Matrix Completion in Almost-Verification Time

Jonathan A. Kelner* Jerry Li† Allen Liu‡ Aaron Sidford§ Kevin Tian¶

Abstract

We give a new framework for solving the fundamental problem of low-rank matrix completion, i.e., approximating a rank-r matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ (where $m \geq n$) from random observations. First, we provide an algorithm which completes $\mathbf{M}$ on 99% of rows and columns under no further assumptions on $\mathbf{M}$ from $\approx mr$ samples and using $\approx mr^2$ time. Then, assuming the row and column spans of $\mathbf{M}$ satisfy additional regularity properties, we show how to boost this partial completion guarantee to a full matrix completion algorithm by aggregating solutions to regression problems involving the observations.

In the well-studied setting where $\mathbf{M}$ has incoherent row and column spans, our algorithms complete $\mathbf{M}$ to high precision from $mr^{2+o(1)}$ observations in $mr^{3+o(1)}$ time (omitting logarithmic factors in problem parameters), improving upon the prior state-of-the-art [JN15] which used $\approx mr^5$ samples and $\approx mr^7$ time. Under an assumption on the row and column spans of $\mathbf{M}$ we introduce (which is satisfied by random subspaces with high probability), our sample complexity improves to an almost information-theoretically optimal $mr^{1+o(1)}$, and our runtime improves to $mr^{2+o(1)}$. Our runtimes have the appealing property of matching the best known runtime to verify that a rank-r decomposition $\mathbf{U}\mathbf{V}^\top$ agrees with the sampled observations. We also provide robust variants of our algorithms that, given random observations from $\mathbf{M}+\mathbf{N}$ with $\|\mathbf{N}\|_{\mathrm{F}} \leq \Delta$, complete $\mathbf{M}$ to Frobenius norm distance $\approx r^{1.5}\Delta$ in the same runtimes as the noiseless setting. Prior noisy matrix completion algorithms [CP10] only guaranteed a distance of $\approx \sqrt{n}\Delta$.

*MIT, kelner@mit.edu. Supported in part by NSF awards CCF-1955217, CCF-1565235, and DMS-2022448.

†Microsoft Research, jerrl@microsoft.com.

‡MIT, cliu568@mit.edu. This work was partially done while working as an intern at Microsoft Research, and was supported in part by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Fellowship.

§Stanford University, sidford@stanford.edu. Supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research award, and a Sloan Research Fellowship.

¶Microsoft Research, tiankevin@microsoft.com.

Contents

1 Introduction

Matrix completion is a fundamental and well-studied problem in both the theory and practice of computer science, machine learning, operations research, and statistics. Broadly, the matrix completion problem asks to recover a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ from a small (i.e., sublinear) number of randomly revealed, and potentially noisy, entries. This problem was originally studied in the context of collaborative filtering [RS05] (see e.g., the Netflix challenge [SN07]) and has since found a myriad of applications in diverse settings such as signal processing [LLR95, SY07], genetics [ND14], social network analysis [MJGP19], and traffic engineering [GC12].

Structural assumptions. In the absence of additional assumptions, matrix completion is impossible. Unless there is structure among the entries of $\mathbf{M}$, then all of $\mathbf{M}$ must be revealed for recovery (as otherwise unrevealed entries can be arbitrary). Correspondingly, there has been a long line of work developing algorithms for matrix completion under different structural assumptions on $\mathbf{M}$. Perhaps the most prevalent and natural assumption placed on $\mathbf{M}$ is that it is low-rank. This assumption is well-motivated for the matrices arising in collaborative filtering or signal processing, for example, as discussed in [CR12]. Furthermore, rank-r matrices $\mathbf{M} \in \mathbb{R}^{m \times n}$ can be represented in $O((m+n)r)$-space simply by storing its rank-r factorization. Consequently, naïve parameter-counting arguments suggest it may be possible to recover $\mathbf{M}$ using $O((m+n)r)$ observations.

However, the assumption that $\mathbf{M}$ is low-rank alone is insufficient to enable algorithms for matrix completion that use $o(mn)$ observations. If $\mathbf{M}$ has a single non-zero entry, then it has rank-1, and yet $\Omega(mn)$ observations are required to recover the nonzero entry (and consequently $\mathbf{M}$) with constant probability. Correspondingly, works on low-rank matrix completion place different additional structural assumptions that preclude such sparse obstacles to solving the problem.

The setting where $\mathbf{M}$ has *incoherent* row and column spans is particularly well-studied [CR12]. A dimension-r subspace of $\mathbb{R}^d$ is μ-incoherent if no projection of a basis vector has squared norm more than $\frac{\mu r}{d}$, i.e., the subspace is well-spread over coordinates; we use "incoherent subspace" without a parameter to mean a $\widetilde{O}(1)$-incoherent subspace.[1] Letting $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a singular value decomposition (SVD) of $\mathbf{M}$, and assuming $\mathbf{U}, \mathbf{V}$ span incoherent subspaces (and an entrywise bound on $\mathbf{U}\mathbf{V}^\top$), [Rec11] refined results of [CT10, CR12, KMO10], and demonstrated that there are polynomial-time algorithms completing $\mathbf{M}$ from $\widetilde{O}((m+n)r)$ observations.

The parameters used in the definition of incoherence are motivated by the fact that they are satisfied with high probability by random rank-r matrices. Consequently, prior work showed that matrix completion is information-theoretically possible so long as the structure of $\mathbf{M}$ is "suitably-random." However, it is perhaps unclear whether incoherence is the correct or best notion of "suitably-random," aside from the post-hoc justification that it allows for efficient matrix completion.

Performance of matrix completion algorithms. Despite a plethora of work on matrix completion when e.g., $\mathbf{M}$ has incoherent row and column spans (discussed below and in greater detail in Section 1.2), many surprisingly fundamental algorithmic questions remain unresolved. A number of key open problems relate to the runtime and robustness of existing matrix completion algorithms.

The aforementioned works of [CT10, CR12, Rec11] developed polynomial-time algorithms for completing a rank-r matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ with $\widetilde{O}(1)$-incoherent row and column spans from a near-optimal number of observations. These algorithms were based on semidefinite programming (SDP) for nuclear norm minimzation. The runtimes of state-of-the-art SDP solvers [JKL+20, HJS+22] have a substantial polynomial overhead over the number of observations, inhibiting their practical

[1] Throughout $\widetilde{O}$ hides polylogarithmic factors in m, n, the inverse failure probability, and the relative accuracy.

application. Motivated by this shortcoming, another line of work [KMO10, Har14, JN15, YPCC16] developed iterative first-order methods, based on alternating minimization or gradient descent, whose runtimes depend linearly on the dimension $\max(m, n)$. However, the state-of-the-art algorithms with such runtime guarantees still incur fairly substantial overheads in problem parameters. Prior to our work, the best runtime for incoherent low-rank matrix completion was by [JN15], whose algorithm ran in time $\widetilde{O}((m+n)r^7)$.[2] A contemporaneous work of [YPCC16] yielded an incomparable runtime of $\widetilde{O}((m+n)r^4\kappa^5)$, where κ is the multiplicative range of $\mathbf{M}$'s singular values.

Another parameterization of the performance of matrix completion algorithms, which is rife with open problems, is the degree to which they can handle noise in the observations. In the setting where $\mathbf{M}$ is low-rank and has incoherent row and column spans, suppose that instead of observing random entries of $\mathbf{M}$, the observations we see are of $\mathbf{M}+\mathbf{N}$ for a noise matrix $\mathbf{N}$ satisfying $\|\mathbf{N}\|_{\mathrm{F}} \leq \Delta$. We are unaware of any information-theoretic barriers to recovering a matrix $\widehat{\mathbf{M}}$ satisfying $\|\widehat{\mathbf{M}}-\mathbf{M}\|_{\mathrm{F}} = O(\Delta)$ with no further assumptions. However, state-of-the-art polynomial-time algorithms are only able to achieve a Frobenius norm recovery guarantee of $O(\sqrt{\min(m,n)}\Delta)$, which loses a dimension-dependent factor. While other matrix completion algorithms in the literature also demonstrate robustness to noise, their guarantees either require additional assumptions on the noise such as sparsity, e.g., [CGJ17, YPCC16], or break down for large Δ, e.g., [KMO09, GAGG13, Har14, HW14].

These open problems regarding the complexity of matrix completion give rise to the following key questions which motivate our work.

1. What type of matrix completion is possible when the only structure is a rank bound?
2. Are there alternative structural assumptions to incoherent subspaces which enable faster algorithms, improved sample complexities, and better noise tolerance?
3. Under the well-studied structural assumption of incoherent subspaces, to what extent can we improve upon the runtimes and error tolerance of existing matrix completion algorithms?

1.1 Our results

We provide a new algorithmic framework for matrix completion and technical tools that address the shortcomings raised by each of Questions 1, 2, and 3. The cornerstone of our framework is a new iterative method that answers Question 1 by obtaining (perhaps surprisingly) nontrivial matrix completion guarantees *with no structural assumptions beyond a rank bound.* We believe this result is of independent interest, and we state it first.

Partial matrix completion without structure. As already noted, fully completing low-rank $\mathbf{M}$ from partial observations is impossible without further assumptions due to the possibility of sparse, large entries. However, when $\mathbf{M}$ is low-rank, such entries are necessarily rare (see Lemma 6 for a formal statement) and thus one could still hope to recover a large portion of $\mathbf{M}$. We demonstrate this in the following theorem (where $\mathbf{M}_{S,T}$ denotes the submatrix indexed by $S \subseteq [m], T \subseteq [n]$).

Theorem 1 (informal, see Corollary 2)**.** *Let $m \geq n$,[3] let $\mathbf{M} \in \mathbb{R}^{m\times n}$ be rank-r, and let $\mathbf{N} \in \mathbb{R}^{m\times n}$ satisfy $\|\mathbf{N}\|_{\mathrm{F}} \leq \Delta$. There is an algorithm which, given $\widetilde{O}(m^{1+o(1)}r)$ random observations from*

[2] A more recent work [CGJ17] claims an improved runtime over [JN15]. However, to obtain this result [CGJ17] assumes a sublinear-time exact singular value decomposition subroutine (which does not currently exist), and it is unclear how to recover the runtime claim of the paper without such an assumption [Che22].

[3] All of our results handle $m \leq n$ symmetrically via transposition, so we often assume $m \geq n$ for ease of exposition.

$\mathbf{M} + \mathbf{N}$, *runs in time* $\widetilde{O}(m^{1+o(1)}r^2)$ *and, with high probability, outputs a rank-*$rm^{o(1)}$ *factorization of* $\widehat{\mathbf{M}} \in \mathbb{R}^{m\times n}$ *so that there exist* $S \subseteq [m]$ *and* $T \subseteq [n]$ *with* $|S| \geq 0.99m$, $|T| \geq 0.99n$, *and*

$$\left\| \left[\mathbf{M} - \widehat{\mathbf{M}}\right]_{S,T} \right\|_{\mathrm{F}} \leq \Delta.$$

In other words, on a very large subset of coordinates, Theorem 1 recovers $\mathbf{M}$ up to the optimal error threshold up to constants. Additionally, since $\approx mr$ samples are information-theoretically necessary to perform nontrivial (full) matrix completion [CT10], the sample complexity of Theorem 1 is almost-optimal. As a corollary, in the case when $\Delta = 0$, Theorem 1 shows that matrix completion can be solved exactly on all but 1% of rows and columns (assuming a bounded bit complexity).

The runtime stated in Theorem 1 has the appealing property that it is what we call *almost-verification time*. Consider the natural problem of verifying a rank-r factorization of $\mathbf{M}$, that is the problem of verifying that $\mathbf{U}\mathbf{V}^\top = \mathbf{M}$ on mr observed entries given an explicit rank-r factorization of $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, for $\mathbf{U} \in \mathbb{R}^{m\times r}, \mathbf{V} \in \mathbb{R}^{n\times r}$. The best known running time for this problem is $O(mr^2)$ (even when using fast multiplication). Up to subpolynomial factors, our runtime in Theorem 1 matches this natural bottleneck to improved runtimes for matrix completion.

The guarantees of Theorem 1 are to the best of our knowledge new, and seem particularly striking in light of the long history of matrix completion algorithms. It is worth noting that there has been work which broadly aims to complete a submatrix from observations. Perhaps the most closely-related result is due to recent, similarly-titled work of [KHK22], which studies a different notion of partial matrix completion. [KHK22] shows that if $\mathbf{M}$ is rank-r and has bounded entries, and the distribution of observed entries is supported on a subset $U \subseteq [m] \times [n]$, then one can recover $\mathbf{M}$ to constant average entrywise error on a subset of $[m] \times [n]$ with cardinality at least $|U|$ (for a suitable relaxed notion of average error). For instance, if the algorithm of [KHK22] is instantiated for $U = [m] \times [n]$, then it outputs $\widehat{\mathbf{M}}$ satisfying $\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\mathrm{F}}^2 \leq \epsilon\|\mathbf{M}\|_{\mathrm{F}}^2$ using $O((m+n)r\epsilon^{-2})$ observations. Notably, their complexity depends inverse-polynomially on the accuracy (and hence inhibits exact completion). In contrast, our Theorem 1 achieves exact completion (albeit only on a large submatrix), and works under the standard, i.i.d. observation model.

Matrix completion beyond incoherence. Equipped with our new partial matrix completion subroutine, we turn to Question 2 and ask under what structural assumptions we can leverage it to solve (full) matrix completion efficiently. Given the generality of our partial matrix completion algorithm, it is natural to ask whether we can first run partial completion, and then recover the matrix on the small subset of rows and columns on which our partial completion method fails.

When analyzing this iterative process of recovering rows and columns of the target matrix which were dropped by our partial completion method, the standard structural assumption of incoherence turns out to be a lossy notion of "suitably-random." Instead, we define a new structural assumption on subspaces which we call *subspace regularity*, that serves as a proxy for randomness.

Definition 1 (Regular subspace)**.** *We say a subspace* $V \subseteq \mathbb{R}^d$ *is* (α, β)-regular *if for all* αd*-sparse* $v \in \mathbb{R}^d$, $\|\mathbf{\Pi}_{V_\perp} v\|_2 \geq \beta \|v\|_2$.

Note that Definition 1 implies $\|\mathbf{\Pi}_V v\|_2 \leq (1-\beta^2)\|v\|_2$, a condition which bears resemblance to incoherence (by bounding the relative weight of any small set of coordinates in the subspace). Intuitively, Definition 1 imposes that the restriction of V to a sufficiently large set of coordinates is still well-conditioned (made formal by Lemma 2). We prove that uniformly random subspaces are (α, β)-regular for constant α, β, with exponentially small failure probability, in Appendix A. Subspace regularity is not directly comparable to incoherence without losing r factors in the parameter

settings (see Fact 2), because a $d \times r$ basis matrix for an incoherent subspace can be entirely supported on an $O(\frac{1}{r})$ fraction of rows. However, Definition 1 is naturally compatible with our partial matrix completion method: roughly speaking, we require that the non-dropped rows and columns (e.g., (S, T) in Theorem 1) are representative enough of the remaining matrix to recover dropped subsets. This representativeness is captured by the conditioning requirement in Definition 1. Our main (full) matrix completion result under subspace regularity is the following.

Theorem 2 (informal, see Corollary 3)**.** *Let $m \geq n$, let $\mathbf{M} \in \mathbb{R}^{m \times n}$ be rank-r and have $(\Omega(1), \Omega(1))$-regular row and column spans, and let $\mathbf{N} \in \mathbb{R}^{m \times n}$ satisfy $\|\mathbf{N}\|_{\mathrm{F}} \leq \Delta$. There is an algorithm which, given $\widetilde{O}(mr^{1+o(1)})$ random observations from $\mathbf{M} + \mathbf{N}$, runs in time $\widetilde{O}(mr^{2+o(1)})$ and with high probability outputs a rank-r factorization of $\widehat{\mathbf{M}} \in \mathbb{R}^{m \times n}$ so that*

$$\left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_{\mathrm{F}} = O\left(r^{1.5+o(1)} \cdot \Delta\right).$$

The sample complexity of Theorem 2 is optimal up to subpolynomial factors [CT10] in the noiseless case (captured by our result by taking $\Delta \to 0$); these subpolynomial factors arise due to iterate rank blowup issues discussed in Section 1.3. Moreover, the algorithm of Theorem 2 runs in almost-verification time for the number of observations. Finally, in the noisy case, $\Delta > 0$, the overhead of Theorem 2's recovery guarantee only scales with the rank r, as opposed to the prior state-of-the-art [CR12] whose overhead scaled polynomially with the problem's dimensionality.

Even under subspace regularity, the "fixing" step used to obtain Theorem 2 we briefly described is quite technically involved. One of the main difficulties is that after running partial matrix completion, we do not necessarily know which rows and columns S, T have been completed. Our fixing algorithm circumvents this issue by carefully finding a small set of rows and columns which approximately span the row and column space of $\mathbf{M}$ in a well-conditioned fashion, satisfying a "representative" condition we state in Definition 6. We then show that we can use these representative rows and columns, alongside held-out random observations of the matrix, to robustly recover the rows and columns that were incorrectly completed by the partial completion algorithm. Putting these pieces together yields a fixing algorithm which recovers the subsets our partial completion method is inaccurate on, but increases error by a $\mathrm{poly}(r)$ factor. By carefully interleaving this fixing operation with repeated applications of our partial completion iterative method, we geometrically decrease the error of our overall algorithm. We give a detailed overview of our approach in Section 1.3.

Matrix completion with incoherence. Finally, we return to Question 3, i.e., matrix completion under the well-studied assumption of incoherence. We demonstrate that a small modification of our algorithm in Theorem 2 implies an analogous result under incoherence. In light of Fact 2 (which converts a subspace incoherence bound into a regularity bound), this is immediate up to $\mathrm{poly}(r)$ losses in the sample complexity and runtime. We give a tighter characterization of the lossiness due to assuming incoherence by introducing Definition 4, which subsumes both subspace regularity and incoherence. Leveraging this characterization, our techniques imply the following result for incoherent matrix completion (losing a single r factor in runtime and samples over Theorem 2).

Corollary 1 (informal, see Corollary 4)**.** *Let $m \geq n$, let $\mathbf{M} \in \mathbb{R}^{m \times n}$ be rank-r and have $\widetilde{O}(1)$-incoherent row and column spans, and let $\mathbf{N} \in \mathbb{R}^{m \times n}$ satisfy $\|\mathbf{N}\|_{\mathrm{F}} \leq \Delta$. There is an algorithm which, given $\widetilde{O}(mr^{2+o(1)})$ random observations from $\mathbf{M} + \mathbf{N}$, runs in time $\widetilde{O}(mr^{3+o(1)})$ and with high probability outputs a rank-r factorization of $\widehat{\mathbf{M}} \in \mathbb{R}^{m \times n}$ so that*

$$\left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_{\mathrm{F}} = O\left(r^{1.5+o(1)} \cdot \Delta\right).$$

Even with this additional r factor overhead, our results compare favorably to existing work on incoherent matrix completion. As mentioned previously, the state-of-the-art runtime for incoherent matrix completion (with polylogarithmic dependence on problem conditioning) was $\widetilde{O}(mr^7)$ [JN15], which our Corollary 1 dramatically improves upon. While the sample complexity of Corollary 1 is a factor of r larger than the sample complexity required by matrix completion algorithms based on semidefinite programming, all incoherent matrix completion methods in the literature which run in time nearly-linear in $m = \max(m, n)$ use $\Omega(mr^2)$ observations (and often more), which we match up to subpolynomial factors. Additionally, none of the existing polynomial-time algorithms (even the slower semidefinite programming approaches!) were known to yield dimension-independent recovery guarantees for noisy incoherent matrix completion. We summarize how Corollary 1 compares to prior work on matrix completion under incoherence below.

Algorithm	Sample complexity	Runtime	Recovery error
[Rec11]	mr	$\Omega(m^\omega)$	N/A
[CP10]	mr	$\Omega(m^\omega)$	$\sqrt{n}\Delta$
[HW14]	mr^9	mr^{13}	$\star$
[SL16]	$mr^7\kappa^4$	$m^2r^6\kappa^4$	N/A
[JN15]	mr^5	mr^7	N/A
[YPCC16]	$mr^2\kappa^4$	$mr^4\kappa^5$	$\star$
Corollary 1	$mr^{2+o(1)}$	$mr^{3+o(1)}$	$r^{1.5+o(1)}\Delta$

Figure 1: Comparison of algorithms for completing rank-r $\mathbf{M} \in \mathbb{R}^{m\times n}$ with $\widetilde{O}(1)$-incoherent row and column spans, assuming $m \geq n$. We let Δ upper bound the (Frobenius norm) noise level, κ denote the multiplicative range of $\mathbf{M}$'s singular values, and hide polylogarithmic factors. For [Rec11, CP10], current SDP solvers with m constraints use $\Omega(m^\omega)$ time [JKL+20, HJS+22]. We use $\star$ to mean additional assumptions are made on the noise beyond a Frobenius norm bound.

1.2 Related work

The literature on matrix completion is vast and a full survey is beyond our scope. For conciseness, we only consider the most relevant work here. Much of the algorithmic work on matrix completion falls into three categories, two of which we have already discussed in some depth. First, there is work on solving matrix completion using SDPs such as nuclear norm minimization, e.g., [CT10, CP10, CR12, Rec11, DC20]. These algorithms typically attain strong statistical guarantees, but have superlinear runtimes in the problem dimensionality. Second, there is the line of work on formally analyzing nonconvex methods such as alternating minimization, e.g., [KMO10, Har14, HW14, JN15, ZWL15, SL16, CGJ17, ZW19]. While these achieve runtimes which are linear in the dimension of the problem, all prior results incurred large polynomial factors of r or other problem parameters in their runtime (and sometimes their sample complexity as well). We also remark that many of these papers consider notions of robust matrix completion, but tend to consider the setting where the noise matrix is sparse as opposed to norm-bounded, which is the setting we consider.

Finally, there is also the line of work on analyzing convex methods such as gradient descent for matrix completion. In many of those works, the objective is to demonstrate the more qualitative result that the optimization landscape for matrix completion has no spurious local minima [SQW15, DSRO15, GLM16, JKN16, ZDG18, ZCZ22]. Consequently, their quantitative guarantees tend to be somewhat loose compared to results using convex programming or nonconvex

methods. Additionally, because these methods are based on gradient descent, they tend to have runtimes which scale polynomially with the condition number of the underlying matrix. In contrast, our algorithms run in time which is polylogarithmic in the condition number. One notable exception is [ZCZ22]; however, this paper only proves local convergence results for their method.

1.3 Overview of approach

In this section, we overview the two main components of our matrix completion algorithms: our iterative method for partial matrix completion (given in Section 3) and our recovery algorithm for the missing row and column subsets which our iterative method fails to give guarantees on (given in Section 4). Throughout this discussion we let $\mathbf{M}^\star := \mathbb{R}^{n\times n}$ be a rank-$r^\star$ matrix which we wish to recover to disambiguate from iterates denoted as $\mathbf{M}$; we also let $m = n$ for simplicity. We delay discussion of the noise-robustness of our matrix completion algorithms to the end of the section.

1.3.1 Partial matrix completion

Short-flat decompositions. Our partial matrix completion algorithm is motivated by a recent approach to sparse recovery developed in [KLL+22]. This approach iteratively makes progress towards recovering a sparse target vector $x^\star$ by taking projected gradient steps. The key observation of [KLL+22] is that in the sparse recovery setting, the gradient of the least-squares objective is decomposable into an ℓ_2-bounded component (the signal direction towards $x^\star$) and an ℓ_∞-bounded component (the noise), termed a "short-flat decomposition." The algorithm of [KLL+22] carefully used truncation onto the set of sparse vectors (which enjoys a bounded ℓ_1-to-ℓ_2 ratio), along with the ℓ_1-ℓ_∞ Hölder's inequality, to bound how much the flat noise component inhibits progress.

We now give a first attempt at executing this strategy for matrix completion, noting that the set of low-rank matrices is a spectral analog of the set of sparse vectors. Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a current iterate, assume it is rank-$r^\star$ (for simplicity), and let $\Omega \subseteq [n] \times [n]$ be a uniformly random set of indices with $|\Omega| \approx pn^2$, where p is the observation probability. Suppose we are promised $\|\mathbf{M} - \mathbf{M}^\star\|_{\mathrm{F}} \le 1$. A natural descent step balancing the goals of making progress towards $\mathbf{M}^\star$ and maintaining that our iterate has low rank takes $\mathbf{D} \leftarrow [\mathbf{M}^\star - \mathbf{M}]_\Omega$ to be the observed difference matrix, lets $\mathbf{G}$ be the rank-$O(r^\star)$ truncation of the SVD of $\mathbf{D}$, and updates $\mathbf{M}' \leftarrow \mathbf{M} + \frac{\eta}{p}\mathbf{G}$ for an appropriate step size $\eta > 0$. If $\mathbf{D}$ sufficiently approximates $\mathbf{M}^\star - \mathbf{M}$ in the operator norm (up to $\approx (r^\star)^{-\frac{1}{2}}$), it is straightforward to adapt arguments of [KLL+22] to show that this step makes substantial progress in decreasing distance to $\mathbf{M}^\star$, e.g., $\|\mathbf{M}' - \mathbf{M}^\star\|_{\mathrm{F}} \le \frac{1}{2}$. The intuition for this argument is that

$$\frac{1}{p}\mathbf{D} = \underbrace{\mathbf{M}^\star - \mathbf{M}}_{:=\mathbf{X}} + \underbrace{\left(\frac{1}{p}[\mathbf{M}^\star - \mathbf{M}]_\Omega - (\mathbf{M}^\star - \mathbf{M})\right)}_{:=\mathbf{Y}}. \tag{1.1}$$

In this decomposition, note that $\mathbf{X}$ is low-rank and exactly in the signal direction $\mathbf{M}^\star - \mathbf{M}$, so if we could remove the influence of $\mathbf{Y}$ then the rank-$2r^\star$ truncation of $\mathbf{X}$ (indeed, even no truncation at all) would exactly take us towards $\mathbf{M}^\star$. Moreover, if we could bound the operator norm of the noise component $\mathbf{Y}$, then applying perturbation arguments such as Weyl's theorem shows that $\mathbf{Y}$ cannot affect the progress direction by too much after truncating $\mathbf{D}$'s SVD. Furthermore, assuming the random samples Ω are independently drawn,[4] it is straightforward to see that $\mathbf{Y}$ is mean-zero, so we can hope to control its operator norm using concentration bounds such as the matrix Bernstein inequality. This argument parallels the strategy of [KLL+22], where we may think of $\mathbf{X}$ as the short progress component and $\mathbf{Y}$ as the flat noise component (each in a singular value sense).

[4] We show how to lift this assumption by splitting samples and using them iteratively as holdouts in Lemma 1.

Bounding the difference matrix. Unfortunately, without further assumptions, the operator norm of $\mathbf{Y}$ may be too large. A hard example is when $\mathbf{M}^\star = uu^\top$ and $\mathbf{M} = vv^\top$ where u, v have entries in $\pm n^{-\frac{1}{2}}$ differing in only one coordinate. In this example, a randomly sampled $\mathbf{Y}$ (after debiasing via rescaling by the inverse sampling probability $\approx n$, as in (1.1)) will have constant rank and operator norm. A natural way to prove an operator norm bound on such a randomly sampled matrix is via the matrix Bernstein inequality, which shows that we obtain the desired bounds if the difference $\mathbf{M} - \mathbf{M}^\star$ has row and column norms bounded by $\approx n^{-\frac{1}{2}}$ and entries bounded by $\approx \sqrt{r^\star} \cdot n^{-1}$ (see Lemma 7); these conditions fail in our hard example as it has one row and column norm which is too large. Nevertheless, Markov's inequality shows that in general, only a constant fraction of rows and columns of the difference matrix can have norms which are too large; these subsets can then be estimated from observations and dropped (carried out in Section 3.1).

This leaves the issue of large entries, a second obstacle for our matrix Bernstein argument. With no assumptions on the row and column spans of $\mathbf{M}^\star$, it is possible that $\mathbf{M} - \mathbf{M}^\star$ has a few large entries missed by our random observations which can ruin our bound on $\mathbf{Y}$. We first show that due to the rank bound on $\mathbf{M} - \mathbf{M}^\star$, these large entries must be localized to small (unknown) subsets of rows and columns (Lemma 6). We then introduce a new measure of progress (Definition 2) where we say two matrices are close if their difference has small Frobenius norm on a large submatrix, which allows us to exclude these small unknown subsets with large entries. Finally, we are able to prove our iterative method makes progress in this modified notion of distance, and thus achieves partial completion. We give a complete statement of the guarantees of our partial matrix completion method in Proposition 1, and demonstrate how to use it recursively to obtain Theorem 1 in Section 3.3.

Mitigating rank blowup. One technical issue which arises in our partial completion method is that, roughly speaking, the rank of our iterate $\mathbf{M}$ increases by a constant factor in each iteration. Our earlier argument relied on a rank bound on $\mathbf{M}$, so this rank blowup is problematic. If our progress measure were $\|\mathbf{M} - \mathbf{M}^\star\|_{\mathrm{F}}$ (i.e., an exact distance bound), we could simply truncate the SVD of $\mathbf{M}$ to project it onto the set of low-rank matrices, which affects our progress by a constant factor. However, our guarantee is with respect to a modified notion of distance, so this does not hold. Instead, we show that we can make substantially more progress by taking slightly more samples, cutting the modified distance measure by a factor of $\approx \exp(\sqrt{\log(r^\star)}) = (r^\star)^{o(1)}$ in each iteration, so that in $\approx \sqrt{\log(r^\star)}$ iterations we have made a polynomial factor progress. This results in only a $(r^\star)^{o(1)}$ factor blowup in the rank of our iterate, and we then apply our fixing procedure (discussed next) to reduce the rank. For our self-contained partial completion result (Theorem 1), which is performed in one shot without a fixing step, the corresponding overhead is a factor of $n^{o(1)}$.

1.3.2 From partial completion to full completion

Finding a representative subset. Our distance measure in our partial completion algorithm (see e.g., Theorem 1) allows for the subsets on which we make progress to be unknown, but this causes issues when used for full completion. Indeed, our partial completion method made no assumptions about the regularity of $\mathbf{M}^\star$, but to recover dropped subsets (as well as subsets excluded by our distance measure) we need to impose structural assumptions. For simplicity in the following discussion, assume $\mathbf{M}^\star$ has $(\Omega(1), \Omega(1))$-regular row and column spans for appropriate constants (Definition 1). We also assume for simplicity that $\mathbf{M}$, the output of our partial completion method, satisfies $[\mathbf{M}]_{A,B} = [\mathbf{M}^\star]_{A,B}$ for $|A|, |B| \geq 0.99n$ exactly, i.e., we have run the partial completion method to high accuracy. Finally, we ignore the effect of explicitly dropped rows and columns, as these can be recovered analogously to the (unknown) excluded subsets in our distance measure.

Our high-level strategy is to identify a set T of $\approx r^\star$ columns of $\mathbf{M}$, such that $\mathbf{M} = \mathbf{M}^\star$ exactly

on these columns, and the column space of $\mathbf{M}^\star_{:T}$ spans the column space of $\mathbf{M}^\star_{:T}$. We call such a set T "representative" with respect to $(\mathbf{M}, \mathbf{M}^\star)$, defined formally in Definition 6 (which includes additional parameters when $\mathbf{M}_{A\times B}$ is only close to $\mathbf{M}^\star_{A\times B}$, rather than exactly equal). We begin with a preprocessing phase in Section 4.1, where we drop any rows and columns upon which we observe empirical errors. This guarantees that on the remaining submatrix, the difference matrix $\mathbf{M}-\mathbf{M}^\star$ has at most $\frac{0.01r^\star}{n}$ nonzero entries per row or column (else they would have been dropped).

We next provide a structural fact that any rank-$r^\star$ matrix with such bounded row and column sparsity must have all of its errors localized to a $1\% \times 1\%$ submatrix (see Lemma 13 for a formal statement which handles noise). In the noiseless case, this fact follows straightforwardly from a Gram-Schmidt argument (Lemma 14). This implies that a majority of the remaining columns of $\mathbf{M}$ and $\mathbf{M}^\star$ (after preprocessing) are actually identical, and are thus valid to include in a representative subset. We further develop a tester for verifying whether a given column $j \in [n]$ should be included in our representative subset, by drawing $\approx r^\star$ random columns of our iterate $\mathbf{M}$ and checking whether column $\mathbf{M}_{:j}$ is contained in the span of these random columns. This test is motivated by the observation that if $\mathbf{M}_{:j}$ contains a sparse error (and hence should not be included), with constant probability our random sample will dodge this error due to our preprocessing step, and hence $\mathbf{M}_{:j}$ will not be contained in its span. By repeating our tester a small number of times, we can ensure the subset of columns we include is representative.

Regression with a representative subset. Once we have determined a representative subset T, it suffices to use our regularity assumptions to argue that $\approx r^\star$ random observations of any column of $\mathbf{M}^\star$ uniquely determine how it can be completed as a linear combination of $\mathbf{M}_{:T} = \mathbf{M}^\star_{:T}$. In the noiseless case, this means that we can simply solve roughly n regression problems in $r^\star \times r^\star$ matrices to fully complete the matrix. Our formal definition of a representative subset contains a quantitative bound ensuring $\mathbf{M}^\star_{:T}$ spans the column space of $\mathbf{M}^\star$ in a well-conditioned manner. This allows for us to argue about the generalization error of our regression subroutines under noise.

We remark that if after our partial completion subroutine, we knew which row and column subsets A, B our iterate was close to $\mathbf{M}^\star$ on, we could directly skip to this regression step for recovering poorly-behaved subsets. Handling the potential of sparse errors on unknown subsets of our iterate in a noise-tolerant way constitutes the bulk of our technical development in Section 4.

1.3.3 Robust matrix completion

Finally, we discuss how our framework extends to the noisy setting in a natural way. In general, our fixing step in Section 4 takes as input $\mathbf{M}$ with the guarantee that $\mathbf{M}$ is $\widetilde{\Delta}$-close to $\mathbf{M}^\star$ on a submatrix (see Definition 2), after excluding an $\frac{\alpha}{2}$-fraction of rows and columns explicitly dropped by our iterative method, and an additional $\frac{\alpha}{2}$-fraction due to our distance measure (where α is a subspace regularity parameter). Assuming $\widetilde{\Delta}$ is sufficiently larger than $\|\mathbf{N}\|_{\mathrm{F}}$, where we receive observations from $\mathbf{M}^\star+\mathbf{N}$ (i.e. $\mathbf{N}$ is the noise), our fixing step learns any excluded rows and columns to a comparable distance to the average undropped row or column, and yields a standard distance guarantee (rather than a partial one). However, this stronger standard distance guarantee comes at the cost of a $\mathrm{poly}(r^\star)$ overhead over the initial distance promise $\widetilde{\Delta}$, and is stated formally in Proposition 3. This overhead is due to lossiness when converting between operator norm distance guarantees (which naturally arises in analyzing the generalization error of our regression step), and Frobenius norm distance guarantees (which our iterative method yields). By ensuring that all matrices encountered throughout are low-rank, this lossiness only has $r^\star$-dependent factors.

Our robust matrix completion results stated in Theorem 2 and Corollary 1 follow by applying the guarantees of Proposition 1 (our partial matrix completion algorithm) and Proposition 3 (our

fixing step) recursively. By running Proposition 1 for a small number of steps to control the blowup of our iterate's rank, and applying Proposition 3 to reduce the rank and recover dropped subsets, we can make multiplicative distance progress towards our noise threshold $\Delta \geq \|\mathbf{N}\|_{\mathrm{F}}$. The error overhead incurred by our algorithms is then due to a final application of Proposition 3.

2 Preliminaries

General notation. Throughout $[n] := \{i \in \mathbb{N} \mid i \leq n\}$. When $S \subseteq T$ and T is clear from context, we let $S^c := T \setminus S$. We say $v \in \mathbb{R}^d$ is s-sparse if it has at most s nonzero entries. Applied to a vector, $\|\cdot\|_p$ is the ℓ_p norm. The Frobenius, operator, and trace norms of a matrix are denoted $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_{\mathrm{tr}}$ and correspond to the 2-norm, ∞-norm, and 1-norm of the singular values of a matrix. The all-zeroes and all-ones vectors of dimension d are denoted $\mathbb{0}_d$ and $\mathbb{1}_d$.

Matrices. Matrices are denoted in boldface. We equip $\mathbb{R}^{m\times n}$ with the inner product $\langle \mathbf{A}, \mathbf{B}\rangle := \mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$. The $d \times d$ identity matrix is denoted $\mathbf{I}_d$, and the all-zero $m \times n$ matrix is denoted $\mathbf{0}_{m\times n}$. The ordered singular values of $\mathbf{M} \in \mathbb{R}^{m\times n}$ with $m \geq n$ are denoted $\{\sigma_i(\mathbf{M})\}_{i\in[n]}$, where σ_1 is largest and σ_n is smallest; when $\mathbf{M} \in \mathbb{R}^{d\times d}$ is symmetric, we similarly define $\{\lambda_i(\mathbf{M})\}_{i\in[d]}$. When i is larger than the rank of $\mathbf{M}$, $\sigma_i(\mathbf{M}) := 0$. The number of nonzero entries of $\mathbf{M}$ is denoted $\mathrm{nnz}(\mathbf{M})$, and the largest absolute value among its entries is denoted $\|\mathbf{M}\|_{\max}$. For $\tau \geq 0$, $\mathbf{M} \in \mathbb{R}^{m\times n}$, we let $\mathbf{M}^{\leq\tau}$ be such that $\mathbf{M}^{\leq\tau}_{ij}$ is the median of $-\tau$,τ, and $\mathbf{M}_{ij}$. We say $\mathbf{M} \in \mathbb{R}^{m\times n}$ is given as a rank-r factorization if we have explicit access to $\mathbf{U} \in \mathbb{R}^{m\times r}$, $\mathbf{V} \in \mathbb{R}^{n\times r}$ with $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$. For symmetric positive semidefinite $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ we use $\mathbf{A} \approx_\epsilon \mathbf{B}$ to denote $\exp(-\epsilon)\mathbf{B} \preceq \mathbf{A} \preceq \exp(\epsilon)\mathbf{B}$. When $\mathbf{A}$ is symmetric positive definite we let $\kappa(\mathbf{A})$ be the ratio of its largest and smallest eigenvalues. We define $\mathcal{T}_{\mathrm{mv}}(\mathbf{M})$ as the amount of time it takes to compute $\mathbf{M}v$ for any v; note $\mathcal{T}_{\mathrm{mv}}(\mathbf{M}) = O(\mathrm{nnz}(\mathbf{M}))$, and if $\mathbf{M} \in \mathbb{R}^{m\times n}$ is given as a rank-r factorization then $\mathcal{T}_{\mathrm{mv}}(\mathbf{M}) = O((m+n)r)$.

Submatrices. For $\mathbf{M} \in \mathbb{R}^{m\times n}$ and subsets $S \subseteq [m]$, $T \subseteq [n]$, the matrix $\mathbf{M}_{S,T}$ denotes the $|S|\times|T|$ submatrix of $\mathbf{M}$ restricted to rows S and columns T. When $A = \{i\}$ for $i \in [m]$, we abbreviate this as $\mathbf{M}_{i,B}$, and similarly define $\mathbf{M}_{A,j}$ for $j \in [n]$. For $\mathbf{M} \in \mathbb{R}^{m\times n}$ we write $\mathbf{M}_{A:}$ as shorthand for $\mathbf{M}_{A,[n]}$ and $\mathbf{M}_{:B}$ for $\mathbf{M}_{[m],B}$. The i^{th} row and j^{th} column of $\mathbf{M}$ are similarly denoted $\mathbf{M}_{i:}$ and $\mathbf{M}_{:j}$. When dimensions are clear, the matrix which is all-zeroes except for a one in the $(i,j)^{\text{th}}$ entry is $\mathbf{E}_{ij}$ and e_i is the i^{th} standard basis vector. We say that $\mathbf{N}$ is a γ-submatrix of $\mathbf{M} \in \mathbb{R}^{m\times n}$ if $\mathbf{N} = \mathbf{M}_{S,T}$ for $S \subseteq [m]$, $T \subseteq [n]$ with $|S| \geq m - \gamma\min(m,n)$ and $|T| \geq n - \gamma\min(m,n)$. We say that $\mathbf{M}$ is s-row column sparse (RCS) if each row and column of $\mathbf{M}$ has as most s nonzero entries. When $\Omega \subseteq [m] \times [n]$ is a set of index pairs, $\mathbf{M}_\Omega$ zeroes out all entries in $\mathbf{M}$ indexed by Ω^c (we similarly define v_Ω for vectors $v \in \mathbb{R}^d$ and $\Omega \subseteq [d]$).

Comparing matrices. We introduce two nonstandard notions of closeness between matrices. These notions will be used primarily in stating the guarantees of our subroutines in Sections 3 and 4 respectively, to deal with subsets or sparse error patterns out of our control.

Definition 2 (Closeness on a submatrix)**.** *We say $\mathbf{M}, \mathbf{M}' \in \mathbb{R}^{m\times n}$ are Δ-close on a γ-submatrix if there exist subsets $A \subseteq [m]$, $B \subseteq [n]$ satisfying $|A| \geq m - \gamma\min(m,n)$, $|B| \geq n - \gamma\min(m,n)$, and*

$$\left\|\left[\mathbf{M} - \mathbf{M}'\right]_{A,B}\right\|_{\mathrm{F}} \leq \Delta.$$

Definition 3 (Closeness away from an RCS matrix)**.** *We say $\mathbf{M}, \mathbf{M}' \in \mathbb{R}^{m\times n}$ are Δ-close away from an s-RCS matrix if $\mathbf{M} - \mathbf{M}' = \mathbf{X} + \mathbf{Y}$, for some $\|\mathbf{X}\|_{\mathrm{F}} \leq \Delta$, and s-RCS $\mathbf{Y}$.*

We note that in Definition 2, the sets A, B are unknown; similarly, in Definition 3, the factorization $\mathbf{X}$, $\mathbf{Y}$ is unknown. Our analysis will only use these definitions as existential statements.

Observation model. For $\mathbf{M} \in \mathbb{R}^{m\times n}$, we specialize the notation $\mathbf{M}_\Omega \leftarrow \mathcal{O}_p(\mathbf{M})$ to mean $\Omega \subset [m] \times [n]$ contains each $(i,j) \in [m] \times [n]$ with probability p (sampled independently), and $\mathbf{M}_\Omega$ is the sum of the observations $\mathbf{M}_{ij}\mathbf{E}_{ij}$ for $(i,j) \in \Omega$. When an algorithm requires the ability to query $\mathbf{M} \in \mathbb{R}^{m\times n}$ with $\mathcal{O}_p$ for various p (specified in the algorithm description), we list the input as $\mathcal{O}_{[0,1]}(\mathbf{M})$, which also gives access to $\mathcal{O}_{[0,1]}(\mathbf{M}_{S,T})$ for $S \subseteq [m]$, $T \subseteq [n]$.

We note that this observation model (querying $\mathcal{O}_p$, possibly multiple times independently) is compatible with the standard model in the literature (which only allows for a one-shot set of realized observations), up to a small loss in parameters. This is made formal through the following lemma, which shows how to simulate K draws from $\mathcal{O}_p$ given one-time access to $\mathcal{O}_{Kp}$.

Lemma 1. *Let $\{p_k\}_{k\in[K]} \in (0,1)$ satisfy $p_k \le p \le \frac{1}{K}$ for all $k \in [K]$, and let $\mathbf{M} \in \mathbb{R}^{m\times n}$. We can simulate sequential access to $\mathcal{O}_{p_k}(\mathbf{M})$ for all $k \in [K]$ with access to $\mathcal{O}_{Kp}(\mathbf{M})$.*

Proof. The probability that the entry is revealed in any of the independent, sequential queries is

$$p_{\text{tot}} := 1 - \prod_{k\in[K]} (1-p_k) \le Kp.$$

The conclusion then follows from two observations. First, letting $q \ge p$ satisfy $1-(1-q)^K = Kp$, if $\mathcal{O}_{Kp}$ reveals an entry we can efficiently simulate how many of K calls to $\mathcal{O}_q$ would have revealed that entry conditioned on at least one call resulting in a reveal. Second, given access to $\mathcal{O}_q$ we can simulate $\mathcal{O}_{p_k}$ for any $p_k \le q$ by rejecting a revealed entry with the appropriate probability. □

In other words, Lemma 1 allows us to draw observations from a matrix a single time, and then split the samples in a way that simulates multiple sequential accesses to the matrix.

Subspaces. For a subspace $V \subseteq \mathbb{R}^d$ of dimension r, we denote its orthogonal complement by $V_\perp$. We let $\mathbf{\Pi}_V \in \mathbb{R}^{d\times d}$ be the projection matrix onto V. We let $\mathbf{B}_V \in \mathbb{R}^{d\times r}$ denote an arbitrary matrix satisfying $\mathbf{B}_V\mathbf{B}_V^\top = \mathbf{\Pi}_V$ and $\mathbf{B}_V^\top\mathbf{B}_V = \mathbf{I}_r$. We say $\mathbf{U\Sigma V}^\top$ is the singular value decomposition (SVD) of $\mathbf{M}$ if $\mathbf{U}, \mathbf{V}$ have orthonormal columns and $\mathbf{\Sigma}$ is nonnegative and diagonal; when this is not unique, we take an arbitrary SVD. We recall our definition of a regular subspace in Definition 1. We will mainly use this definition through the following equivalence.

Lemma 2. *Let $V \subseteq \mathbb{R}^d$ have dimension r, and let $\{b_i\}_{i\in[d]} \subset \mathbb{R}^r$ be rows of an (arbitrary) choice of $\mathbf{B}_V$. V is (α,β)-regular if and only if for every $S \subseteq [d]$ with $|S| \ge (1-\alpha)d$,*

$$\beta^2\mathbf{I}_r \preceq \sum_{i\in S} b_i b_i^\top \preceq \mathbf{I}_r.$$

Proof. First observe that $\|\mathbf{\Pi}_V v\|_2^2 + \|\mathbf{\Pi}_{V_\perp} v\|_2^2 = \|v\|_2^2$ and $\|\mathbf{\Pi}_V v\|_2^2 = v^\top\mathbf{\Pi}_V v = \|\mathbf{B}_V v\|_2^2$ for all $v \in \mathbb{R}^d$. Consequently, V is (α,β) regular if and only if $\|\sum_{i\in[d]} b_i v_i\|_2^2 \le (1-\beta^2)\|v\|_2^2$, for all αd-sparse $v \in \mathbb{R}^d$. This is equivalent to the condition that for all $T \subseteq [d]$ with $|T| \le \alpha d$ and (not necessarily sparse) $v \in \mathbb{R}^d$, $\|\sum_{i\in T} b_i v_i\|_2^2 \le (1-\beta^2)\sum_{i\in T} v_i^2$. Equivalently, for every $T \subseteq [d]$ with $|T| \le \alpha d$, the matrix $\mathbf{B}_{T:}$ must have operator norm $\le \sqrt{1-\beta^2}$, so $\sum_{i\in T} b_i b_i^\top \preceq (1-\beta^2)\mathbf{I}_r$. Since $\sum_{i\in S} b_i b_i^\top = \mathbf{I}_r - \sum_{i\in S^c} b_i b_i^\top$ and $\sum_{i\in[d]} b_i b_i^\top = \mathbf{I}_r$, the result follows. □

We also introduce a notion of a standard subspace in Definition 4, which is more compatible with the aforementioned incoherence assumption in the matrix completion literature. This definition is used to streamline the application of the tools from Section 4.

Definition 4 (Standard subspace)**.** *We say a subspace $V \subseteq \mathbb{R}^d$ of dimension r is* (α, β, μ)-standard *if it is (α, β)-regular and there exists a subset $S \subseteq [d]$ with $|S| \geq (1 - \frac{\alpha}{3})d$ such that for all $i \in S$, $\|\mathbf{\Pi}_V e_i\|_2 \leq \sqrt{\frac{\mu r}{d}}$.*

The following fact is immediate by Markov's inequality, $\|\mathbf{\Pi}_V e_i\|_2 = \|\mathbf{B}_V e_i\|_2$, and $\|\mathbf{B}_V\|_{\mathrm{F}}^2 = r$.

Fact 1. *If a subspace $V \subseteq \mathbb{R}^d$ is (α, β)-regular, then it is $(\alpha, \beta, \frac{3}{\alpha})$-standard.*

Thus, whenever we mention a subspace being (α, β, μ)-standard, we may assume $\mu \leq \frac{3}{\alpha}$. Finally, for comparison to the matrix completion literature, we also give the definition of incoherence which is typically used to parameterize algorithms.

Definition 5 (Incohererent subspace)**.** *We say a subspace $V \subseteq \mathbb{R}^d$ of dimension r is* μ-incoherent *if $\|\mathbf{\Pi}_V e_i\|_2 \leq \sqrt{\frac{\mu r}{d}}$ for all $i \in [d]$.*

The following is then immediate from the characterization in Lemma 2.

Fact 2. *If a subspace $V \subseteq \mathbb{R}^d$ is μ-incoherent, it is $(\frac{3}{4\mu r}, \frac{1}{2}, \mu)$-standard.*

Proof. Note that $\|\sum_{i \in S^c} b_i b_i^\top\|_{\mathrm{op}} \leq |S^c| \max_{i \in S^c} \|b_i\|_2^2$ and apply Weyl's perturbation theorem. □

We introduce the notion of a standard subspace primarily for technical convenience as it captures the parameters of both subspace regularity and incoherence. We will prove a result (Theorem 3) in terms of all of these parameters α, β, μ and then deduce our results for subspace regularity and incoherence by combining Theorem 3 with Fact 1 and Fact 2 respectively.

Concentration. We use the following concentration inequalities and their scalar specializations.

Fact 3 (Matrix Chernoff, Theorem 5.1.1 [Tro15])**.** *Let $\{\mathbf{X}_i\}_{i \in [n]}$ be independent, $d \times d$ positive semidefinite, matrix-valued random variables satisfying $\|\mathbf{X}_i\|_{\mathrm{op}} \leq R$ with probability 1 for all $i \in [n]$, and let $\mathbf{X}$ denote their sum. For any $\epsilon \in (0, 1)$,*

$$\Pr\left[\lambda_{\min}(\mathbf{X}) \leq (1-\epsilon)\lambda_{\min}(\mathbb{E}\mathbf{X})\right] \leq d \exp\left(-\frac{\epsilon^2 \lambda_{\min}(\mathbb{E}\mathbf{X})}{3R}\right),$$
$$\Pr\left[\lambda_{\max}(\mathbf{X}) \geq (1+\epsilon)\lambda_{\max}(\mathbb{E}\mathbf{X})\right] \leq d \exp\left(-\frac{\epsilon^2 \lambda_{\max}(\mathbb{E}\mathbf{X})}{3R}\right).$$

Fact 4 (Matrix Bernstein, Theorem 1.6.2 [Tro15])**.** *Let $\{\mathbf{X}_i\}_{i \in [n]}$ be independent, $d_1 \times d_2$ matrix-valued random variables satisfying $\mathbb{E}\mathbf{X}_i = \mathbf{0}_{d_1 \times d_2}$ and $\|\mathbf{X}_i\|_{\mathrm{op}} \leq R$ with probability 1 for all $i \in [n]$, let $\mathbf{X}$ denote their sum, and let*

$$\sigma^2 := \max\left(\left\|\sum_{i \in [n]} \mathbb{E}\mathbf{X}_i \mathbf{X}_i^\top\right\|_{\mathrm{op}}, \left\|\sum_{i \in [n]} \mathbb{E}\mathbf{X}_i^\top \mathbf{X}_i\right\|_{\mathrm{op}}\right).$$

Then for all $t \geq 0$, $\Pr[\|\mathbf{X}\|_{\mathrm{op}} \geq t] \leq (d_1 + d_2) \exp\left(-\frac{t^2}{2\sigma^2 + \frac{2}{3}Rt}\right)$, so for all $\delta \in (0, 1)$,

$$\Pr\left[\|\mathbf{X}\|_{\mathrm{op}} \geq \max\left(2\sigma\sqrt{\log\left(\frac{d_1+d_2}{\delta}\right)}, \frac{4R}{3}\log\left(\frac{d_1+d_2}{\delta}\right)\right)\right] \leq \delta.$$

3 Partial matrix completion

In this section, we give a novel subroutine for making partial progress towards a target low-rank matrix $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ (whose rank is denoted $r^\star$), from which we can query noisy observations. In particular, the method we develop in this section only assumes the target matrix is low-rank, without any requirement of subspace regularity in the vein of Definition 1. However, our guarantees are with respect to a weaker notion of progress, which involves explicitly dropping or excluding a small number of poorly-behaved rows and columns.

The main result of this section is the following Proposition 1, which gives a guarantee on Algorithm 2 (which builds upon Algorithm 1, a preprocessing subroutine which we explain shortly). Our Algorithm 2 takes as parameters γ_{drop} and γ_{add}, as well as a matrix $\mathbf{M}$ which is Δ-close to $\mathbf{M}^\star$ on a γ-submatrix. It then explicitly drops roughly a γ fraction of rows and columns which it makes no guarantees on, adds γ_{add} to the submatrix parameter, and triples the rank of $\mathbf{M}$. In return, it cuts the distance on a $(\gamma+\gamma_{\text{add}})$-submatrix by a factor of ℓ.

Proposition 1. *Let $\Delta \geq 0$, $\gamma, \gamma_{\text{add}}, \delta \in (0,1)$, and $\ell \geq 1$. Let $\widehat{\mathbf{M}} := \mathbf{M}^\star + \mathbf{N} \in \mathbb{R}^{m\times n}$ for $m \geq n$, $\mathbf{M}^\star$ which is rank-$r^\star$, and $\mathbf{N}$ satisfying $\|\mathbf{N}\|_{\text{F}} \leq \frac{\Delta}{20\ell}$. If rank-$r$ $\mathbf{M} \in \mathbb{R}^{m\times n}$ is Δ-close to $\mathbf{M}^\star$ on a γ-submatrix and given as a rank-r factorization, Algorithm 2 returns $\widetilde{\mathbf{M}} \in \mathbb{R}^{m\times n}$ as a rank-$3(r+r^\star)$ factorization and $S \subseteq [m]$, $T \subseteq [n]$ satisfying the following with probability $\geq 1-\delta$.*

1. *$|S| \geq m - \gamma_{\text{drop}} n$, $|T| \geq (1-\gamma_{\text{drop}})n$, for $\gamma_{\text{drop}} = \max(400\gamma\log(m),\ 10^5\ell^2(\gamma+\gamma_{\text{add}}))$.*
2. *$\widetilde{\mathbf{M}}_{S,T}$ is $\frac{\Delta}{\ell}$-close to $\mathbf{M}^\star_{S,T}$ on a $(\gamma+\gamma_{\text{add}})$-submatrix.*

Algorithm 2 uses $O(mnp(r+r^\star))$ time and one call to $\mathcal{O}_p(\widehat{\mathbf{M}})$ where for a sufficiently large constant,

$$p = O\left(\frac{(r+r^\star)\ell^2}{n}\cdot\frac{\gamma+\gamma_{\text{add}}}{\gamma_{\text{add}}^2}\log^2\left(\frac{m}{\delta}\right)\right).$$

In Section 3.1, we begin by analyzing Algorithm 1 (Filter), a preprocessing step for setting aside roughly a γ_{add} fraction of poorly-behaved rows and columns from empirical observations. In Section 3.2, we then use the control that this preprocessing step affords over the remaining rows and columns to analyze our main iterative step, Algorithm 2 (Descent), and prove Proposition 1. Finally, to illustrate a typical use case of Proposition 1 for partial matrix completion (which reflects its use in our final algorithm), we give a self-contained result in Section 3.3 only relying on recursive use of Algorithm 1, without the use of subspace regularity assumptions.

3.1 Row and column removal

The first step is to remove some rows and columns whose norm in $\widehat{\mathbf{M}}-\mathbf{M}$ is too large. This is useful because we would like to use $\widehat{\mathbf{M}}-\mathbf{M}$ to guide the direction of our steps but we only have partial observations of it. The rows and columns with large norms can ruin the spectral concentration of the empirical observations, so removing them allows us to prove spectral closeness between the empirical and true difference matrices. Before analyzing our removal algorithm, we state a simple concentration inequality we will use in its proof about the error of empirical norm estimates.

Lemma 3. *Let $p, \delta \in (0,1)$, let $v \in \mathbb{R}^d$ have $\|v\|_\infty \leq \tau$ and let $\tilde{v} \in \mathbb{R}^d$ have each entry $\tilde{v}_i$ independently set to v_i with probability p, and 0 otherwise. Then with probability $\geq 1-\delta$,*

$$\left|\|v\|_2^2 - \frac{1}{p}\|\tilde{v}\|_2^2\right| \leq \max\left(\frac{1}{10}\|v\|_2^2,\ \frac{30\tau^2\log\frac{2}{\delta}}{p}\right).$$

Proof. By Fact 3 with (scalar) $x_i \leftarrow \frac{1}{p}\tilde{v}_i^2$, so $\mathbb{E}\sum_{i\in[d]} x_i = \|v\|_2^2$, with probability $\geq 1-\delta$,

$$\left|\|v\|_2^2 - \frac{1}{p}\|\tilde{v}\|_2^2\right| \leq \frac{\tau}{\sqrt{p}}\|v\|_2\sqrt{3\log\frac{2}{\delta}}.$$

The conclusion follows depending on which of $\frac{1}{\sqrt{10}}\|v\|_2$ or $\frac{\tau}{\sqrt{p}}\sqrt{30\log\frac{2}{\delta}}$ is larger. □

We are now ready to state and analyze our removal process, which for logarithmically many iterations simply drops the largest rows and columns of the difference matrix, estimated from empirical observations. Our analysis proceeds in two phases. The goal of the first phase is to decrease the Frobenius norm of the true difference matrix until it is below a certain threshold, which we argue we continually make progress by concentration of the empirical observations. The second phase applies Markov's inequality to bound the number of large rows and columns once the Frobenius norm is below this threshold.

We remark that the assumed upper bound on τ in the following statement is for convenience in simplifying logarithmic terms and is not saturated in our eventual parameter settings (whereas the ρ bound reflects its eventual setting). Further, the parameter γ_{add} will eventually be set to be sufficiently small when iterating upon our algorithm, as it reflects the growth of the number of rows and columns we do not make guarantees on. To build intuiton (following discussion in Section 1.3), the reader may think of $\gamma, \gamma_{\text{add}}$ as small constants, $\tau \approx \Delta\cdot\frac{\sqrt{r}}{n}$, $\rho \approx \Delta\cdot\frac{1}{\sqrt{n}}$, and $p \approx \frac{r}{n}$.

Lemma 4. *Let $\Delta, \tau, \rho \geq 0$ and $\gamma, \gamma_{\text{add}}, p, \delta \in (0,1)$. Assume $\mathbf{M} \in \mathbb{R}^{m\times n}$ is Δ-close to $\widehat{\mathbf{M}}$ on a γ-submatrix, and that $m \geq n$. Finally, assume that*

$$\tau \leq \frac{\Delta n}{\gamma_{\text{add}}},\ \rho \geq \frac{8\Delta}{\sqrt{200\gamma n\log(\frac{m}{\gamma_{\text{add}}})}},\ p \geq 60\tau^2\log\left(\frac{100m}{\delta\gamma_{\text{add}}}\right)\max\left(\frac{\gamma n}{\Delta^2},\ \frac{5}{\rho^2}\right).$$

With probability $\geq 1-\delta$, Algorithm 1 returns $S \subseteq [m]$, $T \subseteq [n]$ satisfying the following.

- $|S| \geq m - \gamma_{\text{drop}}n$, $|T| \geq (1-\gamma_{\text{drop}})n$, *for* $\gamma_{\text{drop}} = 400\gamma\log(m)$.
- *For all $i \in S$, $\left\|[\mathbf{M}-\widehat{\mathbf{M}}]^{\leq\tau}_{i,T}\right\|_2 \leq \rho$, and for all $j \in T$, $\left\|[\mathbf{M}-\widehat{\mathbf{M}}]^{\leq\tau}_{S,j}\right\|_2 \leq \rho$.*
- $\left\|[\mathbf{M}-\widehat{\mathbf{M}}]^{\leq\tau}_{S,T}\right\|_{\text{F}} \leq 2\Delta$.

Proof. Throughout for convenience, we denote

$$\mathbf{D}_t^\star := \left[\mathbf{M}-\widehat{\mathbf{M}}\right]^{\leq\tau}_{S_t\times T_t} \text{ and } \Phi_t := \|\mathbf{D}_t^\star\|_{\text{F}}^2.$$

Also, by applying Lemma 3 with $\delta \leftarrow \frac{\delta\gamma_{\text{add}}}{100m} \leq \frac{\delta}{(m+n)(t_{\max}+1)}$, we assume throughout the proof (giving the failure probability by a union bound) that for all iterations $0 \leq t < t_{\max}$ and all $i \in S_t$, $j \in T_t$,

$$\begin{aligned}\left|r_{i,t} - \|[\mathbf{D}_t^\star]_{i:}\|_2^2\right| &\leq \max\left(\frac{1}{10}\|[\mathbf{D}_t^\star]_{i:}\|_2^2,\ \frac{\Delta^2}{2\gamma n}\right),\\ \left|c_{j,t} - \left\|[\mathbf{D}_t^\star]_{:j}\right\|_2^2\right| &\leq \max\left(\frac{1}{10}\left\|[\mathbf{D}_t^\star]_{:j}\right\|_2^2,\ \frac{\Delta^2}{2\gamma n}\right),\end{aligned} \tag{3.1}$$

as well as (corresponding to the last round of Algorithm 1), for all $i \in S_{t_{\max}}$ and $j \in T_{t_{\max}}$,

$$\begin{aligned} \left| r_i - \left\| \left[\mathbf{D}^\star_{t_{\max}} \right]_{i:} \right\|_2^2 \right| &\le \max\left(\frac{1}{10} \left\| \left[\mathbf{D}^\star_{t_{\max}} \right]_{i:} \right\|_2^2, \frac{\rho^2}{10} \right), \\ \left| c_j - \left\| \left[\mathbf{D}^\star_{t_{\max}} \right]_{:j} \right\|_2^2 \right| &\le \max\left(\frac{1}{10} \left\| \left[\mathbf{D}^\star_{t_{\max}} \right]_{:j} \right\|_2^2, \frac{\rho^2}{10} \right). \end{aligned} \tag{3.2}$$

By definition, $\Phi_0 \le mn\tau^2$, and Φ_t is nonincreasing. Next, consider an iteration t where $\Phi_t \ge 4\Delta^2$. By the closeness assumption, there are $A_t^\star \subseteq S_t$, $B_t^\star \subseteq T_t$ with $|A_t^\star|, |B_t^\star| \le \gamma n$, and

$$\sum_{i \in A_t^\star} \|[\mathbf{D}_t^\star]_{i:}\|_2^2 + \sum_{j \in B_t^\star} \|[\mathbf{D}_t^\star]_{:j}\|_2^2 \ge \frac{3}{4} \|\mathbf{D}_t^\star\|_{\mathrm{F}}^2 = \frac{3}{4}\Phi_t.$$

Now if $\sum_{i \in A_t^\star} \|[\mathbf{D}_t^\star]_{i:}\|_2^2 \ge \frac{3}{8}\Phi_t$, by removing the γn largest rows by $r_{i,t}$, (3.1) yields

$$\Phi_{t+1} \le \left(1 - \frac{4}{5} \cdot \frac{3}{8}\right) \Phi_t + \gamma n \cdot \frac{\Delta^2}{\gamma n} \le \frac{7}{10}\Phi_t + \Delta^2 \le 0.95\Phi_t.$$

Otherwise, $\sum_{j \in B_t^\star} \|[\mathbf{D}_t^\star]_{:j}\|_2^2 \ge \frac{3}{8}\Phi_t$, and so again $\Phi_{t+1} \le 0.95\Phi_t$. Inducting, we thus have

$$\Phi_{t_{\max}} \le 4\Delta^2.$$

Therefore, by Markov's inequality there are at most $\frac{\gamma_{\mathrm{drop}} n}{2}$ rows in $S_{t_{\max}}$ and $\frac{\gamma_{\mathrm{drop}} n}{2}$ columns in $T_{t_{\max}}$ with norm more than $\frac{4\Delta}{\sqrt{\gamma_{\mathrm{drop}} n}} \le \frac{\rho}{2}$ in $\mathbf{D}^\star_{t_{\max}}$. If a row $i \in S_{t_{\max}}$ had norm more than ρ in $\mathbf{D}^\star_{t_{\max}}$, (3.2) ensures it will be removed, and a similar argument holds for columns. Finally, the number of dropped rows and columns in the first $t_{\max}$ iterations is at most $\frac{\gamma_{\mathrm{drop}} n}{2}$ by our parameter choices; here we note that without loss of generality, $\gamma_{\mathrm{add}} \ge \frac{1}{m}$, so $\frac{m}{\gamma_{\mathrm{add}}} \le m^2$. The last condition follows since we showed $\Phi_{t_{\max}} \le 4\Delta^2$ and then dropped entries. □

Lemma 4 does not give control over entries where $\widehat{\mathbf{M}} - \mathbf{M}$ is large. However, below we show that the entries where $\widehat{\mathbf{M}} - \mathbf{M}$ is large must be contained in a small number of rows and columns. We begin by observing a structural fact about entries from distinct rows and columns.

Lemma 5. *Let $\mathbf{M} \in \mathbb{R}^{m \times n}$ be rank-r and let $\{(i_k, j_k)\}_{k \in [K]} \subset [m] \times [n]$ be such that $\{i_k\}_{k \in [K]}$ are distinct and $\{j_k\}_{k \in [K]}$ are distinct. Then $\sum_{k \in [K]} |\mathbf{M}_{i_k, j_k}| \le \|\mathbf{M}\|_{\mathrm{tr}}$.*

Proof. Letting $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be an SVD of $\mathbf{M}$ where columns of $\mathbf{U}$, $\mathbf{V}$ are $\{u_\ell\}_{\ell \in [r]}$, $\{v_\ell\}_{\ell \in [r]}$ respectively,

$$\begin{aligned} \sum_{k \in [K]} |\mathbf{M}_{i_k, j_k}| &\le \sum_{k \in [K]} \sum_{\ell \in [r]} |\sigma_\ell| |u_\ell|_{i_k} |v_\ell|_{j_k} \le \sum_{\ell \in [r]} |\sigma_\ell| \left(\frac{1}{2} \sum_{k \in [K]} |u_\ell|_{i_k}^2 + \frac{1}{2} \sum_{k \in [K]} |v_\ell|_{j_k}^2 \right) \\ &\le \sum_{\ell \in [r]} |\sigma_\ell| \left(\frac{1}{2} \sum_{i \in [m]} |u_\ell|_i^2 + \frac{1}{2} \sum_{j \in [n]} |v_\ell|_j^2 \right) = \|\mathbf{M}\|_{\mathrm{tr}}. \end{aligned}$$

□

Using Lemma 5, we can show that not too many distinct rows and columns of the difference between a pair of low-rank matrices which are close on a submatrix can contain very large entries.

Algorithm 1: $\mathsf{Filter}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), \mathbf{M}, \tau, \rho, \Delta, \gamma, \gamma_{\mathrm{add}}, p, \delta)$

Input: $\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}})$, $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\tau, \rho, \Delta \geq 0$, $\gamma, \gamma_{\mathrm{add}}, p, \delta \in (0,1)$
$S_0 \leftarrow [m], T_0 \leftarrow [n]$
$t_{\max} \leftarrow \lceil 20 \log \frac{mn\tau^2}{4\Delta^2} \rceil$
$\gamma_{\mathrm{drop}} \leftarrow 400\gamma \log(m)$
for $0 \leq t < t_{\max}$ **do**
  $\mathbf{D}_t \leftarrow \mathcal{O}_p([\mathbf{M} - \widehat{\mathbf{M}}]^{\leq\tau}_{S_t,T_t})$
  for $i \in S_t$ **do** $r_{i,t} \leftarrow \frac{1}{p}\|[\mathbf{D}_t]_{i:}\|_2^2$
  for $j \in T_t$ **do** $c_{j,t} \leftarrow \frac{1}{p}\|[\mathbf{D}_t]_{:j}\|_2^2$
  $S_{t+1} \leftarrow S_t \setminus A_t$ where $A_t \subset S_t$ corresponds to the γn indices i with largest $r_{i,t}$
  $T_{t+1} \leftarrow T_t \setminus B_t$ where $B_t \subset T_t$ corresponds to the γn indices j with largest $c_{i,t}$
end
$\mathbf{D} \leftarrow \mathcal{O}_p([\mathbf{M} - \widehat{\mathbf{M}}]^{\leq\tau}_{S_{t_{\max}},T_{t_{\max}}})$
for $i \in S_{t_{\max}}$ **do** $r_i \leftarrow \frac{1}{p}\|\mathbf{D}_{i:}\|_2^2$
for $j \in T_{t_{\max}}$ **do** $c_j \leftarrow \frac{1}{p}\|\mathbf{D}_{:j}\|_2^2$
$S \leftarrow S_{t_{\max}} \setminus A$ where $A \subset S_{t_{\max}}$ corresponds to the $\frac{\gamma_{\mathrm{drop}} n}{2}$ indices i with largest r_i
$T \leftarrow T_{t_{\max}} \setminus B$ where $B \subset T_{t_{\max}}$ corresponds to the $\frac{\gamma_{\mathrm{drop}} n}{2}$ indices j with largest c_j
return (S, T)

Lemma 6. *Assume rank-r $\mathbf{M} \in \mathbb{R}^{m\times n}$ and rank-$r^\star$ $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ are Δ-close on a γ-submatrix, and $m \geq n$. There are sets $A \subseteq [m]$ and $B \subseteq [n]$ such that $\|[\mathbf{M} - \mathbf{M}^\star]_{A,B}\|_{\max} \leq \tau$,*

$$|[m] \setminus A| \leq \gamma n + \frac{\Delta\sqrt{r + r^\star}}{\tau} \text{ and } |[n] \setminus B| \leq \gamma n + \frac{\Delta\sqrt{r + r^\star}}{\tau}.$$

Proof. By assumption, there are $A_0 \subseteq [m]$, $B_0 \subseteq [n]$ with $|A_0| \geq m - \gamma n$, $|B_0| \geq (1-\gamma)n$ and $\|[\mathbf{M} - \mathbf{M}^\star]_{A_0,B_0}\|_{\mathrm{F}} \leq \Delta$. Let $\{(i_k, j_k)\}_{k\in[K]} \subset A_0 \times B_0$ be maximal such that $\{i_k\}_{k\in[K]}$ and $\{j_k\}_{k\in[K]}$ contain no duplicates, and $|[\mathbf{M} - \mathbf{M}^\star]_{i_k,j_k}| \geq \tau$ for all $k \in [K]$. By Lemma 5,

$$K\tau \leq \|[\mathbf{M} - \mathbf{M}^\star]_{A_0,B_0}\|_{\mathrm{tr}} \leq \sqrt{r + r^\star}\,\|[\mathbf{M} - \mathbf{M}^\star]_{A_0,B_0}\|_{\mathrm{F}} \leq \Delta\sqrt{r + r^\star}.$$

So, $K \leq \frac{\Delta\sqrt{r+r^\star}}{\tau}$ and we may set $A \leftarrow A_0 \setminus \{i_k\}_{k\in[K]}$ and $B \leftarrow B_0 \setminus \{j_k\}_{k\in[K]}$. □

3.2 Proof of Proposition 1

We begin by introducing the tools we use to analyze our algorithm which proves Proposition 1. The first is a guarantee on an approximate k-SVD procedure from [MM15].

Proposition 2 (Theorem 1, Theorem 6, [MM15])**.** *Let $\mathbf{M} \in \mathbb{R}^{m\times n}$, $k \in [\min(m,n)]$, and $\epsilon, \delta \in (0,1)$. There is an algorithm $\mathsf{Power}(\mathbf{M}, k, \epsilon, \delta)$ which runs in time*

$$O\left((\mathrm{nnz}(\mathbf{M})k + (m+n)k^2) \cdot \frac{\log\frac{m+n}{\delta}}{\epsilon}\right)$$

and outputs $\mathbf{U} \in \mathbb{R}^{m\times r}$ with orthonormal columns such that, with probability $\geq 1 - \delta$,

$$\left\|(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top)\mathbf{M}\right\|_{\mathrm{op}} \leq (1+\epsilon)\sigma_{k+1}(\mathbf{M}).$$

The second is a bound on the operator norm error of revealing entries independently at random.

Lemma 7. *Let $\mathbf{M} \in \mathbb{R}^{m\times n}$, $p, \delta \in (0,1)$, and suppose $\|\mathbf{M}\|_{\max} \le \tau$ and $\max_{i\in[m]} \|\mathbf{M}_{i:}\|_2 \le \rho$, $\max_{j\in[n]} \|\mathbf{M}_{:j}\|_2 \le \rho$. Let $\widetilde{\mathbf{M}}$ be obtained by including each $(i,j) \in [d]\times[d]$ in a set S with probability p, and setting $\widetilde{\mathbf{M}} = \frac{1}{p}\sum_{(i,j)\in S} \mathbf{M}_{ij}\mathbf{E}_{ij}$. Then with probability $\ge 1-\delta$,*

$$\left\|\mathbf{M} - \widetilde{\mathbf{M}}\right\|_{\text{op}} \le \max\left(\frac{2\rho}{\sqrt{p}}\sqrt{\log\left(\frac{m+n}{\delta}\right)},\ \frac{4\tau}{3p}\log\left(\frac{m+n}{\delta}\right)\right).$$

Proof. For all $(i,j) \in [m]\times[n]$, define the random matrix

$$\mathbf{X}_{(i,j)} := \begin{cases} \left(\frac{1}{p} - 1\right)\mathbf{M}_{ij}\mathbf{E}_{ij} & \text{with probability } p, \\ -\mathbf{M}_{ij}\mathbf{E}_{ij} & \text{with probability } 1-p. \end{cases}$$

By definition, all $\mathbb{E}\mathbf{X}_{(i,j)} = \mathbf{0}_{m\times n}$, and $\sum_{(i,j)\in[m]\times[n]} \mathbf{X}_{(i,j)} = \mathbf{M} - \widetilde{\mathbf{M}}$, so we may apply Fact 4. First of all, clearly it suffices to choose $R = \frac{\tau}{p}$. Further, we bound σ:

$$\sum_{(i,j)\in[m]\times[n]} \left(p\left(\frac{1}{p}-1\right)^2 + 1 - p\right)\mathbf{M}_{ij}^2\mathbf{E}_{ii} = \left(\frac{1}{p}-1\right)\sum_{i\in[m]} \|\mathbf{M}_{i:}\|_2^2\,\mathbf{E}_{ii} \le \frac{\rho^2}{p}$$

and a similar calculation for the other term shows $\sigma = \frac{\rho}{\sqrt{p}}$ suffices. For t in the lemma statement,

$$\Pr\left[\left\|\mathbf{M} - \widetilde{\mathbf{M}}\right\|_{\text{op}} \ge t\right] \le (m+n)\exp\left(-\frac{t^2}{\frac{2\rho^2}{p} + \frac{2\tau t}{3p}}\right) \le \delta.$$

□

The third is a bound on the Frobenius norm of a matrix which is close to an operator norm ball.

Lemma 8. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ satisfy $\|\mathbf{A}\|_{\text{op}} \le a$ and $\|\mathbf{B}\|_{\text{F}} \le b$. If $\mathbf{A}+\mathbf{B}$ is rank-r,*

$$\|\mathbf{A}+\mathbf{B}\|_{\text{F}} \le \sqrt{2(ra^2+b^2)}.$$

Proof. Let the singular values of $\mathbf{M} := \mathbf{A}+\mathbf{B}$ be $\{\sigma_i\}_{i\in[r]}$. By construction, the distance from $\mathbf{M}$ to the set of $m\times n$ matrices with operator norm at most a is bounded by b, and this distance squared is $\sum_{i\in[r]} \mathbf{1}_{\sigma_i\ge a}(\sigma_i - a)^2$, so $\sum_{i\in[r]} \mathbf{1}_{\sigma_i\ge a}(\sigma_i-a)^2 \le b^2$. The conclusion then follows from

$$\|\mathbf{M}\|_{\text{F}}^2 = \sum_{i\in[r]} \sigma_i^2 \le \sum_{i\in[r]} 2(a^2 + \mathbf{1}_{\sigma_i\ge a}(\sigma_i - a)^2) \le 2(ra^2+b^2).$$

□

The last is a simple fact on singular values of a perturbed low-rank matrix.

Lemma 9. *If $\mathbf{A} \in \mathbb{R}^{m\times n}$ is rank-r and $\mathbf{B} \in \mathbb{R}^{m\times n}$ satisfies $\|\mathbf{B}\|_{\text{F}} \le b$, $\sigma_{2r+1}(\mathbf{A}+\mathbf{B}) \le \frac{b}{\sqrt{r}}$.*

Proof. Let $V \subseteq \mathbb{R}^m$ span the image of $\mathbf{A}$, and let $U \subseteq \mathbb{R}^m$ be the top-r left singular vector space of $\mathbf{\Pi}_{V_\perp}\mathbf{B} = \mathbf{\Pi}_{V_\perp}(\mathbf{A}+\mathbf{B})$. Since $\|\mathbf{\Pi}_{V_\perp}\mathbf{B}\|_{\text{F}} \le b$, the largest singular value of $\mathbf{\Pi}_{(U\cup V)_\perp}(\mathbf{A}+\mathbf{B})$ is $\le \frac{b}{\sqrt{r}}$. By the min-max principle for singular values, we have the claim (as $U\cup V$ has dimension-$2r$). □

Algorithm 2: $\mathsf{Descent}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), \mathbf{M}, r^\star, \Delta, \gamma, \gamma_{\mathrm{add}}, \delta, \ell)$

Input: $\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}})$ for $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N} \in \mathbb{R}^{m\times n}$ where $\mathbf{M}^\star$ is rank-$r^\star$ and $\|\mathbf{N}\|_{\mathrm{F}} \le \frac{\Delta}{20\ell}$, $\mathbf{M} \in \mathbb{R}^{m\times n}$ which is Δ-close to $\mathbf{M}^\star$ on a γ-submatrix, given as a rank-r factorization $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\Delta \ge 0$, $\gamma, \gamma_{\mathrm{add}}, \delta \in (0,1)$, $\ell \ge 1$

$(\tau, \rho) \leftarrow \left(\frac{\Delta\sqrt{r+r^\star}}{\gamma_{\mathrm{add}} n}, \ \frac{\Delta}{20\ell\sqrt{(\gamma+\gamma_{\mathrm{add}})n}} \right)$

$(S, T) \leftarrow \mathsf{Filter}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), \mathbf{M}, \tau, \rho, 1.1\Delta, \gamma, \gamma_{\mathrm{add}}, \frac{120000(r+r^\star)\ell^2}{n} \cdot \frac{\gamma+\gamma_{\mathrm{add}}}{\gamma_{\mathrm{add}}^2} \log(\frac{300m}{\delta\gamma_{\mathrm{add}}}), \frac{\delta}{3})$

$\mathbf{X} \leftarrow \mathcal{O}_q([\widehat{\mathbf{M}} - \mathbf{M}]_{S,T}^{\le\tau})$ for $q \leftarrow \frac{15(r+r^\star)\ell \log\frac{6m}{\delta}}{\gamma_{\mathrm{add}} n}$

$\widehat{\mathbf{U}} \leftarrow \mathsf{Power}(\mathbf{X}, 2(r+r^\star), 0.1, \frac{\delta}{3})$ (see Proposition 2)

$(\mathbf{U}', \mathbf{V}') \leftarrow (\mathbf{U}, \mathbf{V})$ with columns of $\widehat{\mathbf{U}}$, $\frac{1}{q}\mathbf{X}^\top\widehat{\mathbf{U}}$ appended respectively

return $(\mathbf{U}', \mathbf{V}', S, T)$

Now we assemble the pieces and prove Proposition 1, our main iterative method guarantee. To a large extent, the proof strategy in Proposition 1 is patterned off the short-flat decomposition analysis of the iterative method in [KLL+22]. Specifically, we show how to decompose the difference matrix (on a large submatrix) into a Frobenius-norm bounded component and an operator-norm bounded component, which allows us to bound the effect of the error on the submatrix via Lemma 8. We restate the result here for convenience to the reader.

Proposition 1. *Let $\Delta \ge 0$, $\gamma, \gamma_{\mathrm{add}}, \delta \in (0,1)$, and $\ell \ge 1$. Let $\widehat{\mathbf{M}} := \mathbf{M}^\star + \mathbf{N} \in \mathbb{R}^{m\times n}$ for $m \ge n$, $\mathbf{M}^\star$ which is rank-$r^\star$, and $\mathbf{N}$ satisfying $\|\mathbf{N}\|_{\mathrm{F}} \le \frac{\Delta}{20\ell}$. If rank-$r$ $\mathbf{M} \in \mathbb{R}^{m\times n}$ is Δ-close to $\mathbf{M}^\star$ on a γ-submatrix and given as a rank-r factorization, Algorithm 2 returns $\widetilde{\mathbf{M}} \in \mathbb{R}^{m\times n}$ as a rank-$3(r+r^\star)$ factorization and $S \subseteq [m]$, $T \subseteq [n]$ satisfying the following with probability $\ge 1-\delta$.*

1. *$|S| \ge m - \gamma_{\mathrm{drop}} n$, $|T| \ge (1-\gamma_{\mathrm{drop}})n$, for $\gamma_{\mathrm{drop}} = \max(400\gamma\log(m),\ 10^5\ell^2(\gamma+\gamma_{\mathrm{add}}))$.*
2. *$\widetilde{\mathbf{M}}_{S,T}$ is $\frac{\Delta}{\ell}$-close to $\mathbf{M}^\star_{S,T}$ on a $(\gamma+\gamma_{\mathrm{add}})$-submatrix.*

Algorithm 2 uses $O(mnp(r+r^\star))$ time and one call to $\mathcal{O}_p(\widehat{\mathbf{M}})$ where for a sufficiently large constant,

$$p = O\left(\frac{(r+r^\star)\ell^2}{n} \cdot \frac{\gamma+\gamma_{\mathrm{add}}}{\gamma_{\mathrm{add}}^2} \log^2\left(\frac{m}{\delta}\right) \right).$$

Proof. Throughout, we denote (in accordance with the guarantees of Lemma 4):

$$\rho := \frac{\Delta}{20\ell\sqrt{(\gamma+\gamma_{\mathrm{add}})n}}, \ \tau := \frac{\Delta\sqrt{r+r^\star}}{\gamma_{\mathrm{add}} n}.$$

We also denote $\mathbf{X}' := \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{X}$, and let

$$\widetilde{\mathbf{M}} := \mathbf{M} + \frac{1}{q}\mathbf{X}' = \mathbf{U}'(\mathbf{V}')^\top$$

be the matrix whose low-rank factorization is the output of Algorithm 2. By the assumed bound on $\mathbf{N}$, $\mathbf{M}$ is 1.1Δ-close to $\widehat{\mathbf{M}}$ on a γ-submatrix, and hence we may apply Lemma 4 with the chosen parameters. Further, since $\mathbf{M}$ is Δ-close to $\mathbf{M}^\star$ on a γ-submatrix, Lemma 6 applied to $[\mathbf{M}-\mathbf{M}^\star]_{S,T}$ produces $A \subseteq S$, $B \subseteq T$ by deleting $\le (\gamma+\gamma_{\mathrm{add}})n$ rows and columns, so $[\mathbf{M}-\mathbf{M}^\star]_{A,B}$ is entrywise

in $[-\tau, \tau]$. We define $\mathbf{M}^\circ$ to be equal to $\mathbf{M}^\star$ on $A \times B$ and equal to $\mathbf{M}$ on $(S \times T) \setminus (A \times B)$, i.e. (where rows and columns are permuted so $A \times B$ is on the top left)

$$\mathbf{M}^\circ_{S,T} = \begin{pmatrix} \mathbf{M}^\star_{A,B} & \mathbf{M}_{A,T\setminus B} \\ \mathbf{M}_{S\setminus A,B} & \mathbf{M}_{S\setminus A,T\setminus B} \end{pmatrix}.$$

We will prove $\|[\mathbf{M}^\circ - \widetilde{\mathbf{M}}]_{S,T}\|_{\mathrm{F}} \le \frac{\Delta}{\ell}$, and then the conclusion follows as $\mathbf{M}^\circ_{S,T} = \mathbf{M}^\star_{S,T}$ except on $(\gamma + \gamma_{\mathrm{add}})n$ rows and columns by Lemma 6, and S, T drop $\le \gamma_{\mathrm{drop}} n$ rows and columns by Lemma 4. To begin, we summarize our strategy. We decompose $[\mathbf{M}^\circ - \widetilde{\mathbf{M}}]_{S,T}$ into three parts:

$$\begin{aligned} \left[\mathbf{M}^\circ - \widetilde{\mathbf{M}}\right]_{S,T} &= \left([\mathbf{M}^\circ - \mathbf{M}]_{S,T} - \left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S,T}\right) \\ &\quad + \left(\left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S,T} - \frac{1}{q}\mathbf{X}_{S,T}\right) + \frac{1}{q}\left[\mathbf{X} - \mathbf{X}'\right]_{S,T}. \end{aligned} \tag{3.3}$$

We will bound each of the terms in (3.3) (the first in Frobenius norm and the latter two in operator norm), and then apply Lemma 8. First, we claim that for all $(i,j) \in A \times B$,

$$\left|[\mathbf{M}^\circ - \mathbf{M}]_{i,j} - [\widehat{\mathbf{M}} - \mathbf{M}]^{\le\tau}_{i,j}\right| \le \left|[\mathbf{M}^\circ - \mathbf{M}]_{i,j} - [\widehat{\mathbf{M}} - \mathbf{M}]_{i,j}\right| = \left|[\mathbf{M}^\circ - \widehat{\mathbf{M}}]_{i,j}\right|.$$

This is because $[\mathbf{M}^\circ - \mathbf{M}]_{A,B}$ is entrywise in $[-\tau, \tau]$ by definition, so projecting an entry of $[\widehat{\mathbf{M}} - \mathbf{M}]_{S,T}$ onto $[-\tau, \tau]$ only decreases the distance. Hence, we bound the first term of (3.3) in $A \times B$ and outside separately: since $[\mathbf{M}^\circ - \mathbf{M}]_{S,T}$ vanishes outside $A \times B$,

$$\begin{aligned} \left\|[\mathbf{M}^\circ - \mathbf{M}]_{S,T} - \left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S,T}\right\|_{\mathrm{F}} &\le \left\|\left[\mathbf{M}^\circ - \widehat{\mathbf{M}}\right]_{A,B}\right\|_{\mathrm{F}} \\ &\quad + \left\|\left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S\setminus A,T}\right\|_{\mathrm{F}} + \left\|\left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S,T\setminus B}\right\|_{\mathrm{F}} \\ &\le \|\mathbf{N}\|_{\mathrm{F}} + \left(\sqrt{|S \setminus A|} + \sqrt{|T \setminus B|}\right)\rho \\ &\le 2\sqrt{(\gamma + \gamma_{\mathrm{add}})n}\rho + \frac{\Delta}{20\ell} \le \frac{\Delta}{10\ell}. \end{aligned} \tag{3.4}$$

Next, by Lemma 7, the entrywise bound on $[\widehat{\mathbf{M}} - \mathbf{M}]^{\le\tau}$ and the row/column bounds on $[\widehat{\mathbf{M}} - \mathbf{M}]_{S,T}$,

$$\begin{aligned} \left\|\left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S,T} - \frac{1}{q}\mathbf{X}_{S,T}\right\|_{\mathrm{op}} &\le \max\left(\frac{2\rho}{\sqrt{q}}\sqrt{\log\left(\frac{3(m+n)}{\delta}\right)},\ \frac{4\tau}{3q}\log\left(\frac{3(m+n)}{\delta}\right)\right) \\ &\le \frac{\Delta}{10\sqrt{r + r^\star}\ell}, \end{aligned} \tag{3.5}$$

with probability $\ge 1 - \frac{\delta}{3}$. Finally, note that

$$\frac{1}{q}\mathbf{X}_{S,T} = [\mathbf{M}^\circ - \mathbf{M}]_{S,T} + \left(\left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S,T} - [\mathbf{M}^\circ - \mathbf{M}]_{S,T}\right) + \left(\frac{1}{q}\mathbf{X}_{S,T} - \left[\widehat{\mathbf{M}} - \mathbf{M}\right]^{\le\tau}_{S,T}\right) \tag{3.6}$$

so it is the sum of a rank-$(r + r^\star)$ matrix, a Frobenius norm bounded matrix (by (3.4)), and an

operator norm bounded matrix (by (3.5)). Therefore,

$$\begin{aligned}\sigma_{2(r+r^\star)+1}\left(\frac{1}{q}\mathbf{X}_{S,T}\right) &\le \left\|\left[\widehat{\mathbf{M}}-\mathbf{M}\right]_{S,T}^{\le\tau}-\frac{1}{q}\mathbf{X}_{S,T}\right\|_{\mathrm{op}} \\ &\quad+\sigma_{2(r+r^\star)+1}\left([\mathbf{M}^\circ-\mathbf{M}]_{S,T}+\left(\left[\widehat{\mathbf{M}}-\mathbf{M}\right]_{S,T}^{\le\tau}-[\mathbf{M}^\circ-\mathbf{M}]_{S,T}\right)\right) \\ &\le \frac{\Delta}{10\sqrt{r+r^\star}\ell}+\frac{\Delta}{10\sqrt{r+r^\star}\ell}\le\frac{\Delta}{5\sqrt{r+r^\star}\ell}.\end{aligned}$$

Above, the first inequality followed by Weyl's perturbation theorem, and the second followed from Lemma 9 and (3.4), (3.5). By Proposition 2 we then have that

$$\left\|\frac{1}{q}\left[\mathbf{X}-\mathbf{X}'\right]_{S,T}\right\|_{\mathrm{op}}\le\frac{1.1\Delta}{5\sqrt{r+r^\star}\ell}, \tag{3.7}$$

with probability $\ge 1-\frac{\delta}{3}$. The decomposition (3.3) shows we can write $[\mathbf{M}^\circ-\widetilde{\mathbf{M}}]_{S,T}$ as the sum of a Frobenius norm bounded matrix (the contribution of (3.4)) and an operator norm bounded matrix (the contributions of (3.5) and (3.7)). Further, since $[\mathbf{M}^\circ-\widetilde{\mathbf{M}}]_{S,T}=[\mathbf{M}^\star-\mathbf{M}]_{A,B}-\frac{1}{q}\mathbf{X}'_{S,T}$ is the sum of a rank-$(r+r^\star)$ matrix and a rank-$2(r+r^\star)$ matrix, it is rank $3(r+r^\star)$. Hence,

$$\left\|\left[\mathbf{M}^\circ-\widetilde{\mathbf{M}}\right]_{S,T}\right\|_{\mathrm{F}}\le\sqrt{6(r+r^\star)}\cdot\left(\frac{1.1\Delta}{5\sqrt{r+r^\star}\ell}+\frac{\Delta}{10\sqrt{r+r^\star}\ell}\right)+\sqrt{2}\cdot\frac{\Delta}{10\ell}\le\frac{\Delta}{\ell}$$

follows by applying Lemma 8 with $\mathbf{A}=[\widehat{\mathbf{M}}-\mathbf{M}]_{S,T}^{\le\tau}-\frac{1}{q}\mathbf{X}'_{S,T}$ and $\mathbf{B}=[\mathbf{M}^\circ-\mathbf{M}]_{S,T}-[\widehat{\mathbf{M}}-\mathbf{M}]_{S,T}^{\le\tau}$. The failure probability comes from a union bound over Lemma 4, Lemma 7, and Proposition 2. We use Lemma 1 to upper bound the reveal probability, since Line 3 requires $O(\log(\frac{m}{\gamma_{\mathrm{add}}}))$ calls to $\mathcal{O}_{q'}$ for the specified q', and Line 4 requires one call to $\mathcal{O}_q$ for $q=O(q')$.

Finally, we discuss runtime. The runtime of Lines 3 and 4 are bottlenecked by computing $O(mnp)$ entries of $\widehat{\mathbf{M}}-\mathbf{M}$, where p is specified in the statement of Proposition 1; a Chernoff bound implies the number of revealed entries will be within a constant factor of its expectation within the failure probability budget. Since $\mathbf{M}$ is given as a rank-r factorization and entries of $\widehat{\mathbf{M}}$ are given, this cost is $O(mnp\cdot r)$. The runtime cost of Power is specified by Proposition 2 to be $O(mnq\cdot(r+r^\star)\log\frac{m}{\delta})$, where $\mathcal{T}_{\mathrm{mv}}(\mathbf{X})=O(\mathrm{nnz}(\mathbf{X}))=O(mnq)$, and this does not dominate. The runtime cost of computing $\mathbf{X}^\top\widehat{\mathbf{U}}$ is $O(mr^2)$ using $\mathbf{X}=\mathbf{U}\mathbf{V}^\top$ and also does not dominate. □

3.3 Partial matrix completion via Descent

In this section, we give a simple recursive application of Descent to give a self-contained result on partial matrix completion. For simplicity, we assume we have an upper bound on the largest singular value of the target matrix $\mathbf{M}^\star$; we will show how to lift this assumption in Section 5.

Corollary 2. *Let $\mathbf{M}^\star\in\mathbb{R}^{m\times n}$ be rank-$r^\star$, $\|\mathbf{M}^\star\|_{\mathrm{op}}\le\sigma$, $m\ge n$, $\delta\in(0,1)$, let $\widehat{\mathbf{M}}=\mathbf{M}^\star+\mathbf{N}$ for $\|\mathbf{N}\|_{\mathrm{F}}\le\Delta$, and let $\ell\ge 1$. Algorithm 3 returns $\mathbf{U}\in\mathbb{R}^{m\times r}$, $\mathbf{V}\in\mathbb{R}^{n\times r}$, and (S,T), for $r=r^\star\mathrm{poly}(\ell)$, such that $[\mathbf{U}\mathbf{V}^\top]_{S,T}$ is $O(\max(\sigma\sqrt{r^\star}\exp(-\log^2(\ell)),\Delta))$-close to $\mathbf{M}^\star$ on an α-submatrix and $|S|\ge m-\alpha n$, $|T|\ge(1-\alpha)n$, with probability $\ge 1-\delta$. Algorithm 3 uses $O(\frac{m(r^\star)^2\mathrm{poly}(\ell)}{\alpha}\log^3(\frac{m}{\delta}))$ time and one call to $\mathcal{O}_p(\widehat{\mathbf{M}})$, where for a sufficiently large constant,*

$$p=O\left(\frac{r^\star\mathrm{poly}(\ell)}{\alpha n}\log^3\left(\frac{m}{\delta}\right)\right).$$

Algorithm 3: PartialMatrixCompletion$(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), r^\star, \sigma, \Delta, \alpha, \delta, \ell)$

Input: $\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}})$ for $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N} \in \mathbb{R}^{m\times n}$ where $\mathbf{M}^\star$ is rank-$r^\star$ satisfying $\|\mathbf{M}^\star\|_{\mathrm{op}} \le \sigma$ and $\|\mathbf{N}\|_{\mathrm{F}} \le \Delta$, $\alpha, \delta \in (0,1)$, $\ell \ge 1$
$\widetilde{\Delta} \leftarrow \sqrt{r^\star}\sigma$
$(\mathbf{U}, \mathbf{V}) \leftarrow (\mathbf{0}_{m\times 0}, \mathbf{0}_{n\times 0})$
$k \leftarrow 0$
$(S, T) \leftarrow ([m], [n])$
$K \leftarrow \lceil \log \ell \rceil$
$\gamma_{\mathrm{add}} \leftarrow \frac{\alpha}{\max(800K^2\log(m), 2\cdot 10^5 \ell^2 K^2)}$
while $\widetilde{\Delta} \ge 20\ell\Delta$ and $k \le K$ **do**
  $(\mathbf{U}, \mathbf{V}, S, T) \leftarrow$ Descent$(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}_{S,T}), [\mathbf{U}\mathbf{V}^\top]_{S,T}, r^\star, \widetilde{\Delta}, \gamma_{\mathrm{add}}k, \gamma_{\mathrm{add}}, \frac{\delta}{2K}, \ell)$
  $\widetilde{\Delta} \leftarrow \frac{\Delta}{\ell}$
  $k \leftarrow k + 1$
end
$k_{\mathrm{freeze}} \leftarrow k$
while $\widetilde{\Delta} \ge 20e\Delta$ and $k - k_{\mathrm{freeze}} \le K$ **do**
  $(\mathbf{U}, \mathbf{V}, S, T) \leftarrow$ Descent$(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}_{S,T}), [\mathbf{U}\mathbf{V}^\top]_{S,T}, r^\star, \widetilde{\Delta}, \gamma_{\mathrm{add}}k, \gamma_{\mathrm{add}}, \frac{\delta}{2K}, e)$
  $\widetilde{\Delta} \leftarrow \frac{\Delta}{e}$
  $k \leftarrow k + 1$
end
return $(\mathbf{U}, \mathbf{V}, S, T)$

Proof. The failure probability follows by applying a union bound to the $2K$ calls to Descent. Next, we claim that throughout the algorithm we maintain the invariant that $\widetilde{\Delta}$ is an overestimate on the distance from $[\mathbf{U}\mathbf{V}^\top]_{S,T}$ to $[\mathbf{M}^\star]_{S,T}$ on a $\gamma_{\mathrm{add}}k$-submatrix; this is clearly true at the beginning of the algorithm, since $\|\mathbf{M}^\star\|_{\mathrm{F}} \le \sqrt{r^\star}\|\mathbf{M}^\star\|_{\mathrm{op}}$. Further, applying Proposition 1 shows that this invariant is preserved in each iteration, which gives the closeness guarantee since $\ell^{-K} \le \exp(-\log^2(\ell))$. We note that the role of the first phase (Lines 8 to 12) is to cut the initial distance estimate by a factor ℓ^{-K}, but is bottlenecked by the requirement that $\widetilde{\Delta} \ge 20\ell\Delta$. To bring this overhead down to a constant factor, we repeat the argument in Lines 14 to 18, but set $\ell = e$.

By our parameter settings Proposition 1 drops $\le \frac{\alpha n}{2K}$ rows and columns in each iteration, giving the lower bounds on $|S|, |T|$. Further, we can inductively apply Proposition 1 to maintain that the rank r of our iterate is bounded by $3^{2K}r^\star = r^\star\mathrm{poly}(\ell)$, since the potential function $r + r^\star$ at most triples each iteration. The bounds on the runtime and p then follow by using Proposition 1 $2K$ times; we recall that we can aggregate the observation probabilities using Lemma 1. □

To briefly interpret Corollary 2, let $\alpha = \frac{1}{200}$, in other words, consider the case where we are willing to give up on recovering a small constant fraction of rows and columns. Further, suppose $\frac{\sigma\sqrt{r^\star}}{\Delta}$ is polynomially bounded in m, i.e. our initial distance estimate is not too far off from our noise level. By balancing terms via setting

$$\ell = \exp\left(\sqrt{\log\left(\frac{\sigma\sqrt{r^\star}}{\Delta}\right)}\right) = \exp\left(O\left(\sqrt{\log(m)}\right)\right) = m^{o(1)},$$

we see that Corollary 2 yields partial matrix completion obtaining the desired noise level Δ up to

constant overhead, on at least 99% of rows and columns. This setting of ℓ also implies that the rank of the iterates of Algorithm 3 is bounded by $r^\star \cdot m^{o(1)}$ throughout. Further, assuming $\delta = \text{poly}(m^{-1})$ for simplicity, the sample complexity of $mnp = O(m^{1+o(1)}r^\star)$ is almost information-theoretically optimal, and the runtime of $O(m^{1+o(1)}(r^\star)^2)$ is almost-verification time. In other words, by giving up on recovering a subconstant fraction of rows and columns, we obtain almost-optimal matrix completion on the remaining submatrix, without any subspace regularity assumptions.

4 Recovering dropped subsets

In this section, we provide the second key ingredient of our framework, an algorithm which recovers rows and columns which were dropped by our iterative method in Section 3. The method of this section takes as input $\mathbf{M}$ which satisfies submatrix closeness to our target rank-$r^\star$ $\mathbf{M}^\star$, and returns a rank-$O(r^\star)$ factorization. This factorization has the appealing property that it satisfies standard Frobenius norm closeness to $\mathbf{M}^\star$ (without any dropped rows or columns), at a cost of a roughly $\text{poly}(r^\star)$ factor increase in the closeness bound. We now state the main export of this section, parameterized by the notion of standard subspaces (Definition 4).

Proposition 3. *Let $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ be rank-$r^\star$ with (α,β,μ)-standard row and column spans, $m \geq n$, $\delta \in (0,1)$ and let $S \subseteq [m]$, $T \subseteq [n]$ have $|S| \geq m - \frac{\alpha n}{9}$, $|T| \geq m - \frac{\alpha n}{9}$. Assume $\mathbf{M} \in \mathbb{R}^{m\times n}$ is given as a rank-r factorization, $r \geq r^\star$, $\mathbf{M}_{S,T}$ is $(\frac{\alpha}{1800\log(m)}, \Delta)$-close to $\mathbf{M}^\star_{S,T}$ on a γ-submatrix, and $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$ for $\|\mathbf{M}^\star\|_{\text{op}} \leq \sigma$, $\|\mathbf{N}\|_{\text{F}} \leq \frac{\Delta}{20}$. Algorithm 10 returns $\mathbf{U} \in \mathbb{R}^{m\times r'}$ and $\mathbf{V} \in \mathbb{R}^{n\times r'}$ satisfying*

$$\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\right\|_{\text{F}} \leq \frac{C_{\text{fix}} r^\star \sqrt{r^\star \log(r^\star)}}{\beta^8}\Delta \text{ and } r' \leq 2r^\star, \tag{4.1}$$

for a universal constant C_{fix}, with probability $\geq 1-\delta$. Algorithm 10 uses $O(\frac{mr^2\mu^2}{\alpha\beta^4}\log^2(\frac{m}{\beta\delta})\log(\frac{m(\sigma+\Delta)}{\Delta\beta\delta}))$ time and one call to $\mathcal{O}_p(\widehat{\mathbf{M}})$ where for a sufficiently large constant,

$$p = O\left(\frac{r\mu\log^2(\frac{m}{\beta})\log(\frac{m}{\delta})}{\alpha\beta^2 n}\right).$$

We prove Proposition 3 in a number of steps organized into subsections, summarized as follows.

1. In Section 4.1, we give an algorithm Sparsify which takes matrices which are close on a submatrix and drops a few more rows and columns, with the guarantee that the resulting submatrices satisfy Definition 3, i.e., they are close away from an RCS matrix.

2. In Section 4.2, we give an algorithm Representative which takes as input a matrix which is close to the target away from an RCS matrix. By repeatedly testing for regression error of our iterate's columns against itself, Representative learns a subset B of columns which are representative of the difference between our iterate and the target, in the sense of Definition 6.

3. In Section 4.3, we first show that Definition 6 implies that the columns indexed by B can be used to effectively approximate dropped rows and columns from observations. We then give an algorithm Complete which learns a low-rank approximation of $\mathbf{M}^\star$ to slightly higher error using B.

4. We put all the pieces together to prove Proposition 3 in Section 4.4.

4.1 Sparsifying errors

The output of the method of Section 3 is $\widetilde{\mathbf{M}}$ that is Δ-close to the true matrix $\mathbf{M}^\star$ on a γ-submatrix (up to dropped subsets). We begin with a postprocessing step which yields finer control over the structure of $\widetilde{\mathbf{M}}-\mathbf{M}^\star$. In particular, we drop some additional rows and columns so that the difference $\widetilde{\mathbf{M}}-\mathbf{M}^\star$ is close away from an s-RCS matrix, for $s \approx \frac{n}{r}$. The algorithm is Algorithm 4, and its statement and analysis are similar to that of Algorithm 1.

We use the following concentration inequality to control the error of empirical estimates.

Lemma 10. *Let $p, \delta \in (0,1)$, $\tau > 0$, $v \in \mathbb{R}^d$, and let $\tilde{v} \in \mathbb{R}^d$ have each entry $\tilde{v}_i$ independently set to v_i with probability p, and 0 otherwise. Then with probability $\geq 1-\delta$,*

$$\left| |\{i \in d \mid |v_i| \geq \tau\}| - \frac{1}{p} |\{i \in d \mid |\tilde{v}_i| \geq \tau\}| \right| \leq \max\left(\frac{1}{10} |\{i \in d \mid |v_i| \geq \tau\}|, \frac{30 \log \frac{2}{\delta}}{p} \right).$$

Proof. Let $x_i \in \{0,1\}$ be a scalar random variable for all $i \in [d]$ which is 1 if $|v_i| \geq \tau$ and let $\tilde{x}_i$ be analogously defined for $\tilde{v}$; clearly $\mathbb{E}\tilde{x}_i = px_i$. Fact 3 shows that with probability $\geq 1-\delta$,

$$\left| \sum_{i \in [d]} x_i - \frac{1}{p} \sum_{i \in [d]} \tilde{x}_i \right| \leq \sqrt{\frac{\sum_{i \in [d]} x_i}{p}} \sqrt{3 \log \frac{2}{\delta}}.$$

The conclusion follows depending on which of $\frac{\sqrt{\sum_{i \in [d]} x_i}}{\sqrt{10}}$ or $\frac{1}{\sqrt{p}} \sqrt{30 \log \frac{2}{\delta}}$ is larger. □

We now use an analogous argument to that of Lemma 4 to analyze Algorithm 4.

Lemma 11. *Let $0 \leq \tau \leq \Delta$ and $\gamma, \gamma_{\text{drop}}, p, \delta \in (0,1)$, and $s \in [n]$. Assume $\mathbf{M} \in \mathbb{R}^{m \times n}$ is Δ-close to $\widehat{\mathbf{M}}$ on a γ-submatrix, and that $m \geq n$. Finally define $s := \frac{16\Delta^2}{\tau^2 \gamma_{\text{drop}} n}$ and assume*

$$\gamma_{\text{drop}} \geq 200\gamma \log(m), \; p \geq \frac{20\gamma_{\text{drop}} n \tau^2}{\Delta^2} \log\left(\frac{100m}{\delta}\right).$$

With probability $\geq 1-\delta$, Algorithm 4 returns $S \subseteq [m]$, $T \subseteq [n]$ satisfying the following.

1. *$|S| \geq m - \gamma_{\text{drop}} n$, $|T| \geq (1-\gamma_{\text{drop}})n$.*
2. *$\mathbf{M}_{S,T}$, $\widehat{\mathbf{M}}_{S,T}$ are 2Δ-close away from an s-RCS matrix.*

Proof. Our first goal is to show that before the application of Filter, every row and column of $[\mathbf{M} - \widehat{\mathbf{M}}]_{S,T}$ has at most s entries larger than τ in magnitude. We will denote

$$\mathbf{D}_t^\star := \left[\mathbf{M} - \widehat{\mathbf{M}}\right]_{S_t \times T_t} \text{ and } \Phi_t := \left| \left\{ (i,j) \in S_t \times T_t \;\middle|\; |[\mathbf{D}_t^\star]_{ij}| \geq \tau \right\} \right|.$$

In other words, Φ_t tracks the number of large entries in $\mathbf{D}_t^\star$, and by definition $\Phi_0 \leq mn$ and Φ_t is nonincreasing. We also denote the exact number of large entries per row and column by

$$r_{i,t}^\star := |\{j \in T_t \mid |[\mathbf{D}_t^\star]_{ij}| \geq \tau\}| \text{ for all } i \in S_t,$$
$$c_{j,t}^\star := |\{i \in S_t \mid |[\mathbf{D}_t^\star]_{ij}| \geq \tau\}| \text{ for all } j \in T_t.$$

Also, as in the proof of Lemma 4, by applying Lemma 10 and a union bound over $\leq 40\log(m)$ iterations (giving the failure probability with a union bound over the call to Filter succeeding), we assume that for all $0 \leq t < t_{\max}$ and all $i \in S_t$, $j \in T_t$,

$$\left|r_{i,t} - r^\star_{i,t}\right| \leq \max\left(\frac{1}{10}r^\star_{i,t},\ \frac{\Delta^2}{2\tau^2}\cdot\frac{1}{\gamma n}\right),$$
$$\left|c_{j,t} - c^\star_{j,t}\right| \leq \max\left(\frac{1}{10}c^\star_{j,t},\ \frac{\Delta^2}{2\tau^2}\cdot\frac{1}{\gamma n}\right),$$

as well as, for all $i \in S_{t_{\max}}, j \in T_{t_{\max}}$,

$$\left|r_i - r^\star_{i,t_{\max}}\right| \leq \max\left(\frac{1}{10}r^\star_{i,t_{\max}},\ \frac{s}{10}\right),$$
$$\left|c_j - c^\star_{j,t_{\max}}\right| \leq \max\left(\frac{1}{10}c^\star_{j,t_{\max}},\ \frac{s}{10}\right).$$

We observe that two matrices which are Δ-close in Frobenius norm have at most $\frac{\Delta^2}{\tau^2}$ entries of the difference with magnitude more than τ. Next, consider an iteration t where $\Phi_t \geq \frac{4\Delta^2}{\tau^2}$. By an analogous argument to Lemma 4, removing the γn largest rows and columns decreases the potential by at least a 0.05 factor, so inducting shows that after $t_{\max}$ iterations,

$$\Phi_{t_{\max}} \leq \frac{4\Delta^2}{\tau^2}.$$

Therefore by Markov's inequality there are at most $\frac{\gamma_{\text{drop}} n}{4}$ rows in $S_{t_{\max}}$ and $\frac{\gamma_{\text{drop}} n}{4}$ columns in $T_{t_{\max}}$ with at least $\frac{16\Delta^2}{\tau^2\gamma_{\text{drop}} n} \leq \frac{s}{2}$ entries larger than τ. In conclusion, if a row or column has more than s entries larger than τ in the last iteration, it will be removed as claimed.

In the last iteration $\mathbf{M}_{S,T}$ and $\widehat{\mathbf{M}}_{S,T}$ are clearly still Δ-close on a γ-submatrix, since we only dropped rows or columns. Filter ensures that by dropping $\frac{\gamma_{\text{drop}} n}{2}$ more rows and columns, the difference matrix truncated at τ has Frobenius norm $\leq 2\Delta$ (see the third guarantee of Lemma 4). Hence, we can take $\mathbf{X}$ to be the truncated difference and $\mathbf{Y}$ to be the sparse errors in Definition 3. $\square$

4.2 Learning a representative subset

4.2.1 Structural properties

In this section, we collect several structural tools which will be helpful in the analysis of our testers. We first provide simple spectral bounds on a randomly subsampled matrix.

Lemma 12. *Let $\delta \in (0,1)$ and let $V \subseteq \mathbb{R}^d$ be (α,β,μ)-standard of dimension r, let $\{b_i\}_{i\in[d]} \subset \mathbb{R}^r$ be rows of an (arbitrary) choice of $\mathbf{B}_V$, and let $S \subset [d]$ have $|S| \geq (1-\frac{\alpha}{2})d$. Let $T \subseteq S$ have each element in S included with probability $p \geq \frac{12\mu r}{\beta^2 d}\log(\frac{2r}{\delta})$. Then with probability $\geq 1-\delta$,*

$$\frac{p\beta^2}{2}\mathbf{I}_r \preceq \sum_{i\in T} b_i b_i^\top \preceq 2p\mathbf{I}_r.$$

Proof. By the assumption on the subspace V, there is a set $A \subset S$ of size at most $\frac{\alpha d}{3}$ such that every row $i \in R := S \setminus A$ satisfies $\|b_i\|_2 \leq \sqrt{\frac{\mu r}{d}}$. For all $i \in R$ define a random matrix

$$\mathbf{X}_i := \begin{cases} b_i b_i^\top & \text{with probability } p, \\ \mathbf{0}_{r\times r} & \text{otherwise.} \end{cases}$$

Algorithm 4: $\mathsf{Sparsify}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), \mathbf{M}, \tau, \Delta, \gamma, \gamma_{\text{drop}}, p, \delta)$

Input: $\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}})$, $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\tau, \Delta \geq 0$, $\gamma, \gamma_{\text{drop}}, p, \delta \in (0,1)$
$S_0 \leftarrow [m], T_0 \leftarrow [n]$
$t_{\max} \leftarrow \lceil 20 \log \frac{mn\tau^2}{4\Delta^2} \rceil$
for $0 \leq t < t_{\max}$ **do**
  $\mathbf{D}_t \leftarrow \mathcal{O}_p([\mathbf{M} - \widehat{\mathbf{M}}]_{S_t,T_t})$
  for $i \in S_t$ **do** $r_{i,t} \leftarrow \frac{1}{p}|\{j \in T_t \mid |[\mathbf{D}_t]_{ij}| \geq \tau\}|$
  for $j \in T_t$ **do** $c_{j,t} \leftarrow \frac{1}{p}|\{i \in S_t \mid |[\mathbf{D}_t]_{ij}| \geq \tau\}|$
  $S_{t+1} \leftarrow S_t \setminus A_t$ where $A_t \subset S_t$ corresponds to the γn indices i with largest $r_{i,t}$
  $T_{t+1} \leftarrow T_t \setminus B_t$ where $B_t \subset T_t$ corresponds to the γn indices j with largest $c_{j,t}$
end
$\mathbf{D} \leftarrow \mathcal{O}_p([\mathbf{M} - \widehat{\mathbf{M}}]_{S_{t_{\max}},T_{t_{\max}}})$
for $i \in S_{t_{\max}}$ **do** $r_i \leftarrow \frac{1}{p}|\{j \in T_{t_{\max}} \mid |\mathbf{D}_{ij}| \geq \tau\}|$
for $j \in T_{t_{\max}}$ **do** $c_j \leftarrow \frac{1}{p}|\{j \in T_{t_{\max}} \mid |\mathbf{D}_{ij}| \geq \tau\}|$
$S \leftarrow S_{t_{\max}} \setminus A$ where $A \subset S_{t_{\max}}$ corresponds to the $\frac{\gamma_{\text{drop}} n}{4}$ indices i with largest r_i
$T \leftarrow T_{t_{\max}} \setminus B$ where $B \subset T_{t_{\max}}$ corresponds to the $\frac{\gamma_{\text{drop}} n}{4}$ indices j with largest c_j
return $\mathsf{Filter}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}_{S,T}), \mathbf{M}_{S,T}, \tau, \infty, \Delta, \gamma, \gamma_{\text{drop}}, 1, p, \frac{\delta}{2})$

We recognize $\sum_{i\in R} \mathbf{X}_i \preceq \sum_{i\in T} b_i b_i^\top$ (since it is restricted to a subset of the rows), and $\|\mathbf{X}_i\|_{\text{op}} \leq \frac{\mu r}{d}$ with probability 1 for all $i \in R$. Moreover, we have $\mathbb{E}\sum_{i\in R} X_i = p\sum_{i\in R} b_i b_i^\top$, and

$$p\beta^2 \mathbf{I}_r \preceq p\sum_{i\in R} b_i b_i^\top \preceq p\mathbf{I}_r,$$

by Lemma 2 since $|R| \geq (1-\alpha)d$. The conclusion follows by applying Fact 3 with $\epsilon = \frac{1}{2}$. □

We also use Proposition 4, an existential variant of Lemma 12 which does not impose a regularity constraint, which can be viewed as a one-sided discrepancy statement potentially of independent interest. We use this terminology because a random S of size $d\lambda$ yields $\sum_{i\in S} b_i b_i^\top = \lambda \mathbf{I}_r$ in expectation, and Proposition 4 matches this up to constant factors in the smallest eigenvalue (a one-sided guarantee). We defer a proof of this claim to Appendix B.

Proposition 4. *Let $\lambda \in [\frac{5600r}{d}, 1)$, let $\mathbf{B} \in \mathbb{R}^{d\times r}$ have orthonormal columns, and denote rows of $\mathbf{B}$ by $\{b_i\}_{i\in[d]} \subset \mathbb{R}^r$. There exists $S \subseteq [d]$ with $|S| \leq d\lambda$ and*

$$\sum_{i\in S} b_i b_i^\top \succeq \frac{\lambda}{8}\mathbf{I}_r.$$

By using Proposition 4, we show that removing a few columns from a pair of matrices which are close away from a RCS matrix induces a submatrix on which they are truly close.

Lemma 13. *Let $\mathbf{M}, \mathbf{M}' \in \mathbb{R}^{m\times n}$ and suppose $\mathbf{M}, \mathbf{M}'$ are Δ-close away from an s-RCS matrix. Further, suppose $\mathbf{M} - \mathbf{M}'$ is rank-r. There is $T \subseteq [n]$ with $|T| \geq n - 5600rs\log m$ such that*

$$\left\|[\mathbf{M} - \mathbf{M}']_{:T}\right\|_{\text{F}} \leq \Delta \cdot \frac{\sqrt{\log m}}{13}.$$

Proof. Let $\mathbf{D} := \mathbf{M} - \mathbf{M}'$ for notational convenience, and let $\mathbf{D} = \mathbf{X} + \mathbf{Y}$ be the decomposition guaranteed by Definition 3. Partition $[m]$ into sets $\{S_j\}_{j\in[k]}$ where $k \le \log m$ as follows. Let S_1 be the set of $i \in [m]$ such that $\|\mathbf{X}_{i:}\|_2 \le \frac{2\Delta}{\sqrt{m}}$ and for $j > 1$, let S_j be the set of $i \in [m]$ with

$$2^{j-1}\frac{\Delta}{\sqrt{m}} < \|\mathbf{X}_{i:}\|_2 \le 2^j \frac{\Delta}{\sqrt{m}}.$$

Now for each $j \in [k]$, consider an SVD of $\mathbf{D}_{S_j:} = \mathbf{U}_j\mathbf{\Sigma}_j\mathbf{V}_j^\top$ and apply Proposition 4 to $\mathbf{U}_j$ with $\lambda_j = \frac{5600r}{|S_j|}$. We obtain $A_j \subseteq S_j$ such that $|A_j| \le 5600r$ and for all $v \in \mathbb{R}^n$,

$$\left\|\mathbf{D}_{A_j:}v\right\|_2 \ge \sqrt{\frac{\lambda_j}{8}}\left\|\mathbf{D}_{S_j:}v\right\|_2. \tag{4.2}$$

Next recall $\mathbf{D} = \mathbf{X} + \mathbf{Y}$ where $\mathbf{Y}$ is s-RCS. For each $i \in [m]$, let $T_i \subset [n]$ be the set of entries on which $\mathbf{Y}_{i:}$ is supported. Let $T = [n] \setminus \bigcup_{i\in A_1\cup\ldots\cup A_k} T_i$ so $|T| \ge n - 5600rs\log m$. Now we can bound $\mathbf{D}_{:T}$ by applying (4.2):

$$\begin{aligned}\|\mathbf{D}_{:T}\|_{\mathrm{F}}^2 &= \sum_{j\in[k]}\left\|\mathbf{D}_{S_j,T}\right\|_{\mathrm{F}}^2 \le \sum_{j\in[k]}\frac{8}{\lambda_j}\left\|\mathbf{D}_{A_j,T}\right\|_{\mathrm{F}}^2 \\ &\le \sum_{j\in[k]}\frac{|S_j||A_j|}{700r}\cdot\frac{2^{2j}\Delta^2}{m} \le \frac{\Delta^2 k}{175} \le \frac{\Delta^2\log m}{175},\end{aligned}$$

where in the second-to-last step, we used that $|S_j| \le \frac{4m}{2^{2j}}$ by Markov's inequality. □

We note that Lemma 13 is a robust variant of a simpler claim, which says that a low-rank matrix with a sparse nonzero pattern must have all of its entries localized to a small submatrix. We provide a proof of this claim for convenience, as we believe it aids in building intuition for our method.

Lemma 14. *Let $\mathbf{D} \in \mathbb{R}^{m\times n}$ be rank-r with $m \ge n$, and suppose $\mathbf{D}$ is $\frac{\alpha}{r}$-RCS for $\alpha \in (0,1)$. There are $A \subseteq [m]$, $B \subseteq [n]$ with $|A| \ge m - \alpha n$, $|B| \ge (1-\alpha)n$ such that $\mathbf{D}_{A,B}$ has no nonzero entries.*

Proof. Consider an iterative process which takes any row of $\mathbf{D}$ with nonzero entries, and orthogonalizes all rows of $\mathbf{D}$ against it. The process terminates after r iterations as $\mathbf{D}$ is rank-r, and the union of the supports of all rows used by the process grows by $\le \frac{\alpha n}{r}$ in each iteration. Hence, the support of all rows is contained in a subset of size αn, and a symmetric argument holds for columns. □

4.2.2 Basic testing

We next analyze properties of a simple algorithm, Test, which solves a regression problem attempting to boundedly combine a set of columns of a matrix to approximate another column. For ease of discussion, we focus on testing columns rather than rows, but a symmetric argument handles both.

Algorithm 5: $\mathsf{Test}(\mathbf{M}, T, j, \phi, \tau)$

Input: $\mathbf{M} \in \mathbb{R}^{m\times n}$, $T \subseteq [n]$, $j \in [n]$, $\phi, \tau \ge 0$
if $\min_{v\in\mathbb{R}^T}\|\mathbf{M}_{:T}v - \mathbf{M}_{:j}\|_2^2 + \frac{\phi^2}{\tau^2}\|v\|_2^2 \le 2\phi^2$ **then return** "True"
else return "False"

If Test returns "True" then we say it has passed, and otherwise we say it has failed. Intuitively, Test simulates testing the value of the following constrained problem:

$$\min_{\substack{v\in\mathbb{R}^T\\ \|v\|_2\le\tau}} \|\mathbf{M}_{:T}v-\mathbf{M}_{:j}\|_2\le\phi,$$

but is easier to compute. We use the following helper claim, which follows from a calculation.

Fact 5. *Let $\mathbf{M}\in\mathbb{R}^{m\times n}$ have SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and have rank r, and let $T\subseteq[n]$ satisfy $\mathbf{V}_{T:}^\top\mathbf{V}_{T:}\succeq\gamma^2\mathbf{I}_r$. Then for some $j\in[n]$, letting $c\in\mathbb{R}^T$ be the vector such that $\mathbf{M}_{:T}c=\mathbf{M}_{:j}$, $\|c\|_2\le\frac{1}{\gamma}\|\mathbf{V}_{j:}\|_2$.*

Fact 5 will be used to show the regression problems we encounter have bounded solutions. Motivated by this, we next specify a set of properties that guarantee Test will pass, combining regularity of a comparison matrix $[\mathbf{M}^\star]_{:T}$ in the sense of Fact 5 with closeness of $\mathbf{M}$, $\mathbf{M}^\star$.

Definition 6. *We say $T\subseteq[n]$ is a (Δ,γ)-representative subset with respect to a pair of matrices $\mathbf{M},\mathbf{M}^\star\in\mathbb{R}^{m\times n}$ if the following properties hold.*

- $\|[\mathbf{M}-\mathbf{M}^\star]_{:T}\|_{\mathrm{F}}\le\Delta$.
- $[\mathbf{V}_\star]_{T:}^\top[\mathbf{V}_\star]_{T:}\succeq\gamma^2\mathbf{I}_{r^\star}$, *where $\mathbf{U}_\star\mathbf{\Sigma}_\star\mathbf{V}_\star^\top$ is an SVD of $\mathbf{M}^\star$ which has rank $r^\star$.*

Lemma 15. *Let $T\subseteq[n]$ contain a $(\frac{\phi}{2\tau},\frac{\theta}{\tau})$-representative subset with respect to $\mathbf{M},\mathbf{M}^\star\in\mathbb{R}^{m\times n}$, for some $\theta\ge0$. Let $j\in[n]$ satisfy $\|[\mathbf{V}_\star]_{j:}\|_2\le\theta$ and $\|[\mathbf{M}-\mathbf{M}^\star]_{:j}\|_2\le\frac{\phi}{2}$, where $\mathbf{U}_\star\mathbf{\Sigma}_\star\mathbf{V}_\star$ is an SVD of $\mathbf{M}^\star$. Then $\mathsf{Test}(\mathbf{M},T,j,\phi,\tau)$ will pass.*

Proof. By Fact 5 and the fact that T is a representative subset with parameter $\frac{\theta}{\tau}$, there is a vector $c\in\mathbb{R}^T$ with $\|c\|_2\le\tau$ and $\mathbf{M}^\star_{:T}c=\mathbf{M}^\star_{:j}$. Using the other property of a representative subset shows

$$\begin{aligned}\|\mathbf{M}_{:T}c-\mathbf{M}_{:j}\|_2&\le\|\mathbf{M}_{:T}c-\mathbf{M}^\star_{:T}c\|_2+\|\mathbf{M}^\star_{:T}c-\mathbf{M}_{:j}\|_2\\&\le\|[\mathbf{M}-\mathbf{M}^\star]_{:T}\|_{\mathrm{F}}\|c\|_2+\|[\mathbf{M}-\mathbf{M}^\star]_{:j}\|_2\le\frac{\phi}{2\tau}\cdot\tau+\frac{\phi}{2}=\phi.\end{aligned}$$

It is then straightforward to check that c attains objective value $2\phi^2$ as desired. □

To complete our analysis of Test we further specify a set of properties that guarantees it will fail. Intuitively our conditions impose that for a small set of coordinates (which cannot significantly affect subspace regularity), the deviation from the underlying matrix $\mathbf{M}^\star$ on a particular column restricted to those coordinates is substantially larger than it should be.

Lemma 16. *Let $\mathbf{M},\mathbf{M}^\star\in\mathbb{R}^{m\times n}$, $m\ge n$, and $R\subset S\subseteq[m]$, $T\subseteq[n]$, $j\in[n]$ satisfy the following.*

1. $\|[\mathbf{M}-\mathbf{M}^\star]_{S,T}\|_{\mathrm{F}}\le\frac{\phi}{\tau}$.
2. $\left\|[\mathbf{M}-\mathbf{M}^\star]_{S\setminus R,j}\right\|_2\le\phi$.
3. $\|[\mathbf{M}-\mathbf{M}^\star]_{R,j}\|_2\ge\frac{7\phi}{\beta}$.
4. $|S|\ge m-\frac{\alpha n}{2}$ *and* $|R|\le\frac{\alpha n}{2}$.

Then if the column span of $\mathbf{M}^\star$ is (α,β)-regular, $\mathsf{Test}(\mathbf{M},T,j,\phi,\tau)$ will fail.

Proof. Assume for contradiction that Test passes, and let $v \in \mathbb{R}^T$ be the solution. This implies

$$\|v\|_2 \le \sqrt{2}\tau,\ \|\mathbf{M}_{:T}v - \mathbf{M}_{:j}\|_2 \le \sqrt{2}\phi. \tag{4.3}$$

We next observe that

$$\begin{aligned}\left\|\mathbf{M}^\star_{S\setminus R,T}v - \mathbf{M}^\star_{S\setminus R,j}\right\|_2 &\le \left\|\mathbf{M}^\star_{S\setminus R,T}v - \mathbf{M}_{S\setminus R,j}\right\|_2 + \phi \\ &\le \left\|[\mathbf{M}-\mathbf{M}^\star]_{S\setminus R,T}v\right\|_2 + \left\|\mathbf{M}_{S\setminus R,T}v - \mathbf{M}_{S\setminus R,j}\right\|_2 + \phi \le 4\phi,\end{aligned}$$

where the first line used Item 2, and the second used Item 1, $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_{\mathrm{F}}$, the bounds in (4.3), and $1 + 2\sqrt{2} \le 4$. We further have

$$\begin{aligned}\left\|\mathbf{M}^\star_{R,T}v - \mathbf{M}^\star_{R,j}\right\|_2 &\ge \frac{7\phi}{\beta} - \left\|\mathbf{M}^\star_{R,T}v - \mathbf{M}_{R,j}\right\|_2 \\ &\ge \frac{7\phi}{\beta} - \|[\mathbf{M}-\mathbf{M}^\star]_{R,T}v\|_2 - \|\mathbf{M}_{R,T}v - \mathbf{M}_{R,j}\|_2 \ge \frac{4\phi}{\beta},\end{aligned}$$

where the first line used Item 3, and the second line followed similarly to the previous calculation. Finally, note that the column span of $\mathbf{M}^\star_{S:}$ is a $(\frac{\alpha}{2}, \beta)$-regular subspace by the size bound on S from Item 4, and $u := \mathbf{M}^\star_{S,T}v - \mathbf{M}^\star_{S,j}$ is an element of this subspace. However, we have proven $\left\|u_{S\setminus R}\right\|_2 \le \beta \|u\|_2$ by combining the above displays, a contradiction to Lemma 2. □

We now give a consequence of Lemma 16 that handles the case of a randomly chosen subset T.

Lemma 17. *Let $\mathbf{M}, \mathbf{M}^\star \in \mathbb{R}^{m\times n}$ be Δ-close away from an s-RCS matrix, and let $\mathbf{X}+\mathbf{Y}$ be the decomposition in Definition 3. Let $T \subset [n]$ have each element included independently with probability p, and suppose*

$$\|\mathbf{Y}_{:j}\|_2 \ge \frac{100}{\beta}\left(\|\mathbf{X}_{:j}\|_2 + \tau\sqrt{p}\Delta + \phi\right),$$

for some $j \in [n]$. Finally suppose $s \le \frac{\alpha \min(m,n)}{2}$, and $p \le \frac{\alpha}{1000s}$. Then with probability at least 0.9 over the randomness of T, $\mathsf{Test}(\mathbf{M}, T, j, \phi, \tau)$ fails.

Proof. For all $k \in [n]$ let S_k be the support of the $\mathbf{Y}_{:k}$ satisfying $|S_k| \le s$, and let $S := [m]\setminus\bigcup_{k\in T} S_k$. We first condition on the following three events, each of which happens with probability at least 0.99 by Markov's inequality, giving the failure probability via a union bound.

1. $\|\mathbf{X}_{:T}\|_{\mathrm{F}} \le 10\sqrt{p}\Delta$.
2. $|T| \le 100pn$.
3. $\left\|\mathbf{Y}_{S\cap S_j,j}\right\|_2 \ge 0.9\|\mathbf{Y}_{:j}\|_2$.

To see that the last event holds with probability 0.99, we used Markov's inequality and

$$\mathbb{E}\left[\left\|\mathbf{Y}_{S_j\setminus S,j}\right\|_2^2\right] = \mathbb{E}\left[\left\|\mathbf{Y}_{\bigcup_{k\in T} S_k,j}\right\|_2^2\right] \le ps\|\mathbf{Y}_{:j}\|_2^2 \le \frac{1}{1000}\|\mathbf{Y}_{:j}\|_2^2,$$

because for each $i \in S_j$, at most s other columns of $\mathbf{Y}$ have $i \in S_k$ by assumption. Under these events, we now prove Test fails by applying Lemma 16 with parameters ϕ', τ where

$$\phi' := 10\tau\sqrt{p}\Delta + \|\mathbf{X}_{:j}\|_2 + \phi.$$

We will use $R = S \cap S_j$, so Item 4 of Lemma 16 follows from the assumed bound on $s \geq |R|$, and that $|[m] \setminus S| \leq 100pns \leq \frac{\alpha n}{2}$. Item 1 follows because $[\mathbf{M} - \mathbf{M}^\star]_{S,T} = \mathbf{X}_{S,T}$ (as $\mathbf{Y}_{S,T}$ is zero by definition), and $\|\mathbf{X}_{S,T}\|_{\mathrm{F}} \leq \|\mathbf{X}_{:T}\|_{\mathrm{F}} \leq \frac{\phi'}{\tau}$. Item 2 follows because

$$\left\|[\mathbf{M} - \mathbf{M}^\star]_{S \setminus R, j}\right\|_2 = \left\|\mathbf{X}_{S \setminus R, j}\right\|_2 \leq \|\mathbf{X}_{:j}\|_2 \leq \phi'.$$

Finally, Item 3 follows because

$$\left\|[\mathbf{M} - \mathbf{M}^\star]_{S \cap S_j, j}\right\|_2 \geq \left\|\mathbf{Y}_{S \cap S_j, j}\right\|_2 - \|\mathbf{X}_{:j}\|_2 \geq \frac{7\phi'}{\beta}.$$

We note that as $\phi \leq \phi'$, Test failing with parameter ϕ' implies Test with parameter ϕ also fails. □

4.2.3 Finding a representative subset

In this section, we finally analyze our main algorithm, Representative, for finding a representative subset of columns of an iterate $\mathbf{M}$ in the sense of Definition 6, assuming $\mathbf{M}$ is close to $\mathbf{M}^\star$ away from an RCS matrix. We showed in Lemma 15 that this ensures good regression error on completing other columns of $\mathbf{M}$. This property will be used with the representative subset we return in the next Section 4.3 to complete our current matrix (including rows and columns we dropped).

Algorithm 6: Representative$(\mathbf{M}, \phi, p)$

Input: $\mathbf{M} \in \mathbb{R}^{m \times n}$, $\phi \geq 0$, $p \in (0, 1)$
Sample $B_0 \subseteq [n]$ by independently including each $k \in [n]$ with probability p
$\mathsf{count} \leftarrow \mathbb{0}_{B_0}$
$t_{\max} \leftarrow \lceil 40 \log(mn) \rceil$
$\tau \leftarrow \frac{1}{\sqrt{40 \log(r)}}$
for $t \in [t_{\max}]$ **do**
  Sample $T \subseteq [n]$ by independently including each $k \in [n]$ with probability p
  for $j \in B_0$ **do**
    if $\mathsf{Test}(\mathbf{M}, T, j, \phi, \tau)$ **then** $\mathsf{count}_j \leftarrow \mathsf{count}_j + 1$
  end
end
$B \leftarrow B_0$ with all $j \in B_0$ satisfying $\mathsf{count}_j \leq \frac{1}{2} t_{\max}$ removed
return B

Lemma 18. *Let $\mathbf{M} \in \mathbb{R}^{m \times n}$ be given as a rank-r factorization and Δ-close to $\mathbf{M}^\star \in \mathbb{R}^{m \times n}$ away from an s-RCS matrix, where $\mathbf{M}^\star$ is rank-$r^\star$ with (α, β, μ)-standard row and column spans. If $m \geq n$, $r \geq r^\star$, and*

$$s \leq \min\left(\frac{\alpha \min(m,n)}{10^5 r \log(mn)}, \frac{\alpha}{1000p}\right), \ \phi = \Delta\sqrt{p \log m}, \ p \geq \frac{40\mu r^\star}{\beta^2 n} \log(r^\star),$$

then letting $B \subseteq [n]$ be the subset output by Algorithm 6 with parameters $(\mathbf{M}, \phi, p)$, B is a $(\widetilde{\Delta}, \gamma)$-representative subset with respect to $\mathbf{M}, \mathbf{M}^\star$ with probability at least 0.9, for

$$\widetilde{\Delta} := \Delta \cdot 1000p\sqrt{\frac{n}{\beta^2}}, \ \gamma := \sqrt{\frac{p\beta^2}{2}}.$$

Proof. Throughout we follow the parameter settings of τ and $t_{\max}$ in Algorithm 6. We begin by proving the second condition in Definition 6. By Lemma 13 and the upper bound on s, there is a subset $Q \subseteq [n]$ with $|Q| \geq (1 - \frac{\alpha}{10})n$ such that

$$\left\|[\mathbf{M} - \mathbf{M}^\star]_{:Q}\right\|_{\mathrm{F}} \leq \Delta \cdot \frac{\sqrt{\log m}}{13}. \tag{4.4}$$

Further, let $\mathbf{U}_\star \mathbf{\Sigma}_\star \mathbf{V}_\star^\top$ be an SVD of $\mathbf{M}^\star$, and let $Q' \subseteq Q$ be the indices j satisfying both

$$\|[\mathbf{M} - \mathbf{M}^\star]_{:j}\|_2 \leq \Delta \cdot \sqrt{\frac{\log m}{\alpha n}} \leq \frac{\phi}{2}, \ \left\|[\mathbf{V}_\star]_{j:}\right\|_2 \leq \theta := \sqrt{\frac{2\mu r}{n}}.$$

By Markov's inequality and the definition of a regular subspace, we have that $|Q'| \geq (1 - \frac{\alpha}{2})n$. We next claim that every index in $Q' \cap B_0$ will be included in B with probability at least 0.99. It suffices to prove that the conditions of Lemma 15 are met with probability at least 0.9, and then a Chernoff bound shows a majority of the $t_{\max}$ tests will pass with probability $\geq 1 - \frac{1}{100n}$ for each $j \in Q' \cap B_0$. To see this, by Markov's inequality and (4.4), with probability at least 0.99 we have

$$\left\|[\mathbf{M} - \mathbf{M}^\star]_{:T \cap Q}\right\|_{\mathrm{F}} \leq \Delta\sqrt{p \log m} \leq \frac{\phi}{2\tau},$$

and by Lemma 12 with $S \leftarrow Q$, with probability at least 0.99,

$$[\mathbf{V}_\star]_{T\cap Q:}^\top [\mathbf{V}_\star]_{T\cap Q:} \succeq \frac{p\beta^2}{2}\mathbf{I}_{r^\star} \succeq \frac{\theta^2}{\tau^2}\mathbf{I}_{r^\star}.$$

Clearly T contains $T \cap Q$ so the conditions of Lemma 15 are all met with probability 0.9. Therefore, conditioning that $B \supseteq Q' \cap B_0$, and since B_0 is independently sampled from Q', applying Lemma 12 once more with $S \leftarrow Q'$ shows the first condition of Definition 6 is met with probability 0.95.

Now we verify the first of the desired conditions in Definition 6. Let $\mathbf{M} - \mathbf{M}^\star = \mathbf{X} + \mathbf{Y}$ be the promised decomposition from Definition 3, and note that the given bound on s shows the preconditions of Lemma 17 are met. Therefore for every $j \in B$, a Chernoff bound shows that with probability 0.99, the contrapositive of Lemma 17 holds, i.e.

$$\|\mathbf{Y}_{:j}\|_2^2 \leq \frac{30000}{\beta^2}\left(\|\mathbf{X}_{:j}\|_2^2 + \tau^2 p\Delta^2 + \phi^2\right) \leq \frac{30000}{\beta^2}\left(\|\mathbf{X}_{:j}\|_2^2 + 1.5\phi^2\right),$$

where we used $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and the lower bound on ϕ. Summing the above display over $j \in B$ and using $\|u+v\|_2^2 \leq 2\|u\|_2^2 + 2\|v\|_2^2$ with $u \leftarrow \mathbf{X}_{:j}$ and $v \leftarrow \mathbf{Y}_{:j}$, we have

$$\|[\mathbf{M} - \mathbf{M}^\star]_{:B}\|_{\mathrm{F}}^2 \leq \frac{90000pn\phi^2}{\beta^2} + \frac{60000pn}{\beta^2}\|\mathbf{X}_{:B}\|_{\mathrm{F}}^2, \tag{4.5}$$

since $|B| \leq |B_0| \leq 2pn$ with probability at least 0.99 by a Chernoff bound. Finally, the conclusion follows since $\|\mathbf{X}_{:B}\|_{\mathrm{F}}^2 \leq 30p\Delta^2$ with probability at least 0.97 by Markov's inequality, and we union bound over these two events and the prior failure probabilities. □

4.3 Filling in the matrix

4.3.1 Completing columns with a representative subset

In this section, we show that given a representative subset of columns (in the setting of Lemma 18), we can efficiently learn coefficients completing the rest of our iterate $\mathbf{M}$ as combinations of the subset via observations from $\mathbf{M}^\star$. We begin by proving several helper regularity bounds which will allow us to argue that the regression problems we solve are well-conditioned with good probability. Specifically, we analyze the regularity of a (truncated) span of our representative columns.

Lemma 19. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be rank-r with SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, and let $\mathbf{B} \in \mathbb{R}^{m\times n}$ satisfy $\|\mathbf{A}-\mathbf{B}\|_{\mathrm{F}} \le \Delta$. For some $\theta \in (0,1)$ let $\mathbf{B}'$ be the matrix obtained by taking an SVD of $\mathbf{B}$ and dropping singular values smaller than $\frac{\Delta}{\theta}$. Let $\mathbf{U}'\mathbf{\Sigma}'(\mathbf{V}')^\top$ be an SVD of $\mathbf{B}'$. Then the following statements hold.*

1. *$\mathbf{U}'$ has rank at most $2r$.*
2. *$\left\|(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top)u\right\|_2 \le \theta$ for all unit vectors u in the column span of $\mathbf{U}'$.*
3. *$\|\mathbf{A}-\mathbf{B}'\|_{\mathrm{F}} \le \frac{4\sqrt{r}\Delta}{\theta}$.*

Proof. To see the first claim, Lemma 9 (overloading the application with $\mathbf{B} \leftarrow \mathbf{B}-\mathbf{A}$) shows that $\mathbf{B}$ has at most $2r$ singular values more than $\frac{\Delta}{\sqrt{r}}$, so $\mathbf{B}'$ is rank at most $2r$. We move onto the second claim: let $u \in \mathbb{R}^m$ be in the column span of $\mathbf{U}'$. We bound

$$\begin{aligned}\left\|\mathbf{B}^\top\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u\right\|_2 &\ge \left\|\mathbf{V}'\mathbf{\Sigma}'(\mathbf{U}')^\top\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u\right\|_2 \\ &\ge \frac{\Delta}{\theta}\left\|(\mathbf{U}')^\top\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u\right\|_2 \\ &\ge \frac{\Delta}{\theta}u^\top\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u = \frac{\Delta}{\theta}\left\|\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u\right\|_2^2,\end{aligned}$$

where the first inequality followed since $\mathbf{B}' = \mathbf{U}'\mathbf{\Sigma}'(\mathbf{V}')^\top$ drops singular values from $\mathbf{B}$, the second used orthonormality of $\mathbf{V}'$ and our lower bound on $\mathbf{\Sigma}'$, the third used that u is contained in the column span of $\mathbf{U}'$, and the last used that $\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top$ is a projector. On the other hand,

$$\left\|\mathbf{B}^\top\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u\right\|_2 = \left\|(\mathbf{A}-\mathbf{B})^\top\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u\right\|_2 \le \Delta\left\|\left(\mathbf{I}_m - \mathbf{U}\mathbf{U}^\top\right)u\right\|_2$$

where we used $\mathbf{A} = \mathbf{U}\mathbf{U}^\top\mathbf{A}$. The above two displays yield the second claim. To see the third,

$$\left\|\mathbf{B}'-\mathbf{A}\right\|_{\mathrm{op}} \le \left\|\mathbf{B}'-\mathbf{B}\right\|_{\mathrm{op}} + \|\mathbf{B}-\mathbf{A}\|_{\mathrm{F}} \le \frac{2\Delta}{\theta}.$$

Since $\mathbf{B}'-\mathbf{A}$ is rank at most $3r$, the conclusion follows from $\|\mathbf{B}'-\mathbf{A}\|_{\mathrm{F}} \le \sqrt{3r}\,\|\mathbf{B}'-\mathbf{A}\|_{\mathrm{op}}$. □

Applying Lemma 19 then yields a regularity bound on a truncated SVD of our iterate.

Lemma 20. *Let $B \subseteq [n]$ be $(\widetilde{\Delta},\gamma)$-representative with respect to $\mathbf{M}, \mathbf{M}^\star \in \mathbb{R}^{m\times n}$, and assume $\mathbf{M}^\star$ is rank-$r^\star$ with SVD $\mathbf{U}_\star\mathbf{\Sigma}_\star\mathbf{V}_\star^\top$ and (α,β,μ)-standard row and column spans. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be an SVD of $\mathbf{M}_{:B}$ after dropping singular values smaller than $\frac{2\widetilde{\Delta}}{\beta}$. Then the following statements hold.*

1. *$\mathbf{U}$ has rank at most $2r^\star$.*
2. *The column span of $\mathbf{U}$ is $(\alpha,\frac{\beta}{2})$-regular.*
3. *There is a matrix $\mathbf{Y} \in \mathbb{R}^{r\times n}$, where $\mathbf{U} \in \mathbb{R}^{m\times r}$, satisfying*

$$\|\mathbf{U}\mathbf{Y} - \mathbf{M}^\star\|_{\mathrm{F}} \le \frac{8\sqrt{r^\star}\widetilde{\Delta}}{\gamma\beta}.$$

Proof. We are in the setting of Lemma 19 with $\mathbf{A} \leftarrow \mathbf{M}^\star_{:B}$, $\mathbf{B} \leftarrow \mathbf{M}_{:B}$, and $\theta \leftarrow \frac{\beta}{2}$, so the first claim follows. The second claim in Lemma 19 shows any unit u in the column span of $\mathbf{U}$ can be

decomposed as $u = v + w$ where v is the projection of u into the column space of $\mathbf{U}_\star$ and $\|w\|_2 \le \frac{\beta}{2}$. Since v is in the column span of $\mathbf{U}_\star$, Lemma 2 shows that for any $S \subseteq [m]$ with $|S| \ge (1-\alpha)m$,

$$\|v_S\|_2 \ge \beta \|v\|_2 \implies \|u_S\|_2 \ge \|v_S\|_2 - \|w_S\|_2 \ge \beta - \frac{\beta}{2} = \frac{\beta}{2},$$

proving the desired regularity of the column span of $\mathbf{U}$ via Lemma 2. To see the last claim, representativeness of B shows that by taking $\mathbf{Z} = [\mathbf{V}_\star]_{B:}([\mathbf{V}_\star]_{B:}^\top [\mathbf{V}_\star]_{B:})^{-1}\mathbf{V}_\star^\top \in \mathbb{R}^{|B|\times n}$,

$$\|\mathbf{Z}\|_{\mathrm{op}} = \sqrt{\lambda_1(\mathbf{Z}\mathbf{Z}^\top)} = \sqrt{\lambda_1\left(([\mathbf{V}_\star]_{B:}^\top [\mathbf{V}_\star]_{B:})^{-1}\right)} \le \frac{1}{\gamma}.$$

Further, this $\mathbf{Z}$ satisfies $\mathbf{M}^\star = \mathbf{M}^\star_{:B}\mathbf{Z}$. Hence for $\mathbf{Y} = \mathbf{\Sigma}\mathbf{V}^\top\mathbf{Z}$, we have the desired

$$\|\mathbf{U}\mathbf{Y} - \mathbf{M}^\star\|_{\mathrm{F}} = \left\|\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top - \mathbf{M}^\star_{:B}\right)\mathbf{Z}\right\|_{\mathrm{F}} \le \frac{1}{\gamma}\left\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top - \mathbf{M}^\star_{:B}\right\|_{\mathrm{F}} \le \frac{8\sqrt{r^\star}\widetilde{\Delta}}{\gamma\beta}.$$

Above we used the last claim of Lemma 19 and, for any $\mathbf{A}$ with $|B|$ columns,

$$\|\mathbf{A}\mathbf{Z}\|_{\mathrm{F}}^2 = \left\langle \mathbf{A}^\top\mathbf{A}, \mathbf{Z}\mathbf{Z}^\top\right\rangle \le \frac{1}{\gamma^2}\left\langle \mathbf{A}^\top\mathbf{A}, \mathbf{I}_{|B|}\right\rangle = \frac{1}{\gamma^2}\|\mathbf{A}\|_{\mathrm{F}}^2.$$

□

We further require one helper claim on regression error from noisy observations.

Lemma 21. *Let $v = \mathbf{U}y + \xi$ for $\mathbf{U} \in \mathbb{R}^{m\times r}$ with orthonormal columns. Suppose $\mathbf{U}_{A:}^\top\mathbf{U}_{A:} \succeq \lambda^2\mathbf{I}_r$ for $A \subseteq [m]$. Then for $c^\star := \arg\min_{c\in\mathbb{R}^r}\|\mathbf{U}_{A:}c - v_{A:}\|_2$, and any $\widehat{c} \in \mathbb{R}^r$ with $\|\widehat{c} - c^\star\|_2 \le \Delta$,*

$$\|\mathbf{U}\widehat{c} - v\|_2 \le \|\xi\|_2 + \frac{2}{\lambda}\|\xi_{A:}\|_2 + \Delta.$$

Proof. Because setting $c = y$ attains error $\|\xi_{A:}\|_2$, we must have $\|\mathbf{X}_{A:}c^\star - v_{A:}\|_2 \le \|\xi_{A:}\|_2$. The conclusion follows from the assumption on A and the triangle inequality:

$$\begin{aligned}
\|\mathbf{U}\widehat{c} - v\|_2 &\le \|\mathbf{U}y - v\|_2 + \|\mathbf{U}(c^\star - y)\|_2 + \|\mathbf{U}(c^\star - \widehat{c})\|_2 \\
&\le \|\xi\|_2 + \frac{1}{\lambda}\|\mathbf{U}_{A:}(\widehat{c} - y)\|_2 + \Delta \\
&\le \|\xi\|_2 + \frac{1}{\lambda}\|\mathbf{U}_{A:}\widehat{c} - v_{A:}\|_2 + \frac{1}{\lambda}\|\mathbf{U}_{A:}y - v_{A:}\|_2 + \Delta \le \|\xi\|_2 + \frac{2}{\lambda}\|\xi_{A:}\|_2 + \Delta.
\end{aligned}$$

□

Finally, we state a standard result on the runtime of well-conditioned linear regression.

Proposition 5 ([Nes83])**.** *Let $\mathbf{A} \in \mathbb{R}^{d\times r}$ have full column rank, let $b \in \mathbb{R}^d$, and let*

$$x^\star := \arg\min_{x\in\mathbb{R}^r}\|\mathbf{A}x - b\|_2^2.$$

There is an algorithm $\mathsf{AGD}(\mathbf{A}, b, x_0, N)$ which outputs $x \in \mathbb{R}^r$ in time $O(\mathcal{T}_{\mathrm{mv}}(\mathbf{A})\cdot N)$ satisfying $\|x - x^\star\|_2 \le \Delta$, if

$$N \ge \sqrt{\kappa(\mathbf{A}^\top\mathbf{A})}\log\left(\frac{2\kappa(\mathbf{A}^\top\mathbf{A})\|x_0 - x^\star\|_2^2}{\Delta^2}\right).$$

Algorithm 7: $\mathsf{Complete}(\mathcal{O}_p(\widehat{\mathbf{M}}), \mathbf{M}_{:B}, r^\star, B, \Delta, \widetilde{\Delta}, \sigma, \alpha, \beta)$

1 **Input:** $\mathcal{O}_p(\widehat{\mathbf{M}})$ for $p \in (0,1)$ and $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N} \in \mathbb{R}^{m \times n}$ where $\mathbf{M}^\star$ is rank-$r^\star$, $\|\mathbf{N}\|_{\mathrm{F}} \le \Delta$, $\mathbf{M}_{:B} \in \mathbb{R}^{m \times |B|}$, $B \subseteq [n]$, $\widetilde{\Delta}, \sigma \ge 0$, $\alpha, \beta \in (0,1)$

2 $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \leftarrow$ SVD of $\mathbf{M}_{:B}$ with singular values smaller than $\frac{2\widetilde{\Delta}}{\beta}$ dropped, for $\mathbf{U} \in \mathbb{R}^{m \times r'}$

3 $\widehat{\mathbf{V}} \leftarrow \mathbf{0}_{n \times r'}$

4 **if** $r' > 2r^\star$ **then return** $(\mathbf{U}, \widehat{\mathbf{V}})$

5 $R \leftarrow \{i \in [m] \mid \|\mathbf{U}_{i:}\|_2^2 \ge \frac{2r'}{\alpha n}\}$

6 $N \leftarrow \lceil \frac{4}{\beta} \log(\frac{3 \cdot 10^5 r^\star (\Delta^2 + \widetilde{\Delta}^2 + \sigma^2) n}{p \gamma^2 \beta^6 \Delta^2}) \rceil$

7 **for** $j \in [n]$ **do** $S_j \leftarrow A_j \setminus R$ where $A_j \subseteq [m]$ corresponds to revealed entries of $\widehat{\mathbf{M}}_{:j}$

8 **for** $j \in [n]$ **do** $\widehat{\mathbf{V}}_{:j} \leftarrow \mathsf{AGD}(\mathbf{U}_{S_j:}, \widehat{\mathbf{M}}_{S_j,j}, \mathbb{0}_{r'}, N)$ (see Proposition 5)

9 **return** $(\mathbf{U}, \widehat{\mathbf{V}})$

We now analyze our subroutine for learning coefficients with respect to a representative subset.

Lemma 22. *Following notation of Algorithm 7, suppose B is $(\widetilde{\Delta}, \gamma)$-representative with respect to $\mathbf{M}^\star, \mathbf{M} \in \mathbb{R}^{m \times n}$, $\mathbf{M}^\star$ has (α, β)-regular row and column spans, $\|\mathbf{M}^\star\|_{\mathrm{op}} \le \sigma$, $\|\mathbf{N}\|_{\mathrm{F}} \le \Delta$, and $p \ge \frac{500 r^\star}{\alpha \beta^2 n} \log(n)$. Then with probability at least 0.9 over the randomness of $\mathcal{O}_p(\widehat{\mathbf{M}})$,*

$$\left\|\mathbf{U}\widehat{\mathbf{V}}^\top - \mathbf{M}^\star\right\|_{\mathrm{F}} \le \frac{200}{\beta^2} \cdot \left(\Delta + \frac{\sqrt{r^\star}\widetilde{\Delta}}{\gamma}\right) \text{ and } r' \le 2r^\star.$$

Proof. By Lemma 20, whenever B is $(\widetilde{\Delta}, \gamma)$-representative, the algorithm never terminates on Line 4, and the column space of $\mathbf{U}$ is $(\alpha, \frac{\beta}{2})$-regular (and hence $(\alpha, \beta, \frac{3}{\alpha})$-standard). We condition on the following two events, each of which holds with probability at least 0.95, giving the failure probability by a union bound. First, by an application of Fact 3 analogous to its use in proving Lemma 12, since $|R| \le \frac{\alpha n}{2}$, we have for all $j \in [n]$ simultaneously,

$$\frac{p\beta^2}{8}\mathbf{I}_{r'} \preceq \sum_{i \in S_j} u_i u_i^\top \preceq 2p\mathbf{I}_{r'}. \tag{4.6}$$

Second, let $\mathbf{N}' = \mathbf{M}^\star - \mathbf{U}\mathbf{Y}$ be the difference matrix from Lemma 20, so that $\widehat{\mathbf{M}} = \mathbf{U}\mathbf{Y} + \mathbf{N} + \mathbf{N}'$ and $\mathbf{N}'$ is independent of $\mathcal{O}_p(\widehat{\mathbf{M}})$. We will condition on the following via Markov's inequality:

$$\sum_{j \in [n]} \left\|\left[\mathbf{N} + \mathbf{N}'\right]_{S_j,j}\right\|_2^2 \le 20p \left\|\mathbf{N} + \mathbf{N}'\right\|_{\mathrm{F}}^2. \tag{4.7}$$

Under these events, Lemma 20 also proves $\|\mathbf{N}'\|_{\mathrm{F}} \le \frac{8\sqrt{r^\star}\widetilde{\Delta}}{\gamma\beta}$, and by orthonormality of $\mathbf{U}$,

$$\|\mathbf{Y}\|_{\mathrm{F}} = \|\mathbf{U}\mathbf{Y}\|_{\mathrm{F}} \le \|\mathbf{M}^\star\|_{\mathrm{F}} + \left\|\mathbf{N}'\right\|_{\mathrm{F}} \le \sigma\sqrt{r^\star} + \frac{8\sqrt{r^\star}\widetilde{\Delta}}{\gamma\beta}. \tag{4.8}$$

Finally, for all $j \in [n]$ we bound the error of AGD. Let $c_j^\star$ minimize $\|\mathbf{U}_{S_j:}c - \widehat{\mathbf{M}}_{S_j,j}\|_2^2$, and for simplicity let $\mathbf{A}_j := \mathbf{U}_{S_j:}$ and $b_j := \widehat{\mathbf{M}}_{S_j,j}$. By Lemma 21 with $y \leftarrow \mathbf{Y}_{:j}$ and $\xi \leftarrow [\mathbf{N} + \mathbf{N}']_{:j}$,

$$\left\|\mathbf{A}_j c_j^\star - b_j\right\|_2^2 \le 2\left\|\left[\mathbf{N} + \mathbf{N}'\right]_{:j}\right\|_2^2 + \frac{64}{p\beta^2}\left\|\left[\mathbf{N} + \mathbf{N}'\right]_{S_j,j}\right\|_2^2, \tag{4.9}$$

where we used the lower bound $\lambda^2 = \frac{p\beta^2}{8}$ in (4.6). Further, by integrating the lower bound in (4.6),

$$\begin{aligned}\left\|c_j^\star - \mathbf{Y}_{S_j,j}\right\|_2^2 &\le \frac{8}{p\beta^2}\left\|\mathbf{A}_j(c_j^\star - \mathbf{Y}_{S_j,j})\right\|_2^2 \\ &= \frac{8}{p\beta^2}\left(\left\|\mathbf{A}_j\mathbf{Y}_{S_j,j} - b_j\right\|_2^2 - \left\|\mathbf{A}_j c_j^\star - b_j\right\|_2^2\right) \\ &\le \frac{8}{p\beta^2}\left\|\mathbf{U}\mathbf{Y} - \widehat{\mathbf{M}}\right\|_{\mathrm{F}}^2 \le \frac{8}{p\beta^2}\left\|\mathbf{N} + \mathbf{N}'\right\|_{\mathrm{F}}^2,\end{aligned}$$

so plugging in (4.8) gives the crude bound

$$\left\|c_j^\star\right\|_2^2 \le \frac{24}{p\beta^2}\left\|\mathbf{N} + \mathbf{N}'\right\|_{\mathrm{F}}^2 + 3\sigma^2 r^\star + \frac{192 r^\star \widetilde{\Delta}^2}{\gamma^2\beta^2} \le \frac{3300 r^\star}{p\gamma^2\beta^4}\left(\Delta^2 + \widetilde{\Delta}^2 + \sigma^2\right). \tag{4.10}$$

Therefore, by combining (4.9), (4.10), the condition number bound in (4.6), and Proposition 5, running for N iterations yields $\widehat{c}_j := \widehat{\mathbf{V}}_{:j}$ satisfying $\|\widehat{c}_j - c_j^\star\|_2 \le \frac{\Delta}{\sqrt{2n}}$, so by Lemma 21 once more,

$$\left\|\mathbf{U}\widehat{\mathbf{V}}_{:j} - \widehat{\mathbf{M}}_{:j}\right\|_2^2 \le 3\left\|\left[\mathbf{N} + \mathbf{N}'\right]_{:j}\right\|_2^2 + \frac{96}{p\beta^2}\left\|\left[\mathbf{N} + \mathbf{N}'\right]_{S_j,j}\right\|_2^2 + \frac{\Delta^2}{n}.$$

The conclusion follows by summing over all columns and using (4.7) which we conditioned on. □

4.3.2 Geometric aggregation

In this section, we give an aggregation technique for boosting the constant error guarantees of earlier sections. We begin with an approximation algorithm for the distance between low-rank matrices.

Algorithm 8: LowRankDist$(\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{Z}, \delta)$

Input: $\mathbf{U}, \mathbf{W} \in \mathbb{R}^{m\times r}$, $\mathbf{V}, \mathbf{Z} \in \mathbb{R}^{n\times r}$, $\delta \in (0,1)$
$d \leftarrow \lceil 1000 \log \frac{m}{\delta} \rceil$
Sample $\mathbf{Q} \in \mathbb{R}^{d\times m}$ with independently random unit vector rows in $\mathbb{R}^m$
$\widetilde{\mathbf{D}} \leftarrow \frac{1}{\sqrt{d}}(\mathbf{Q}\mathbf{U}\mathbf{V}^\top - \mathbf{Q}\mathbf{W}\mathbf{Z}^\top)$
return $\|\widetilde{\mathbf{D}}\|_{\mathrm{F}}$

Lemma 23. *Let $\mathbf{M}, \mathbf{M}' \in \mathbb{R}^{m\times n}$ be given as rank-r factorizations $\mathbf{U}\mathbf{V}^\top$, $\mathbf{W}\mathbf{Z}^\top$ respectively. For any $\delta \in (0,1)$, LowRankDist$(\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{Z})$ returns a value V such that with probability $\ge 1-\delta$,*

$$|V - \left\|\mathbf{M} - \mathbf{M}'\right\|_{\mathrm{F}}| \le 0.1\left\|\mathbf{M} - \mathbf{M}'\right\|_{\mathrm{F}}.$$

The runtime of the algorithm is $O((m+n)r\log\frac{m}{\delta})$.

Proof. First, letting $\mathbf{D} := \mathbf{U}\mathbf{V}^\top - \mathbf{W}\mathbf{Z}^\top$, standard guarantees on Johnson-Lindenstrauss sketches [DG03] guarantee that with probability at least $1-\delta$,

$$\left|\|\mathbf{D}\|_{\mathrm{F}}^2 - \|\widetilde{\mathbf{D}}\|_{\mathrm{F}}^2\right| \le 0.1\|\mathbf{D}\|_{\mathrm{F}}^2 \implies \left|\|\mathbf{D}\|_{\mathrm{F}} - \|\widetilde{\mathbf{D}}\|_{\mathrm{F}}\right| \le 0.1\|\mathbf{D}\|_{\mathrm{F}},$$

since multiplying by $d^{-\frac{1}{2}}\mathbf{Q}$ preserves all row norms of $\mathbf{D}$ up to a 0.1 factor with this probability. Finally, we can explicitly compute $\widetilde{\mathbf{D}}$ and return its Frobenius norm in time $O((m+n)rd)$. □

Algorithm 9: Aggregate$(\{\mathbf{M}_i\}_{i\in[k]}, \Delta, \delta)$

Input: $\{\mathbf{M}_i\}_{i\in[k]} \subset \mathbb{R}^{m\times n}$ each given as rank-r factorizations $\{\mathbf{U}_i\mathbf{V}_i^\top\}_{i\in[k]}$, $\Delta \geq 0$ such that $\|\mathbf{M}_i - \mathbf{M}^\star\|_{\mathrm{F}} \leq \Delta$ for an unknown $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ and at least $0.51k$ of the $i \in [k]$, $\delta \in (0,1)$
for $(i,j) \in [k]\times[k]$ **do** $d_{ij} \leftarrow$ LowRankDist$(\mathbf{U}_i, \mathbf{V}_i, \mathbf{U}_j, \mathbf{V}_j, \frac{\delta}{k^2})$
for $i \in [k]$ **do**
  if $d_{ij} \leq 2.2\Delta$ for at least $0.51k$ distinct $j \in [k]$ **then return** i
end

Leveraging Lemma 23, we give our approximation-tolerant geometric aggregation technique. The algorithm is identical to Algorithm 4 of [KLL+22] other than our use of approximate distance computations, but we provide an analysis of this modification here for completeness.

Lemma 24. *Under the input assumptions of* Aggregate*, with probability* $\geq 1-\delta$*, an index* i *is returned in time* $O((m+n)rk^2 \log\frac{mk}{\delta})$ *satisfying*

$$\|\mathbf{M}_i - \mathbf{M}^\star\|_{\mathrm{F}} \leq 4\Delta. \tag{4.11}$$

Proof. We condition on all calls to LowRankDist returning a pairwise distance up to 0.1 error, giving the failure probability and runtime via an application of Lemma 23. To prove (4.11), let

$$T := \{i \in [k] \mid \|\mathbf{M}_i - \mathbf{M}^\star\|_{\mathrm{F}} \leq \Delta\}.$$

Note that any $i \in T$ passes the check on Line 4 by the triangle inequality, so the algorithm will return. Further, any index $i \in [k]$ with $\|\mathbf{M}_i - \mathbf{M}^\star\|_{\mathrm{F}} \geq 4\Delta$ will fail the check on Line 4 by the triangle inequality, since its (approximate) distance to any $i \in T$ is too large. □

4.4 Proof of Proposition 3

We now put all the pieces together in Algorithm 10, and prove Proposition 3.

Proposition 3. *Let* $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ *be rank-*$r^\star$ *with* (α,β,μ)*-standard row and column spans,* $m \geq n$*,* $\delta \in (0,1)$ *and let* $S \subseteq [m]$*,* $T \subseteq [n]$ *have* $|S| \geq m - \frac{\alpha n}{9}$*,* $|T| \geq m - \frac{\alpha n}{9}$*. Assume* $\mathbf{M} \in \mathbb{R}^{m\times n}$ *is given as a rank-*r *factorization,* $r \geq r^\star$*,* $\mathbf{M}_{S,T}$ *is* $(\frac{\alpha}{1800\log(m)}, \Delta)$*-close to* $\mathbf{M}^\star_{S,T}$ *on a* γ*-submatrix, and* $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$ *for* $\|\mathbf{M}^\star\|_{\mathrm{op}} \leq \sigma$*,* $\|\mathbf{N}\|_{\mathrm{F}} \leq \frac{\Delta}{20}$*. Algorithm 10 returns* $\mathbf{U} \in \mathbb{R}^{m\times r'}$ *and* $\mathbf{V} \in \mathbb{R}^{n\times r'}$ *satisfying*

$$\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\right\|_{\mathrm{F}} \leq \frac{C_{\mathrm{fix}} r^\star\sqrt{r^\star\log(r^\star)}}{\beta^8}\Delta \text{ and } r' \leq 2r^\star, \tag{4.1}$$

for a universal constant C_{fix}*, with probability* $\geq 1-\delta$*. Algorithm 10 uses* $O(\frac{mr^2\mu^2}{\alpha\beta^4}\log^2(\frac{m}{\beta\delta})\log(\frac{m(\sigma+\Delta)}{\Delta\beta\delta}))$ *time and one call to* $\mathcal{O}_p(\widehat{\mathbf{M}})$ *where for a sufficiently large constant,*

$$p = O\left(\frac{r\mu\log^2(\frac{m}{\beta})\log(\frac{m}{\delta})}{\alpha\beta^2 n}\right).$$

Proof. First, by applying Lemma 11 with $\gamma_{\mathrm{drop}} = \frac{\alpha}{9}$ and $\Delta \leftarrow 1.05\Delta$ (to account for the error due to $\mathbf{N}$), with probability $\geq 1-\frac{\delta}{6}$ we have that $|S'| \geq m - \frac{\alpha n}{3}$ and $|T'| \geq (1-\frac{\alpha}{3})n$, and that $\mathbf{M}_{S',T'}$ and $\widehat{\mathbf{M}}_{S',T'}$ are 2.2Δ-close away from an s-RCS matrix (accounting for $\mathbf{N}$ again), for

$$s := \frac{\alpha\beta^2 n}{15\cdot 10^4\mu r\log m}.$$

Algorithm 10: $\mathsf{Fix}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), \mathbf{M}, r^\star, \sigma, S, T, \Delta, \alpha, \beta, \mu, \delta)$

Input: $\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}})$ for $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N} \in \mathbb{R}^{m\times n}$ where $\mathbf{M}^\star$ is rank-$r^\star$, $\|\mathbf{M}^\star\|_{\mathrm{op}} \le \sigma$ and $\|\mathbf{N}\|_{\mathrm{F}} \le \frac{\Delta}{20}$, $\mathbf{M} \in \mathbb{R}^{m\times n}$ given as a rank-r factorization, $S \subseteq [m]$, $T \subseteq [n]$, $\Delta, \mu \ge 0$, $\alpha, \beta, \delta \in (0,1)$

$p \leftarrow \frac{4.8\cdot 10^5 \mu r \log(m)\log(\frac{600m}{\delta})}{\alpha\beta^2 n}$

$(S', T') \leftarrow \mathsf{Sparsify}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}_{S,T}), \mathbf{M}_{S,T}, \frac{120\sqrt{15\mu r\log(m)}}{\alpha\beta n}\Delta, 1.05\Delta, \frac{\alpha}{1800\log(m)}, \frac{\alpha}{9}, p, \frac{\delta}{6})$

$K \leftarrow \lceil 10\log\frac{6}{\delta}\rceil$

for $k \in [K]$ **do**

  $q \leftarrow \frac{750\mu r^\star}{\beta^2 n}\log(n)$, $q' \leftarrow \frac{750 r^\star}{\alpha\beta^2 n}\log(n)$

  $B_k \leftarrow \mathsf{Representative}(\mathbf{M}_{S',T'}, \frac{14\log(m)}{\beta\sqrt{n}}\cdot\Delta, q)$

  $(\mathbf{U}_k, \mathbf{V}_k) \leftarrow \mathsf{Complete}(\mathcal{O}_{q'}(\widehat{\mathbf{M}}_{S':}), \mathbf{M}_{S',B}, r^\star, B, \frac{\Delta}{20}, \frac{88000\mu r^\star\log(r^\star)}{\beta^3\sqrt{n}}\cdot\Delta, \sigma, \frac{2\alpha}{3}, \beta)$

end

$k^\star \leftarrow \mathsf{Aggregate}(\{\mathbf{U}_k\mathbf{V}_k^\top\}_{k\in[K]}, \frac{10^8 r^\star\sqrt{\log(r^\star)}}{\beta^5}\Delta, \frac{\delta}{6})$

$(\mathbf{U}, \mathbf{V}) \leftarrow (\mathbf{U}_{k^\star}, \mathbf{V}_{k^\star})$

for $k \in [K]$ **do**

  $(\mathbf{V}_k, \mathbf{U}_k) \leftarrow \mathsf{Complete}(\mathcal{O}_{q'}(\widehat{\mathbf{M}}^\top), \mathbf{V}\mathbf{U}^\top, r^\star, S', \frac{\Delta}{20}, \frac{4\cdot 10^8 r^\star\sqrt{\log(r^\star)}}{\beta^5}\Delta, \sigma, \frac{2\alpha}{3}, \beta)$

end

$k^\star \leftarrow \mathsf{Aggregate}(\{\mathbf{U}_k\mathbf{V}_k^\top\}_{k\in[K]}, \frac{10^{10} r^\star\sqrt{r^\star\log(r^\star)}}{\beta^8}\Delta, \frac{\delta}{6})$

return $(\mathbf{U}_{k^\star}, \mathbf{V}_{k^\star})$

Condition on this event for the remainder of the proof. Next, consider one run $k \in [K]$ of the loop from Line 12 to Line 9. It is straightforward to check that for $p \leftarrow \frac{40\mu r^\star\log(r^\star)}{\beta^2 n}$ and $\phi \leftarrow \frac{14\log(m)}{\beta\sqrt{n}}$, the preconditions of Lemma 18 are met because we have 2.2Δ-closeness between $\mathbf{M}_{S',T'}$ and $\mathbf{M}^\star_{S',T'}$ away from an s-RCS matrix, and $\mathbf{M}^\star_{S',T'}$ has $(\frac{2\alpha}{3}, \beta, \mu)$-standard row and column spans. Therefore, with probability ≥ 0.9, B_k is $(\widetilde{\Delta}, \gamma)$-representative with respect to $\mathbf{M}_{S',T'}$ and $\mathbf{M}^\star_{S',T'}$ for

$$\widetilde{\Delta} := \frac{88000\mu r^\star\log(r^\star)}{\beta^3\sqrt{n}}\Delta,\ \gamma := \sqrt{\frac{q\beta^2}{2}}.$$

Under this event, Lemma 22 shows Complete returns a rank-r' factorization $(\mathbf{U}_k, \mathbf{V}_k)$ satisfying

$$\left\|\mathbf{U}_k\mathbf{V}_k^\top - \mathbf{M}^\star\right\|_{\mathrm{F}} \le \frac{10^8 r^\star\sqrt{\log(r^\star)}}{\beta^5}\Delta,$$

with probability ≥ 0.9, and guarantees $r' \le 2r^\star$. Therefore this occurs with probability ≥ 0.8 for each independent run $k \in [K]$. A Chernoff bound shows the preconditions of Aggregate are met with probability $\ge 1 - \frac{\delta}{6}$, and then with probability $\ge 1 - \frac{\delta}{6}$, Lemma 24 implies that on Line 11,

$$\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star_{S':}\right\|_{\mathrm{F}} \le \Delta' := \frac{4\cdot 10^8 r^\star\sqrt{\log(r^\star)}}{\beta^5}\Delta.$$

Next, note that S' is a (Δ', β)-representative subset with respect to any extension of $\mathbf{V}\mathbf{U}^\top$ to $\mathbb{R}^{n\times m}$ and $(\mathbf{M}^\star)^\top$, by subspace regularity and Lemma 2. An analogous argument to the above shows that

with probability $\geq 1 - \frac{\delta}{3}$, applying Complete and Aggregate with the given parameters yields (4.1). Union bounding over all these events, we have a failure probability of $1 - \frac{5\delta}{6}$. We condition on one last event with failure probability $\frac{\delta}{6}$ via standard Chernoff bounds: that the total number of observed entries in Sparsify, and the total number of sampled rows and columns in calls to Representative and Complete, are within constant factors of their expectations.

Regarding the choice of p in the statement, note that the only subroutines which require observations are Sparsify and Complete, and our bound then follows from our parameter choices and Lemma 1 (the dominant term is the $O(\log(\frac{m}{\beta}))$ observation calls used by Sparsify). Finally, we discuss runtime. There are four components to bound: Sparsify, Representative, Complete, and Aggregate. The runtime bottleneck of Sparsify is computing $O(pmn)$ observations $O(\log \frac{m}{\beta})$ times, where each observation takes time $O(r)$ to compute by our low-rank factorization. The runtime of Representative is dominated by $O(\log(m))$ calls to Test, and each call solves a regression problem in a $O(m) \times O(nq)$ matrix, which is within the required budget. The cost of Complete is dominated by running AGD for $O(\frac{1}{\beta} \log(\frac{m(\Delta+\sigma)}{\beta\Delta}))$ iterations for each column, and the total number of nonzero entries among all regression matrices is $O(mnr^\star q')$, assuming $r' \leq 2r^\star$. We remark that in the second application of Complete, we need to take an SVD of an $n \times \Theta(m)$ matrix, but its row space is given as an orthonormal basis, so we may apply Lemma 25 to perform this efficiently. Finally, by an application of Lemma 24, the calls to Aggregate do not dominate the runtime. □

Lemma 25. *Let $\mathbf{M} = \mathbf{U}\mathbf{V}^\top \in \mathbb{R}^{m \times n}$ be given as a rank-r factorization and suppose $\mathbf{U} \in \mathbb{R}^{m \times r}$ has orthonormal columns and $\mathbf{V} \in \mathbb{R}^{n \times r}$. We can compute an SVD of $\mathbf{M}$ in time $O((m+n)r^2)$.*

Proof. Let an SVD be $\mathbf{Z}\mathbf{\Sigma}\mathbf{W}^\top$. The right singular vectors $\mathbf{W}$ are an $n \times r$ matrix with orthonormal columns corresponding to the nonzero eigenvalues of $\mathbf{V}\mathbf{V}^\top$, and we can compute these in the given time by forming $\mathbf{V}^\top\mathbf{V}$, performing eigendecomposition, and multiplying by $\mathbf{V}$. This also yields the diagonal matrix $\mathbf{\Sigma}$. We can then directly compute $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top\mathbf{W}\mathbf{\Sigma}^{-1}$ within the allotted time. □

5 Matrix completion algorithms

5.1 Estimating the operator norm

Our algorithms in Section 4, as well as computation of an initial distance bound, require an estimate on $\|\mathbf{M}^\star\|_{\mathrm{op}}$. We give a simple algorithm for performing this estimation under a boundedness assumption on the noise. We then justify that this noise boundedness assumption is without loss of generality, up to a small overhead in our recovery guarantee.

Algorithm 11: EstimateOpNorm($\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), p, \delta$)

Input: $\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}})$, $p, \delta \in (0,1)$
$T \leftarrow \lceil 20 \log \frac{1}{\delta} \rceil$
for $t \in [T]$ **do**
  $s_t \leftarrow \sqrt{\frac{32}{p\beta^2}} \left\| \mathcal{O}_p(\widehat{\mathbf{M}}) \right\|_{\mathrm{F}}$
end
return median$(\{s_t\}_{t \in [T]})$

Lemma 26. *Assume* $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$ *where* $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ *is rank-*$r^\star$ *with* (α,β,μ)*-standard row and column spans, and* $m \geq n$*. If* $\|\mathbf{N}\|_{\mathrm{F}} \leq \frac{\beta}{10}\|\mathbf{M}^\star\|_{\mathrm{F}}$ *and* $p \geq \frac{30\mu r^\star}{\beta^2 n}\log(n)$*, Algorithm 11 returns a value* V *such that with probability* $\geq 1-\delta$*,* $\|\mathbf{M}^\star\|_{\mathrm{op}} \leq V \leq 2\sqrt{n}\,\|\mathbf{M}^\star\|_{\mathrm{op}}$*.*

Proof. Consider one independent run of the loop in Algorithm 11, and let Ω be the observed entries. With probability at least $\frac{2}{3}$, by Markov's inequality we have

$$\|\mathbf{N}_\Omega\|_{\mathrm{F}}^2 \leq \frac{p\beta^2}{100}\|\mathbf{M}^\star\|_{\mathrm{F}}^2,$$

where we used the assumption on $\|\mathbf{N}\|_{\mathrm{F}}$. Further, let $S_j \subseteq [m]$ be the observed entries in column j for all $j \in [n]$, and let $\mathbf{U}_\star\mathbf{\Sigma}_\star\mathbf{V}_\star^\top$ be an SVD of $\mathbf{M}^\star$. With probability at least $\frac{1}{15n}$ for each $j \in [n]$, by an analogous argument to the lower bound in (4.6) (since adding outer products of rows can only increase the smallest eigenvalue), we have that $\left\|[\mathbf{U}_\star]_{S_j:}v\right\|_2^2 \geq \frac{p\beta^2}{8}\|v\|_2^2$ for all $v \in \mathbb{R}^{r^\star}$. Therefore, by a union bound on this event over all $j \in [n]$ we have with probability at least $\frac{14}{15}$,

$$\|\mathbf{M}_\Omega^\star\|_{\mathrm{F}}^2 = \sum_{j\in[n]}\left\|[\mathbf{M}^\star]_{S_j,j}\right\|_2^2 \geq \frac{p\beta^2}{8}\sum_{j\in[n]}\left\|\mathbf{M}^\star_{:j}\right\|_2^2 = \frac{p\beta^2}{8}\|\mathbf{M}^\star\|_{\mathrm{F}}^2.$$

Combining the above two displays and taking a union bound implies that in each independent run, with probability at least $\frac{3}{5}$, we have

$$\frac{32}{p\beta^2}\left\|\widehat{\mathbf{M}}_\Omega\right\|_{\mathrm{F}}^2 \geq \frac{32}{p\beta^2}\left(\frac{1}{2}\|\mathbf{M}_\Omega^\star\|_{\mathrm{F}}^2 - 2\|\mathbf{N}\|_{\mathrm{F}}^2\right) \geq \|\mathbf{M}^\star\|_{\mathrm{F}}^2,$$

where we applied $(a+b)^2 \geq \frac{1}{2}a^2 - b^2$ entrywise to $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$. Applying a Chernoff bound then implies the median estimate over the runs satisfies the above display with probability $\geq 1-\frac{\delta}{2}$, which gives the upper bound on $\|\mathbf{M}^\star\|_{\mathrm{op}} \leq \|\mathbf{M}^\star\|_{\mathrm{F}} \leq V$. For the lower bound,

$$\left\|\widehat{\mathbf{M}}_\Omega\right\|_{\mathrm{F}}^2 \leq 2\|\mathbf{M}_\Omega^\star\|_{\mathrm{F}}^2 + 2\|\mathbf{N}_\Omega\|_{\mathrm{F}}^2 \leq 3\|\mathbf{M}^\star\|_{\mathrm{F}}^2 \leq 3r^\star\|\mathbf{M}^\star\|_{\mathrm{op}}^2,$$

for each independent run with probability at least $\frac{3}{5}$ by conditioning on the same event on $\mathbf{N}$ as before. A similar Chernoff bound and $\frac{96r^\star}{p\beta^2} \leq 4n$ then yields the upper bound on V. □

Remark 1. *In the regime* $\|\mathbf{N}\|_{\mathrm{F}} \geq \frac{\beta}{10}\|\mathbf{M}^\star\|_{\mathrm{F}}$*, the revealed matrix* $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$ *can equivalently be written as* $\widehat{\mathbf{M}} = \mathbf{0}_{m\times n} + (\mathbf{M}^\star + \mathbf{N})$*, where we treat* $\mathbf{0}_{m\times n}$ *as the target low-rank matrix and* $(\mathbf{M}^\star + \mathbf{N})$ *as the noise. This only increases the target noise level by a* $\frac{11}{\beta}$ *factor.*

5.2 Main result

We are now ready to state our main meta-result for matrix completion.

Theorem 3. *Let* $\mathbf{M}^\star \in \mathbb{R}^{m\times n}$ *be rank-*$r^\star$ *with* (α,β,μ)*-row and column spans,* $m \geq n$*,* $\delta \in (0,1)$*, and let* $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$ *for* $\|\mathbf{N}\|_{\mathrm{F}} \leq \Delta$*. Algorithm 12 returns* $\mathbf{U} \in \mathbb{R}^{m\times r^\star}$ *and* $\mathbf{V} \in \mathbb{R}^{n\times r^\star}$ *satisfying* $\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\right\|_{\mathrm{F}} \leq \frac{(r^\star)^{1.5+o(1)}}{\beta^9}\Delta$*, with probability* $\geq 1-\delta$*. Algorithm 12 uses*

$$O\left(\frac{m(r^\star)^{2+o(1)}\mu^2}{\alpha\beta^{4+o(1)}}\cdot\left(\log^6\left(\frac{m}{\alpha\beta\delta}\right)\log\left(\frac{m\|\mathbf{M}^\star\|_{\mathrm{op}}}{\Delta\beta\delta}\right) + \log^{2.5}\left(\frac{m}{\alpha\beta\delta}\right)\log^2\left(\frac{m\|\mathbf{M}^\star\|_{\mathrm{op}}}{\Delta\beta\delta}\right)\right)\right)$$

Algorithm 12: $\mathsf{MatrixCompletion}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), r^\star, \alpha, \beta, \mu, \Delta, \delta)$

Input: $\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}})$, $r^\star \in \mathbb{N}$, $\mu, \Delta \geq 0$, $\alpha, \beta, \delta \in (0,1)$
$\Delta \leftarrow \frac{11\Delta}{\beta}$
$\sigma \leftarrow \mathsf{EstimateOpNorm}(\widehat{\mathbf{M}}, \frac{30\mu r^\star}{\beta^2 n}\log(n), \frac{\delta}{4})$
$\ell \leftarrow \exp(\sqrt{\log(r^\star \beta^{-1})})$
$K \leftarrow \frac{1}{\log \ell} \cdot \log(2C_{\mathrm{fix}} r^\star \sqrt{r^\star \log(r^\star)}\beta^{-8})$
$\widetilde{\Delta} \leftarrow \sqrt{r^\star}\sigma$
$(\mathbf{U}, \mathbf{V}) \leftarrow (\mathbf{0}_{m\times 0}, \mathbf{0}_{n\times 0})$
$k \leftarrow 0$
$(S, T) \leftarrow ([m], [n])$
$N \leftarrow K \log_2(\frac{\widetilde{\Delta}}{20\ell\Delta})$
$\gamma_{\mathrm{add}} \leftarrow \frac{\alpha}{9\cdot 10^5 \log(\frac{m}{\alpha\beta})\ell^2 K^2}$
while $\widetilde{\Delta} \geq 20\ell\Delta$ **do**
  $(\mathbf{U}, \mathbf{V}, S, T) \leftarrow \mathsf{Descent}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}_{S,T}), [\mathbf{U}\mathbf{V}^\top]_{S,T}, r^\star, \widetilde{\Delta}, \gamma_{\mathrm{add}} k, \gamma_{\mathrm{add}}, \frac{\delta}{4N}, \ell)$
  $\widetilde{\Delta} \leftarrow \frac{\Delta}{\ell}$
  $k \leftarrow k + 1$
  if $k = K$ **then**
    $(\mathbf{U}, \mathbf{V}) \leftarrow \mathsf{Fix}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), \mathbf{U}\mathbf{V}^\top, r^\star, \sigma, S, T, \widetilde{\Delta}, \alpha, \beta, \mu, \frac{\delta}{4N})$
    $\widetilde{\Delta} \leftarrow C_{\mathrm{fix}} r^\star \sqrt{r^\star \log(r^\star)}\widetilde{\Delta}$
    $(S, T) \leftarrow ([m], [n])$
    $k \leftarrow 0$
  end
end
$(\mathbf{U}, \mathbf{V}^\top) \leftarrow$ top $r^\star$ components of an SVD of $\mathsf{Fix}(\mathcal{O}_{[0,1]}(\widehat{\mathbf{M}}), \mathbf{U}\mathbf{V}^\top, r^\star, \sigma, S, T, \widetilde{\Delta}, \alpha, \beta, \mu, \frac{\delta}{4})$ sorted by the corresponding singular value
return $(\mathbf{U}, \mathbf{V})$

time and one call to $\mathcal{O}_p(\widehat{\mathbf{M}})$ *where for a sufficiently large constant,*

$$p = O\left(\frac{(r^\star)^{1+o(1)}\mu}{\alpha\beta^{2+o(1)}n} \cdot \log^6\left(\frac{m}{\alpha\beta\delta}\right)\log\left(\frac{n\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}}{\Delta}\right)\right).$$

Proof. By Remark 1 and the guarantees of Lemma 26, our estimate σ is an upper bound on $\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}$ with probability at least $1 - \frac{\delta}{4}$; we condition on this for the remainder of the proof. This also implies that our initial estimate $\widetilde{\Delta}$ is a valid overestimate of $\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\right\|_{\mathrm{F}} \leq \sqrt{r^\star}\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}$ at the beginning of the algorithm. We next claim that throughout the algorithm, $[\mathbf{U}\mathbf{V}^\top]_{S,T}$ and $\mathbf{M}^\star_{S,T}$ are $\widetilde{\Delta}$-close on a $\gamma_{\mathrm{add}}k$-submatrix. This invariant is preserved every time we call $\mathsf{Descent}$ (assuming it succeeds), by Proposition 1. Further, our parameter settings imply the preconditions of Proposition 3 are met whenever it is called: it is straightforward to check that the γ_{drop} parameter in Proposition 1 is bounded by $\frac{\alpha}{9K}$, so that after K steps, at most an $\frac{\alpha}{9}$ fraction of rows and columns are dropped, and the submatrix parameter is at most $\gamma_{\mathrm{add}}K \leq \frac{\alpha}{1800\log(m)}$. Hence, every time we call Fix (assuming it succeeds) the invariant is also preserved, by the guarantees of Proposition 3.

The above argument also shows that every time the loop in Lines 12 to 22 is executed, $\widetilde{\Delta}$ is

decreased by a factor of $\ell^K \cdot (C_{\text{fix}} r^\star \sqrt{r^\star \log(r^\star)}) = 2$, by combining the guarantees of Proposition 1 (K times) and Proposition 3 (once). This implies that the number of times the loop is executed is at most N. By union bounding over all N calls to Descent and Fix, the last call to Fix, and the first call to EstimateOpNorm, this gives the failure probability; we condition on all of these calls succeeding for the remainder of the proof. When the algorithm exits the loop and before Fix is called for the last time, the closeness parameter (on a submatrix) is bounded by $20\ell\Delta$, so the distance bound follows from Proposition 3 and since we increased Δ by a $\frac{1}{\beta}$ factor at the start of the algorithm. Finally, we note that because the top-$r^\star$ truncation of the output's SVD minimizes the projection to rank-$r^\star$ matrices by Frobenius norm, the distance to $\mathbf{M}^\star$ (which is rank-$r^\star$) can at most double.

Further, note that throughout the algorithm, we can inductively apply Proposition 1 to maintain that the rank r of our iterate is bounded by $3^{k+1} r^\star = (r^\star)^{1+o(1)} \beta^{-o(1)}$, since the potential function $r + r^\star$ at most triples each iteration, and whenever k is reset to 0, Proposition 3 guarantees that $r \le 2r^\star$. The bounds on the runtime and p then follow by combining Propositions 1 and 3 (at most $N+1$ times) with Lemma 26, where we apply Lemma 1 to aggregate the observation probabilities. To handle the runtime of the final SVD and truncation, it suffices to use Lemma 25. □

By combining Theorem 3 with Facts 1 and 2, we then obtain the following results.

Corollary 3. *Let $\mathbf{M}^\star \in \mathbb{R}^{m \times n}$ be rank-$r^\star$ with $(\Omega(1), \Omega(1))$-regular row and column spans, $m \ge n$, $\delta \in (0,1)$, and let $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$ for $\|\mathbf{N}\|_{\mathrm{F}} \le \Delta$. Algorithm 12 returns $\mathbf{U} \in \mathbb{R}^{m \times r^\star}$ and $\mathbf{V} \in \mathbb{R}^{n \times r^\star}$ satisfying $\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\right\|_{\mathrm{F}} \le (r^\star)^{1.5+o(1)}\Delta$, with probability $\ge 1 - \delta$. Algorithm 12 uses*

$$O\left(m(r^\star)^{2+o(1)} \cdot \left(\log^6\left(\frac{m}{\delta}\right) \log\left(\frac{m\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}}{\Delta\delta}\right) + \log^{2.5}\left(\frac{m}{\delta}\right) \log^2\left(\frac{m\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}}{\Delta\delta}\right)\right)\right)$$

time and one call to $\mathcal{O}_p(\widehat{\mathbf{M}})$ where for a sufficiently large constant,

$$p = O\left(\frac{(r^\star)^{1+o(1)}}{n} \cdot \log^6\left(\frac{m}{\delta}\right) \log\left(\frac{n\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}}{\Delta}\right)\right).$$

Corollary 4. *Let $\mathbf{M}^\star \in \mathbb{R}^{m \times n}$ be rank-$r^\star$ with μ-incoherent row and column spans, $m \ge n$, $\delta \in (0,1)$, and let $\widehat{\mathbf{M}} = \mathbf{M}^\star + \mathbf{N}$ for $\|\mathbf{N}\|_{\mathrm{F}} \le \Delta$. Algorithm 12 returns $\mathbf{U} \in \mathbb{R}^{m \times r^\star}$ and $\mathbf{V} \in \mathbb{R}^{n \times r^\star}$ satisfying $\left\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\right\|_{\mathrm{F}} \le (r^\star)^{1.5+o(1)}\Delta$, with probability $\ge 1 - \delta$. Algorithm 12 uses*

$$O\left(m(r^\star)^{3+o(1)}\mu^3 \cdot \left(\log^6\left(\frac{m}{\delta}\right) \log\left(\frac{m\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}}{\Delta\delta}\right) + \log^{2.5}\left(\frac{m}{\delta}\right) \log^2\left(\frac{m\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}}{\Delta\delta}\right)\right)\right)$$

time and one call to $\mathcal{O}_p(\widehat{\mathbf{M}})$ where for a sufficiently large constant,

$$p = O\left(\frac{(r^\star)^{2+o(1)}\mu^2}{n} \cdot \log^6\left(\frac{m}{\delta}\right) \log\left(\frac{n\left\|\mathbf{M}^\star\right\|_{\mathrm{op}}}{\Delta}\right)\right).$$

Acknowledgements

We thank Yeshwanth Cherapanamjeri for communications on the prior work [CGJ17].

References

[CGJ17] Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly optimal robust matrix completion. In *International Conference on Machine Learning*, pages 797–805. PMLR, 2017.

[Che22] Yeshwanth Cherapanamjeri. Personal communication, 2022.

[CP10] Emmanuel J. Candès and Yaniv Plan. Matrix completion with noise. *Proc. IEEE*, 98(6):925–936, 2010.

[CR12] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

[CT10] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[DC20] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274–7301, 2020.

[DG03] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[DSRO15] Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International conference on machine learning*, pages 2332–2341. PMLR, 2015.

[FM99] Alan M. Frieze and Michael Molloy. Splitting an expander graph. *J. Algorithms*, 33(1):166–172, 1999.

[GAGG13] Suriya Gunasekar, Ayan Acharya, Neeraj Gaur, and Joydeep Ghosh. Noisy matrix completion using alternating minimization. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013*, volume 8189 of *Lecture Notes in Computer Science*, pages 194–209. Springer, 2013.

[GC12] Gonca Gürsun and Mark Crovella. On traffic matrix completion in the internet. In *Proceedings of the 12th ACM SIGCOMM Internet Measurement Conference, IMC '12*, pages 399–412. ACM, 2012.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

[Har14] Moritz Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660. IEEE, 2014.

[HJS+22] Baihe Huang, Shunhua Jiang, Zhao Song, Runzhou Tao, and Ruizhe Zhang. Solving SDP faster: A robust IPM framework and efficient implementation. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022*, pages 233–244. IEEE, 2022.

[HW14] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 638–678. JMLR.org, 2014.

[JKL+20] Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. A faster interior point method for semidefinite programming. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 910–918, 2020.

[JKN16] Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *Advances in Neural Information Processing Systems*, 29, 2016.

[JN15] Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1007–1034, 2015.

[KHK22] Varun Kanade, Elad Hazan, and Adam Tauman Kalai. Partial matrix completion. *arXiv preprint arXiv:2208.12063*, 2022.

[KLL+22] Jonathan A. Kelner, Jerry Li, Allen Liu, Aaron Sidford, and Kevin Tian. Semi-random sparse recovery in nearly-linear time. *CoRR*, abs/2203.04002, 2022.

[KMO09] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009*, pages 952–960. Curran Associates, Inc., 2009.

[KMO10] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[LLR95] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15:215–245, 1995.

[LM00] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[MJGP19] Gunjan Mahindre, Anura P. Jayasumana, Kelum Gajamannage, and Randy Paffenroth. On sampling and recovery of topology of directed social networks – a low-rank matrix completion based approach. In *2019 IEEE 44th Conference on Local Computer Networks (LCN)*, pages 324–331, 2019.

[MM15] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 1396–1404, 2015.

[MSS15] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families ii: Mixed characteristic polynomials and the kadison–singer problem. *Annals of Mathematics*, 182(1):327–350, 2015.

[ND14] Nagarajan Natarajan and Inderjit S. Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):60–68, 2014.

[Nes83] Yurii Nesterov. A method for solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

[Rec11] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[RH17] Philippe Rigollet and Jan-Christian Hütter. *High-Dimensional Statistics*. 2017.

[RS05] Jason D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In Luc De Raedt and Stefan Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005)*, volume 119 of *ACM International Conference Proceeding Series*, pages 713–719. ACM, 2005.

[SL16] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[SN07] ACM SIGKDD and Netflix. Proceedings of kdd cup and workshop. https://www.cs.uic.edu/ liub/KDD-cup-2007/proceedings.html, 2007. Accessed: 2023-04-01.

[SQW15] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[Sri13] Nikhil Srivastava. Discrepancy, graphs, and the kadison-singer problem. https://windowsontheory.org/2013/07/11/discrepancy-graphs-and-the-kadison-singer-conjecture-2/, 2013. Accessed: 2023-03-14.

[SY07] Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming, Series B*, (109):367—-384, 2007.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, 2015.

[Ver16] Roman Vershynin. *High-Dimensional Probability, An Introduction with Applications in Data Science*. 2016.

[YPCC16] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. *Advances in neural information processing systems*, 29, 2016.

[ZCZ22] Jialun Zhang, Hong-Ming Chiu, and Richard Y Zhang. Accelerating sgd for highly ill-conditioned huge-scale online matrix completion. *Advances in Neural Information Processing Systems*, 35:37549–37562, 2022.

[ZDG18] Xiao Zhang, Simon Du, and Quanquan Gu. Fast and sample efficient inductive matrix completion via multi-phase procrustes flow. In *International Conference on Machine Learning*, pages 5756–5765. PMLR, 2018.

[ZW19] Shuai Zhang and Meng Wang. Correction of corrupted columns through fast robust hankel matrix completion. *IEEE Transactions on Signal Processing*, 67(10):2580–2594, 2019.

[ZWL15] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. *Advances in Neural Information Processing Systems*, 28, 2015.

A Regularity of random subspaces

In this section, we prove that uniformly random subspaces of $\mathbb{R}^d$ of dimension r are $(\Omega(1), \Omega(1))$-regular with exponentially small failure probability, when $\frac{d}{r}$ is at least a sufficiently large constant. This latter condition is not typically restrictive, as the regime of interest in matrix completion is where $r = o(\min(m, n))$ (otherwise, it is information-theoretically necessary to reveal at least a constant fraction of the matrix, limiting the runtime gains of matrix completion algorithms). Our main helper tool is the following standard concentration bound on the spectra of Wishart matrices.

Lemma 27. *Let $\mathbf{G} \in \mathbb{R}^{d \times r}$ have independent entries $\sim \mathcal{N}(0, 1)$, and assume $\frac{d}{r}$ is sufficiently large. For a universal constant C, $\kappa(\mathbf{G}^\top \mathbf{G}) \le 3$ with probability $\ge 1 - \exp(-Cd)$ (where recall $\kappa(\mathbf{A})$ denotes the condition number of a matrix $\mathbf{A}$).*

Proof. For shorthand let $\mathbf{K} := \mathbf{G}^\top \mathbf{G} \in \mathbb{R}^{r \times r}$, where $\mathbb{E}\mathbf{K} = d\mathbf{I}_r$. Letting N be a maximal 0.1-net of the unit ball in $\mathbb{R}^r$, Lemma 1.18 of [RH17] shows $|N| \le \exp(4r)$. By Exercise 4.3.3 of [Ver16],

$$\|\mathbf{K} - \mathbf{I}_r\|_{\text{op}} \le 1.25 \max_{v \in N} \left| v^\top (\mathbf{K} - \mathbf{I}_r) v \right|,$$

so it suffices to prove that with the desired probability, we simultaneously have $|v^\top \mathbf{K} v - d| \le 0.4d$ for all $v \in N$. For any $v \in N$, $v^\top \mathbf{K} v$ is a chi-squared random variable with d degrees of freedom, so

$$\Pr[|v^\top \mathbf{K} v - d| > 0.4d] \le \exp(-2Cd)$$

for $C \ge \frac{1}{80}$, by Lemma 1 of [LM00]. The conclusion follows from a union bound for $4r \le Cd$. □

Corollary 5. *Let $V \subseteq \mathbb{R}^d$ be a uniformly random subspace of dimension r, where $\frac{d}{r}$ is sufficiently large. For universal constants α and γ, V is $(\alpha, \frac{1}{3})$-regular with probability $\ge 1 - \exp(-\gamma d)$.*

Proof. By the characterization in Lemma 2 (and following its notation), it suffices to prove that for every $S \subset [d]$ with $|S| = \lceil (1-\alpha)d \rceil$, we have $\kappa(\sum_{i \in S} b_i b_i^\top) \le 9$, since taking larger S can only improve the condition number. Let α be a sufficiently small constant such that

$$\binom{d}{d - \lceil (1-\alpha)d \rceil} \le \exp\left(\frac{Cd}{3}\right),$$

which exists following the estimate $\binom{d}{k} \le (\frac{ed}{k})^k$. By rotational symmetry, it suffices to consider $\mathbf{B}_V = \mathbf{K}^{-\frac{1}{2}} \mathbf{G}$, following the notation of Lemma 2. In this case we further have

$$\sum_{i \in S} b_i b_i^\top = \mathbf{K}^{-\frac{1}{2}} \mathbf{K}_S \mathbf{K}^{-\frac{1}{2}} \text{ for } \mathbf{K}_S := \mathbf{G}_{S:}^\top \mathbf{G}_{S:}.$$

Finally, with probability at least $1 - (\exp(-Cd) + \exp(-\frac{Cd}{3})) \ge 1 - \exp(-\gamma d)$ for an appropriate constant γ, $\mathbf{K}$ and $\mathbf{K}_S$ (for all $|S| = \lceil (1-\alpha)d \rceil$ simultaneously) satisfy the conclusion of Lemma 27. Therefore, the claim follows from Lemma 28:

$$\kappa\left(\mathbf{K}^{-\frac{1}{2}} \mathbf{K}_S \mathbf{K}^{-\frac{1}{2}}\right) \le \kappa\left(\mathbf{K}^{-1}\right) \kappa\left(\mathbf{K}_S\right) = \kappa\left(\mathbf{K}\right) \kappa\left(\mathbf{K}_S\right) \le 9.$$

□

Lemma 28. *For any positive definite $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$, $\kappa(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}}) \le \kappa(\mathbf{A}) \kappa(\mathbf{B})$.*

Proof. It suffices to take a ratio of the following bounds:

$$\lambda_1\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\right) = \max_{\|u\|_2=1} u^\top \mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}u \le \lambda_1(\mathbf{A}) \max_{\|v\|_2=1} v^\top \mathbf{B}v = \lambda_1(\mathbf{A})\lambda_1(\mathbf{B}),$$
$$\lambda_d\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\right) = \min_{\|u\|_2=1} u^\top \mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}u \ge \lambda_d(\mathbf{A}) \min_{\|v\|_2=1} v^\top \mathbf{B}v = \lambda_d(\mathbf{A})\lambda_d(\mathbf{B}).$$

□

B One-sided matrix discrepancy bound

In this section, we prove Proposition 4, a one-sided matrix discrepancy bound. Our proof follows from a straightforward application of the resolution of the Kadison-Singer conjecture from Marcus, Spielman and Srivastava. In particular, we use the following result restated from [MSS15].

Proposition 6 (Specialization of Corollary 1.5, [MSS15])**.** *For any $t \in \mathbb{N}$ and $\{u_i\}_{i\in[m]} \subset \mathbb{R}^r$ such that $\sum_{i\in[m]} u_i u_i^\top = \mathbf{I}_r$ and $\|u_i\|_2^2 \le \delta$ for all $i \in [m]$, there is a partition $\{S_j\}_{j\in[t]}$ of $[m]$ with*

$$\left\|\sum_{i\in S_j} u_i u_i^\top\right\|_{\mathrm{op}} \le \left(\frac{1}{\sqrt{t}} + \sqrt{\delta}\right)^2 = \frac{1}{t}\left(1+\sqrt{\delta t}\right)^2 \text{ for all } j \in [t].$$

As a corollary of Proposition 6 we have the following result on splitting an approximation of a multiple of the identity into two pieces; we will later apply this procedure recursively.

Corollary 6. *For any $\{v_i\}_{i\in[m]} \subset \mathbb{R}^r$ and $\lambda > 0$ such that $\sum_{i\in[m]} v_i v_i^\top \approx_\epsilon \lambda \mathbf{I}_r$ and $\|v_i\|_2^2 \le \delta$ for all $i \in [m]$, and $\epsilon \in (0, \frac{1}{4})$, $\delta \in (0, \frac{\lambda}{100})$, there exists a partition $\{S_1, S_2\}$ of $[m]$ such that for all $j \in [2]$,*

$$\sum_{i\in S_j} u_i u_i^\top \approx_{\epsilon+5\sqrt{\delta/\lambda}} \frac{\lambda}{2}\mathbf{I}_r.$$

Proof. Let $\mathbf{M} := \sum_{i\in[m]} v_i v_i^\top$ and let $u_i = \mathbf{M}^{-\frac{1}{2}} v_i$. Note that

$$\sum_{i\in[m]} u_i u_i^\top = \mathbf{M}^{-\frac{1}{2}}\left(\sum_{i\in[m]} v_i v_i^\top\right)\mathbf{M}^{-\frac{1}{2}} = \mathbf{I}_r,$$

and

$$\|u_i\|_2^2 = v_i^\top \mathbf{M}^{-1} v_i \le \frac{1}{\lambda}\exp(\epsilon)\|v_i\|_2^2 \le \frac{\delta}{\lambda}\exp(\epsilon).$$

Applying Proposition 6 to $\{u_i\}_{i\in[m]}$ with $t = 2$ yields a partition $\{S_1, S_2\}$ of $[d]$ such that

$$\begin{aligned}\left\|\sum_{i\in S_j} u_i u_i^\top\right\|_{\mathrm{op}} &\le \frac{1}{2}\left(1+\sqrt{\frac{2\delta}{\lambda}\exp(\epsilon)}\right)^2 = \frac{1}{2}\left(1+2\sqrt{\frac{2\delta}{\lambda}\exp(\epsilon)}+\frac{2\delta}{\lambda}\exp(\epsilon)\right)\\ &\le \frac{1}{2}\left(1+\sqrt{\frac{12\delta}{\lambda}}\right) \le \frac{1}{2}\exp\left(\sqrt{\frac{12\delta}{\lambda}}\right), \text{ for all } j \in [2], \end{aligned} \tag{B.1}$$

where we used that $2\sqrt{2\exp(\frac{1}{4})} + 2\exp(\frac{1}{4}) \cdot \frac{1}{10} \leq \sqrt{12}$. Consequently, for all $x \in \mathbb{R}^r$ we have

$$x^\top \left(\sum_{i \in S_1} u_i u_i^\top\right) x = \|x\|_2^2 - x^\top \left(\sum_{j \in S_2} u_j u_j^\top\right) x \geq \|x\|_2^2 \left(1 - \left\|\sum_{j \in S_2} u_j u_j^\top\right\|_{\text{op}}\right)$$
$$\geq \|x\|_2^2 \left(1 - \frac{1}{2}\left(1 + \sqrt{\frac{12\delta}{\lambda}}\right)\right) = \|x\|_2^2 \cdot \frac{1}{2}\left(1 - \sqrt{\frac{12\delta}{\lambda}}\right).$$

Using $\sqrt{12\delta/\lambda} \leq \frac{1}{2}$, $1 - x \geq \exp(-x - x^2)$ for all $x \in [0, \frac{1}{2}]$, and $\sqrt{12} + 1.2 \leq 5$ we have that

$$1 - \sqrt{\frac{12\delta}{\lambda}} \geq \exp\left(-\sqrt{\frac{\delta}{\lambda}}\left(\sqrt{12} + 12\sqrt{\frac{\delta}{\lambda}}\right)\right) \geq \exp\left(-4\sqrt{\frac{\delta}{\lambda}}\right).$$

Combining with (B.1) then yields $\sum_{i \in S_1} u_i u_i^\top \approx_{5\sqrt{\delta/\lambda}} \frac{1}{2}\mathbf{I}_r$. Since $u_i = \mathbf{M}^{-\frac{1}{2}} v_i$ and $\mathbf{M} \approx_\epsilon \mathbf{I}_r$ the result follows for S_1, and the result for S_2 is symmetric. □

Applying Corollary 6 repeatedly then yields the following result on splitting a decomposition of the identity into smaller pieces, inspired by procedures described in [FM99, Sri13].

Corollary 7. *For any $k \in \mathbb{N}$ and $\{u_i\}_{i \in [m]} \in \mathbb{R}^r$ such that $\sum_{i \in [m]} u_i u_i^\top = \mathbf{I}_r$ and $\|u_i\|_2^2 \leq \delta \leq \frac{1}{1400 \cdot 2^k}$ for all $i \in [m]$, there exists a partition $\{S_j\}_{j \in [2^k]}$ of $[m]$ such that*

$$\sum_{i \in S_j} u_i u_i^\top \approx_{13\sqrt{\delta 2^k}} \frac{1}{2^k}\mathbf{I}_r \text{ for all } j \in [2^\ell].$$

Proof. We prove the result by induction to show that for all $\ell \in [k]$, under the given assumptions we can find a partition $\{S_j^{(\ell)}\}_{j \in [2^\ell]}$ of $[m]$ such that for all $j \in [2^\ell]$,

$$\sum_{i \in S_j^{(\ell)}} u_i u_i^\top \approx_{\epsilon_\ell} \frac{1}{2^\ell}\mathbf{I}_r \text{ where } \epsilon_\ell := \sum_{i \in [\ell - 1]} 5\sqrt{\delta 2^i}.$$

This suffices to prove the result as

$$\epsilon_\ell = 5\sqrt{\delta} \sum_{i \in [\ell-1]} \left(\sqrt{2}\right)^i = 5\sqrt{\delta} \cdot \left(\frac{\left(\sqrt{2}\right)^\ell - 1}{\sqrt{2} - 1}\right) \leq 13\sqrt{\delta 2^\ell}.$$

The base case $\ell = 0$ clearly holds as in this case $2^\ell = 1$, $\epsilon_\ell = 0$, and $\sum_{i \in [d]} u_i u_i^\top = \mathbf{I}_r$. Next, suppose that the claim holds for some $\ell \in [k-1]$. Since $2^\ell \leq \frac{1}{2800\delta}$ and $13\sqrt{2800^{-1}} \leq \frac{1}{4}$ we can apply Corollary 6 2^ℓ times, where in each application $\{v_i\}_{i \in [m]}$ is set to the u_i in some $S_j^{(\ell)}$ and λ is set to $\frac{1}{2^\ell}$. The resulting sets partition $[m]$ into $2^{\ell+1}$ pieces that have the desired properties. □

Leveraging Corollary 7 and a standard, natural splitting argument we prove our main result.

Proposition 4. *Let $\lambda \in [\frac{5600r}{d}, 1)$, let $\mathbf{B} \in \mathbb{R}^{d \times r}$ have orthonormal columns, and denote rows of $\mathbf{B}$ by $\{b_i\}_{i \in [d]} \subset \mathbb{R}^r$. There exists $S \subseteq [d]$ with $|S| \leq d\lambda$ and*

$$\sum_{i \in S} b_i b_i^\top \succeq \frac{\lambda}{8}\mathbf{I}_r.$$

Proof. Let $k \in \mathbb{N}$ be such that $\frac{1}{2^{k+1}} \leq \frac{\lambda}{4} \leq \frac{1}{2^k}$, and let $\{u_i\}_{i\in[m]}$ be formed by replacing every b_i with $\alpha_i := \lceil \|b_i\|_2^2 \cdot \frac{d}{r} \rceil$ copies of $\frac{1}{\sqrt{\alpha_i}} b_i$. Note that each $\|u_i\|_2^2 \leq \delta := \frac{r}{d}$ and

$$m = \sum_{i\in[d]} \alpha_i \leq d + \frac{d}{r} \sum_{i\in[d]} \|b_i\|_2^2 = 2d.$$

Now since $\delta \leq \frac{\lambda}{5600} \leq \frac{1}{1400\cdot 2^k}$, we can apply Corollary 7 to the $\{u_i\}_{i\in[m]}$, and let $T \subseteq [m]$ be the smallest cardinality set in the output partition. This set satisfies $|T| \leq \frac{2d}{2^k} \leq d\lambda$, and

$$\lambda_{\min}\left(\sum_{i\in S} u_i u_i^\top\right) \geq \exp\left(-13\sqrt{\frac{r}{d}\cdot 2^k}\right)\frac{1}{2^k} \geq \frac{1}{2^{k+1}} \geq \frac{\lambda}{8}.$$

Finally, letting $S \subseteq [d]$ consist of all indices of a b_i associated with one of the u_i indexed by T, we have $\sum_{i\in[S]} b_i b_i^\top \succeq \sum_{i\in T} u_i u_i^\top$ and $|S| \leq |T|$ since $b_i b_i^\top$ is the sum of all associated $u_i u_i^\top$. □