experiments, shown in Figure 2. From left to right, the experiments are as follows: (1) we sample $k = 2$ observations per row and vary the number of rows m from 10,000 to 1,000,000, (2) $k = 30$ with m between 100 and 10,000 (where fewer rows are necessary because k is large), (3) $m = 10{,}000$ with k between 2 and 40, and (4) $m = 500{,}000$ with k between 2 and 40.

In these experiments, our algorithm produces the most accurate rowspace estimates for all parameter ranges evaluated on. The algorithm is especially strong compared to other methods when m is large or k is small, which are the settings that motivate our study. Full matrix completion performs well when m is small and k is large, i.e. the setting with X closer to square and with more observations, which is when we expect full completion to be feasible. However, even in this setting, our algorithm performs competitively with or outperforms full matrix completion.

5.3. 1000genomes experiments

Next, we display similar plots for the 1000genomes dataset experiments, shown in Figure 3, where from left to right the experiments are as follows: (1) we observe $k = 2$ entries per row with the number of rows m between 100,000 and 1,500,000, (2) $k = 10$ with m between 100,000 and 1,500,000, (3) $m = 100{,}000$ with k between 2 and 40, and (4) $m = 1{,}500{,}000$ with k between 2 and 40.[7]

In these experiments, our algorithm is again most accurate for most parameter settings and is strongest relative to other methods when m is large or k is small. However, it is outperformed by full matrix completion when m is small and k is large, which is when we expect full completion to be possible. For the $m = 1{,}500{,}000$ setting, which is most representative of the ratio m/d we might see in practice for this setting, our algorithm is most accurate for all values of k evaluated on and is able to recover the rowspace reliably with as few as 5 observations per row.

Finally, we visualize the recovered column factors using TSNE (van der Maaten & Hinton, 2008), which projects the factors into two dimensions while attempting to preserve similarity structure (Figure 1). Visually, our algorithm recovers the most accurate factors by far and is almost identical to the ground truth. Factorization without the diagonal produces the next best estimates but is unable to separate the EUR, AMR, and SAS clusters. Meanwhile, full matrix completion is unable to recover coherent clusters because there are too few observations per row.

5.4. Dependence on rank

Finally, we perform synthetic experiments to verify the r^2 dependence in our derived sample complexity of $m = \Omega(r^2 d \log d)$. Specifically, for each rank r, we sample i.i.d. Gaussian factors with $d = 200$ columns and $k = 2$ observations per row, and we perform binary search over m to achieve some target error.[8] The result is shown in Figure 4, where we plot $\log m$ versus $\log r$. The points lie almost exactly on a line with slope 2, suggesting a dependence of $m \propto r^2$ and confirming our derived bounds.

6. Related Work

While low-rank matrix completion and factorization have a long history of research, here we touch on just a few threads of work most directly related to our paper.

[7] We set $r = 10$ for each algorithm and compute the evaluation metric (Equation 4) with respect to the rank 10 SVD of the original matrix. On the other hand, the ground truth TSNE plot is produced with respect to the full SVD of the original matrix.

[8] For each candidate m in the binary search, we average the loss over 20 runs and accept if $\frac{1}{r}\|\hat{Q}R - Q\|_F^2 \in 0.1 \pm 0.02$. The search starts with a range of $m \in (0, 4e6)$.

Matrix completion. Matrix completion is the problem of estimating a low-rank matrix after observing a subset of its entries, and a variety of methods have been proposed and analyzed for this problem, including nuclear norm minimization (Candès & Recht, 2009; Candes & Plan, 2010), SVD with trimming (Keshavan et al., 2009), alternating minimization (Hardt, 2014), and non-convex gradient descent (Ge et al., 2016; Jin et al., 2016). To rule out certain matrices whose recovery is impossible, these papers propose various kinds of incoherence assumptions: for example, Candès & Recht (2009) make assumptions on the leverage scores of X, while Negahban & Wainwright (2012) make assumptions about the "spikiness" of X. We adopt the assumptions and analysis framework of Negahban & Wainwright (2012), who prove additive Frobenius error bounds under the assumption that the maximum entry of the underlying matrix is bounded.

Unbalanced noisy matrix factorization. A recent series of works explores the problem of noisy matrix factorization for matrices with high aspect ratio (i.e. many more rows than columns). In this problem setting, the matrix is fully observed but with entry-wise additive noise. Feldman (2021), Donoho & Feldman (2022), and Montanari & Wu (2022) study the asymptotics of this problem, characterizing the Bayes optimal error of recovering the singular vectors as the number of rows and columns $m, d \to \infty$. Broadly speaking, in contrast with past works which took $m, d \to \infty$ with the ratios d/m and m/d remaining bounded, these papers consider cases where $d/m \to \infty$ and $d/m \to 0$, finding regimes where recovery of the left singular values is possible but not the right singular vectors (and vice versa). While their setting and results are very different from ours, they share the commonality of studying cases where only "one-sided" recovery is possible.

Subspace and covariance estimation from partial observations. The papers most directly related to our setting are Lounici (2014), Gonen et al. (2016), Montanari & Sun (2018), and Cai et al. (2021). Gonen et al. (2016) consider the problem of subspace estimation from partial observations: given m partially observed vectors of dimension d, sampled i.i.d. from a bounded distribution, their goal is to recover the rank-r subspace that those vectors lie in. As their algorithm, they perform factorization on $P_E(X)^T P_E(X)$ with the diagonal rescaled. However, they make no incoherence assumptions, making matrix completion inapplicable. Therefore, they find that a sample size of $m = \Omega((d/k)^2 r)$ is both necessary and sufficient to recover the subspace (where $k \geq 2$ is the average number of observations per vector). Lounici (2014) study a similar setting, but with the weaker assumption that the vectors are sampled from a sub-Gaussian distribution, and they prove a similar sample complexity of $m = \Omega((d/k)^2 r \log d)$. Montanari & Sun (2018) and Cai et al. (2021) also study subspace estimation from partial observations. Similar to the above papers, both algorithms involve factoring $P_E(X)^T P_E(X)$ with the diagonal rescaled, but they adopt incoherence assumptions to establish sample complexity bounds no longer quadratic in d. As their sampling distribution, they assume n observations uniformly chosen from the md entries, and they both prove similar sample complexities of $n = \Omega(r\sqrt{dm}\ \mathrm{polylog}(dm))$ (please see the original papers for the full results). For k observations per row on average, the number of rows required then becomes $m = \Omega(r^2(d/k)\ \mathrm{polylog}(d))$.

One difference in setting between our paper and the aforementioned papers is that they focus on subspace estimation, so it suffices to show that $P_E(X)^T P_E(X)$ (after rescaling) is close to $X^T X$ in operator norm. In contrast, because we are also interested in completing $X^T X$ and recovering the column factors, we show error bounds in Frobenius norm, which requires more accurate estimation of $X^T X$.

The other main difference between our paper and the aforementioned papers is that they consider the setting where each entry of X is observed independently with probability p (or equivalently, that the n observations are uniform over the matrix). Under this model, even if p is small enough such that there are two observations per row on average, some rows might still have larger numbers of observations. In contrast, we show that the column factors can be recovered even if *all* of the rows have only two observations. From a theoretical standpoint, this setting is strictly harder than the Bernoulli observation setting because given the latter, we can keep the rows with at least two observations (which is satisfied by roughly $3/4$ of the rows for $d \gg 2$), subsample to two per row, and use our analysis to produce the same sample complexity (up to constant factors).

7. Conclusion and Future Directions

One limitation of our result is that it only applies to two observations per row; therefore, a fruitful direction could involve extending it to more general cases, like k observations per row or other sampling patterns. Empirically, we hope that our paper can inspire work on datasets that were previously too sparsely annotated for full matrix completion, but might be amenable to our algorithm.

Acknowledgements

We thank John Thickstun, Tae Kyu Kim, and anonymous reviewers for their helpful comments and discussion. Gregory Valiant is supported by a Simons Investigator Award, and NSF awards CCF-1704417 and CCF-1813049. Steven Cao is supported by the NSF Graduate Research Fellowship. Experiments were run on the Stanford NLP cluster.

References

Cai, C., Li, G., Chi, Y., Poor, H. V., and Chen, Y. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49(2):944 – 967, 2021. doi: 10.1214/20-AOS1986. URL https://doi.org/10.1214/20-AOS1986.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. doi: 10.1109/JPROC.2009.2035722.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009. ISSN 1615-3383. doi: 10.1007/s10208-009-9045-5. URL https://doi.org/10.1007/s10208-009-9045-5.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. ISSN 00361429. URL http://www.jstor.org/stable/2949580.

Donoho, D. L. and Feldman, M. J. Optimal eigenvalue shrinkage in the semicircle limit, 2022. URL https://arxiv.org/abs/2210.04488.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sep 1936. ISSN 1860-0980. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1):D941–D947, 10 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz836. URL https://doi.org/10.1093/nar/gkz836.

Feldman, M. J. Spiked singular values and vectors under extreme aspect ratios, 2021. URL https://arxiv.org/abs/2104.15127.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2981–2989, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Gonen, A., Rosenbaum, D., Eldar, Y. C., and Shalev-Shwartz, S. Subspace learning with partial information. *Journal of Machine Learning Research*, 17(52):1–21, 2016. URL http://jmlr.org/papers/v17/14-443.html.

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., and Bernstein, M. S. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445423. URL https://doi.org/10.1145/3411764.3445423.

Hardt, M. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 651–660, 2014. doi: 10.1109/FOCS.2014.75.

Jin, C., Kakade, S. M., and Netrapalli, P. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4527–4535, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Keshavan, R. H., Oh, S., and Montanari, A. Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pp. 324–328, 2009. doi: 10.1109/ISIT.2009.5205567.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics. Springer Berlin, Heidelberg, 1991. doi: 10.1007/978-3-642-20212-4.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10(1):387–406, 2009. doi: 10.1146/annurev.genom.9.081307.164242. URL https://doi.org/10.1146/annurev.genom.9.081307.164242. PMID: 19715440.

Lounici, K. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029 – 1058, 2014. doi: 10.3150/12-BEJ487. URL https://doi.org/10.3150/12-BEJ487.

Mirsky, L. SYMMETRIC GAUGE FUNCTIONS AND UNITARILY INVARIANT NORMS. *The Quarterly Journal of Mathematics*, 11(1):50–59, 01 1960. ISSN 0033-5606. doi: 10.1093/qmath/11.1.50. URL https://doi.org/10.1093/qmath/11.1.50.

Montanari, A. and Sun, N. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425, 2018. doi: https://doi.org/10.1002/cpa.21748. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21748.

Montanari, A. and Wu, Y. Fundamental limits of low-rank matrix estimation with diverging aspect ratios, 2022. URL https://arxiv.org/abs/2211.00488.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13 (53):1665–1697, 2012. URL http://jmlr.org/papers/v13/negahban12a.html.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012. ISSN 08834237. URL http://www.jstor.org/stable/41714783.

Powers, R. T. and Størmer, E. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1–33, Mar 1970. ISSN 1432-0916. doi: 10.1007/BF01645492. URL https://doi.org/10.1007/BF01645492.

Schonemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Shin, P. J., Larson, P. E. Z., Ohliger, M. A., Elad, M., Pauly, J. M., Vigneron, D. B., and Lustig, M. Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion. *Magnetic Resonance in Medicine*, 72(4): 959–970, 2014. doi: https://doi.org/10.1002/mrm.24997. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.24997.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the davis—kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. ISSN 00063444. URL http://www.jstor.org/stable/43908537.

A. Proofs

A.1. Notation

Letting A and B represent an arbitrary matrices, we'll use the following notation:

- $\|A\|_{\text{max}}$ – maximum of the absolute values of the entries of A
- $\|A\|_{\text{op}}$ – operator norm
- $\|A\|_{\text{nuc}}$ – nuclear norm
- $\|A\|_F$ – Frobenius norm
- $\langle A, B\rangle = \text{tr}(A^T B)$ – matrix inner product
- For a square matrix $C \in \mathbb{R}^{d\times d}$, we'll use $P_{\text{diag}}(C)$ to represent the matrix C with the off-diagonal terms set to zero, and $P_{\text{off-diag}}(C) = C - P_{\text{diag}}(C)$ to represent the matrix C with the diagonal terms set to zero.
- We'll use $E_{i,j} \in \mathbb{R}^{d\times d}$ to represent the matrix with 1 in entry (i, j) and 0 elsewhere, and $\tilde{E}_{i,j}$ to refer to the symmetric mask $\frac{1}{2}(E_{i,j} + E_{j,i})$.
- $f \lesssim g$ will denote that there exists a universal constant c such that $f \leq cg$.
- $[d]$ will denote the set $\{1, 2, ..., d\}$.

Throughout the proofs, when there are long chains of inequalities, we will sometimes box the terms that change from line to line for ease of reading.

A.2. Setup and main result

The problem setup is as follows: from a rank-r matrix $X = UV^T \in \mathbb{R}^{m\times d}$ with $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{d\times r}$, we randomly observe two entries per row, which we can represent as indices $(a(1), b(1)), ..., (a(m), b(m))$ drawn i.i.d. uniformly from the set $\{(i, j) : i, j \in [d],\ i \neq j\}$. We wish to estimate the matrix

$$\Theta^* = \frac{1}{m}X^T X = V\bar{S}V^T$$
$$\bar{S} = \frac{1}{m}U^T U = \frac{1}{m}\sum_{i=1}^{m} u_i u_i^T \in \mathbb{R}^{r\times r},$$

where $u_i \in \mathbb{R}^r$ is the ith row of U. Our estimator for Θ^* minimizes a squared loss with a nuclear norm regularizer, along with constraints on the maximum off-diagonal and diagonal entries, and is given as follows:

$$\hat{\Theta} \in \underset{\|\Theta\|_{\text{max}}\leq\alpha}{\text{argmin}}\ \mathcal{L}(\Theta) + \lambda\|\Theta\|_{\text{nuc}}, \tag{5}$$

where the squared loss $\mathcal{L}(\Theta)$ is given by

$$\mathcal{L}(\Theta) = \frac{1}{4m}\sum_{i=1}^{m}(\Theta_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})^2 + (\Theta_{b(i),a(i)} - X_{i,b(i)}X_{i,a(i)})^2$$
$$+ \frac{1}{4m}\sum_{i=1}^{m}(\Theta_{a(i),a(i)} - X^2_{i,a(i)})^2 + (\Theta_{b(i),b(i)} - X^2_{i,b(i)})^2. \tag{6}$$

Given the assumption that $\|X\|^2_{\text{max}} \leq \alpha$, we will prove the following error bound:

Theorem A.1. *Let $\hat{\Theta}$ be the solution of the optimization problem defined in Equation 5, where λ is set to $16\alpha\sqrt{\frac{\log d+\delta}{dm}}$. Also, suppose that X is rank r with $\|X\|^2_{max} \leq \alpha$, and $m \geq d(\log d + \delta)$. Then, with probability $\geq 1 - 3e^{-\delta}$, we have that*

$$\frac{1}{d^2}\|\hat{\Theta} - \Theta^*\|^2_F \lesssim \alpha^2\frac{rd(\log d + \delta)}{m}.$$

A.3. Rowspace recovery

From Theorem A.1, we can first derive a rowspace recovery result, where our goal is to estimate the subspace spanned by the rows of X. In particular, the rowspace of X is equal to the rowspace of $\Theta^* = \frac{1}{m}X^TX$, and we have an error bound for our estimator $\hat{\Theta}$. Therefore, we can use the rowspace of $\hat{\Theta}$ as our estimator and use standard perturbation theory to bound the rowspace estimation error. Specifically, letting the rank SVD of Θ^* be given by $Q\Lambda Q^T$ for $Q \in \mathbb{R}^{d\times r}$, we can think of rowspace recovery as estimating the right-side singular vectors Q up to rotation. Then, we can define rowspace recovery error as $\min_{R\in\mathbb{R}^{r\times r}:R^TR=I_r} \|\hat{Q}R - Q\|_F^2$ and produce the following error bound:

Corollary A.2 (Right-side singular vector recovery)**.** *Let $\hat{\Theta}$ be the solution of the optimization problem defined in Equation 5, where λ is set to $16\alpha\sqrt{\frac{\log d+\delta}{dm}}$, and let $\hat{Q} \in \mathbb{R}^{d\times r}$ be the top r singular vectors of $\hat{\Theta}$. Also, suppose that X is rank r with $\|X\|_{max}^2 \le \alpha$ and $m \ge d(\log d + \delta)$, and let $Q \in \mathbb{R}^{d\times r}$ denote the right-side singular vectors of X. Then, letting σ_r be the rth singular value of $\Theta^* = \frac{1}{m}X^TX$, we have that with probability $\ge 1 - 3e^{-\delta}$,*

$$\min_{R\in\mathbb{R}^{r\times r}:R^TR=I_r} \|\hat{Q}R - Q\|_F^2 \lesssim \left(\frac{d\alpha}{\sigma_r}\right)^2 \frac{rd(\log d + \delta)}{m}.$$

Note that while the additive error in Theorem A.1 scales with the magnitude of X, the rowspace recovery error is scale-invariant. Therefore, while the bound in Theorem A.1 scales with α, the bound in Corollary A.2 is in terms of the quantity $\frac{d\alpha}{\sigma_r}$. This quantity can be thought of as capturing both the incoherence and condition number of X: for example, it is large for "spiky" matrices, where rowspace recovery is impossible (e.g. dm for the matrix with 1 in a single entry and 0 otherwise), and it is small for "incoherent" matrices (e.g. 1 for the all-ones matrix). The corollary directly follows from Theorem 2 of Yu et al. (2015), a variant of the Davis-Kahan theorem (Davis & Kahan, 1970):

Lemma A.3 (Theorem 2 of Yu et al. (2015))**.** *Let $A, \hat{A} \in \mathbb{R}^{d\times d}$ be symmetric matrices with eigenvalues $\lambda_1 \ge ... \ge \lambda_d$ and $\hat{\lambda}_1 \ge ... \ge \hat{\lambda}_d$. Fixing some $1 \le r \le d$, suppose that $\lambda_r - \lambda_{r+1} > 0$, where $\lambda_{d+1} := -\infty$. Let $V = (v_1, ..., v_r) \in \mathbb{R}^{d\times r}$ and $\hat{V} = (\hat{v}_1, ..., \hat{v}_r) \in \mathbb{R}^{d\times r}$ have orthonormal columns satisfying $Av_i = \lambda_i v_i$ and $\hat{A}\hat{v}_i = \hat{\lambda}_i\hat{v}_i$. Then,*

$$\min_{R\in\mathbb{R}^{r\times r}:R^TR=I_r} \|\hat{V}R - V\|_F \le \frac{2^{3/2}\|\hat{A} - A\|_F}{\lambda_r - \lambda_{r+1}}.$$

Proof of Corollary A.2. First, letting the SVD of X be given by $P\Sigma Q^T$, we have that the SVD of Θ^* is given by $Q\Lambda Q^T$ (for $\Lambda = \frac{1}{m}\Sigma^T\Sigma \in \mathbb{R}^{r\times r}$). Then, by Theorem 2 of Yu et al. (2015), we have that

$$\min_{R\in\mathbb{R}^{r\times r}:R^TR=I_r} \|\hat{Q}R - Q\|_F^2 \le 8\frac{\|\hat{\Theta} - \Theta^*\|_F^2}{\sigma_r(\Theta^*)^2}.$$

The corollary then immediately follows from Theorem A.1. □

A.4. Column factor recovery

In Corollary A.2, a σ_r factor appears in the denominator because the algorithm is tasked with recovering all r singular directions, even if some singular directions have low weight (i.e. σ_r is small). However, in many applications, we often care only about recovering the singular directions with high weight, as the low weight singular directions have little effect on the data. Therefore, from Theorem A.1 we can also derive a column factor recovery result, which can be thought of as a weighted version of the rowspace recovery result. In particular, our goal here is to recover $Q\Lambda^{1/2} \in \mathbb{R}^{d\times r}$ up to rotation, which can be thought of as containing column factors, or r-dimensional vector representations for each column in X. Then, using the top r singular values and vectors of $\hat{\Theta}$ for our estimate, we can produce the following error bound:

Corollary A.4 (Column factor recovery)**.** *Let $\hat{\Theta}$ be the solution of the optimization problem defined in Equation 5, where λ is set to $16\alpha\sqrt{\frac{\log d+\delta}{dm}}$, and let $\hat{Q} \in \mathbb{R}^{d\times r}$ and $\hat{\Lambda} \in \mathbb{R}^{r\times r}$ be the top r singular vectors and singular values of $\hat{\Theta}$. Also, suppose that X is rank r with $\|X\|_{max}^2 \le \alpha$ and $m \ge d(\log d + \delta)$, and let $\Theta^* = Q\Lambda Q^T$ be the SVD of $\Theta^* = \frac{1}{m}X^TX$, where $Q \in \mathbb{R}^{d\times r}$ and $\Lambda \in \mathbb{R}^{r\times r}$. Then, we have that with probability $\ge 1 - 3e^{-\delta}$,*

$$\min_{R\in\mathbb{R}^{r\times r}:R^TR=I_r} \frac{1}{d}\|\hat{Q}\hat{\Lambda}^{1/2}R - Q\Lambda^{1/2}\|_F^2 \lesssim \alpha\sqrt{\frac{r^2d(\log d + \delta)}{m}}.$$

Note that unlike Corollary A.2, this bound does not depend on σ_r because the singular vectors are weighted by their corresponding singular values. The proof uses Powers-Størmer (Powers & Størmer, 1970) and proceeds as follows:

Lemma A.5 (Powers-Størmer)**.** *For positive semidefinite matrices A and B, we have that*

$$\|A - B\|_F^2 \leq \|A^2 - B^2\|_{nuc}.$$

Proof of Corollary A.4. First, by the Powers-Størmer inequality (Powers & Størmer, 1970), we have that

$$\|\hat{Q}\hat{\Lambda}^{1/2}\hat{Q}^T - Q\Lambda^{1/2}Q^T\|_F^2 \leq \|\hat{\Theta}_r - \Theta^*\|_{\text{nuc}},$$

where $\hat{\Theta}_r = \hat{Q}\hat{\Lambda}\hat{Q}^T$ is the rank-r truncated version of $\hat{\Theta}$. Next, we have that

$$\begin{aligned}
\|\hat{\Theta}_r - \Theta^*\|_{\text{nuc}} &\overset{(i)}{\leq} \sqrt{2r}\|\hat{\Theta}_r - \Theta^*\|_F \\
&\overset{(ii)}{\leq} \sqrt{2r}(\|\hat{\Theta} - \hat{\Theta}_r\|_F + \|\hat{\Theta} - \Theta^*\|_F) \\
&\overset{(iii)}{\leq} 2\sqrt{2r}\|\hat{\Theta} - \Theta^*\|_F \\
&\overset{(iv)}{\lesssim} \alpha\sqrt{r}d\sqrt{\frac{rd(\log d + \delta)}{m}}.
\end{aligned}$$

where (i) follows from $\hat{\Theta}_r - \Theta^*$ being at most rank $2r$, (ii) from triangle, (iii) from the Eckart-Young-Mirsky theorem (Eckart & Young, 1936; Mirsky, 1960), which states that $\hat{\Theta}_r$ is the closest rank-r matrix to $\hat{\Theta}$ in any unitarily invariant norm, and (iv) from applying Theorem A.1. Then, it suffices to show that

$$\min_{R \in \mathbb{R}^{r\times r}: R^T R = I_r} \|\hat{Q}\hat{\Lambda}^{1/2}R - Q\Lambda^{1/2}\|_F^2 \leq \|\hat{Q}\hat{\Lambda}^{1/2}\hat{Q}^T - Q\Lambda^{1/2}Q^T\|_F^2.$$

To show this inequality, we can choose a particular rotation R as follows:

$$\begin{aligned}
\|\hat{Q}\hat{\Lambda}^{1/2}R - Q\Lambda^{1/2}\|_F^2 &= \text{tr}(\Lambda) + \text{tr}(\hat{\Lambda}) - 2\text{tr}(R^T\hat{\Lambda}^{1/2}\hat{Q}^T Q\Lambda^{1/2}) \\
&\overset{(i)}{=} \text{tr}(\Lambda) + \text{tr}(\hat{\Lambda}) - 2\|\hat{\Lambda}^{1/2}\hat{Q}^T Q\Lambda^{1/2}\|_{\text{nuc}} \\
&\overset{(ii)}{=} \text{tr}(\Lambda) + \text{tr}(\hat{\Lambda}) - 2\|\hat{Q}\hat{\Lambda}^{1/2}\hat{Q}^T Q\Lambda^{1/2}Q\|_{\text{nuc}} \\
&\overset{(iii)}{\leq} \text{tr}(\Lambda) + \text{tr}(\hat{\Lambda}) - 2\text{tr}(\hat{Q}\hat{\Lambda}^{1/2}\hat{Q}^T Q\Lambda^{1/2}Q) \\
&= \|\hat{Q}\hat{\Lambda}^{1/2}\hat{Q}^T - Q\Lambda^{1/2}Q^T\|_F^2
\end{aligned}$$

where (i) follows from choosing $R = AB^T$ for A and B given by the SVD $\hat{\Lambda}^{1/2}\hat{Q}^T Q\Lambda^{1/2} = ASB^T$, (ii) follows from the fact that $\|C\|_{\text{nuc}} = \|Q_1 C Q_2^T\|_{\text{nuc}}$ for any Q_1, Q_2 such that $Q_1^T Q_1 = I$ and $Q_2^T Q_2 = I$, and (iii) follows from the fact that $\text{tr}(C) \leq |\text{tr}(C)| \leq \|C\|_{\text{nuc}}$ for any matrix C, proving the desired result. □

A.5. Proof outline

In this section, we outline the proof of Theorem A.1. Our analysis uses restricted strong convexity arguments, as described in Negahban & Wainwright (2012), Negahban et al. (2012), and Wainwright (2019). For completeness, we reproduce parts of their analysis; such lemmas will also be marked with their source. The proof proceeds as follows: letting Δ denote the error $\hat{\Theta} - \Theta^*$, by the optimality of $\hat{\Theta}$ we can write

$$\begin{aligned}
0 &\geq \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) + \lambda(\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}}) \\
&\geq \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \lambda\|\Delta\|_{\text{nuc}},
\end{aligned}$$

where the second line follows from reverse triangle. If $\mathcal{L}$ were strongly convex with parameter τ, then we would have $0 \geq \mathcal{L}(\Theta^*+\Delta) - \mathcal{L}(\Theta^*) - \lambda\|\Delta\|_{\text{nuc}} \geq \langle\nabla\mathcal{L}(\Theta^*), \Delta\rangle + \tau\|\Delta\|_F^2 - \lambda\|\Delta\|_{\text{nuc}} \geq -\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}\|\Delta\|_{\text{nuc}} + \tau\|\Delta\|_F^2 - \lambda\|\Delta\|_{\text{nuc}}$, where the second inequality follows from strong convexity and the third from Holder's inequality. Then, bounding $\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}$ (Lemma A.6) and $\|\Delta\|_{\text{nuc}}$ (Lemma A.8) and rearranging would result in a bound for the error $\|\Delta\|_F^2$. However, $\mathcal{L}$ is not

strongly convex: if we do not restrict to low-rank matrices, then there are multiple matrices that agree with the incomplete observations. Therefore, we'll instead show a form of restricted strong convexity: in particular, we'll show (in Lemma A.7) that the quantity $\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle$ concentrates around $\frac{1}{d^2}\|\Delta\|_F^2$, but with deviation terms depending on $\|\Delta\|_{\text{nuc}}$ and $\|\Delta\|_{\max}$. Therefore, recovery will depend on $\|\Delta\|_{\text{nuc}}$ being small, which follows from the condition that the regularization strength λ is large enough (Lemma A.8), and the entries of Δ being bounded, which follows by our assumption that $\|X\|_{\max}^2 \leq \alpha$.

We'll conclude this outline by providing caricatures of each of the lemmas and showing how they come together to produce the desired bound.

(a) Section A.6 (operator norm bound):

$$\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}} \lesssim \alpha \sqrt{\frac{\log d}{dm}}.$$

(b) Section A.7 (restricted strong convexity):

$$\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle \geq \frac{1}{d^2}\|\Delta\|_F^2 - c\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}}.$$

(c) Section A.8 (decomposability):

$$\text{If } \lambda \geq 2\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}}, \text{ then } \|\Delta\|_{\text{nuc}} \lesssim \sqrt{r}\|\Delta\|_F.$$

Then, from these lemma caricatures, we can first apply restricted strong convexity as follows:

$$\begin{aligned}
0 &\geq \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \lambda\|\Delta\|_{\text{nuc}} \\
&\overset{(i)}{\geq} \boxed{\langle \nabla \mathcal{L}(\Theta^*), \Delta \rangle + \frac{1}{d^2}\|\Delta\|_F^2 - c\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}}} - \lambda\|\Delta\|_{\text{nuc}} \\
&\overset{(ii)}{\geq} \boxed{-\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}}\|\Delta\|_{\text{nuc}}} + \frac{1}{d^2}\|\Delta\|_F^2 - c\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - \lambda\|\Delta\|_{\text{nuc}} \\
&\overset{(iii)}{\geq} \boxed{-\frac{\lambda}{2}\|\Delta\|_{\text{nuc}}} + \frac{1}{d^2}\|\Delta\|_F^2 - c\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - \lambda\|\Delta\|_{\text{nuc}}
\end{aligned}$$

where (i) follows from Lemma A.7 (restricted strong convexity), (ii) from Holder's inequality, and (iii) from choosing $\lambda \geq 2\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}}$. Next, we can combine the terms with λ and use our upper bound on $\|\Delta\|_{\text{nuc}}$, producing

$$\begin{aligned}
&= \frac{1}{d^2}\|\Delta\|_F^2 - \|\Delta\|_{\text{nuc}}\left(c\alpha\sqrt{\frac{\log d}{dm}} + \frac{3}{2}\lambda\right) \\
&\overset{(iv)}{=} \frac{1}{d^2}\|\Delta\|_F^2 - \|\Delta\|_{\text{nuc}}\left(c\alpha\sqrt{\frac{\log d}{dm}} + \boxed{c'\alpha\sqrt{\frac{\log d}{dm}}}\right) \\
&\overset{(v)}{\geq} \frac{1}{d^2}\|\Delta\|_F^2 - \boxed{\sqrt{r}\|\Delta\|_F}\, c''\alpha\sqrt{\frac{\log d}{dm}}
\end{aligned}$$

where (iv) follows from choosing $\lambda = O(\|\nabla \mathcal{L}(\Theta^*)\|_{\text{op}})$ and Lemma A.6 (operator norm bound), and (v) from Lemma A.8 (nuclear norm bound). Finally, rearranging produces $\frac{1}{d}\|\Delta\|_F \lesssim \alpha\sqrt{\frac{rd\log d}{m}}$ as desired. Note that this calculation is not a proof; please see Section A.9 for the full proof.

A.6. Operator norm bound

In this section, we use concentration arguments to upper bound the operator norm $\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}$ with high probability (where the randomness is over the sampled indices). Recall that $\mathcal{L}$ is defined in Equation 6; a quick calculation reveals that

$$\begin{aligned}\nabla\mathcal{L}(\Theta^*) = &\frac{1}{m}\sum_{i=1}^{m}(\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})\tilde{E}_{a(i),b(i)} \\ &+ \frac{1}{2m}\sum_{i=1}^{m}(\Theta^*_{a(i),a(i)} - X^2_{i,a(i)})E_{a(i),a(i)} \\ &+ \frac{1}{2m}\sum_{i=1}^{m}(\Theta^*_{b(i),b(i)} - X^2_{i,b(i)})E_{b(i),b(i)},\end{aligned}$$

where we use $\tilde{E}_{a(i),b(i)}$ to denote the symmetric mask $\frac{1}{2}(E_{a(i),b(i)} + E_{b(i),a(i)})$.

Lemma A.6. *Given matrices $X = UV^T \in \mathbb{R}^{m\times d}$ and $\Theta^* = \frac{1}{m}X^TX \in \mathbb{R}^{d\times d}$, suppose that X is bounded by $\|X\|^2_{max} \leq \alpha$. Also, let $(a(1), b(1)), ..., (a(m), b(m))$ denote indices sampled i.i.d. uniformly from the set $\{(i,j) : i,j \in [d], i \neq j\}$. Then, for $\mathcal{L}$ defined in Equation 6, we have that*

$$\|\nabla\mathcal{L}(\Theta^*)\|_{op} \leq 8\alpha\sqrt{\frac{\log d + \delta}{dm}}$$

with probability $\geq 1 - e^{-\delta}$, for $m \geq d(\log d + \delta)$.

Proof. We'll first divide the bound into three parts, such that by triangle we have that

$$\begin{aligned}\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}} \leq &\left\|\frac{1}{m}\sum_{i=1}^{m}(\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})\tilde{E}_{a(i),b(i)}\right\|_{\text{op}} \\ &+ \left\|\frac{1}{2m}\sum_{i=1}^{m}(\Theta^*_{a(i),a(i)} - X^2_{i,a(i)})E_{a(i),a(i)}\right\|_{\text{op}} \\ &+ \left\|\frac{1}{2m}\sum_{i=1}^{m}(\Theta^*_{b(i),b(i)} - X^2_{i,b(i)})E_{b(i),b(i)}\right\|_{\text{op}}.\end{aligned}$$

Then, we can bound each part separately with high probability and then use the union bound to bound their sum. The bound for each term will proceed as follows: first, writing the sum as $\|\frac{1}{m}\sum_{i=1}^{m} Q_i\|_{\text{op}}$ for ease of notation, we can use Markov's inequality to produce the Chernoff bound

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m} Q_i\right\|_{\text{op}} \geq t\right) \leq \mathbb{E}\left[\exp\left\{\xi\left\|\sum_{i=1}^{m} Q_i\right\|_{\text{op}}\right\}\right] e^{-\xi m t}.$$

Next, we'll use symmetrization to bound

$$\mathbb{E}\left[\exp\left\{\xi\left\|\sum_{i=1}^{m} Q_i\right\|_{\text{op}}\right\}\right] \leq \mathbb{E}\left[\exp\left\{2\xi\left\|\sum_{i=1}^{m}\varepsilon_i\tilde{Q}_i\right\|_{\text{op}}\right\}\right],$$

where ε_i are i.i.d. Radamacher random variables (i.e. uniform over the set $\{-1, +1\}$). Finally, we can bound the moments $\mathbb{E}(\varepsilon_i\tilde{Q}_i)^{2n}$ to bound this expectation, leading to a matrix Bernstein bound.

Radamacher symmetrization: we'll first apply the symmetrization argument to the first term; the other two terms proceed similarly. First, note that by the definitions of $X = UV^T$ and $\Theta^* = \frac{1}{m}X^TX$ we can write

$$X_{i,a(i)}X_{i,b(i)} - \Theta^*_{a(i),b(i)} = \left\langle E_{a(i),b(i)}, V\left(u_iu_i^T - \frac{1}{m}\sum_{i=1}^{m} u_iu_i^T\right)V^T\right\rangle,$$

where u_i denotes the ith row of U. Substituting this expression into the expectation $\mathbb{E}\left[\exp\left\{\xi \left\|\sum_{i=1}^m Q_i\right\|_{\text{op}}\right\}\right]$, we have

$$\mathbb{E}\left[\exp\left\{\xi \left\|\sum_{i=1}^m (\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})\tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\right\}\right]$$
$$= \mathbb{E}\left[\exp\left\{\xi \left\|\sum_{i=1}^m \left\langle E_{a(i),b(i)}, V\left(u_iu_i^T - \frac{1}{m}\sum_{i=1}^m u_iu_i^T\right)V^T\right\rangle \tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\right\}\right].$$

Next, because the random indices are drawn i.i.d., this expectation is invariant to the u_i's being permuted with each other (i.e. sampling a random permutation σ and setting $u_i' = u_{\sigma(i)}$). Therefore, we can take an expectation with respect to sampling a random permutation σ while also replacing $\frac{1}{m}\sum_{i=1}^m u_iu_i^T = \mathbb{E}_{\tilde{\sigma}} u_{\tilde{\sigma}(i)}u_{\tilde{\sigma}(i)}^T$, resulting in

$$= \mathbb{E}_{a(i),b(i)}\boxed{\mathbb{E}_\sigma}\left[\exp\left\{\xi \left\|\sum_{i=1}^m \langle E_{a(i),b(i)}, V\left(\boxed{u_{\sigma(i)}u_{\sigma(i)}^T - \mathbb{E}_{\tilde{\sigma}}[u_{\tilde{\sigma}(i)}u_{\tilde{\sigma}(i)}^T]}\right)V^T\rangle \tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\right\}\right],$$

where $\mathbb{E}_Z$ denotes taking the expectation with respect to Z. At this point, we can apply the definition of the operator norm and proceed with standard symmetrization arguments, resulting in the following chain of inequalities:

Replacing operator norm $\|C\|_{\text{op}}$ with $\sup_{\|z\|_2=1}\langle z, Cz\rangle$:

$$= \mathbb{E}_{a(i),b(i)}\mathbb{E}_\sigma\left[\exp\left\{\xi \sup_{\|z\|_2=1}\left\langle z, \sum_{i=1}^m \langle E_{a(i),b(i)}, V\left(u_{\sigma(i)}u_{\sigma(i)}^T - \mathbb{E}_{\tilde{\sigma}}[u_{\tilde{\sigma}(i)}u_{\tilde{\sigma}(i)}^T]\right)V^T\rangle \tilde{E}_{a(i),b(i)}z\right\rangle\right\}\right]$$

Pulling out the expectation via $\Phi(\sup_{g\in\mathcal{G}}\mathbb{E}|g(X)|) \le \mathbb{E}\Phi(\sup_{g\in\mathcal{G}}|g(X)|)$ for $\Phi = \exp$ convex and non-decreasing:

$$\le \mathbb{E}_{a(i),b(i)}\mathbb{E}_\sigma\boxed{\mathbb{E}_{\tilde{\sigma}}}\left[\exp\left\{\xi \sup_{\|z\|_2=1}\left\langle z, \sum_{i=1}^m \langle E_{a(i),b(i)}, V\left(u_{\sigma(i)}u_{\sigma(i)}^T - u_{\tilde{\sigma}(i)}u_{\tilde{\sigma}(i)}^T\right)V^T\rangle \tilde{E}_{a(i),b(i)}z\right\rangle\right\}\right]$$

We can insert Radamacher random variables ε because σ and $\tilde{\sigma}$ are i.i.d.:

$$= \boxed{\mathbb{E}_\varepsilon}\mathbb{E}_{a(i),b(i)}\mathbb{E}_\sigma\mathbb{E}_{\tilde{\sigma}}\left[\exp\left\{\xi \sup_{\|z\|_2=1}\left\langle z, \sum_{i=1}^m \langle E_{a(i),b(i)}, V\boxed{\varepsilon_i}\left(u_{\sigma(i)}u_{\sigma(i)}^T - u_{\tilde{\sigma}(i)}u_{\tilde{\sigma}(i)}^T\right)V^T\rangle \tilde{E}_{a(i),b(i)}z\right\rangle\right\}\right]$$

Splitting the sum via Jensen's inequality:

$$\le \mathbb{E}_{\varepsilon,a(i),b(i)}\mathbb{E}_\sigma\left[\frac{1}{2}\exp\left\{2\xi \sup_{\|z\|_2=1}\left\langle z, \sum_{i=1}^m \varepsilon_i\langle E_{a(i),b(i)}, V\left(u_{\sigma(i)}u_{\sigma(i)}^T\right)V^T\rangle \tilde{E}_{a(i),b(i)}z\right\rangle\right\}\right]$$
$$+ \mathbb{E}_{\varepsilon,a(i),b(i)}\mathbb{E}_{\tilde{\sigma}}\left[\frac{1}{2}\exp\left\{2\xi \sup_{\|z\|_2=1}\left\langle z, \sum_{i=1}^m \varepsilon_i\langle E_{a(i),b(i)}, V\left(u_{\tilde{\sigma}(i)}u_{\tilde{\sigma}(i)}^T\right)V^T\rangle \tilde{E}_{a(i),b(i)}z\right\rangle\right\}\right]$$

Removing σ and $\tilde{\sigma}$ by again applying invariance of the expectation to permutation:

$$= \mathbb{E}_{\varepsilon,a(i),b(i)}\left[\exp\left\{2\xi \sup_{\|z\|_2=1}\left\langle z, \sum_{i=1}^m \varepsilon_i\langle E_{a(i),b(i)}, Vu_iu_i^TV^T\rangle \tilde{E}_{a(i),b(i)}z\right\rangle\right\}\right]$$
$$= \mathbb{E}\left[\exp\left\{2\xi \left\|\sum_{i=1}^m \varepsilon_i X_{i,a(i)}X_{i,b(i)}\tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\right\}\right].$$

We can proceed in exactly the same way for the diagonal terms, resulting in the following inequalities:

$$\mathbb{E}\left[\exp\left\{\xi\left\|\sum_{i=1}^{m}(\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})\tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\right\}\right] \leq \mathbb{E}\left[\exp\left\{2\xi\left\|\sum_{i=1}^{m}\varepsilon_i X_{i,a(i)}X_{i,b(i)}\tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\right\}\right],$$

$$\mathbb{E}\left[\exp\left\{\xi\left\|\sum_{i=1}^{m}\frac{1}{2}(\Theta^*_{a(i),a(i)} - X^2_{i,a(i)})E_{a(i),a(i)}\right\|_{\text{op}}\right\}\right] \leq \mathbb{E}\left[\exp\left\{2\xi\left\|\sum_{i=1}^{m}\frac{1}{2}\varepsilon_i X^2_{i,a(i)}E_{a(i),a(i)}\right\|_{\text{op}}\right\}\right].$$

Bounding moments: at this point, we can apply standard matrix Bernstein arguments. First note that for symmetric independent random matrices Q_i, we have

$$\mathbb{E}e^{2\xi\|\sum_i Q_i\|_{\text{op}}} \overset{(i)}{=} \mathbb{E}\|e^{2\xi\sum_i Q_i}\|_{\text{op}} \overset{(ii)}{\leq} \mathbb{E}\text{tr}(e^{2\xi\sum_i Q_i}) = \text{tr}(\mathbb{E}e^{2\xi\sum_i Q_i}) \overset{(iii)}{\leq} \text{tr}(e^{\sum_i \log \mathbb{E}e^{2\xi Q_i}}),$$

where (i) follows from the spectral mapping theorem, (ii) from the fact that the matrix exponential $e^Q = \sum_{k=0}^{\infty}\frac{Q^k}{k!}$ is positive semidefinite, and (iii) from Lemma 6.13 of Wainwright (2019). Therefore, it suffices to bound each $\mathbb{E}e^{2\xi Q_i}$. For ease of notation, we'll define the following random matrices, which are symmetric:

$$R_i = \frac{1}{2}\varepsilon_i X_{i,a(i)}X_{i,b(i)}(E_{a(i),b(i)} + E_{b(i),a(i)})$$
$$S_i = \varepsilon_i X^2_{i,a(i)}E_{a(i),a(i)}.$$

Then, to bound $\mathbb{E}e^{(2\xi R_i)}$ and $\mathbb{E}e^{(2\xi S_i)}$, we can bound $\mathbb{E}R_i^{2n}$ and $\mathbb{E}S_i^{2n}$ (note that the odd moments are zero because ε_i is symmetric around the origin). Using our assumption that $\|X\|^2_{\max} \leq \alpha$, we can compute these moments as follows:

$$\begin{aligned}\mathbb{E}R_i^{2n} &= \frac{1}{2^{2n}}X^{2n}_{i,a(i)}X^{2n}_{i,b(i)}\frac{2}{d}I_d \\ &\preceq \alpha^{2n}\frac{1}{d}I_d \\ \mathbb{E}S_i^{2n} &= (X^2_{i,a(i)})^{2n}\frac{1}{d}I_d \\ &\preceq \alpha^{2n}\frac{1}{d}I_d,\end{aligned}$$

so R_i and S_i both satisfy the matrix Bernstein condition with $b = \alpha$ and $\text{var}(R_i) \preceq \alpha^2\frac{1}{d}I_d$. Then, by a matrix Bernstein bound (see, e.g., Lemma 6.11 of Wainwright (2019)), we have

$$\begin{aligned}\mathbb{E}e^{2\xi R_i} &\preceq \exp\left\{\frac{2\xi^2\text{var}(R_i)}{1 - b|\xi|}\right\} \text{ for all } |\xi| < 1/b \\ &\preceq \exp\left\{\frac{2\xi^2\alpha^2 I_d}{d(1-\alpha|\xi|)}\right\} \text{ for all } |\xi| < 1/\alpha,\end{aligned}$$

with the same inequality holding for S_i. Substituting into the original inequality, for all $|\xi| < 1/\alpha$ we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}(\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})\tilde{E}_{a(i),b(i)}\right\|_{\text{op}} \geq t\right) &\leq \text{tr}(e^{\sum_i \log \mathbb{E}e^{2\xi R_i}})e^{-\xi m t} \\ &\leq \text{tr}\left(\exp\left\{\frac{2m\xi^2\alpha^2 I_d}{d(1-\alpha|\xi|)}\right\}\right)e^{-\xi m t} \\ &\leq d\exp\left\{\frac{2m\xi^2\alpha^2}{d(1-\alpha|\xi|)}\right\}e^{-\xi m t},\end{aligned}$$

where the last line follows from the fact that $\text{tr}(e^R) \leq de^{\|R\|_{\text{op}}}$ for symmetric matrices $R \in \mathbb{R}^{d\times d}$. Setting $\xi = \frac{t}{4\alpha^2/d + \alpha t}$ produces the bound

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}(\Theta^*_{a(i),b(i)} - X_{i,a(i)}X_{i,b(i)})\tilde{E}_{a(i),b(i)}\right\|_{\text{op}} \leq t\right) \geq 1 - d\exp\left\{-\frac{mt^2}{8\alpha^2/d + 2\alpha t}\right\},$$

with the same bound holding for the second and third terms. Finally, for all three bounds to hold simultaneously with probability $\geq 1 - e^{-\delta}$, we can set

$$t = 4\max\left(2\alpha\sqrt{\frac{\log d + \delta}{dm}}, \alpha\frac{\log d + \delta}{m}\right),$$

which is dominated by the first term for $m \geq d(\log d + \delta)$. □

A.7. Restricted strong convexity

In this section, we will lower bound the quantity $\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla\mathcal{L}(\Theta^*), \Delta\rangle$ with high probability (where the randomness is over the sampled indices). In particular, we'll show that this quantity concentrates around $\frac{1}{d^2}\|\Delta\|_F^2$ through careful analysis of the diagonal and off-diagonal terms, along with peeling arguments similar to those in Theorem 10.17 of Wainwright (2019) and Theorem 1 of Negahban & Wainwright (2012). Recall that $\mathcal{L}$ is defined in Equation 6; a quick calculation reveals that

$$\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla\mathcal{L}(\Theta^*), \Delta\rangle = \frac{1}{2m}\sum_{i=1}^{m}[\Delta^2_{a(i),b(i)} + \Delta^2_{b(i),a(i)}] + [\Delta^2_{a(i),a(i)} + \Delta^2_{b(i),b(i)}].$$

For a matrix $\Delta \in \mathbb{R}^{d\times d}$, we'll use $P_{\text{diag}}(\Delta)$ to refer to Δ with the off-diagonal terms set to zero, and $P_{\text{off-diag}}(\Delta) = \Delta - P_{\text{diag}}(\Delta)$ to refer to Δ with the diagonal set to zero.

Lemma A.7. *Let $(a(1), b(1)), ..., (a(m), b(m))$ be random indices sampled i.i.d. uniformly from the set $\{(i,j) : i,j \in [d], i \neq j\}$. Also, let $m \geq d\log d$. Then, for universal constants c_1, c_2, and c_3, we have that for $\mathcal{L}$ defined in Equation 6, the following bound holds uniformly for all matrices $\Delta \in \mathbb{R}^{d\times d}$, with probability $\geq 1 - 2e^{-\delta}$:*

$$\begin{aligned}\mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla\mathcal{L}(\Theta^*), \Delta\rangle \geq\ & \frac{1}{d^2}\|\Delta\|_F^2 \\ & - c_1\|\Delta\|_{max}\|\Delta\|_{nuc}\sqrt{\frac{\log d}{dm}} \\ & - c_2\|\Delta\|_{max}\|\Delta\|_F\sqrt{\frac{\delta}{dm}} \\ & - c_3\|\Delta\|_{max}^2\frac{\delta}{m}.\end{aligned}$$

Proof. To show this bound, we'll show the following two bounds for the off-diagonal and diagonal terms, which each hold with probability $\geq 1 - e^{-\delta}$ (and therefore together with probability $\geq 1 - 2e^{-\delta}$):

(a) **Off-diagonal terms**: with probability $\geq 1 - e^{-\delta}$, the following holds uniformly for all $\Delta \in \mathbb{R}^{d\times d}$:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}[\Delta^2_{a(i),b(i)} + \Delta^2_{b(i),a(i)}] \geq\ & \frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2 \\ & - c_1\|P_{\text{off-diag}}(\Delta)\|_{\max}\|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} \\ & - c_2\|P_{\text{off-diag}}(\Delta)\|_{\max}\|P_{\text{off-diag}}(\Delta)\|_F\sqrt{\frac{\delta}{d(d-1)m}} \\ & - c_3\|P_{\text{off-diag}}(\Delta)\|_{\max}^2\frac{\delta}{m}.\end{aligned}$$

(b) **Diagonal terms**: with probability $\geq 1 - e^{-\delta}$, the following holds uniformly for all $\Delta \in \mathbb{R}^{d \times d}$:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}[\Delta_{a(i)}^2 + \Delta_{b(i)}^2] \geq \frac{1}{d}\|P_{\text{diag}}(\Delta)\|_F^2 \\ - c_1\|P_{\text{diag}}(\Delta)\|_{\max}\|P_{\text{diag}}(\Delta)\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} \\ - c_2\|P_{\text{diag}}(\Delta)\|_{\max}\|P_{\text{diag}}(\Delta)\|_F\sqrt{\frac{\delta}{dm}} \\ - c_3\|P_{\text{diag}}(\Delta)\|_{\max}^2\frac{\delta}{m}.\end{aligned}$$

Note that the lemma follows from adding these two claims because we have the following inequalities:

(i) **Max**: the max of the entire matrix bounds the max of subsets of the matrix, so $\|P_{\text{diag}}(\Delta)\|_{\max} \leq \|\Delta\|_{\max}$ and $\|P_{\text{off-diag}}(\Delta)\|_{\max} \leq \|\Delta\|_{\max}$.

(ii) **Nuclear norm**: the P_{diag} operator reduces nuclear norm, so $\|P_{\text{diag}}(\Delta)\|_{\text{nuc}} \leq \|\Delta\|_{\text{nuc}}$ and $\|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}} = \|\Delta - P_{\text{diag}}(\Delta)\|_{\text{nuc}} \leq 2\|\Delta\|_{\text{nuc}}$.

(iii) **Frobenius norm terms**: because $\|P_{\text{off-diag}}(\Delta)\|_F^2 + \|P_{\text{diag}}(\Delta)\|_F^2 = \|\Delta\|_F^2$, we have

$$\frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2 + \frac{1}{d}\|P_{\text{diag}}(\Delta)\|_F^2 \geq \frac{1}{d^2}\|P_{\text{off-diag}}(\Delta)\|_F^2 + \frac{1}{d^2}\|P_{\text{diag}}(\Delta)\|_F^2 = \frac{1}{d^2}\|\Delta\|_F^2,$$

lower bounding the first term, and we also have

$$\begin{aligned}\|P_{\text{off-diag}}(\Delta)\|_F\sqrt{\frac{1}{d(d-1)}} + \|P_{\text{diag}}(\Delta)\|_F\sqrt{\frac{1}{d}} &\leq \sqrt{2\|P_{\text{off-diag}}(\Delta)\|_F^2\frac{1}{d(d-1)} + 2\|P_{\text{diag}}(\Delta)\|_F^2\frac{1}{d}} \\ &\leq \sqrt{2\|P_{\text{off-diag}}(\Delta)\|_F^2\frac{1}{d} + 2\|P_{\text{diag}}(\Delta)\|_F^2\frac{1}{d}} \\ &= \sqrt{\frac{2}{d}}\|\Delta\|_F,\end{aligned}$$

which upper bounds the deviation.

Off-diagonal term bound: We'll start by bounding the off-diagonal terms; the proof for the diagonal terms will proceed similarly. First, note that the inequality is scale-invariant, so WLOG we can assume $\|P_{\text{off-diag}}(\Delta)\|_{\max} = \alpha$. Fixing some D and ρ, let $\mathbb{Q}(D, \rho)$ denote the set

$$\mathbb{Q}(D, \rho) = \{\Delta \in \mathbb{R}^{d \times d} : \|P_{\text{off-diag}}(\Delta)\|_{\max} = \alpha,\ \|P_{\text{off-diag}}(\Delta)\|_F \leq D,\ \|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}} \leq \rho\}.$$

and let $Z(D, \rho)$ denote the largest deviation in $\mathbb{Q}(D, \rho)$, or

$$Z(D, \rho) = \sup_{\Delta \in \mathbb{Q}(D,\rho)}\left|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}[\Delta_{a(i),b(i)}^2 + \Delta_{b(i),a(i)}^2] - \frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2\right|.$$

We'll first produce a high probability upper bound on $Z(D, \rho)$ for fixed D and ρ, and we'll then use a peeling argument to produce a high probability bound for general D and ρ.

Bound: First, note that each summand has expectation

$$\mathbb{E}\left[\langle \Delta, \tilde{E}_{a(i),b(i)}\rangle^2\right] = \frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2,$$

so $Z(D,\rho)$ is an empirical process of the form $\sup_{g\in\mathcal{G}}|\frac{1}{m}\sum_{i=1}^m g(X_i)-\mathbb{E}g(X)|$. Furthermore, we have that each term is uniformly bounded $\frac{1}{2}[\Delta^2_{a(i),b(i)}+\Delta^2_{b(i),a(i)}]\le\alpha^2$, with uniformly bounded variance:

$$\begin{aligned}\text{var}\left(\langle\Delta,\tilde{E}_{a(i),b(i)}\rangle^2\right)&=\mathbb{E}\langle\Delta,\tilde{E}_{a(i),b(i)}\rangle^4-\mathbb{E}\langle\Delta,\tilde{E}_{a(i),b(i)}\rangle^2\\&\le\alpha^2\mathbb{E}\langle\Delta,\tilde{E}_{a(i),b(i)}\rangle^2-\mathbb{E}\langle\Delta,\tilde{E}_{a(i),b(i)}\rangle^2\\&=(\alpha^2-1)\frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2\\&\le\alpha^2\frac{1}{d(d-1)}D^2.\end{aligned}$$

Therefore, by a functional Bernstein inequality (Theorem 3.27 of Wainwright (2019)), we have

$$\mathbb{P}\left(Z(D,\rho)\ge 2\mathbb{E}Z(D,\rho)+2\sigma\sqrt{\frac{\delta'}{m}}+2b\frac{\delta'}{m}\right)\le e^{-\delta'}$$

where $\sigma=\alpha D\sqrt{\frac{1}{d(d-1)}}$ and $b=\alpha^2$. Next, to bound the expectation $\mathbb{E}Z(D,\rho)$, we have that

$$\begin{aligned}\mathbb{E}\sup_{\Delta\in\mathbb{Q}(D,\rho)}\left|\frac{1}{m}\sum_{i=1}^m\langle\tilde{E}_{a(i),b(i)},\Delta\rangle^2-\mathbb{E}[\langle\tilde{E}_{a(i),b(i)},\Delta\rangle^2]\right|&\overset{(i)}{\le}2\mathbb{E}\sup_{\Delta\in\mathbb{Q}(D,\rho)}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\langle\tilde{E}_{a(i),b(i)},\Delta\rangle^2\right|\\&\overset{(ii)}{=}2\mathbb{E}\sup_{\Delta\in\mathbb{Q}(D,\rho)}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\langle\tilde{E}_{a(i),b(i)},P_{\text{off-diag}}(\Delta)\rangle^2\right|\\&\overset{(iii)}{\le}4\alpha\mathbb{E}\sup_{\Delta\in\mathbb{Q}(D,\rho)}\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\langle\tilde{E}_{a(i),b(i)},P_{\text{off-diag}}(\Delta)\rangle\right|\\&\overset{(iv)}{\le}4\alpha\mathbb{E}\sup_{\Delta\in\mathbb{Q}(D,\rho)}\left|\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}}\right|\\&\le4\alpha\rho\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\\&\overset{(v)}{\le}64\alpha\rho\sqrt{\frac{\log d}{dm}},\end{aligned}$$

where (i) follows from Radamacher symmetrization, (ii) follows from the fact that $a(i)\neq b(i)$ so only off-diagonal terms of Δ are sampled, (iii) follows from the fact that $\|P_{\text{off-diag}}(\Delta)\|_{\max}\le\alpha$ and the Ledoux-Talagrand contraction inequality for Radamacher processes (see (5.61) in Wainwright (2019) or Section 4.2 in Ledoux & Talagrand (1991)), and (iv) follows from Holder's inequality. To show (v), note that each $\varepsilon_i\tilde{E}_{a(i),b(i)}$ is mean zero with operator norm 1 and variance $\text{var}(\varepsilon_i\tilde{E}_{a(i),b(i)})=\frac{1}{d}I_d$, so by matrix Bernstein (Theorem 6.17 of Wainwright (2019)) we have

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\tilde{E}_{a(i),b(i)}\right\|_{\text{op}}\ge t\right)\le 2d\exp\left\{-\frac{mt^2}{2(\frac{1}{d}+t)}\right\}.$$

Then, for $m\ge d\log d$, we can integrate to bound the expectation by $16\sqrt{\frac{\log d}{dm}}$: in particular, by Exercise 2.8(a) of Wainwright (2019), we have that

$$\mathbb{P}(Z\ge t)\le Ce^{-\frac{t^2}{2(\nu^2+bt)}}\implies\mathbb{E}Z\le2\nu(\sqrt{\pi}+\sqrt{\log C})+4b(1+\log C)$$

where we have $C = 2d$, $\nu^2 = \frac{1}{dm}$, and $b = \frac{1}{m}$, resulting in

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i \tilde{E}_{a(i),b(i)}\right\|_{\text{op}} \leq 2\frac{1}{\sqrt{dm}}(\sqrt{\pi}+\sqrt{\log(2d)}) + 4\frac{1}{m}(1+\log(2d))$$
$$\leq 16\sqrt{\frac{\log d}{dm}},$$

where the second inequality follows from combining terms and using the fact that the first term is larger when $m \geq d\log d$. Putting everything together, we have that

$$\mathbb{P}\left(Z(D,\rho) \leq 128\alpha\rho\sqrt{\frac{\log d}{dm}} + 4\alpha D\sqrt{\frac{\delta'}{d(d-1)m}} + 4\alpha^2\frac{\delta'}{m}\right) \geq 1 - e^{-\delta'}. \tag{7}$$

Peeling: given this bound, we can use a peeling argument to extend to general D and ρ. Our approach will be to cover all possible Δ with the sets

$$\mathbb{Q}_{k,\ell} = \{\Delta \in \mathbb{R}^{d\times d} : \|P_{\text{off-diag}}(\Delta)\|_{\max} = \alpha,\ \alpha 2^{k-1} \leq \|P_{\text{off-diag}}(\Delta)\|_F \leq \alpha 2^k,\ \alpha 2^{\ell-1} \leq \|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}} \leq \alpha 2^\ell\}.$$

The idea is that we can use the union bound to ensure that the uniform bound in the first part (Equation 7) applies to each set $\mathbb{Q}_{k,\ell}$. Then, because $\mathbb{Q}_{k,\ell}$ captures values of D and ρ up to factors of 2, we can produce the desired inequality for general Δ while only losing constant factors in the deviation terms.

The first step is to bound the number of such sets that we need, which we can do by upper- and lower-bounding the Frobenius and nuclear norms with respect to α:

$$\alpha = \|P_{\text{off-diag}}(\Delta)\|_{\max} \leq \|P_{\text{off-diag}}(\Delta)\|_F \leq d\|P_{\text{off-diag}}(\Delta)\|_{\max} = d\alpha$$
$$\alpha = \|P_{\text{off-diag}}(\Delta)\|_{\max} \leq \|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}} \leq d^{3/2}\|P_{\text{off-diag}}(\Delta)\|_{\max} = d^{3/2}\alpha.$$

Therefore, it suffices to have $k = 1, 2, ..., \lceil \log d \rceil$ and $\ell = 1, 2, ..., \lceil (3/2)\log d \rceil$. By the union bound, the probability that the bound in Equation 7 holds for all of the sets $\mathbb{Q}_{k,\ell}$ is at least $\geq 1 - \lceil (3/2)\log d \rceil \lceil \log d \rceil 2\exp\{-\delta'\}$, which we can bound by $\geq 1 - e^{-\delta}$ for $\delta \geq \log 6 + \log\log d$ by setting $\delta' = 3\delta$.

Then, for any specific Δ, letting k, ℓ be the indices such that $\Delta \in \mathbb{Q}_{k,\ell}$, we have that

$$\frac{1}{m}\sum_{i=1}^{m}\langle \Delta, \tilde{E}_{a(i),b(i)}\rangle^2 \overset{(i)}{\geq} \frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2 - c_1\alpha\rho\sqrt{\frac{\log d}{dm}} - c_2\alpha D\sqrt{\frac{\delta}{d(d-1)m}} - c_3\alpha^2\frac{\delta}{m}$$
$$= \frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2 - c_1\alpha(\alpha 2^\ell)\sqrt{\frac{\log d}{dm}} - c_2\alpha(\alpha 2^k)\sqrt{\frac{\delta}{d(d-1)m}} - c_3\alpha^2\frac{\delta}{m}$$
$$\overset{(ii)}{\geq} \frac{1}{d(d-1)}\|P_{\text{off-diag}}(\Delta)\|_F^2 - 2c_1\alpha\|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}}$$
$$- 2c_2\alpha\|P_{\text{off-diag}}(\Delta)\|_F\sqrt{\frac{\delta}{d(d-1)m}} - c_3\alpha^2\frac{\delta}{m},$$

where (i) follows from the fact that $\mathbb{Q}_{k,\ell} \subseteq \mathbb{Q}(D,\rho)$ for $D = \alpha 2^k, \rho = \alpha 2^\ell$, and (ii) follows from the fact that $\alpha 2^{k-1} \leq \|P_{\text{off-diag}}(\Delta)\|_F$ and $\alpha 2^{\ell-1} \leq \|P_{\text{off-diag}}(\Delta)\|_{\text{nuc}}$.

Diagonal terms: the bound for the diagonal terms proceeds in the same way, but with slightly different quantities. As before, we'll assume that $\|P_{\text{diag}}(\Delta)\|_{\max} = \alpha$. Fixing D and ρ, we'll define the set $\mathbb{Q}(D,\rho)$ as

$$\mathbb{Q}(D,\rho) = \{\Delta \in \mathbb{R}^{d\times d} : \|P_{\text{diag}}(\Delta)\|_{\max} = \alpha,\ \|P_{\text{diag}}(\Delta)\|_F \leq D,\ \|P_{\text{diag}}(\Delta)\|_{\text{nuc}} \leq \rho\},$$

and we'll define $Z(D,\rho)$ as

$$Z(D,\rho) = \sup_{\Delta\in\mathbb{Q}(D,\rho)}\left|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}[\Delta^2_{a(i),a(i)} + \Delta^2_{b(i),b(i)}] - \frac{1}{d}\|P_{\text{diag}}(\Delta)\|_F^2\right|,$$

where we have that each term has expectation

$$\mathbb{E}\left[\frac{1}{2}[\Delta^2_{a(i),a(i)} + \Delta^2_{b(i),b(i)}]\right] = \frac{1}{d}\|P_{\text{diag}}(\Delta)\|_F^2.$$

We again have that each term is uniformly bounded by α^2 and has uniformly bounded variance, but this time with $\alpha^2 \frac{1}{d} D^2$. The bound for $\mathbb{E}Z(D,\rho) \leq 64\alpha\rho\sqrt{\frac{\log d}{dm}}$ proceeds exactly as before, producing the functional Bernstein inequality

$$\mathbb{P}\left(Z(D,\rho) \leq 128\alpha\rho\sqrt{\frac{\log d}{dm}} + 4\alpha D\sqrt{\frac{\delta'}{dm}} + 4\alpha^2\frac{\delta'}{m}\right) \geq 1 - e^{-\delta'}.$$

Finally, we can proceed with exactly the same peeling argument as before, leading to the following bound holding uniformly with probability $\geq 1 - e^{-\delta}$:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}[\Delta^2_{a(i),a(i)} + \Delta^2_{b(i),b(i)}] \geq \frac{1}{d}\|P_{\text{diag}}(\Delta)\|_F^2 - 2c_1\alpha\|P_{\text{diag}}(\Delta)\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}}$$
$$- 2c_2\alpha\|P_{\text{diag}}(\Delta)\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m},$$

which completes the proof. □

A.8. Decomposability

In this section, we show that for large enough regularization strength λ, the error $\Delta = \hat{\Theta} - \Theta^*$ must have small nuclear norm. The arguments in this section proceed exactly as in Proposition 9.13 of Wainwright (2019), which we reproduce here for completeness. To aid our analysis of the nuclear norm regularizer, we'll first define the subspaces $\mathbb{M}$ and $\bar{\mathbb{M}}$ of $\mathbb{R}^{d\times d}$ as follows (where $\mathbb{S}^\perp$ denotes the orthogonal complement of a subspace $\mathbb{S}$), following Negahban et al. (2012):

$$\mathbb{M} = \{\Theta : \text{ rowspace}(\Theta) \subseteq \text{rowspace}(\Theta^*),\ \text{colspace}(\Theta) \subseteq \text{colspace}(\Theta^*)\}$$
$$\bar{\mathbb{M}}^\perp = \{\Theta : \text{ rowspace}(\Theta) \subseteq \text{rowspace}(\Theta^*)^\perp,\ \text{colspace}(\Theta) \subseteq \text{colspace}(\Theta^*)^\perp\}.$$

From these subspaces, we can define $\mathbb{M}^\perp$ and $\bar{\mathbb{M}}$ accordingly as their orthogonal complements. As an example, if Θ^* is the rank-r matrix with the $r \times r$ identity in the top left corner and zeros otherwise, then we have

$$\mathbb{M} = \begin{bmatrix}\Gamma_{r\times r} & 0\\ 0 & 0\end{bmatrix}$$
$$\mathbb{M}^\perp = \begin{bmatrix}0 & \Gamma_{r\times(d-r)}\\ \Gamma_{(d-r)\times r} & \Gamma_{(d-r)\times(d-r)}\end{bmatrix}$$
$$\bar{\mathbb{M}}^\perp = \begin{bmatrix}0 & 0\\ 0 & \Gamma_{(d-r)\times(d-r)}\end{bmatrix}$$
$$\bar{\mathbb{M}} = \begin{bmatrix}\Gamma_{r\times r} & \Gamma_{r\times(d-r)}\\ \Gamma_{(d-r)\times r} & 0\end{bmatrix},$$

where each $\Gamma_{a\times b}$ represents an arbitrary matrix in $\mathbb{R}^{a\times b}$. With these subspaces defined, we have the following two facts: (1) for any $A \in \mathbb{M}$ and $B \in \bar{\mathbb{M}}^\perp$, we have $\|A + B\|_{\text{nuc}} = \|A\|_{\text{nuc}} + \|B\|_{\text{nuc}}$, i.e. the nuclear norm is decomposable with respect to $(\mathbb{M}, \bar{\mathbb{M}})$ (Negahban et al., 2012), and (2) if Θ^* is rank r, then all matrices in $\mathbb{M}$ are at most rank r and all matrices in $\bar{\mathbb{M}}$ are at most rank $2r$. Broadly speaking, the proof will proceed by using the optimality of $\hat{\Theta}$ (with large enough regularization strength λ) to bound $\|\Delta\|_{\text{nuc}}$, which involves projecting Δ onto these subspaces and using decomposability of the nuclear norm.

Lemma A.8. *[Proposition 9.13 of Wainwright (2019)] Let $\Delta = \hat{\Theta} - \Theta^*$ denote the error, where $\hat{\Theta}$ solves the optimization problem defined in Equation 5. Also, suppose that the regularization strength λ is at least*

$$\lambda \geq 2\|\nabla\mathcal{L}(\Theta^*)\|_{op}.$$

Then, we have

$$\|\Delta\|_{nuc} \leq 4\sqrt{2r}\|\Delta\|_F.$$

Proof. First, note that we can project the error Δ onto orthogonal subspaces $\bar{\mathbb{M}}$ and $\bar{\mathbb{M}}^\perp$ as follows:

$$\begin{aligned}
\|\Delta\|_{\text{nuc}} &= \|\Delta_{\bar{\mathbb{M}}^\perp} + \Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}} \\
\text{(triangle)} \quad &\leq \|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} + \|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}} \\
&\overset{(i)}{\leq} \|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} + \sqrt{2r}\|\Delta_{\bar{\mathbb{M}}}\|_F \\
&\leq \|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} + \sqrt{2r}\|\Delta\|_F,
\end{aligned}$$

where (i) follows from the fact that any matrix in $\bar{\mathbb{M}}$ is at most rank $2r$. Therefore, at this point it suffices to bound $\|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}}$. By the optimality of $\hat{\Theta}$, we have

$$\begin{aligned}
0 &\geq L(\Theta^* + \Delta) - L(\Theta^*) + \lambda\left(\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}}\right) \\
&\overset{(i)}{\geq} \langle \nabla L(\Theta^*), \Delta\rangle + \lambda\left(\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}}\right) \\
&\overset{(ii)}{\geq} -\|\nabla L(\Theta^*)\|_{\text{op}}\|\Delta\|_{\text{nuc}} + \lambda\left(\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}}\right) \\
&\overset{(iii)}{\geq} -\frac{\lambda}{2}\|\Delta\|_{\text{nuc}} + \lambda\left(\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}}\right),
\end{aligned}$$

where (i) follows from the convexity of $\mathcal{L}$, (ii) from Holder's inequality, and (iii) from our assumption that $\lambda \geq 2\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}$. Next, we'll project the error Δ onto $\bar{\mathbb{M}}$ and $\bar{\mathbb{M}}^\perp$ to expand the second term as follows:

$$\begin{aligned}
\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}} &= \|\Theta^* + \Delta_{\bar{\mathbb{M}}^\perp} + \Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}} \\
&\overset{(i)}{\geq} \|\Theta^* + \Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} - \|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}} \\
&\overset{(ii)}{=} \|\Theta^*\|_{\text{nuc}} + \|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} - \|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}} \\
&= \|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} - \|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}},
\end{aligned}$$

where (i) follows from reverse triangle, and (ii) follows from decomposability and the fact that $\Theta^* \in \mathbb{M}$. Substituting the expression for the second term, we have that

$$\begin{aligned}
0 &\geq -\frac{\lambda}{2}\|\Delta\|_{\text{nuc}} + \lambda(\|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} - \|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}}) \\
&\overset{(i)}{\geq} -\frac{\lambda}{2}(\|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} + \|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}}) + \lambda(\|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} - \|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}}) \\
&= \frac{\lambda}{2}(\|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} - 3\|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}}),
\end{aligned}$$

where (i) follows from the triangle inequality. Rearranging, we have that $\|\Delta_{\bar{\mathbb{M}}^\perp}\|_{\text{nuc}} \leq 3\|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}}$. Putting everything together, we have that $\|\Delta\|_{\text{nuc}} \leq 4\|\Delta_{\bar{\mathbb{M}}}\|_{\text{nuc}} \leq 4\sqrt{2r}\|\Delta\|_F$ as desired. □

A.9. Proof of theorem

Finally, we can put everything together to prove Theorem A.1, which we reproduce here:

Theorem. *Let $\hat{\Theta}$ be the solution of the optimization problem defined in Equation 5, where λ is set to $16\alpha\sqrt{\frac{\log d+\delta}{dm}}$. Also, suppose that X is rank r with $\|X\|_{max}^2 \leq \alpha$, and $m \geq d(\log d + \delta)$. Then, with probability $\geq 1 - 3e^{-\delta}$, we have that*

$$\frac{1}{d^2}\|\hat{\Theta} - \Theta^*\|_F^2 \lesssim \alpha^2 \frac{rd(\log d + \delta)}{m}.$$

Proof. By our setting of $\lambda = 16\alpha\sqrt{\frac{\log d+\delta}{dm}}$, we can apply Lemma A.6 (operator norm bound) to show that

$$\frac{\lambda}{2} \geq \|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}$$

with probability $\geq 1-e^{-\delta}$. We'll also condition on the high-probability bound in Lemma A.7 (restricted strong convexity), which holds with probability $\geq 1-2e^{-\delta}$ (so by the union bound, both hold with probability $\geq 1-3e^{-\delta}$):

$$\mathcal{L}(\Theta^*+\Delta)-\mathcal{L}(\Theta^*)-\langle\nabla\mathcal{L}(\Theta^*),\Delta\rangle \geq \frac{1}{d^2}\|\Delta\|_F^2 - c_1\|\Delta\|_{\max}\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - c_2\|\Delta\|_{\max}\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\|\Delta\|_{\max}^2\frac{\delta}{m}.$$

First, note that we have $\|\Delta\|_{\max} \leq \|\hat{\Theta}\|_{\max} + \|\Theta^*\|_{\max} \leq 2\alpha$, where $\|\hat{\Theta}\|_{\max} \leq \alpha$ is a constraint of the optimization problem (Equation 5) and $\|\Theta^*\|_{\max} \leq \alpha$ follows from our assumption that $\|X\|_{\max}^2 \leq \alpha$:

$$\begin{aligned}\max_{ij}|\Theta^*_{ij}| &= \max_{ij}\left|\frac{1}{m}\sum_{k=1}^m X_{ki}X_{kj}\right| \\ &\leq \max_{ij}\frac{1}{m}\sum_{k=1}^m |X_{ki}||X_{kj}| \\ &\leq \frac{1}{m}\sum_{k=1}^m \alpha.\end{aligned}$$

Then, given these lemmas, we can prove the result as follows: first, by the optimality of $\hat{\Theta}$, we have

$$\begin{aligned}
0 &\geq \mathcal{L}(\Theta^*+\Delta)-\mathcal{L}(\Theta^*)+\lambda(\|\Theta^*+\Delta\|_{\text{nuc}}-\|\Theta^*\|_{\text{nuc}}) \\
&\overset{(i)}{\geq} \boxed{\langle\nabla\mathcal{L}(\Theta^*),\Delta\rangle + \frac{1}{d^2}\|\Delta\|_F^2 - c_1\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - c_2\alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m}} \\
&+ \lambda(\|\Theta^*+\Delta\|_{\text{nuc}}-\|\Theta^*\|_{\text{nuc}}) \\
&\overset{(ii)}{\geq} \boxed{-\|\nabla\mathcal{L}(\Theta^*)\|_{\text{op}}\|\Delta\|_{\text{nuc}}} + \frac{1}{d^2}\|\Delta\|_F^2 - c_1\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - c_2\alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m} \\
&+ \lambda(\|\Theta^*+\Delta\|_{\text{nuc}}-\|\Theta^*\|_{\text{nuc}}) \\
&\overset{(iii)}{\geq} \boxed{-\frac{\lambda}{2}\|\Delta\|_{\text{nuc}}} + \frac{1}{d^2}\|\Delta\|_F^2 - c_1\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - c_2\alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m} \\
&+ \lambda(\|\Theta^*+\Delta\|_{\text{nuc}}-\|\Theta^*\|_{\text{nuc}}),
\end{aligned}$$

where (i) follows from Lemma A.7 (restricted strong convexity), (ii) from Holder's inequality, and (iii) from our setting of λ and Lemma A.6 (operator norm bound). Next, we can combine terms involving λ and use our bound on the nuclear norm (Lemma A.8), producing

$$\begin{aligned}
&\overset{(i)}{\geq} -\frac{\lambda}{2}\|\Delta\|_{\text{nuc}} + \frac{1}{d^2}\|\Delta\|_F^2 - c_1\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - c_2\alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m} + \boxed{\lambda(-\|\Delta\|_{\text{nuc}})} \\
&= \frac{1}{d^2}\|\Delta\|_F^2 - c_1\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - c_2\alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m} - \boxed{\frac{3}{2}\lambda\|\Delta\|_{\text{nuc}}} \\
&\overset{(ii)}{\geq} \frac{1}{d^2}\|\Delta\|_F^2 - c_1\alpha\|\Delta\|_{\text{nuc}}\sqrt{\frac{\log d}{dm}} - c_2\alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m} - c_4\|\Delta\|_{\text{nuc}}\boxed{\alpha\sqrt{\frac{\log d+\delta}{dm}}} \\
&\overset{(iii)}{\geq} \frac{1}{d^2}\|\Delta\|_F^2 - c_1\alpha\boxed{\sqrt{r}\|\Delta\|_F}\sqrt{\frac{\log d}{dm}} - c_2\alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}} - c_3\alpha^2\frac{\delta}{m} - c_4\boxed{\sqrt{r}\|\Delta\|_F}\alpha\sqrt{\frac{\log d+\delta}{dm}},
\end{aligned}$$

where (i) follows from reverse triangle, (ii) from our setting of λ, and (iii) from Lemma A.8 (nuclear norm bound). At this point, we can rearrange to produce the bound

$$\frac{1}{d^2}\|\Delta\|_F^2 \lesssim \max\left(\alpha\sqrt{r}\|\Delta\|_F\sqrt{\frac{\log d}{dm}},\ \alpha\|\Delta\|_F\sqrt{\frac{\delta}{dm}},\ \alpha^2\frac{\delta}{m},\ \sqrt{r}\|\Delta\|_F\alpha\sqrt{\frac{\log d+\delta}{dm}}\right),$$

which we can simplify into

$$\frac{1}{d}\|\Delta\|_F \lesssim \max\left(\alpha\sqrt{\frac{rd\log d}{m}},\ \alpha\sqrt{\frac{d\delta}{m}},\ \alpha\sqrt{\frac{\delta}{m}},\ \alpha\sqrt{\frac{rd(\log d+\delta)}{m}}\right)$$
$$\lesssim \alpha\sqrt{\frac{rd(\log d+\delta)}{m}},$$

completing the proof. □

B. Experiments and hyperparameters

Experiments were run on TITAN RTX and RTX 3090 GPUs with 24 gigabytes of memory; in all experiments we set the random seed to zero. Optimization was done via Adam (Kingma & Ba, 2015) with lr $= 1\text{e}{-10},\ \beta = (0.9, 0.999)$, and 10,000 steps.