Randomly Initialized Alternating Least Squares: Fast Convergence for Matrix Sensing

Kiryung Lee and Dominik Stöger[*][†]

April 26, 2022

We consider the problem of reconstructing rank-one matrices from random linear measurements, a task that appears in a variety of problems in signal processing, statistics, and machine learning. In this paper, we focus on the Alternating Least Squares (ALS) method. While this algorithm has been studied in a number of previous works, most of them only show convergence from an initialization close to the true solution and thus require a carefully designed initialization scheme. However, random initialization has often been preferred by practitioners as it is model-agnostic. In this paper, we show that ALS with random initialization converges to the true solution with ε-accuracy in $O(\log n + \log(1/\varepsilon))$ iterations using only a near-optimal amount of samples, where we assume the measurement matrices to be i.i.d. Gaussian and where by n we denote the ambient dimension. Key to our proof is the observation that the trajectory of the ALS iterates only depends very mildly on certain entries of the random measurement matrices. Numerical experiments corroborate our theoretical predictions.

1. Introduction

1.1. Alternating minimization and low-rank matrix recovery problems

Suppose we are given observations of the form

$$y_i = \langle A_i, X_\star \rangle_F := \mathrm{Tr}\left(A_i^\top X_\star\right), \quad 1 \le i \le m \tag{1}$$

[*] Corresponding author (email: Dominik.Stoeger@ku.de).

[†] KL is with the Department of Electrical and Computer Engineering at the Ohio State University. DS is with the Department of Mathematics and the Mathematical Institute for Machine Learning and Data Science (MIDS) at KU Eichstätt-Ingolstadt. KL was supported in part by NSF CAREER Award CCF 19-43201.

with known measurement matrices $\{A_i\}_{i=1}^m \subset \mathbb{R}^{n_1 \times n_2}$ and our goal is to estimate an unknown low-rank matrix $X_\star \in \mathbb{R}^{n_1 \times n_2}$, i.e., $\operatorname{rank}(X_\star) = r \ll \min\{n_1; n_2\}$. This problem is ubiquitous in many applications such as matrix completion, blind deconvolution, and phase retrieval. We refer to [1] for a comprehensive overview. Different approaches to this problem have been established in the literature ranging from convex methods such as nuclear norm minimization to non-convex methods based on matrix factorization such as gradient descent and alternating minimization.

The method we want to consider in this paper is the Alternating Least Squares (ALS) method. That is, we consider the non-convex loss function

$$f(U, V) := \frac{1}{2m} \sum_{i=1}^{m} \Big(y_i - \langle A_i, UV^\top \rangle_F \Big)^2 , \tag{2}$$

where $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$, and we alternate between updating U and V, i.e.,

$$\begin{aligned} U_{t+1} &= \underset{U \in \mathbb{R}^{n_1 \times r}}{\operatorname{argmin}} \; f(U, V_t), \\ V_{t+1} &= \underset{V \in \mathbb{R}^{n_2 \times r}}{\operatorname{argmin}} \; f(U_{t+1}, V). \end{aligned} \tag{3}$$

In each step, one needs to solve a linear least-squares problem, which can be achieved efficiently via the conjugate gradient method, see, e.g., [2].

For low-rank matrix recovery, the ALS method has first been proposed in [3]. Later, it was shown that given an initialization close to the ground truth, the ALS method converges linearly to the ground truth solution using a near-optimal amount of samples for the Matrix Sensing and Matrix Completion problem [4]. Moreover, it was shown that such an initialization can be constructed via a so-called spectral initialization.

However, while the ALS method is popular among practitioners, they often use a random initialization for the ALS method instead of a spectral initialization, see, e.g., [5]. One advantage is that random initialization is model-agnostic in contrast to spectral methods. However, despite its importance in practice, the convergence of ALS from random initialization remains poorly understood. Existing theory either shows convergence starting from spectral initialization [4, 6] or with resampling, i.e., that for each iteration fresh samples are used, see, e.g., [7, 8].

1.2. Our contribution

In this paper, we show that, if the A_i's are i.i.d. Gaussian measurement matrices and if $X_\star \in \mathbb{R}^{n_1 \times n_2}$ is a rank-one matrix, then ALS with random initialization converges to the ground truth in $O(\frac{\log n_2 + \log(1/\varepsilon)}{\log \log n_2})$ iterations to ε-accuracy using only a near-optimal amount of measurements. Note that the scenario that the ground truth matrix $X_\star$ is a rank-one matrix indeed appears in many applications such as Blind Deconvolution and Phase Retrieval. To the best of our knowledge, this is the first result in the literature that shows that the ALS iterates for low-rank matrix recovery converge to the true solution starting from random initialization (without resampling at each iteration).

In our analysis, we establish that the convergence of ALS can be separated into two distinct phases. In the first phase, we show that, starting from an initialization that is near-orthogonal to the ground truth, the angle between the true solution and the ALS-iterates is decreasing. More precisely, we show that the cosine of this angle is growing at a geometric rate. As soon as our signal is aligned closely enough with the ground truth signal, we enter the second phase. In this phase, our iterates converge linearly to the ground truth. All of this is corroborated by numerical experiments, see Figure 1, which indeed confirm that there is a sharp phase transition between those two phases.

We note that linear convergence in the second phase can essentially be deduced from the aforementioned previous work [4]. Hence, the key difficulty in proving convergence of ALS from random initialization lies in rigorously establishing the fact that the alignment of the iterates with the true signal is increasing in the first phase. One major obstacle is that there exist many saddle points in minimization of the quadratic loss in (2), see [9]. In particular, it is not clear whether the iterates of ALS can avoid such saddle points.

Our analysis establishes that, with high probability, ALS does not get stuck in saddle points in the first phase. For that, we will show that, in the first phase, the ALS iterates are nearly independent of certain entries in the measurement matrices A_i. This allows us to make much stronger statements than what would be possible by, for example, solely relying on the loss landscape of f. To establish the "near-independence" of the iterates to certain entries of the measurement matrices, we will construct an appropriate (virtual) auxiliary sequence. Our construction is inspired by the use of auxiliary sequences in [10] to show convergence of gradient descent from a random initialization in the phase retrieval problem. However, since the ALS method behaves quite differently than gradient descent, the resulting proofs are also quite different.

We believe that the insights and proof techniques developed in this paper will also pave the way for understanding the convergence of ALS starting from random initialization in scenarios where the rank of the underlying signal is larger than one or where more structured measurement matrices are used, for example, in the problem of Blind Deconvolution.

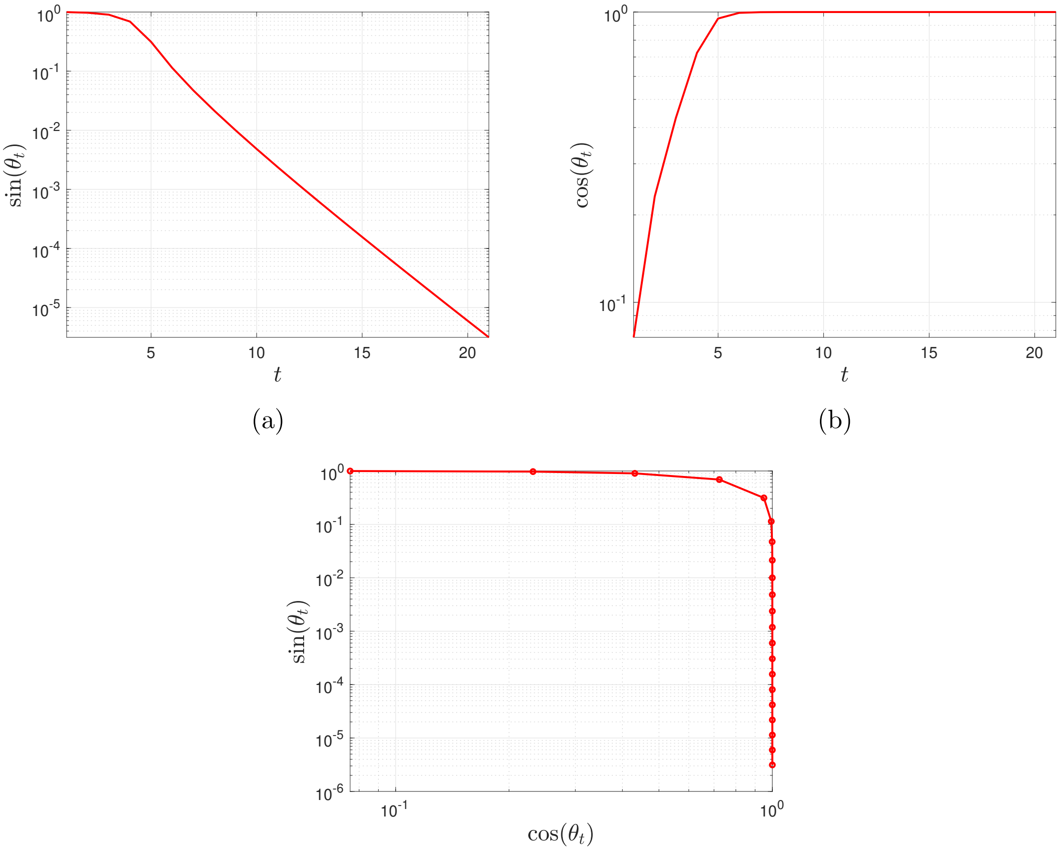
2. Problem formulation

We consider the problem of estimating a rank-one matrix $X_\star \in \mathbb{R}^{n_1 \times n_2}$ from m random linear measurements given by equation (1). In the following, we are going to assume that the measurement matrices $A_i \in \mathbb{R}^{n_1 \times n_2}$ are independent copies of a random matrix with i.i.d. entries following the standard normal distribution $\mathcal{N}(0,1)$.

We define the linear measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ by

$$\mathcal{A}(X) := \left(\frac{1}{\sqrt{m}} \langle A_i, X \rangle_F \right)_{i \in [m]}, \tag{4}$$

where $\langle A_i, X \rangle_F = \operatorname{Tr}(A_i^\top X)$ denotes the Frobenius inner product between $A_i \in \mathbb{R}^{n_1 \times n_2}$ and $X \in \mathbb{R}^{n_1 \times n_2}$ and $[m] := \{1, 2, \dots, m\}$. Since $X_\star$ is a rank-one matrix, we can assume



(a)

(b)

(c)

Figure 1.: Evolution of the iterates by randomly initialized ALS: The size of the ground truth matrix $X_0 = u_\star v_\star^\top$ is given by $n_1 = n_2 = 256$. The number of measurements is $m = 3(n_1 + n_2)$. By t we count the iterates. The estimation error is measured by the angle θ_t between v_t and $v_\star$. (a) $\sin(\theta_t)$ vs number of iterations t; (b) $\cos(\theta_t)$ vs number of iterations t; (c) $\sin(\theta_t)$ vs $\cos(\theta_t)$.

without loss of generality that $X_\star = u_\star v_\star^\top$ for some $u_\star \in \mathbb{R}^{n_1}$ and $v_\star \in \mathbb{R}^{n_2}$. This implies that the equation (1) can be equivalently written as

$$y = \mathcal{A}\left(u_\star v_\star^\top\right).$$

Moreover, note that using this notation equation (2) can be written equivalently as

$$\underset{u,v}{\text{minimize}}\, f(u,v) := \frac{1}{2}\left\|y - \mathcal{A}\left(uv^\top\right)\right\|^2. \tag{5}$$

We will consider a solution to (5) by an Alternating Least Squares (ALS) method given in Algorithm 1.

```
Algorithm 1: Alternating Least Squares
  Input: linear measurement operator $\mathcal{A}: \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, observation vector
         $y \in \mathbb{R}^m$, random initialization $v_0 \in \mathbb{R}^{n_2}$
  for $t = 1, 2, \dots$ do
      $u_{t+1/2} = \mathrm{argmin}_u \|y - \mathcal{A}(uv_t^\top)\|^2$
      $u_{t+1} = u_{t+1/2}/\|u_{t+1/2}\|$
      $v_{t+1/2} = \mathrm{argmin}_v \|y - \mathcal{A}(u_{t+1}v^\top)\|^2$
      $v_{t+1} = v_{t+1/2}/\|v_{t+1/2}\|$
  end
```

Note that compared to (3) there is an additional normalization step in Algorithm 1. However, we have added it only for the sake of convergence analysis and this normalization step is not required for the reconstruction of $X_\star$.

3. Main result

Our main result states that if the initialization vector $v_0 \in \mathbb{R}^{n_2}$ is chosen at random from the sphere with uniform distribution, then ALS converges to the true solution with high probability.

Theorem 1 (Convergence of ALS). *Let $u_\star \in \mathbb{R}^{n_1} \setminus \{0\}$ and $v_\star \in \mathbb{R}^{n_2} \setminus \{0\}$. Let $\mathcal{A}: \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be the measurement operator as defined in (4), where $A_1, \dots, A_m \in \mathbb{R}^{n_1 \times n_2}$ are independent copies of a random matrix whose entries are i.i.d. following $\mathcal{N}(0,1)$. Let the observations in $y \in \mathbb{R}^m$ be given by $y = \mathcal{A}\left(u_\star v_\star^\top\right)$. Let $v_0 \in \mathbb{R}^{n_2}$ be a random initialization vector sampled from the unit sphere with the uniform distribution. Then there exists an absolute constant $C > 0$ such that if the number of measurements m satisfies*

$$m \ge C \max(n_1, n_2) \log^4 n_2, \tag{6}$$

then with probability at least $1 - O(\min(n_1,n_2)^{-1})$ the following holds. For every $\varepsilon > 0$, after

$$t \ge C\left(\frac{\log n_2}{\log\log n_2} + \frac{\log(1/\varepsilon)}{\log\log n_2}\right) \tag{7}$$

iterations, the estimates v_t and u_t from Algorithm 1 satisfy

$$\max\left\{\sin\left(\angle(u_t, u_\star)\right); \sin\left(\angle(v_t, v_\star)\right)\right\} \le \varepsilon.$$

There are a few remarks in order regarding Theorem 1. We first note that the required sample-complexity (6) is optimal up to log-factors. Indeed, the numbers of degrees of freedom of the unknown rank-one matrix $u_\star v_\star^\top \in \mathbb{R}^{n_1 \times n_2}$ is $n_1 + n_2 - 1$ and hence we need to have at least at the order of $\max\{n_1; n_2\}$ measurements in order to recover the underlying ground-truth matrix (see also [11]).

An upper bound on the number of iterations to achieve ε-accuracy is given by inequality (7). As already mentioned in the introduction, our proof shows that convergence can be separated into two distinct phases. Moreover, as it will become clear from our proof the two summands in (7) can be attributed to Phase 1 and Phase 2 as follows

$$\underbrace{C\,\frac{\log n_2}{\log\log n_2}}_{\text{Phase 1}} + \underbrace{C\,\frac{\log(1/\varepsilon)}{\log\log n_2}}_{\text{Phase 2}}. \tag{8}$$

Since the initialization vector v_0 is sampled from the sphere with uniform distribution, we expect that

$$|\langle v_0, v_\star\rangle| \approx 1/\sqrt{n_2}.$$

Hence, we start with an initialization with is near-orthogonal to the ground truth. However, as the (8) shows we only need $O(\frac{\log n_2}{\log\log n_2})$ iterations to obtain an iterate which is closely aligned with the ground truth. After that, we enter the second phase. In this phase, ALS converges linearly to the ground truth as can be seen from the corresponding upper bound on the number of iterations $O\left(\frac{\log(1/\varepsilon)}{\log\log n_2}\right)$.

At the end, we stress again that crucially all of this is proven without the need for sample splitting, i.e., for each ALS step the same measurements are used.

4. Related Work and Discussion

There has been a flurry of work on low-rank matrix recovery over the last fifteen years. For this reason, we will only provide a selective overview of the topic, highlighting the results which are most relevant to our work. In fact, many different algorithmic approaches have been proposed for the low-rank matrix recovery problem. The nuclear norm minimization approach [12] has been studied for Matrix Completion in [13, 14, 15], for Phase Retrieval in [16, 17, 18], for Robust PCA in [19], and for Blind Deconvolution in [20] as well as its extension to the Blind Demixing problem in [21, 22]. We refer also to the overview article [23] for further pointers to the literature. Several other approaches, which have been proposed in the literature, are the projected gradient method [24], the iterative greedy algorithm [25], and the Iteratively Reweighted Least Squares (IRLS) algorithm [26, 27].

In recent years, there has been a flurry of work on non-convex approaches based on matrix factorization due to their small memory footprint and their low computational burden. These approaches can roughly be categorized as first-order methods based on gradient descent, e.g. [28, 29], and as methods based on alternating least squares (ALS) [4], which is also the method studied in this paper. We refer to [30] for an overview of non-convex approaches based on matrix factorization.

Non-convex gradient descent: Non-convex methods based on gradient descent have been studied for the Matrix Sensing problem [29], for Blind Deconvolution [31, 32], its extension to Blind Demixing [33] as well as for the Phase Retrieval problem [28, 34]. However, all of these papers above only guarantee local convergence for gradient descent. That is, convergence is only guaranteed if one picks initialization in a neighborhood of the true solution. In most of these works, such an initialization is constructed via a so-called spectral initialization.

To obtain more insights into the global convergence properties of non-convex gradient descent based on matrix factorization people started to analyse the landscape of the loss function. More precisely, this line of research tries to show that the landscape is benign in the sense that (i) all local minima are in fact global minima and (ii) saddle-points have at least one direction of strictly negative curvature. For the matrix sensing problem [9], for the phase retrieval problem [35], and for the matrix completion problem [36, 37] it has been shown that the landscape of the loss function is benign. In [38] it has been shown that properties (i) and (ii) already imply convergence of gradient descent to a global minimum. However, [39] provides an example, that shows that this property does not rule out exponentially slow convergence. In particular, this means that properties (i) and (ii) do not guarantee convergence in polynomial time. Motivated by this, in [10] the authors showed that in the Phase Retrieval problem with Gaussian measurement vectors gradient descent converges to the ground truth starting from random initialization by using a near-optimal amount of iterations and measurements. In the case of symmetric low-rank matrix sensing, this was also shown in [40, 41]. However, these results require a random initialization which is chosen sufficiently small. For the asymmetric scenario, similar results [42, 43] have only recently been obtained for the population loss case. It remains an open problem to show an analogous result in the finite sample case.

Alternating Least Squares: In general, ALS has been widely used in a broad class of applications including low-rank approximation of data [44] and imaging [45]. In the context of low-rank matrix recovery, ALS approaches are arguably less well studied than methods based on gradient descent. There are several papers that study ALS (or some variants) for the matrix completion problem. However, these works either require fresh samples at each other iteration [7, 8, 46] or they show local convergence starting from a spectral initialization [6].

In [47, 48, 49], the authors propose to use alternating minimization combined with a projection step to recover a rank-one matrix with sparse entries from linear random measurements. However, their analysis requires an initialization close to the ground truth, which is a major bottleneck in the analysis. It is an interesting avenue for future work to see whether our analysis can also be extended to this algorithm.

For the phase retrieval problem, the Error Reduction (ER) algorithm has been proposed [50, 51]. While this method can be interpreted as an alternating minimization method, it is different from the ALS algorithm studied in this paper. Local convergence from a spectral initialization for the ER algorithm, in a setting where the measurements are Gaussian, has been first established in [52], the analysis in this paper requires fresh samples for each iteration. This assumption has been removed by Waldspurger in [53], which showed local converge without sample splitting. Convergence from a random initialization has been established in [54], however using a (suboptimal) sample size at the order of $n^{3/2}$.

The above discussion illustrates that our understanding of global convergence of non-convex methods in low-rank matrix recovery is still in its infancy. This paper contributes to this line of research by establishing the first convergence result from random initialization for the ALS method.

Auxiliary sequences: As already discussed in the introduction, in this paper, we construct a (virtual) auxiliary sequence to establish mild dependence of our ALS iterates on certain entries of the measurement matrices. For optimization tasks, such auxiliary sequences appeared before in [55], where the authors used a slightly different construction (leave-one-out sequences) to establish that the iterates depend only weakly on the individual measurements. In [56, 57], the authors used leave-one-out sequences to show that gradient converges fast to the global optimum, when initialized in a local neighborhood, in several low-rank matrix recovery problems. In [58], leave-one-out sequences were used to improve bounds for the required sample complexity of the nuclear norm minimization approach in matrix completion.

5. Proof ideas and auxiliary sequences

In this section, we illustrate the main ideas for proving Theorem 1. We will also introduce some necessary notation. Moreover, we will define a (virtual) auxiliary sequence, which will be a key ingredient in our proof.

5.1. Notation

Without loss of generality, we assume throughout the proof that $\|u_\star\| = \|v_\star\| = 1$. Furthermore, we set $n := \max(n_1, n_2)$. Moreover, the following shorthand notations will be used throughout this section. We consider the orthogonal decomposition of u_t given by $u_t = u_t^{\|} + u_t^{\perp}$, where $u_t^{\|} := \mu_t u_\star$ with $\mu_t := \langle u_\star, u_t \rangle$ and $u_t^{\perp} := u_t - u_t^{\|}$ denote the projection of u_t into the subspace spanned by $u_\star$ and its orthogonal complement. Consequently, $\|u_t^{\|}\|$ and $\|u_t^{\perp}\|$ respectively correspond to the cosine and sine of the angle between u_t and $u_\star$. These will be used as metrics for convergence. Similarly, v_t is decomposed as $v_t = v_t^{\|} + v_t^{\perp}$, where $v_t^{\|} := \lambda_t v_\star$ with $\lambda_t := \langle v_\star, v_t \rangle$ and $v_t^{\perp} := v_t - v_t^{\|}$. In an analogous fashion, we set $\mu_{t+1/2} = \langle u_\star, u_{t+1/2} \rangle$. Then we have that $u_{t+1/2} = u_{t+1/2}^{\|} + u_{t+1/2}^{\perp}$, where $u_{t+1/2}^{\|} := \mu_{t+1/2} u_\star$ and $u_{t+1/2}^{\perp} := \mu_{t+1/2} - u_{t+1/2}^{\|}$.

By $C > 0$ we denote an absolute numerical constant, whose value may change from line to line.

5.2. First-order necessary conditions

Suppose that $v_t \in \mathbb{R}^{n_2}$ is given and that u_{t+1} is calculated via Algorithm 1. Then it must hold that

$$\nabla_u f\left(u_{t+1/2}, v_t\right) = 0.$$

By explicitly calculating the gradient it follows that

$$\left[\mathcal{A}^* \mathcal{A}\left(u_{t+1/2} v_t^\top - u_\star v_\star^\top\right)\right] v_t = 0.$$

Note that by using $\|v_t\| = 1$ this expression can be rearranged as

$$u_{t+1/2} = \langle v_t, v_\star \rangle u_\star + \left[(\mathrm{Id} - \mathcal{A}^* \mathcal{A})\left(u_{t+1/2} v_t^\top - u_\star v_\star^\top\right)\right] v_t. \tag{9}$$

This identity will be used frequently in our analysis.

5.3. Analysis in population loss

To gain some intuition, we first consider the scenario where the number of samples m is going to infinite, i.e., the population loss scenario. Note that since $A_1, \ldots, A_m$ are independent copies of a random matrix with i.i.d. standard Gaussian entries, it follows that in the scenario the measurement operator $\mathcal{A}$ is isotropic, i.e., $\mathbb{E}\left[\mathcal{A}^* \mathcal{A}\right] = \mathrm{Id}$. Hence, it follows from Equation (9) that in this case

$$u_{t+1/2} = \langle v_\star, v_t \rangle u_\star. \tag{10}$$

This implies that a single step of Algorithm 1 exactly recovers $u_\star$ up to a scale factor (under the assumption that $\langle v_\star, v_t \rangle \neq 0$). The update on $v_{t+1/2}$ that follows will provide $u_{t+1} v_{t+1/2}^\top = u_\star v_\star^\top$. In other words, ALS from any nondegenerate initialization converges in a single iteration.

5.4. Analysis in the finite-sample scenario

At the sample level, the normal equation in (9) deviates from the population-level equation (10) by the factor $\left[(\mathrm{Id} - \mathcal{A}^* \mathcal{A})\left(u_{t+1/2} v_t^\top - u_\star v_\star^\top\right)\right] v_t$. For this reason, we do not expect that one iteration will recover the signal as in the population loss scenario. Nevertheless, in the first convergence phase we aim to show that

$$\|u_{t+1}^{\|}\| = \frac{\|u_{t+1/2}^{\|}\|}{\|u_{t+1/2}\|} \gg \|v_t^{\|}\|, \tag{11}$$

meaning that the iterates become more aligned with the ground truth in each iteration. To show this, we first decompose $u_{t+1/2}$ into its parallel and its perpendicular part, i.e., $u_{t+1/2} = u_{t+1/2}^{\|} + u_{t+1/2}^{\perp}$. We obtain that

$$u_{t+1/2}^{\|} = \langle v_t, v_\star \rangle u_\star + \langle u_\star, \left[(\mathrm{Id} - \mathcal{A}^* \mathcal{A})\left(u_{t+1/2} v_t^\top - u_\star v_\star^\top\right)\right] v_t \rangle u_\star \tag{12}$$

and

$$u_{t+1/2}^{\perp} = \left(\mathrm{Id} - u_\star u_\star^\top\right) \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top\right)\right] v_t.$$

A standard approach to deal with the deviation term is to invoke the well-known Restricted Isometry Property (RIP), see Section 6.1 as well as Lemma 9, which yields

$$\|u_{t+1/2}^{\perp}\| \leq \frac{\delta}{1-\delta} \|v_t^{\perp}\| \tag{13}$$

as well as

$$\left\|u_{t+1/2}^{\|} - \langle v_\star, v_t\rangle u_\star\right\| \leq \frac{\delta}{1-\delta} \|v_t^{\perp}\| \tag{14}$$

for a RIP-constant $0 < \delta < 1$. While inequality (13) will turn out to be sufficient to show (11), inequality (14) will not suffice. The reason is that ideally we would like to have that

$$\|u_{t+1/2}^{\|}\| \approx \|v_t^{\|}\| = |\langle v_\star, v_t\rangle|. \tag{15}$$

However, this does not follow from (14). The reason for this is that we start from random initialization, which yields that $v_0 \in \mathbb{R}^{n_2}$ is almost orthogonal to the ground truth $v_\star$ in the sense that $\|v_0^{\|}\| = |\langle v_0, v_\star\rangle| \approx 1/\sqrt{n_2}$ (and, consequently $\|v_0^{\perp}\|$ is very close to 1).

In particular, this implies that (14) is rather vacuous. Hence, we need to find other approaches to deal with the expression

$$\left|\langle u_\star, \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top\right)\right] v_t\rangle u_\star\right| \tag{16}$$

in (12). Note that we obtained inequality (14) via the Restricted Isometry Property (RIP), which is a uniform bound, i.e., it holds for all vectors $u_{t+1/2} \in \mathbb{R}^{n_1}$ and $v_t \in \mathbb{R}^{n_2}$. In particular, it may be suboptimal for particular choices of v_t and $u_{t+1/2}$. For example, assume for a moment that v_t and $u_{t+1/2}$ would be independent of the measurement operator $\mathcal{A}$ (which of course is not the case). Under this assumption we could hope to derive much stronger concentration bounds than what could be obtained by a uniform estimate induced by the Restricted Isometry Property.

The key insight is that we can indeed establish that $u_{t+1/2}$ and v_t are *nearly independent* of certain entries of the measurement matrices $\{A_i\}_{i=1}^m$, which will allow us to go beyond the suboptimal estimates obtained via the Restricted Isometry Property.

More precisely, to show this *near-independence*, we introduce a new set of measurement matrices $\{\tilde{A}_i\}_{i=1}^m$, which are obtained by substituting partial entries of the original measurement matrices as independent copies. This allows us to define a new measurement operator $\tilde{\mathcal{A}}$, which is constructed using the new measurement matrices $\{\tilde{A}_i\}_{i=1}^m$. Then an auxiliary sequence of estimates $(\tilde{u_t}, \tilde{v_t})$ is obtained from the ALS algorithm starting from the same random initialization v_0, but replacing $\mathcal{A}$ with the new measurement operator $\tilde{\mathcal{A}}$. For a detailed and precise description of the construction of this auxiliary sequence, we refer to the next subsection.

Next, we are going to establish that the trajectory of the auxiliary sequence will stay close to the trajectory the original sequence. Using this property, we expect that we can replace the expression (16) by

$$\left|\langle u_\star, \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A})\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top - u_\star v_\star^\top\right)\right]\tilde{v}_t\rangle u_\star\right|$$

as we expect those terms to be nearly the same. By leveraging that $\tilde{u}_t$ and $\tilde{v}_t$ are independent of certain entries of $\{A_i\}_{i=1}^m$, we can now derive much stronger estimates for the above expression than what would be possible by solely relying on the RIP. These estimates allow us to show (15), from which we can in turn deduce (11). By inductively repeating these arguments we obtain that our iterates become more and more aligned with the ground truth signal until we enter the second convergence phase.

To show convergence in the second phase we then rely on well-known estimates induced by the Restricted Isometry Property of the measurement operator $\mathcal{A}$.

5.5. Auxiliary sequences

As our measurements follow a rotation-invariant distribution, we can assume without loss of generality that $u_\star = e_1 \in \mathbb{R}^{n_1}$ and $v_\star = e_1 \in \mathbb{R}^{n_2}$. Here, with a slight abuse of notation, e_1 denotes the first standard basis vector such that the first entry is 1 and the other entries are 0. The ambient dimension will be clear from the context. We introduce an auxiliary measurement operator $\tilde{\mathcal{A}}$, which is defined by

$$\tilde{\mathcal{A}}\left(X\right) := \left(\frac{1}{\sqrt{m}}\langle \tilde{A}_i, X\rangle_F\right)_{i\in[m]}$$

with the matrix $\tilde{A}_i$ given by

$$(\tilde{A}_i)_{j,k} := \begin{cases} (A_i)_{j,k} & \text{if } (j \neq 1 \text{ and } k \neq 1) \text{ or } (j,k) = (1,1)\,, \\ (\hat{A}_i)_{j,k} & \text{else,} \end{cases}$$

where $(\hat{A}_i)_{j,k}$ are independent copies of $(A_i)_{j,k}$. We observe that it follows directly from the definition of the operator that

$$y = \tilde{\mathcal{A}}\left(u_\star v_\star^\top\right) = \mathcal{A}\left(u_\star v_\star^\top\right).$$

For our analysis we will need the following auxiliary sequences $\{\tilde{u}_t\}_t$ and $\{\tilde{v}_t\}_t$. They are computed via the same algorithm as $\{u_t\}$ and $\{v_t\}$ except that the measurement operator $\mathcal{A}$ is replaced by $\tilde{\mathcal{A}}$. We set $\tilde{v}_0 = v_0$, that is, the auxiliary sequences start from the same initialization. Then for $t \geq 0$, the auxiliary sequences are iteratively updated by alternating least squares in the following four steps: Given $\tilde{v}_t$, $\tilde{u}_t$, the updates are computed via

$$\tilde{u}_{t+1/2} := \operatorname*{argmin}_{u\in\mathbb{R}^{n-1}} \left\|y - \tilde{\mathcal{A}}\left(u\tilde{v}_t{}^\top\right)\right\|^2, \qquad \tilde{u}_{t+1} := \frac{\tilde{u}_{t+1/2}}{\|\tilde{u}_{t+1/2}\|},$$

$$\tilde{v}_{t+1/2} := \operatorname*{argmin}_{v\in\mathbb{R}^{n_2}} \left\|y - \tilde{\mathcal{A}}\left(\tilde{u}_t v^\top\right)\right\|^2, \qquad v_{t+1} := \frac{\tilde{v}_{t+1/2}}{\|\tilde{v}_{t+1/2}\|}.$$

Let $\tilde{f}:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\to\mathbb{R}$ be defined by

$$\tilde{f}\left(u,v\right):=\frac{1}{2}\left\|y-\tilde{\mathcal{A}}\left(uv^\top\right)\right\|^2.$$

Then its gradients with respect to u and v are respectively given by

$$\begin{aligned}\nabla_u\tilde{f}\left(u,v\right)&=\tilde{\mathcal{A}}^*\left(\tilde{\mathcal{A}}\left(uv^\top\right)-y\right)v,\\ \nabla_v\tilde{f}\left(u,v\right)&=\left[\tilde{\mathcal{A}}^*\left(\tilde{\mathcal{A}}\left(uv^\top\right)-y\right)\right]^\top u.\end{aligned}$$

We will now introduce some additional definitions, which will ease the notation in our proofs. For each $i\in[m]$, we consider the decomposition $A_i=D_i+O_i$, where

$$\begin{aligned}D_i&=u_\star u_\star^\top A_iv_\star v_\star^\top+(I_{n_1}-u_\star u_\star^\top)A_i(I_{n_2}-v_\star v_\star^\top),\\ O_i&=u_\star u_\star^\top A_i(I_{n_2}-v_\star v_\star^\top)+\left(I_{n_1}-u_\star u_\star^\top\right)A_iv_\star v_\star^\top.\end{aligned}$$

Moreover, we set

$$\tilde{O}_i=u_\star u_\star^\top\tilde{A}_i(I_{n_2}-v_\star v_\star^\top)+\left(I_{n_1}-u_\star u_\star^\top\right)\tilde{A}_iv_\star v_\star^\top.$$

We observe that it follows directly from these definitions that for all $i\in[m]$

$$\begin{aligned}A_i&=D_i+O_i,\\ \tilde{A}_i&=D_i+\tilde{O}_i.\end{aligned}$$

This allows us define the following linear operators

$$\begin{aligned}\mathcal{D}\left(X\right)&:=\left(\frac{1}{\sqrt{m}}\langle D_i,X\rangle_F\right)_{i\in[m]},\\ \mathcal{O}\left(X\right)&:=\left(\frac{1}{\sqrt{m}}\langle O_i,X\rangle_F\right)_{i\in[m]},\\ \tilde{\mathcal{O}}\left(X\right)&:=\left(\frac{1}{\sqrt{m}}\langle\tilde{O}_i,X\rangle_F\right)_{i\in[m]}.\end{aligned}$$

Note that it follows immediately from these definitions that $\mathcal{A}$ and $\tilde{\mathcal{A}}$ can be decomposed as

$$\mathcal{A}=\mathcal{D}+\mathcal{O}\quad\text{and}\quad\tilde{\mathcal{A}}=\mathcal{D}+\tilde{\mathcal{O}}.$$

Throughout the proof we need to show that the original sequence and the true sequence stay close to each other. For that, we will establish that the inequalities

$$\max\left\{\|u_{t+1}^{\|}-\tilde{u}_{t+1}^{\|}\|;\|u_{t+1}^{\perp}-\tilde{u}_{t+1}^{\perp}\|\right\}\le c_{2t+1}\|u_{t+1}^{\|}\|$$

and

$$\max\left\{\|v_{t+1}^{\|}-\tilde{v}_{t+1}^{\|}\|;\|v_{t+1}^{\perp}-\tilde{v}_{t+1}^{\perp}\|\right\}\le c_{2t+2}\|v_{t+1}^{\|}\|$$

hold (see Lemma 13), where c_t is defined as

$$c_t:=\left(1+\frac{1}{\log n_2}\right)^t-1 \tag{17}$$

for any natural number t. Note that this implies that in the first few iterations, where $\|u_{t+1}^{\|}\|$, respectively $\|v_{t+1}^{\|}\|$, is small, the original iterates and the iterates from the auxiliary sequence are close to each other. In particular, this shows that, in the beginning, the ALS trajectories (or the virtual trajectories) do depend only mildly on $\{O_i\}_{i=1}^m$, respectively $\{\tilde{O}_i\}_{i=1}^m$.

As already noted in Section 4, in [10] an auxiliary sequence with similar properties has been constructed for the analysis of gradient descent for the phase retrieval problem. However, as the algorithms under consideration are quite different, the proofs which show that the auxiliary sequences stay close too each other are quite different. As it turns out, a key difficulty in our proof lies in showing that the auxiliary sequence and the original sequence are still close after the normalization step (see Lemma 13 and its proof in Appendix B.4).

6. Proof of Theorem 1

In this section, we will provide the details for the proof of Theorem 1. We first list several concentration inequalities, which will be used throughout the proof. They are consequences of the Restricted Isometry Property (RIP) of the measurement operator $\mathcal{A}$ and also of the near-independence of auxiliary sequences from the measurement matrices. Then the main proof arguments will be built upon these results.

6.1. Concentration inequalities

We proceed with the proof of Theorem 1 under a set of events on $\mathcal{A}$ and $\tilde{\mathcal{A}}$, which hold with high probability. These events are stated in Lemmas 3, 4, and 5, whose proofs are deferred to the appendix. First note that the linear operator $\mathcal{A}$ satisfies the restricted isometry property.

Lemma 1 (A special case of [59, Theorem 2.3])**.** *Let $\mathcal{A}$ be the linear operator defined in* (4)*. There exists a numerical constant C_0 such that if*

$$m\ge C_0\delta^{-2}\max(n_1,n_2),$$

then with probability at least $1-O\left(\exp(-cm)\right)$

$$(1-\delta)\|Z\|_F^2\le\|\mathcal{A}\left(Z\right)\|^2\le(1+\delta)\|Z\|_F^2 \tag{18}$$

holds for all matrices $Z\in\mathbb{R}^{n_1\times n_2}$ with rank at most 4*.*

The following results, whose proof is deferred to Appendix A.1, are direct consequences of the restricted isometry property and will be used throughout the remainder of the proof.

Lemma 2. *Suppose that $\mathcal{A}$ satisfies the restricted isometry property in* (18) *with constant $\delta > 0$. Then for all $u \in \mathbb{R}^{n_1}, v \in \mathbb{R}^{n_2}$, we have*

$$\left\| \mathcal{O}^* \mathcal{D} \left(uv^\top \right) \right\| \leq \delta \|u\| \|v\|, \tag{19}$$

$$\left\| \mathcal{D}^* \mathcal{O} \left(uv^\top \right) \right\| \leq \delta \|u\| \|v\|, \tag{20}$$

and

$$\| \left(\mathcal{O}^* \mathcal{O} - \mathcal{P}_{\mathcal{O}} \right) \left(uv^\top \right) \| \leq \delta \|uv^\top\|_F, \tag{21}$$

where the orthogonal projection $\mathcal{P}_{\mathcal{O}} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ is defined as

$$\mathcal{P}_{\mathcal{O}} \left(Z \right) = u_\star u_\star^\top Z \left(I_{n_2} - v_\star v_\star^\top \right) + \left(I_{n_1} - u_\star u_\star^\top \right) Z v_\star v_\star^\top.$$

Moreover, if $\langle u_1 v_1^\top, u_2 v_2^\top \rangle = 0$ holds, then we have that

$$\left| \langle \mathcal{A} \left(u_1 v_1^\top \right), \mathcal{A} \left(u_2 v_2^\top \right) \rangle \right| \leq \delta \|u_1 v_1^\top\|_F \|u_2 v_2^\top\|_F. \tag{22}$$

By construction, $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{O}}$ satisfy the same properties in Lemmas 1 and 2.

Next, recall that $u_\star = e_1$. We will also use the following standard concentration result, whose proof can be found in Appendix A.2.

Lemma 3. *With probability at least $1 - \mathcal{O} \left(\exp \left(-c \min \left\{ m; \min(n_1, n_2) \right\} \right) \right)$ it holds that*

$$\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] u_\star \right\| \leq 4 \sqrt{\frac{n_1}{m}} \tag{23}$$

and

$$\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} \tilde{O}_i \right] u_\star \right\| \leq 4 \sqrt{\frac{n_1}{m}}. \tag{24}$$

Finally, by construction the auxiliary sequences are independent from the off-diagonal blocks of the measurement matrices. Therefore we obtain the following lemmas, which are proved in Appendices A.3 and A.4.

Lemma 4. *Let $T \in \mathbb{N}$ and let $\eta > 0$. With probability at least $1 - \eta^{-1} - \mathcal{O} \left(\exp \left(-cm \right) \right)$, it holds for all $t \in [T]$ that*

$$\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] \tilde{v_t}^\perp \right\| \lesssim \sqrt{\frac{\log T + \log \eta}{m}} \cdot \|\tilde{v_t}^\perp\| \tag{25}$$

and

$$\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} \tilde{O}_i \right] v_t^\perp \right\| \lesssim \sqrt{\frac{\log T + \log \eta}{m}} \cdot \|\tilde{v_t}^\perp\|. \tag{26}$$

Lemma 5. *Let $T \in \mathbb{N}$ and let $\eta > 0$. With probability at least $1 - \eta^{-1}$ it holds for all $t \in [T]$ simultaneously that*

$$\frac{1}{m}\Big|\sum_{i=1}^{m}\langle O_i^\top e_i, \tilde{v}_t\rangle\langle D_i, \tilde{u}_{t+1/2}^{\perp}(\tilde{v}_t{}^{\perp})^\top\rangle\Big| \lesssim \sqrt{\frac{\log T + \log \eta}{m}} \cdot \left\|\mathcal{A}\left(\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v}_t{}^{\perp}\right)^\top\right)\right\| \tag{27}$$

and

$$\frac{1}{m}\Big|\sum_{i=1}^{m}\langle \tilde{O}_i{}^\top e_i, \tilde{v}_t\rangle\langle D_i, u_{t+1/2}^{\perp}(v_t^{\perp})^\top\rangle\Big| \lesssim \sqrt{\frac{\log T + \log \eta}{m}} \cdot \left\|\mathcal{A}\left(u_{t+1/2}^{\perp}\left(v_t^{\perp}\right)^\top\right)\right\|. \tag{28}$$

The inequalities in Lemmas 3, 4, and 5 together with the RIP of the measurement operators $\mathcal{A}$ and $\tilde{\mathcal{A}}$ imply the following inequalities in Lemma 6, Lemma 7, and Lemma 8. The proofs are also deferred to Appendices A.5, A.6, and A.7

Lemma 6. *Suppose that eqs.* (23) *to* (26) *hold. Furthermore, suppose that both $\mathcal{A}$ and $\tilde{\mathcal{A}}$ satisfy the RIP with constant $\delta > 0$. Then it holds that*

$$\left\|\left[\left(\mathcal{A}^*\mathcal{A} - \tilde{\mathcal{A}}^*\tilde{\mathcal{A}}\right)\left(u_\star v_\star^\top\right)\right]\tilde{v}_t\right\| \lesssim \left(\sqrt{\frac{\log T + \log \eta}{m}} + \sqrt{\frac{n_1}{m}}\|\tilde{v}_t{}^{\parallel}\|\right) + \delta\|v_t - \tilde{v}_t\|.$$

Lemma 7. *Suppose that eqs.* (23) *to* (28) *hold. Moreover, suppose that the measurement operators $\mathcal{A}$ and $\tilde{\mathcal{A}}$ satisfy RIP with constant $\delta > 0$ and that we have $\|u_{t+1/2}\| \leq 2$ as well as $\|\tilde{u}_{t+1/2}\| \leq 2$. Then it holds that*

$$\begin{aligned}&\left\|\left[\left(\mathcal{A}^*\mathcal{A} - \tilde{\mathcal{A}}^*\tilde{\mathcal{A}}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\right]\tilde{v}_t\right\| \\ \lesssim&\sqrt{\frac{\log T + \log \eta}{m}} + \delta\|v_t^{\parallel}\| + \left(\delta + \sqrt{\frac{n_1}{m}}\right)\|\tilde{v}_t{}^{\parallel}\| + \delta\|\tilde{v}_t - v_t\| + \delta\|\tilde{u}_{t+1/2} - u_{t+1/2}\|.\end{aligned}$$

Lemma 8. *Suppose that eqs.* (25) *to* (28) *hold. Furthermore, suppose that the measurement operator $\mathcal{A}$ satisfies RIP with constant $\delta > 0$ and that $\|\tilde{u}_{t+1/2}\| \leq 2$. Then there exists an absolute constant $C > 0$ for which it holds that*

$$\left|\langle\mathcal{A}\left(u_\star\left(v_t^{\perp}\right)^\top\right), \mathcal{A}\left(u_\star v_\star^\top\right)\rangle\right| \leq \delta\|v_t^{\perp} - \tilde{v}_t{}^{\perp}\| + C\sqrt{\frac{\log T + \log \eta}{m}} \tag{29}$$

and

$$\begin{aligned}&\left|\langle\mathcal{A}\left(u_\star\left(v_t^{\perp}\right)^\top\right), \mathcal{A}\left(u_{t+1/2}^{\perp}\left(v_t^{\perp}\right)^\top\right)\rangle\right| \\ &\leq \delta\|u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp}\| + 2\delta\|v_t^{\perp} - \tilde{v}_t{}^{\perp}\| + C\sqrt{\frac{\log T + \log \eta}{m}}.\end{aligned} \tag{30}$$

Remark 1. *The inequalities in Lemmas 3 to 8 will be used to analyze the update of u_t to u_{t+1} given v_t by the normal equation in* (9) *(by the least-squares minimization step and by the normalization step in Algorithm 1). To analyze the ALS update from v_t to v_{t+1} given u_{t+1} we will need analogous inequalities in order to be able to analyze these updates. Due to symmetry of the problem, the statements and proofs of these analogous results can be obtained in an analogous way. For this reason, to keep the presentation concise we omit the statements and proofs of analogous versions of these lemmas.*

6.2. Phase 1: From random initialization to a local neighborhood of the ground truth

Since the initialization vector $v_0 \in \mathbb{R}^{n_2}$ is chosen from the sphere with uniform distribution, with probability at least $1 - \mathcal{O}\left(n_2^{-1}\right)$, the random initialization $v_0 \in \mathbb{R}^{n_2}$ satisfies

$$\|v_0^{\|}\| = |\langle v_0, v_\star \rangle| \geq \frac{1}{2\sqrt{n_2 \log n_2}}. \tag{31}$$

Then the following proposition illustrates the convergence properties of the ALS iterates $\{u_t\}_t$ and $\{v_t\}_t$ to a neighborhood of $v_\star$ in Phase 1.

Proposition 1. *There exists a numerical constant $c > 0$ for which the following holds. Suppose that*

i) *$\mathcal{A}$ and $\tilde{\mathcal{A}}$ satisfy RIP with constant $\delta = \frac{c}{4 \log n_2}$.*

ii) *$m \geq \delta^{-2} (n_1 + n_2) \log n_2 \log |T|$, where $T = \left\lceil \frac{\log n_2}{4 \log \log n_2} \right\rceil$.*

iii) *eqs.* (23) *and* (24) *hold.*

iv) *eqs.* (25) *to* (28) *hold for all $t \in [T]$ with $\eta = n_2$.*

v) *v_0 satisfies* (31).

vi) *Analogous inequalities of iii) and iv) hold for updating v_t to v_{t+1} given u_{t+1} (see Remark 1) with $\eta = n_2$.*

Then for every $t \in [T]$ it holds that

$$\|v_t^{\|}\| \geq (\log n_2)^{2t} \cdot \|v_0^{\|}\| \geq \frac{(\log n_2)^{2t}}{2\sqrt{n_2 \log n_2}} \tag{32}$$

and

$$\max\left\{\|v_t^{\|} - \tilde{v}_t^{\|}\|; \|v_t^{\perp} - \tilde{v}_t^{\perp}\|\right\} \leq c_{2t} \|v_t^{\|}\|, \tag{33}$$

where c_t is defined in (17) *until we have that*

$$\min\left\{\|v_t^{\|}\|, \|u_t^{\|}\|\right\} \geq \frac{c}{\log n_2}. \tag{34}$$

Proof of Proposition 1. It suffices to only consider the case when the initialization vector $v_0 \in \mathbb{R}^{n_2}$ does not satisfy (34). Otherwise there is nothing to prove.

We are going to show by induction that (32) and (33) hold until condition (34) is fulfilled. In particular, note that by our choice of T this immediately implies that (34) holds for some $t \leq T$. For the base case, observe that for $t = 0$ the two inequalities in (32) and (33) are satisfied since we have $v_0 = \tilde{v}_0$ by definition and since we assume that inequality (31) holds.

For the induction step, suppose that the statements hold for some natural number t with $t \leq T$. Then we will show that the statements also hold for $t + 1$ whenever (34) is not yet satisfied. To this end, we first show that the estimation error and the norm of the next least-squares update $u_{t+1/2}$ are upper-bounded as shown in the following lemma. It is proved in Appendix B.1.

Lemma 9. *Suppose that $\mathcal{A}$ satisfies RIP for $0 < \delta < 1$ and $\|v_t\| = 1$. Then it holds that*

$$\left\|u_{t+1/2} - \langle v_\star, v_t \rangle u_\star\right\| \leq \frac{\delta}{1-\delta} \|v_t^\perp\|. \tag{35}$$

In particular, it follows that

$$\|u_{t+1/2}^\perp\| \leq \frac{\delta}{1-\delta} \|v_t^\perp\|. \tag{36}$$

Moreover, for $\delta \leq \frac{1}{2}$, we have that

$$\|u_{t+1/2}\| \leq 2. \tag{37}$$

Analogously, since $\tilde{\mathcal{A}}$ also satisfies the RIP with the same constant δ and since $\|\tilde{v}_t\| = 1$ holds, we also have

$$\left\|\tilde{u}_{t+1/2} - \langle v_\star, \tilde{v}_t \rangle u_\star\right\| \leq \frac{\delta}{1-\delta} \|\tilde{v}_t{}^\perp\|, \tag{38}$$

$$\|\tilde{u}_{t+1/2}^\perp\| \leq \frac{\delta}{1-\delta} \|\tilde{v}_t{}^\perp\|, \tag{39}$$

$$\|\tilde{u}_{t+1/2}\| \leq 2. \tag{40}$$

Given the upper estimates in eqs. (37) and (40), the next lemma, proven in Appendix B.2, shows that the distances between the least-square updates of the original and auxiliary sequences stay close each other.

Lemma 10. *Under the hypothesis of Proposition 1, suppose that eqs.* (37) *and* (40) *hold. Let $t \in \mathbb{N}$ and assume furthermore that the inequalities* (32) *and* (33) *hold. Then there exists an absolute constant $C_1 > 0$ for which the followings hold:*

$$\|u_{t+1/2}^{\|} - \tilde{u}_{t+1/2}^{\|}\| \leq (c_{2t} + C_1\delta(1 + c_{2t}))\, \|v_t^{\|}\|, \tag{41}$$

$$\|u_{t+1/2}^\perp - \tilde{u}_{t+1/2}^\perp\| \leq C_1\delta\,(1 + c_{2t})\, \|v_t^{\|}\|. \tag{42}$$

The upper estimates in (41) and (42) imply that $\|u_{t+1/2}^{\|}\|$ is close to $\|v_t^{\|}\|$, which is stated in the following lemma, see Appendix B.3.

Lemma 11. *Under the hypothesis of Proposition 1, suppose that eqs.* (37), (41) *and* (42) *hold. Moreover, let $t \in \mathbb{N}$ and assume that the inequalities* (32) *and* (33) *hold. Then there exists an absolute constant $C_2 > 0$ for which the followings hold:*

$$(1 - C_2\delta(1 + c_{2t}))\left\|u_{t+1/2}^{\|}\right\| \leq \|v_t^{\|}\| \leq (1 + C_2\delta\,(1 + c_{2t}))\left\|u_{t+1/2}^{\|}\right\|. \tag{43}$$

Remark 2. *Later on, we will in fact only use the upper bound on* $\|u^{\|}_{t+1/2}\|$ *in inequality* (43)*. As there is no additional effort required in proving the lower bound as well, we also decided to include it in this manuscript.*

Moreover, since $T = \left\lceil \frac{\log n_2}{4 \log\log n_2} \right\rceil$, it follows that c_{2t} is bounded from above by an absolute constant for all $t \leq T$, which is formally stated in the following lemma.

Lemma 12. *Then for all* $t \leq 2T \leq \left\lceil \frac{\log n_2}{2 \log\log n_2} \right\rceil + 1$ *it holds that* c_t *defined in* (17) *satisfies* $c_t \leq C$ *for an absolute constant* C*.*

Proof. For all $t \leq 2T$ we have

$$\begin{aligned} c_t + 1 &= \exp\left(t \log\left(1 + \frac{1}{\log n_2} \right) \right) \\ &\overset{(a)}{\leq} \exp\left(t / \log n_2 \right) \\ &\leq \exp\left(2T / \log n_2 \right) \\ &\leq \exp\left(\frac{1}{2 \log\log n_2} + \frac{2}{\log n_2} \right) \\ &\leq C, \end{aligned}$$

where (a) follows from the elementary inequality $\log(1+x) \leq x$ for $x > 0$. □

Hence, for sufficiently small $c > 0$, (43) implies that

$$\frac{1}{2} \|v^{\|}_t\| \leq \|u^{\|}_{t+1/2}\|. \tag{44}$$

The next lemma, proved in Appendix B.4, shows that the original and auxiliary sequences stay close in ℓ_2-distance under the conditions derived above.

Lemma 13. *Under the hypothesis of Proposition 1, suppose that eqs.* (41) *to* (43) *hold. Moreover, suppose that* $c > 0$ *is chosen small enough (smaller than an absolute constant depending only on* C_1, C_2, C_3*). Then it follows that*

$$\max\left\{ \|u^{\|}_{t+1} - \tilde{u}^{\|}_{t+1}\|; \|u^{\perp}_{t+1} - \tilde{u}^{\perp}_{t+1}\| \right\} \leq c_{2t+1} \cdot \|u^{\|}_{t+1}\|. \tag{45}$$

We further proceed with the following lemma, which shows how the estimation error propagates with the normalization. The proof is provided in Appendix B.5.

Lemma 14. *Suppose that* $\|v_t\| = 1$ *and that for fixed* $t \in \mathbb{N}$ *and real numbers* $0 < \beta < \alpha < 1$ *it holds that*

$$\|u^{\|}_{t+1/2}\|^2 \geq \alpha \|v^{\|}_t\|^2, \tag{46}$$

$$\|u^{\perp}_{t+1/2}\|^2 \leq \beta \|v^{\perp}_t\|^2. \tag{47}$$

Then, whenever $v_t^{||} \neq 0$, it holds that

$$\|u_{t+1}^{||}\|^2 \geq \frac{\alpha \|v_t^{||}\|^2}{\beta + (\alpha - \beta)\, \|v_t^{||}\|^2} \geq \frac{\|v_t^{||}\|^2}{\frac{\beta}{\alpha} + \|v_t^{||}\|^2} \tag{48}$$

and, moreover,

$$\|u_{t+1}^{\perp}\|^2 \leq \frac{\beta}{\alpha \|v_t^{||}\|^2} \cdot \|v_t^{\perp}\|^2. \tag{49}$$

Note that due to (36) with $\delta < \frac{1}{2}$ and due to (44) the assumptions in Lemma 14 are satisfied with $\alpha = \frac{1}{4}$ and $\beta = 4\delta^2$. Therefore, with $\delta = \frac{c}{4 \log n_2}$ and $\|v_t^{||}\| \leq \frac{c}{\log n_2}$ we obtain that

$$\|u_{t+1}^{||}\|^2 \geq \frac{\|v_t^{||}\|^2}{16\delta^2 + \|v_t^{||}\|^2} \geq \frac{2 \log n_2}{c} \|v_t^{||}\|^2 \geq \left(\frac{2 \log n_2}{c} \right)^{2t+1} \|v_0^{||}\|. \tag{50}$$

Since we have shown (45) and (50) this finishes the induction step for u_{t+1}. With exactly the same reasoning we can then prove the inequalities

$$\|v_{t+1}^{||}\| \geq \left(\frac{2 \log n_2}{c} \right)^{2t+2} \|v_0^{||}\|, \tag{51}$$

$$\max \left\{ \|v_{t+1}^{||} - \tilde{v}_{t+1}^{||}\|; \|v_{t+1}^{\perp} - \tilde{v}_{t+1}^{\perp}\| \right\} \leq c_{2t+2} \|v_{t+1}^{||}\|.$$

This shows inequalities (32) and (33) for $t+1$. Note that by choosing $c < \frac{1}{2}$ inequality (51) implies (32). This completes the induction step. □

6.3. Phase 2: Linear convergence by RIP

We enter the second phase as soon as the iterates are sufficiently aligned with the ground truth solution, that is when condition (34) is satisfied. Once we enter the second phase, our iterates converge linearly to the ground truth as it is shown by the next proposition, which describes the second phase.

Proposition 2. *There exists a numerical constant $c' > 0$ for which the following holds. Suppose that $\mathcal{A}$ satisfies RIP with constant $\delta = \frac{c'}{8 \log n_2}$ and either $\|v_{\hat{t}}^{||}\| > \frac{c'}{\log n_2}$ or $\|u_{\hat{t}}^{||}\| > \frac{c'}{\log n_2}$ for some $\hat{t} \in \mathbb{N}$. Then it holds that for all $t > \hat{t}$*

$$\|u_{t+1}^{\perp}\| \leq \frac{1}{2} \left(\frac{1}{2 \log n_2} \right)^{2(t-\hat{t})} \|v_{\hat{t}}^{\perp}\| \quad \text{and} \quad \|v_{t+1}^{\perp}\| \leq \frac{1}{2} \left(\frac{1}{2 \log n_2} \right)^{2(t-\hat{t})+1} \|v_{\hat{t}}^{\perp}\|. \tag{52}$$

Proof. Due to the symmetry of the argument, we may assume without loss of generality that

$$\|v_{\hat{t}}^{||}\| > \frac{c'}{\log n_2} = 8\delta. \tag{53}$$

Next we show that

$$\|u^{\perp}_{\hat{t}+1}\| \leq \frac{1}{3}\,\|v^{\perp}_{\hat{t}}\|. \tag{54}$$

By choosing the absolute constant c' small enough, we may assume that $\delta < \frac{1}{2}$. Then by Lemma 9 and the RIP of $\mathcal{A}$ we have

$$\left\|u_{\hat{t}+1/2} - \langle v_\star, v_{\hat{t}}\rangle u_\star\right\| \leq \frac{\delta}{1-\delta}\|v^{\perp}_{\hat{t}}\| \leq 2\delta\|v^{\perp}_{\hat{t}}\|.$$

This implies

$$\|u^{\perp}_{\hat{t}+1/2}\| \leq 2\delta\|v^{\perp}_{\hat{t}}\|$$

as well as

$$\left|\|u^{\|}_{\hat{t}+1/2}\| - \|v^{\|}_{\hat{t}}\|\right| = \left|\|u^{\|}_{\hat{t}+1/2}\| - \|\langle v_\star, v_{\hat{t}}\rangle u_\star\|\right| \leq \left\|u_{\hat{t}+1/2} - \langle v_\star, v_{\hat{t}}\rangle u_\star\right\| \leq 2\delta\|v^{\perp}_{\hat{t}}\| \leq 2\delta, \tag{55}$$

where in the last inequality we used that $\|v^{\perp}_{\hat{t}}\| \leq \|v_{\hat{t}}\| = 1$. In particular, the inequality in (55) implies that

$$\|u^{\|}_{\hat{t}+1/2}\| \geq \|v^{\|}_{\hat{t}}\| - 2\delta \geq \frac{3}{4}\,\|v^{\|}_{\hat{t}}\|,$$

where the last inequality follows from (53). Hence, setting $\alpha = \frac{9}{16}$ and $\beta = 4\delta^2$, Lemma 14 and (53) yield

$$\|u^{\perp}_{\hat{t}+1}\|^2 \leq \frac{\beta}{\alpha\|v^{\|}_{\hat{t}}\|^2} \cdot \|v^{\perp}_{\hat{t}}\|^2 \leq \frac{1}{9}\,\|v^{\perp}_{\hat{t}}\|^2. \tag{56}$$

This shows (54). Next, one can show by induction that for $t > \hat{t}$

$$\|u^{\perp}_{t+1}\| \leq \left(\frac{c'}{2\sqrt{2}\log n_2}\right)\|v^{\perp}_t\| \quad \text{and} \quad \|v^{\perp}_{t+1}\| \leq \left(\frac{c'}{2\sqrt{2}\log n_2}\right)\|u^{\perp}_{t+1}\|.$$

The proof of these inequalities is analogous to the proof of (54) except that in (56) we can use the estimate $\|v^{\|}_t\|^2 \geq \frac{8}{9}$ due to $\|v^{\perp}_t\|^2 \leq \|u^{\perp}_{t-1}\|^2 \leq \frac{1}{9}$ instead of the weaker estimate $\|v^{\|}_t\| \geq \frac{c'}{\log n_2}$. Finally, one can choose c' small so that (52) is satisfied. □

6.4. Finishing the proof of Theorem 1

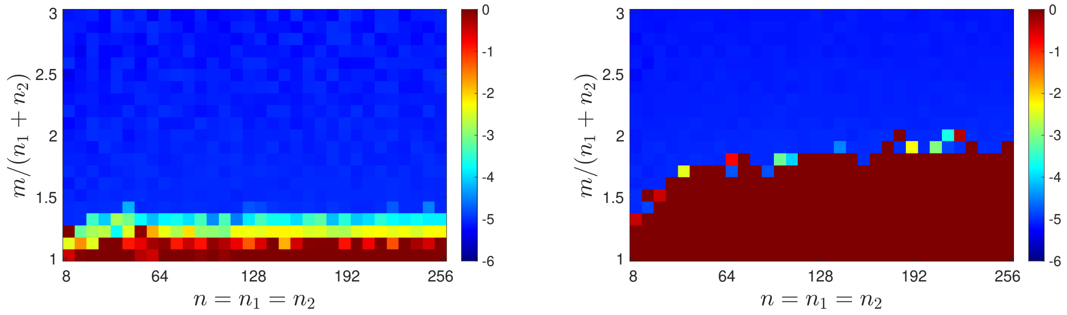
We deduce from (32) in Proposition 1 that Phase 1 is completed after

$$\hat{t} \lesssim \frac{\log n_2}{\log\log n_2} \tag{57}$$

iterations. Next, one observes immediately by a direct calculation that $\sin(\angle(u_t, u_\star)) = \|u^{\perp}_t\|$ and $\sin(\angle(v_t, v_\star)) = \|v^{\perp}_t\|$. Moreover, one obtains from inequalities in (52) of Proposition 2 that after

$$t - \hat{t} \lesssim \frac{\log(1/\varepsilon)}{\log\log n_2}$$

iterations it holds that $\max\left\{\|u^{\perp}_t\|; \|v^{\perp}_t\|\right\} \leq \varepsilon$. Together with (57) this finishes the proof of Theorem 1.



(a) ALS from spectral initialization

(b) ALS from random initialization

Figure 2.: Phase transition of reconstruction error

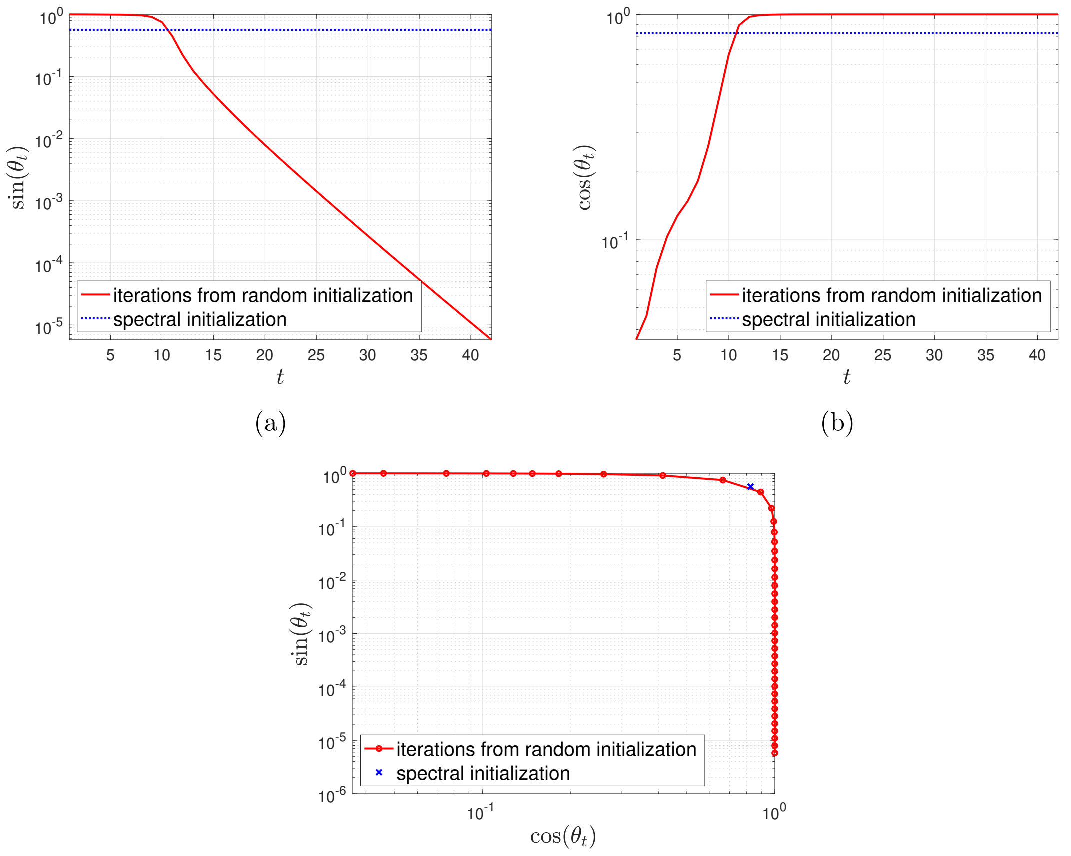
7. Numerical experiments

We present a set of Monte Carlo simulations to compare the theoretical bound in Theorem 1 to the empirical performance of ALS from random initialization. According to the assumptions of Theorem 1, the measurement matrices were generated as independent copies of a random matrix with i.i.d. standard Gaussian entries. Observations were obtained without noise. In the first experiment, we compare the performance of ALS methods respectively from random initialization and from spectral initialization. Figure 2 plots the phase transition of the reconstruction error in this experiment. We vary the matrix size from 8 to 256 while the oversampling factor $m/(n_1+n_2)$ is between 1 and 3. As shown in Figures 2a and 2b, ALS from spectral initialization has larger success regime so that the reconstruction is achieved from fewer observations. In these plots, we displayed the median of the normalized reconstruction error over 100 random trials. Figure 2b shows that compared to ALS from spectral initialization, the phase transition for ALS from random initialization occurs at a higher oversampling factor. The amount of excess observations scales as a poly-log of the matrix size, which coincides with the result in Theorem 1.

Although the main result in Theorem 1 is restricted to the rank-1 case, empirically, ALS from random initialization continues to work at a small oversampling factor when the rank of the unknown matrix becomes larger. We conducted the same experiment in Figure 1 in the rank-r case, which is plotted in Figure 3. One can observe that the same phase transition in Theorem 1 occurs in the rank-5 case.

8. Discussion

We have shown that ALS from random initialization converges to the rank-one ground-truth matrix in the low-rank matrix sensing setting (with high probability). In our analysis, we observed that the trajectory of the iteration can be separated into two



(a)

(b)

(c)

Figure 3.: Evolution of the estimation by randomly initialized ALS over iteration (rank-5 case): $n_1 = n_2 = 256$, $r = 5$, $m = 2r(n_1 + n_2 - r)$. The principal angle between subspaces spanned by $\hat{U}$ and U_t is denoted by θ_t. (a) $\sin\theta_t$ vs t; (b) $\cos\theta_t$ vs t; (c) $\sin\theta_t$ vs $\cos\theta_t$.

distinct phases: in the first one, the iterates converge from random initialization to a local neighborhood in $O(\log n / \log\log n)$ iterations. In the second phase, the iterates converge linearly to the ground truth. This is aligned with our numerical experiments, where a sharp phase transition is visible.

We expect that the convergence analysis in this paper will shed light on the convergence of ALS starting from random initialization in more general settings. For example, empirically, ALS from random initialization was shown to be successful if the ground truth has a rank higher than one. It would be interesting to see whether our analysis can be extended to this setting. Moreover, it would be interesting to examine the scenario when the measurement matrices are more structured such as in the Matrix Completion problem.

Moreover, our result requires a sample size at least in the order of $n \log^4 n$, whereas, for example, approaches based on convex relaxation such as nuclear-norm minimization only need in the order of n samples. It would be interesting to examine whether it is possible to remove the additional log-factors in our result.

References

[1] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[2] L. N. Trefethen and D. I. Bau, *Numerical linear algebra.* Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, 1997.

[3] J. P. Haldar and D. Hernando, "Rank-constrained solutions to linear matrix equations using powerfactorization," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 584–587, 2009.

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

[5] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, "Matrix completion and low-rank SVD via fast alternating least squares," *J. Mach. Learn. Res.*, vol. 16, pp. 3367–3402, 2015.

[6] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[7] M. Hardt, "Understanding alternating minimization for matrix completion," in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science.* IEEE, 2014, pp. 651–660.

[8] M. Hardt and M. Wootters, "Fast matrix completion without the condition number," in *Proceedings of The 27th Conference on Learning Theory*, ser.

Proceedings of Machine Learning Research, M. F. Balcan, V. Feldman, and C. Szepesvári, Eds., vol. 35. Barcelona, Spain: PMLR, 13–15 Jun 2014, pp. 638–678. [Online]. Available: https://proceedings.mlr.press/v35/hardt14a.html

[9] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[10] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Math. Program.*, vol. 176, no. 1, pp. 5–37, 2019.

[11] M. Kech and F. Krahmer, "Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems," *SIAM J. Appl. Algebra Geom.*, vol. 1, no. 1, pp. 20–37, 2017.

[12] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[13] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[14] E. J. Candès and T. Tao, "The power of convex relaxation: near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[15] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[16] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[17] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Appl. Comput. Harmon. Anal.*, vol. 39, no. 2, pp. 277–299, 2015.

[18] D. Gross, F. Krahmer, and R. Kueng, "A partial derandomization of phaselift using spherical designs," *J. Fourier Anal. Appl.*, vol. 21, no. 2, pp. 229–266, 2015.

[19] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[20] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[21] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: algorithms and performance bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4497–4520, 2017.

[22] P. Jung, F. Krahmer, and D. Stöger, "Blind demixing and deconvolution at near-optimal rate," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 704–727, 2018.

[23] T. Fuchs, D. Gross, P. Jung, F. Krahmer, R. Kueng, and D. Stöger, "Proof methods for robust low-rank matrix recovery," *arXiv preprint arXiv:2106.04382*, 2021.

[24] P. Jain, R. Meka, and I. Dhillon, "Guaranteed rank minimization via singular value projection," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[25] K. Lee and Y. Bresler, "ADMiRA: Atomic decomposition for minimum rank approximation," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4402–4416, 2010.

[26] M. Fornasier, H. Rauhut, and R. Ward, "Low-rank matrix recovery via iteratively reweighted least squares minimization," *SIAM J. Optim.*, vol. 21, no. 4, pp. 1614–1640, 2011.

[27] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Mach. Learn. Res.*, vol. 13, pp. 3441–3473, 2012.

[28] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[29] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *International Conference on Machine Learning.* PMLR, 2016, pp. 964–973.

[30] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5239–5269, 2019.

[31] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Appl. Comput. Harmon. Anal.*, vol. 47, no. 3, pp. 893–934, 2019.

[32] W. Huang and P. Hand, "Blind deconvolution by a steepest descent algorithm on a quotient manifold," *SIAM J. Imaging Sci.*, vol. 11, no. 4, pp. 2757–2785, 2018.

[33] S. Ling and T. Strohmer, "Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing," *Inf. Inference*, vol. 8, no. 1, pp. 1–49, 2019.

[34] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Commun. Pure Appl. Math.*, vol. 70, no. 5, pp. 822–883, 2017.

[35] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Found. Comput. Math.*, vol. 18, no. 5, pp. 1131–1198, 2018.

[36] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[37] J. Chen and X. Li, "Model-free nonconvex matrix completion: local minima analysis and applications in memory-efficient kernel PCA," *J. Mach. Learn. Res.*, vol. 20, p. 39, 2019, id/No 142.

[38] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Math. Program.*, vol. 176, no. 1-2 (B), pp. 311–337, 2019.

[39] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, "Gradient descent can take exponential time to escape saddle points," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[40] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," in *Conference On Learning Theory.* PMLR, 2018, pp. 2–47.

[41] D. Stöger and M. Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[42] T. Ye and S. S. Du, "Global convergence of gradient descent for asymmetric low-rank matrix factorization," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[43] L. Jiang, Y. Chen, and L. Ding, "Algorithmic regularization in model-free overparametrized asymmetric matrix factorization," *arXiv preprint arXiv:2203.02839*, 2022.

[44] P. M. Kroonenberg and J. De Leeuw, "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, vol. 45, no. 1, pp. 69–97, 1980.

[45] J. A. O'Sullivan and J. Benac, "Alternating minimization algorithms for transmission tomography," *IEEE Transactions on Medical Imaging*, vol. 26, no. 3, pp. 283–297, 2007.

[46] T. Zhao, Z. Wang, and H. Liu, "A nonconvex optimization framework for low rank matrix estimation," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/39461a19e9eddfb385ea76b26521ea48-Paper.pdf

[47] K. Lee, Y. Wu, and Y. Bresler, "Near-optimal compressed sensing of a class of sparse low-rank matrices via sparse power factorization," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1666–1698, 2018.

[48] K. Lee, Y. Li, M. Junge, and Y. Bresler, "Blind recovery of sparse signals from subsampled convolution," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 802–821, 2017.

[49] J. Geppert, F. Krahmer, and D. Stöger, "Sparse power factorization: balancing peakiness and sample complexity," *Adv. Comput. Math.*, vol. 45, no. 3, pp. 1711–1728, 2019.

[50] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[51] R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[52] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4814–4826, 2015.

[53] I. Waldspurger, "Phase retrieval with random Gaussian sensing vectors by alternating projections," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3301–3312, 2018.

[54] T. Zhang, "Phase retrieval by alternating minimization with random initialization," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4563–4573, 2020.

[55] Y. Zhong and N. Boumal, "Near-optimal bounds for phase synchronization," *SIAM J. Optim.*, vol. 28, no. 2, pp. 989–1016, 2018.

[56] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan, "Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization," *SIAM J. Optim.*, vol. 30, no. 4, pp. 3098–3121, 2020.

[57] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," *IEEE Trans. Inf. Theory*, vol. 67, no. 3, pp. 1928–1950, 2021.

[58] L. Ding and Y. Chen, "Leave-one-out approach for matrix completion: primal and dual analysis," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 7274–7301, 2020.

[59] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[60] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, ser. Appl. Numer. Harmon. Anal. New York, NY: Birkhäuser/Springer, 2013.

[61] R. Vershynin, *High-dimensional probability: An introduction with applications in data science.* Cambridge university press, 2018, vol. 47.

A. Proofs of concentration inequalities

A.1. Proof of Lemma 2

The inequality in (22) is well known (see, e.g., [60, Exercise 6.24]). In fact, since we assumed the RIP to hold for all matrices of rank at most 4 in (18), we even obtain the stronger statement that

$$\left|\langle \mathcal{A}\left(Z_1\right), \mathcal{A}\left(Z_2\right)\rangle - \langle Z_1, Z_2\rangle_F\right| \le \delta \|Z_1\|_F \cdot \|Z_2\|_F. \tag{58}$$

for all matrices Z_1 and Z_2 of rank at most 2 (see [60, Section 6].

We are going to derive the other inequalities in (19), (20), and (21) from (58). For that, we note first that

$$\mathcal{O}^*\mathcal{D} = \mathcal{P}_{\mathcal{O}}\mathcal{A}^*\mathcal{A}(\mathrm{Id} - \mathcal{P}_{\mathcal{O}}).$$

Then there exist $\hat{x} \in \mathbb{R}^{n_1}$ and $\hat{y} \in \mathbb{R}^{n_2}$ with $\|\hat{x}\| = \|\hat{y}\| = 1$ such that

$$\left\|\mathcal{O}^*\mathcal{D}\left(uv^\top\right)\right\| = \left|\langle \hat{x}\hat{y}^\top, \mathcal{O}^*\mathcal{D}\left(uv^\top\right)\rangle_F\right|.$$

Then it follows that the left-hand side of (19) is upper-bounded by

$$\begin{aligned}
\left\|\mathcal{O}^*\mathcal{D}\left(uv^\top\right)\right\| &= \left|\langle \hat{x}\hat{y}^\top, \mathcal{O}^*\mathcal{D}\left(uv^\top\right)\rangle_F\right| \\
&= \left|\langle \hat{x}\hat{y}^\top, \mathcal{P}_{\mathcal{O}}\mathcal{A}^*\mathcal{A}(\mathrm{Id} - \mathcal{P}_{\mathcal{O}})\left(uv^\top\right)\rangle_F\right| \\
&= \left|\langle \mathcal{A}\mathcal{P}_{\mathcal{O}}\left(\hat{x}\hat{y}^\top\right), \mathcal{A}(\mathrm{Id} - \mathcal{P}_{\mathcal{O}})\left(uv^\top\right)\rangle\right| \\
&\overset{\text{(a)}}{\le} \delta\|\mathcal{P}_{\mathcal{O}}(\hat{x}\hat{y}^\top)\|_F \cdot \|(\mathrm{Id} - \mathcal{P}_{\mathcal{O}})(uv^\top)\|_F \\
&\le \delta\|\hat{x}\hat{y}^\top\|_F \cdot \|uv^\top\|_F \\
&= \delta\|u\| \cdot \|v\|,
\end{aligned}$$

where (a) is due to (58) and the fact that $\mathcal{P}_{\mathcal{O}}(\hat{x}\hat{y}^\top$ and $(\mathrm{Id} - \mathcal{P}_{\mathcal{O}})\left(uv^\top\right)$ have rank at most 2 each. This proves inequality (19). Inequality (20) can be derived in an analogous way.

In order to show inequality (21), we again note there is $\tilde{x} \in \mathbb{R}^{n_1}$ and $\tilde{y} \in \mathbb{R}^{n_2}$ with $\|\tilde{x}\| = \|\tilde{y}\| = 1$ such that

$$\left\|\left(\mathcal{O}^*\mathcal{O} - \mathcal{P}_{\mathcal{O}}\right)\left(uv^\top\right)\right\| = \left|\langle \tilde{x}\tilde{y}^\top, \left(\mathcal{O}^*\mathcal{O} - \mathcal{P}_{\mathcal{O}}\right)\left(uv^\top\right)\rangle_F\right|$$

holds. From $\mathcal{O} = \mathcal{A}\mathcal{P}_{\mathcal{O}}$ it follows that

$$\left\|\left(\mathcal{O}^*\mathcal{O} - \mathcal{P}_{\mathcal{O}}\right)\left(uv^\top\right)\right\| = \left|\langle \mathcal{A}\left(\mathcal{P}_O(\tilde{x}\tilde{y}^\top)\right), \mathcal{A}\left(\mathcal{P}_{\mathcal{O}}\left(uv^\top\right)\right)\rangle - \langle \mathcal{P}_O(\tilde{x}\tilde{y}^\top), \mathcal{P}_{\mathcal{O}}\left(uv^\top\right)\rangle_F\right|.$$

Then it follows from (58) that

$$\begin{aligned}
\left\|\left(\mathcal{O}^*\mathcal{O} - \mathcal{P}_{\mathcal{O}}\right)\left(uv^\top\right)\right\| &\le \delta\|\mathcal{P}_O(\tilde{x}\tilde{y}^\top)\|_F \cdot \|\mathcal{P}_O(uv^\top)\|_F \\
&\le \delta\|u\| \cdot \|v\|.
\end{aligned}$$

This finishes the proof.

A.2. Proof of Lemma 3

Note that the first entry of the vector $\sum_{i=1}^{m}(A_i)_{1,1}O_ie_1 \in \mathbb{R}^{n_1}$ vanishes. Conditioned on $\{(A_i)_{1,1}\}_{i=1}^{m}$, all other entries are i.i.d. random variables with distribution $\mathcal{N}(0, \sum_{i=1}^{m} |(A_i)_{1,1}|^2)$. In particular, this implies that conditioned on $\{(A_i)_{1,1}\}_{i=1}^{m}$ with probability at least $1 - \mathcal{O}\left(\exp\left(-cn_1\right)\right)$ we have that

$$\left\| \sum_{i=1}^{m} \left(A_i\right)_{1,1} O_i e_1 \right\| \le 2\sqrt{n_1 \sum_{i=1}^{m} \left(A_i\right)_{1,1}^2}. \tag{59}$$

This is the standard concentration of the norm of a Gaussian vector (see, e.g. [61, Theorem 3.1.1]). Similarly, it holds with probability at least $1 - \mathcal{O}\left(\exp\left(-cm\right)\right)$ that

$$\sum_{i=1}^{m} |\left(A_i\right)_{1,1}|^2 \le 2m. \tag{60}$$

Inserting inequality (60) into inequality (59) provides the first assertion in Lemma 3. The second assertion can be obtained analogously.

A.3. Proof of Lemma 4

We prove only the first assertion. The proof for the second assertion is analogous. We first note that by the concentration of the norm of Gaussian vector (e.g., [61, Theorem 3.1.1]), it holds with probability at least $1 - \mathcal{O}\left(\exp\left(-cm\right)\right)$ that

$$\sum_{i=1}^{m} |\left(A_i\right)_{1,1}|^2 \le 2m. \tag{61}$$

In the following we will proceed conditioned on this event. Since by definition the first entry of $\tilde{v_t}^{\perp}$ vanishes, only the first entry of $O_i\tilde{v_t}^{\perp}$ is non-zero due to the structure of the matrix O_i. In particular, we have that

$$O_i\tilde{v_t}^{\perp} = \langle O_i^{\top}e_1, \tilde{v_t}^{\perp}\rangle e_1.$$

This implies that

$$\frac{1}{m}\left\| \left[\sum_{i=1}^{m} \left(A_i\right)_{1,1} O_i\right] \tilde{v_t}^{\perp} \right\| = \frac{1}{m}\left\| \sum_{i=1}^{m} \langle O_i^{\top}e_1, \tilde{v_t}^{\perp}\rangle \left(A_i\right)_{1,1} e_1 \right\| = \frac{1}{m}\left| \sum_{i=1}^{m} \langle O_i^{\top}e_1, \tilde{v_t}^{\perp}\rangle \left(A_i\right)_{1,1} \right|.$$

We observe that $\tilde{v_t}$ and $\left(A_i\right)_{1,1}$ are independent of O_i for all $i \in [m]$ due to their definitions. Hence, conditioned on $\{(A_i)_{1,1}\}_{i=1}^{m}$ and $\tilde{v_t}$ it holds that

$$\langle O_i^{\top}e_1, \tilde{v_t}^{\perp}\rangle \left(A_i\right)_{1,1} \sim \mathcal{N}\left(0, |\left(A_i\right)_{1,1}| \|\tilde{v_t}^{\perp}\|\right), \quad \text{for all } i \in [m]$$

and, hence,

$$\frac{1}{m}\sum_{i=1}^{m}\langle O_i^\top e_1, \tilde{v}_t{}^\perp\rangle\,(A_i)_{1,1} \sim \mathcal{N}\left(0, \frac{1}{m}\sqrt{\sum_{i=1}^{m}(A_i)_{1,1}^2\|\tilde{v}_t{}^\perp\|}\right).$$

In particular, conditioned on $\{(A_i)_{1,1}\}_{i=1}^m$ and $\tilde{v}_t$ we obtain by a union bound that with probability $1-\eta^{-1}$ it holds for all $t\in[T]$ simultaneously that

$$\frac{1}{m}\Big|\sum_{i=1}^{m}\langle O_i^\top e_1, \tilde{v}_t{}^\perp\rangle\,(A_i)_{1,1}\Big| \lesssim \frac{\sqrt{\log T+\log\eta}}{m}\cdot\sqrt{\sum_{i=1}^{m}\Big((A_i)_{1,1}\Big)^2}\cdot\|\tilde{v}_t{}^\perp\|.$$

By inserting (61) into the above inequality and by integrating over all events $\{(A_i)_{1,1}\}_{i=1}^m$, which satisfy (61), the first assertion in Lemma 4 is obtained. The second assertion in Lemma 4 is obtained analogously.

A.4. Proof of Lemma 5

We note that $\left\{\langle D_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F\tilde{v}_t{}^\perp\right\}_{i=1}^m$ is independent from $\left\{O_i^\top e_1\right\}_{i=1}^m$. This implies that conditioned on $\left\{\langle D_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F\tilde{v}_t{}^\perp\right\}_{i=1}^m$ we have

$$\sum_{i=1}^{m}\langle O_i^\top e_i, \tilde{v}_t\rangle\langle D_i, \tilde{u}_{t+1/2}^\perp(\tilde{v}_t{}^\perp)^\top\rangle_F \sim \sqrt{\sum_{i=1}^{m}\langle D_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F^2\|\tilde{v}_t{}^\perp\|^2}\cdot\mathcal{N}(0,1)$$

Hence, we obtain that conditioned on $\left\{\langle D_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F\tilde{v}_t{}^\perp\right\}_{i=1}^m$ with probability $1-\eta^{-1}$ it holds for all $t\in[T]$ simultaneously that

$$\begin{aligned}
\frac{1}{m}\Big|\sum_{i=1}^{m}\langle O_i^\top e_i, \tilde{v}_t\rangle_F\langle D_i, \tilde{u}_{t+1/2}^\perp(\tilde{v}_t{}^\perp)^\top\rangle_F\Big| &\lesssim \frac{\sqrt{\log T+\log\eta}}{m}\cdot\sqrt{\sum_{i=1}^{m}\langle D_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F^2\|\tilde{v}_t{}^\perp\|^2}\\
&= \sqrt{\frac{\log T+\log\eta}{m}}\cdot\|\tilde{v}_t{}^\perp\|\cdot\sqrt{\frac{1}{m}\sum_{i=1}^{m}\langle A_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F^2}\\
&= \sqrt{\frac{\log T+\log\eta}{m}}\cdot\|\tilde{v}_t{}^\perp\|\cdot\left\|\mathcal{A}\left(\tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\right)\right\|\\
&\le \sqrt{\frac{\log T+\log\eta}{m}}\cdot\left\|\mathcal{A}\left(\tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\right)\right\|.
\end{aligned}$$

This finishes the proof of the first assertion. The second assertion is obtained analogously.

A.5. Proof of Lemma 6

Recall without loss of generality that we assumed $u_\star = e_1$ and $v_\star = e_1$. This implies that we have

$$\left[\left(\mathcal{A}^* \mathcal{A} \right) u_\star v_\star^\top \right] \tilde{v}_t = \frac{1}{m} \left(\sum_{i=1}^m A_i \langle A_i, u_\star v_\star^\top \rangle_F \right) \tilde{v}_t = \frac{1}{m} \left(\sum_{i=1}^m A_i \left(A_i \right)_{1,1} \right) \tilde{v}_t$$

and

$$\left[\left(\tilde{\mathcal{A}}^* \tilde{\mathcal{A}} \right) u_\star v_\star^\top \right] \tilde{v}_t = \frac{1}{m} \left(\sum_{i=1}^m \tilde{A}_i \langle \tilde{A}_i, u_\star v_\star^\top \rangle_F \right) \tilde{v}_t = \frac{1}{m} \left(\sum_{i=1}^m \tilde{A}_i \left(A_i \right)_{1,1} \right) \tilde{v}_t.$$

Then it follows that

$$\left[\left(\mathcal{A}^* \mathcal{A} - \tilde{\mathcal{A}}^* \tilde{\mathcal{A}} \right) \left(u_\star v_\star^\top \right) \right] \tilde{v}_t = \frac{1}{m} \left[\sum_{i=1}^m \left(A_i \right)_{1,1} \left(A_i - \tilde{A}_i \right) \right] \tilde{v}_t.$$

In order to proceed recall that we have decomposition $A_i = D_i + O_i$ and $\tilde{A}_i = D_i + \tilde{O}_i$ for all $i \in [m]$. This implies that $A_i - \tilde{A}_i = O_i - \tilde{O}_i$. Hence, we obtain that

$$\begin{aligned}
&\left\| \left[\left(\mathcal{A}^* \mathcal{A} - \tilde{\mathcal{A}}^* \tilde{\mathcal{A}} \right) \left(u_\star v_\star^\top \right) \right] \tilde{v}_t \right\| \\
=& \frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} \left(O_i - \tilde{O}_i \right) \right] \tilde{v}_t \right\| \\
\le& \frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] \tilde{v}_t \right\| + \frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} \tilde{O}_i \right] \tilde{v}_t \right\| \\
\le& \underbrace{\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] \tilde{v}_t \right\|}_{=:(a)} + \underbrace{\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} \tilde{O}_i \right] \left(v_t - \tilde{v}_t \right) \right\|}_{=:(b)} + \underbrace{\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} \tilde{O}_i \right] v_t \right\|}_{=:(c)}.
\end{aligned} \tag{62}$$

We estimate the three summands in the right-hand side of (62) individually.

Estimating (a)**:** In order to upper-bound the first summand (a) we note that by the triangle inequality it holds that

$$\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] \tilde{v}_t \right\| \le \frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] \tilde{v}_t{}^\perp \right\| + \frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] \tilde{v}_t{}^\parallel \right\|. \tag{63}$$

Then (23) and (25) respectively imply that

$$\frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] \tilde{v}_t{}^\parallel \right\| = \frac{1}{m} \left\| \left[\sum_{i=1}^m \left(A_i \right)_{1,1} O_i \right] e_1 \right\| \cdot \| \tilde{v}_t{}^\parallel \| \le 4 \sqrt{\frac{n_1}{m}} \| \tilde{v}_t{}^\parallel \|$$

and

$$\frac{1}{m}\left\| \left[\sum_{i=1}^{m} (A_i)_{1,1}\, O_i\right] \tilde{v}_t{}^{\perp}\right\| \lesssim \sqrt{\frac{\log T + \log \eta}{m}}.$$

Plugging in these two estimates into (63) provides

$$\frac{1}{m}\left\| \left[\sum_{i=1}^{m} (A_i)_{1,1}\, O_i\right] \tilde{v}_t\right\| \lesssim \sqrt{\frac{\log T}{m}} + \sqrt{\frac{n_1}{m}}\|\tilde{v}_t{}^{\|}\|.$$

Estimating (b)**:** It follows from the restricted isometry property that

$$\frac{1}{m}\left\| \left[\sum_{i=1}^{m} (A_i)_{1,1}\, \tilde{O}_i\right] (v_t - \tilde{v}_t)\right\| = \left\| \left[\left(\tilde{\mathcal{O}}^*\mathcal{D}\right)\left(u_\star v_\star^\top\right)\right](v_t - \tilde{v}_t)\right\| \leq \delta \|v_t - \tilde{v}_t\|,$$

where in the last line we used Lemma 2.

Estimating (c)**:** By an analogous argument as for the first summand (a) we obtain for the third summand (c) that

$$\frac{1}{m}\left\| \left[\sum_{i=1}^{m} (A_i)_{1,1}\, \tilde{O}_i\right] v_t\right\| \lesssim \sqrt{\frac{\log T + \log \eta}{m}} + \sqrt{\frac{n_1}{m}}\|\tilde{v}_t{}^{\|}\|.$$

Hence, by summing up these estimates we have shown that

$$\left\| \left[\left(\mathcal{A}^*\mathcal{A} - \tilde{\mathcal{A}}^*\tilde{\mathcal{A}}\right)\left(u_\star v_\star^\top\right)\right] \tilde{v}_t\right\| \leq C\left(\sqrt{\frac{\log T}{m}} + \sqrt{\frac{n_1}{m}}\|\tilde{v}_t{}^{\|}\|\right) + \delta\|v_t - \tilde{v}_t\|,$$

which finishes the proof.

A.6. Proof of Lemma 7

It follows from $\mathcal{A} = \mathcal{D} + \mathcal{O}$ and $\mathcal{A} = \mathcal{D} + \tilde{\mathcal{O}}$ that

$$\begin{aligned} \mathcal{A}^*\mathcal{A} - \tilde{\mathcal{A}}^*\tilde{\mathcal{A}} &= (\mathcal{D} + \mathcal{O})^*(\mathcal{D} + \mathcal{O}) - \left(\mathcal{D} + \tilde{\mathcal{O}}\right)^*\left(\mathcal{D} + \tilde{\mathcal{O}}\right) \\ &= \mathcal{D}^*\mathcal{O} + \mathcal{O}^*\mathcal{D} + \mathcal{O}^*\mathcal{O} - \mathcal{D}^*\tilde{\mathcal{O}} - \tilde{\mathcal{O}}^*\mathcal{D} - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}} \\ &= \mathcal{D}^*\left(\mathcal{O} - \tilde{\mathcal{O}}\right) + \left(\mathcal{O} - \tilde{\mathcal{O}}\right)^*\mathcal{D} + \left(\mathcal{O}^*\mathcal{O} - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}}\right). \end{aligned} \tag{64}$$

Using decomposition in (64) and the triangle inequality we obtain that

$$\begin{aligned} &\left\| \left[\left(\mathcal{A}^*\mathcal{A} - \tilde{\mathcal{A}}^*\tilde{\mathcal{A}}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\right]\tilde{v}_t\right\| \\ &\leq \underbrace{\left\| \left[\left(\mathcal{D}^*\left(\mathcal{O} - \tilde{\mathcal{O}}\right)\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\right]\tilde{v}_t\right\|}_{=:(I)} + \underbrace{\left\| \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}}\right)^*\mathcal{D}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\right]\tilde{v}_t\right\|}_{=:(II)} \\ &+ \underbrace{\left\| \left[\left(\mathcal{O}^*\mathcal{O} - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}}\right)\right]\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\tilde{v}_t\right\|}_{=:(III)}. \end{aligned}$$

We estimate these three summands separately.

Bounding (I)**:** Note that

$$\begin{aligned}
&\left\| \left[\left(\mathcal{D}^* \left(\mathcal{O} - \tilde{\mathcal{O}} \right) \right) \left(\tilde{u}_{t+1/2} \tilde{v_t}^\top \right) \right] \tilde{v_t} \right\| \\
&\overset{(a)}{=} \left\| \left[\left(\mathcal{D}^* \left(\mathcal{O} - \tilde{\mathcal{O}} \right) \right) \left(\tilde{u}^{\|}_{t+1/2} (\tilde{v_t}^{\perp})^\top + \tilde{u}^{\perp}_{t+1/2} (\tilde{v_t}^{\|})^\top \right) \right] \tilde{v_t} \right\| \\
&\overset{(b)}{\leq} \left\| \left[\left(\mathcal{D}^* \left(\mathcal{O} - \tilde{\mathcal{O}} \right) \right) \left(\tilde{u}^{\|}_{t+1/2} (\tilde{v_t}^{\perp})^\top \right) \right] \tilde{v_t} \right\| + \left\| \left[\left(\mathcal{D}^* \left(\mathcal{O} - \tilde{\mathcal{O}} \right) \right) \left(\tilde{u}^{\perp}_{t+1/2} (\tilde{v_t}^{\|})^\top \right) \right] \tilde{v_t} \right\| \\
&\overset{(c)}{\leq} 2\delta \left(\| \tilde{u}^{\|}_{t+1/2} \| \| \tilde{v_t}^{\perp} \| + \| \tilde{u}^{\perp}_{t+1/2} \| \| \tilde{v_t}^{\|} \| \right) \\
&\overset{(d)}{\leq} 2\delta \left(\| \tilde{u}^{\|}_{t+1/2} \| + 2 \| \tilde{v_t}^{\|} \| \right),
\end{aligned}$$

where equality (a) follows from the definition of $\mathcal{O}$ and $\tilde{\mathcal{O}}$; Inequality (b) follows from the triangle inequality; Inequality (c) is due to Lemma 2 and the assumption that $\|\tilde{v_t}\| = 1$; Inequality (d) follows from $\|\tilde{v_t}\| = 1$ and $\|\tilde{u}_{t+1/2}\| \leq 2$.

Bounding (II)**:** By definition of $\mathcal{D}$ we have that

$$\left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(\tilde{u}_{t+1/2} \tilde{v_t}^\top \right) \right] \tilde{v_t} = \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(\tilde{u}^{\|}_{t+1/2} (\tilde{v_t}^{\|})^\top + \tilde{u}^{\perp}_{t+1/2} (\tilde{v_t}^{\perp})^\top \right) \right] \tilde{v_t}.$$

Hence by the triangle inequality it follows that

$$\begin{aligned}
&\left\| \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(\tilde{u}_{t+1/2} \tilde{v_t}^\top \right) \right] \tilde{v_t} \right\| \\
&\leq \underbrace{\left\| \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(\tilde{u}^{\|}_{t+1/2} (\tilde{v_t}^{\|})^\top \right) \right] \tilde{v_t} \right\|}_{=:(\S)} + \underbrace{\left\| \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(\tilde{u}^{\perp}_{t+1/2} \left(\tilde{v_t}^{\perp} \right)^\top \right) \right] \tilde{v_t} \right\|}_{=:(\S\S)}.
\end{aligned}$$

Estimating $(\S)$**:** In order to bound the first term we note that

$$\left\| \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(\tilde{u}^{\|}_{t+1/2} (\tilde{v_t}^{\|})^\top \right) \right] \tilde{v_t} \right\| = \| \tilde{u}^{\|}_{t+1/2} \| \cdot \| \tilde{v_t}^{\|} \| \cdot \left\| \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(u_\star v_\star^\top \right) \right] \tilde{v_t} \right\|.$$

Moreover note that

$$\left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(u_\star v_\star^\top \right) \right] \tilde{v_t} = \frac{1}{m} \left[\sum_{i=1}^{m} (A_i)_{1,1} \left(O_i - \tilde{O}_i \right) \right] \tilde{v_t}.$$

Note that this is exactly the term, which appeared already in the inequality chain (62). Hence, by exactly the same argument, since we assumed that eqs. (23) to (26) hold, we then obtain that

$$\begin{aligned}
&\left\| \left[\left(\left(\mathcal{O} - \tilde{\mathcal{O}} \right)^* \mathcal{D} \right) \left(\tilde{u}^{\|}_{t+1/2} (\tilde{v_t}^{\|})^\top \right) \right] \tilde{v_t} \right\| \\
&\lesssim \| \tilde{u}^{\|}_{t+1/2} \| \cdot \| v_t^{\|} \| \cdot \left(\sqrt{\frac{\log T + \log \eta}{m}} + \sqrt{\frac{n_1}{m}} \| \tilde{v_t}^{\|} \| + \delta \| v_t - \tilde{v_t} \| \right).
\end{aligned} \tag{65}$$

Estimating (§§)**:** In order to bound term (2) we note that

$$\left[\left(\left(\mathcal{O}-\tilde{\mathcal{O}}\right)^{*}\mathcal{D}\right)\left(\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^{\top}\right)\right]\tilde{v}_t=\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^{\top}\rangle_F\left(O_i-\tilde{O}_i\right)\tilde{v}_t.$$

Due to the triangle inequality it follows that

$$\begin{aligned}\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^{\top}\rangle_F\left(O_i-\tilde{O}_i\right)\tilde{v}_t\Big\| \leq\ & \underbrace{\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^{\top}\rangle O_i\tilde{v}_t\Big\|}_{=:(a)}\\ &+\underbrace{\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^{\top}\rangle_F\tilde{O}_i\left(\tilde{v}_t-v_t\right)\Big\|}_{=:(b)}\\ &+\underbrace{\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}-v_t^{\perp}\right)^{\top}\rangle_F\tilde{O}_i v_t\Big\|}_{=:(c)}\\ &+\underbrace{\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\left(\tilde{u}_{t+1/2}^{\perp}-u_{t+1/2}^{\perp}\right)\left(v_t^{\perp}\right)^{\top}\rangle_F\tilde{O}_i v_t\Big\|}_{=:(d)}\\ &+\underbrace{\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,u_{t+1/2}^{\perp}\left(v_t^{\perp}\right)^{\top}\rangle_F\tilde{O}_i v_t\Big\|.}_{=:(e)}\end{aligned}$$

We will estimate the summands individually.

Estimating (b)**,** (c)**, and** (d) **:** By the consequences of RIP in Lemma 2, the term (b) is upper-bounded by

$$\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^{\top}\rangle_F O_i\left(\tilde{v}_t-v_t\right)\Big\|\leq\delta\|\tilde{u}_{t+1/2}^{\perp}\|\cdot\|\tilde{v_t}^{\perp}\|\cdot\|\tilde{v}_t-v_t\|\leq 2\delta\|\tilde{v}_t-v_t\|,$$

where we used $\|\tilde{u}_{t+1/2}^{\perp}\|\leq 2$ and $\|\tilde{v_t}^{\perp}\|\leq 1$. Similarly we obtain that

$$\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}-v_t^{\perp}\right)^{\top}\rangle_F\tilde{O}_i v_t\Big\|\leq 2\delta\|\tilde{v_t}^{\perp}-v_t^{\perp}\|\leq 2\delta\|\tilde{v}_t-v_t\|$$

and

$$\Big\|\frac{1}{m}\sum_{i=1}^{m}\langle D_i,\left(\tilde{u}_{t+1/2}^{\perp}-u_{t+1/2}^{\perp}\right)\left(v_t^{\perp}\right)^{\top}\rangle_F\tilde{O}_i v_t\Big\|\leq\delta\|\tilde{u}_{t+1/2}^{\perp}-u_{t+1/2}^{\perp}\|\leq\delta\|\tilde{u}_{t+1/2}-u_{t+1/2}\|.$$

Estimating (a)**:** By the triangle inequality it holds that

$$\begin{aligned} &\Big\| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F O_i \tilde{v_t} \Big\| \\ \leq &\Big\| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F O_i \tilde{v_t}^{\perp} \Big\| + \Big\| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F O_i \tilde{v_t}^{\|} \Big\|. \end{aligned} \tag{66}$$

We estimate the two summands individually. Note that from the definition of O_i and $\tilde{v_t}^{\perp}$ it follows that only the first entry of $O_i\tilde{v_t}^{\perp}$ is non-zero. It follows that

$$\Big\| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F O_i \tilde{v_t}^{\perp} \Big\| = \Big| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F \langle O_i^\top e_i, \tilde{v_t}^{\perp}\rangle \Big|$$

Hence, it follows from (27) that

$$\begin{aligned} \Big\| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F O_i \tilde{v_t}^{\perp} \Big\| &\lesssim \sqrt{\frac{\log T + \log \eta}{m}} \cdot \|\mathcal{A}\left(\tilde{u}_{t+1/2}^{\perp}(\tilde{v_t}^{\perp})^\top\right)\| \\ &\lesssim \sqrt{\frac{\log T + \log \eta}{m}}, \end{aligned}$$

where in the second inequality we used the RIP of $\mathcal{A}$ as well as the assumption $\|\tilde{u}_{t+1/2}\| \leq 2$. This provides an upper bound on the first summand of the right-hand side in (66). In order to bound the second summand we first choose a vector $u \in \mathbb{C}^{n_1}$ that satisfies $\|u\| = 1$, $\langle u, u_\star \rangle = 0$, and

$$\Big\| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F O_i \tilde{v_t}^{\|} \Big\| = \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F \langle O_i \tilde{v_t}^{\|}, u \rangle.$$

Such a vector exists due to the definitions of O_i and $\tilde{v_t}^{\|}$ and the fact that the vector $O_i\tilde{v_t}^{\|}$ is orthogonal to $u_\star$. Hence, we obtain that

$$\begin{aligned} \Big\| \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F O_i \tilde{v_t}^{\|} \Big\| &= \frac{1}{m}\sum_{i=1}^m \langle D_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F \langle O_i, u\left(\tilde{v_t}^{\|}\right)^\top \rangle_F \\ &\overset{(i)}{=} \frac{1}{m}\sum_{i=1}^m \langle A_i, \tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top \rangle_F \langle A_i, u\left(\tilde{v_t}^{\|}\right)^\top \rangle_F \\ &\overset{(ii)}{=} \langle \mathcal{A}\left(\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top\right), \mathcal{A}\left(u\left(\tilde{v_t}^{\|}\right)^\top\right) \rangle \\ &\overset{(iii)}{\leq} \delta \|\tilde{u}_{t+1/2}^{\perp}\left(\tilde{v_t}^{\perp}\right)^\top\|_F \cdot \|u\left(\tilde{v_t}^{\|}\right)^\top\|_F \\ &\leq \delta \|\tilde{u}_{t+1/2}^{\perp}\| \cdot \|\tilde{v_t}^{\perp}\| \cdot \|u\| \cdot \|\tilde{v_t}^{\|}\| \\ &\overset{(iv)}{\leq} 2\delta \|\tilde{v_t}^{\|}\|, \end{aligned}$$

where the identity (i) follows from our choice of u and the definition of D_i and O_i; Equation (ii) follows from the definition of $\mathcal{A}$; Inequality (iii) is due to the consequences of RIP in Lemma 2; Inequality (iv) is obtained by $\|\tilde{u}^{\perp}_{t+1/2}\| \leq 2$, $\|u\| = 1$ and $\|\tilde{v}_t{}^{\perp}\| \leq 1$. Hence, we have shown that

$$(a) \lesssim \sqrt{\frac{\log T + \log \eta}{m}} + \delta \|\tilde{v}_t{}^{\|}\|.$$

Estimating (e)**:** We can upper-bound this term in an analogous way to term (a), which yields that

$$(e) \lesssim \sqrt{\frac{\log T + \log \eta}{m}} + \delta \|v_t^{\|}\|.$$

Summing up terms yields that

$$\begin{aligned}
(\S\S) &= (a) + (b) + (c) + (d) + (e) \\
&\lesssim \sqrt{\frac{\log T + \log \eta}{m}} + \delta\|\tilde{v}_t{}^{\|}\| + \delta\|\tilde{v}_t - v_t\| + \delta\|\tilde{u}_{t+1/2} - u_{t+1/2}\|\delta\|v_t^{\|}\| \qquad (67) \\
&\lesssim \sqrt{\frac{\log T + \log \eta}{m}} + \delta\|\tilde{v}_t{}^{\|}\| + \delta\|\tilde{v}_t - v_t\| + \delta\|\tilde{u}_{t+1/2} - u_{t+1/2}\|.
\end{aligned}$$

By combining (65) and (67), we obtain

$$\begin{aligned}
(II) &= (\S) + (\S\S) \\
&\lesssim \|\tilde{u}^{\|}_{t+1/2}\|\|v_t^{\|}\| \left(\sqrt{\frac{\log T + \log \eta}{m}} + \sqrt{\frac{n_1}{m}}\|\tilde{v}_t{}^{\|}\| + \delta\|v_t - \tilde{v}_t\| \right) \\
&+ \sqrt{\frac{\log T + \log \eta}{m}} + \delta\|\tilde{v}_t{}^{\|}\| + \delta\|\tilde{v}_t - v_t\| + \delta\|\tilde{u}_{t+1/2} - u_{t+1/2}\| \\
&\lesssim \sqrt{\frac{\log T + \log \eta}{m}} + \left(\delta + \sqrt{\frac{n_1}{m}}\right)\|\tilde{v}_t{}^{\|}\| + \delta\|\tilde{v}_t - v_t\| + \delta\|\tilde{u}_{t+1/2} - u_{t+1/2}\|.
\end{aligned}$$

Bounding (III)**:** Observe that

$$\begin{aligned}
&\left\| \left[\left(\mathcal{O}^*\mathcal{O} - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}} \right) \right] \left(\tilde{u}_{t+1/2}\tilde{v}_t{}^{\top} \right) \tilde{v}_t \right\| \\
&\overset{(a)}{=} \left\| \left[\left(\mathcal{O}^*\mathcal{O} - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}} \right) \left(\tilde{u}^{\|}_{t+1/2}(\tilde{v}_t{}^{\perp})^{\top} + \tilde{u}^{\perp}_{t+1/2}(\tilde{v}_t{}^{\|})^{\top} \right) \right] \tilde{v}_t \right\| \\
&\leq \left\| \left[\left(\mathcal{O}^*\mathcal{O} - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}} \right) \left(\tilde{u}^{\|}_{t+1/2}(\tilde{v}_t{}^{\perp})^{\top} \right) \right] \tilde{v}_t \right\| + \left\| \left[\left(\mathcal{O}^*\mathcal{O} - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}} \right) \left(\tilde{u}^{\perp}_{t+1/2}(\tilde{v}_t{}^{\|})^{\top} \right) \right] \tilde{v}_t \right\| \\
&\leq \left\| \left[\left(\mathcal{O}^*\mathcal{O} - P_O \right) \left(\tilde{u}^{\|}_{t+1/2}(\tilde{v}_t{}^{\perp})^{\top} \right) \right] \tilde{v}_t \right\| + \left\| \left[\left(P_O - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}} \right) \left(\tilde{u}^{\|}_{t+1/2}(\tilde{v}_t{}^{\perp})^{\top} \right) \right] \tilde{v}_t \right\| \\
&\quad + \left\| \left[\left(\mathcal{O}^*\mathcal{O} - P_O \right) \left(\tilde{u}^{\perp}_{t+1/2}(\tilde{v}_t{}^{\|})^{\top} \right) \right] \tilde{v}_t \right\| + \left\| \left[\left(P_O - \tilde{\mathcal{O}}^*\tilde{\mathcal{O}} \right) \left(\tilde{u}^{\perp}_{t+1/2}(\tilde{v}_t{}^{\|})^{\top} \right) \right] \tilde{v}_t \right\| \\
&\overset{(b)}{\leq} 2\delta \left(\|\tilde{u}^{\|}_{t+1/2}\| \cdot \|\tilde{v}_t{}^{\perp}\| + \|\tilde{u}^{\perp}_{t+1/2}\| \cdot \|\tilde{v}_t{}^{\|}\| \right) \\
&\overset{(c)}{\leq} 4\delta \left(\|\tilde{u}^{\|}_{t+1/2}\| + \|\tilde{v}_t{}^{\|}\| \right),
\end{aligned}$$

where the identity (a) follows from the definition of $\mathcal{O}$ and $\tilde{\mathcal{O}}$; Inequality (b) is due to Lemma 2; Inequality (c) follows from $\|\tilde{v}_t\| = 1$ and $\|\tilde{u}_{t+1/2}\| \leq 2$.

Finally, by combining the upper estimates of (I), (II), and (III), we obtain

$$
\begin{aligned}
&\left\| \left[\left(\mathcal{A}^*\mathcal{A} - \tilde{\mathcal{A}}^*\tilde{\mathcal{A}} \right) \left(\tilde{u}_{t+1/2} \tilde{v}_t{}^\top \right) \right] \tilde{v}_t \right\| \\
\leq &(I) + (II) + (III) \\
&\lesssim \delta \left(\|\tilde{u}_{t+1/2}^{\|}\| + \|v_t^{\|}\| \right) \\
&\quad + \left(\sqrt{\frac{\log T + \log \eta}{m}} + \delta + \sqrt{\frac{n_1}{m}} \right) \|\tilde{v}_t{}^{\|}\| + \delta \|\tilde{v}_t - v_t\| + \delta \|\tilde{u}_{t+1/2} - u_{t+1/2}\| \\
&\quad + \delta \left(\|\tilde{u}_{t+1/2}^{\|}\| + \|v_t^{\|}\| \right) \\
\lesssim &\sqrt{\frac{\log T + \log \eta}{m}} + \left(\delta + \sqrt{\frac{n_1}{m}} \right) \|\tilde{v}_t{}^{\|}\| + \delta \left\|\tilde{u}_{t+1/2}^{\|}\right\| + \delta \|\tilde{v}_t - v_t\| + \delta \|\tilde{u}_{t+1/2} - u_{t+1/2}\|.
\end{aligned}
$$

This completes the proof.

A.7. Proof of Lemma 8

The RIP of $\mathcal{A}$ provides

$$
\begin{aligned}
&\left| \langle \mathcal{A} \left(u_\star \left(v_t^\perp \right)^\top \right), \mathcal{A} \left(u_\star v_\star^\top \right) \rangle \right| \\
&\leq \left| \langle \mathcal{A} \left(u_\star \left(v_t^\perp - \tilde{v}_t{}^\perp \right)^\top \right), \mathcal{A} \left(u_\star v_\star^\top \right) \rangle \right| + \left| \langle \mathcal{A} \left(u_\star \left(\tilde{v}_t{}^\perp \right)^\top \right), \mathcal{A} \left(u_\star v_\star^\top \right) \rangle \right| \\
&\leq \delta \|v_t^\perp - \tilde{v}_t{}^\perp\| + \left| \langle \mathcal{A} \left(u_\star \left(\tilde{v}_t{}^\perp \right)^\top \right), \mathcal{A} \left(u_\star v_\star^\top \right) \rangle \right|.
\end{aligned}
\tag{68}
$$

The second term in the right-hand side of (68) is rewritten as

$$\begin{aligned}
\left|\langle \mathcal{A}\left(u_\star\left(\tilde{v}_t{}^\perp\right)^\top\right), \mathcal{A}\left(u_\star v_\star^\top\right)\rangle\right| &= \frac{1}{m}\Big|\sum_{i=1}^m \langle A_i, u_\star\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F \langle A_i, u_\star v_\star^\top\rangle_F\Big| \\
&= \frac{1}{m}\Big|\sum_{i=1}^m \langle A_i, u_\star\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F \left(A_i\right)_{1,1}\Big| \\
&= \frac{1}{m}\Big|\sum_{i=1}^m \langle O_i, u_\star\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F \left(A_i\right)_{1,1}\Big| \\
&= \frac{1}{m}\Big|\langle \sum_{i=1}^m \left(A_i\right)_{1,1} O_i, u_\star\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F\Big| \\
&= \frac{1}{m}\Big|\langle \sum_{i=1}^m \left(A_i\right)_{1,1} O_i \tilde{v}_t{}^\perp, u_\star\rangle\Big| \\
&= \frac{1}{m}\|\sum_{i=1}^m \left(A_i\right)_{1,1} O_i \tilde{v}_t{}^\perp\|.
\end{aligned}$$

Hence, the assumption in (25) implies

$$\left|\langle \mathcal{A}\left(u_\star\left(v_t^\perp\right)^\top\right), \mathcal{A}\left(u_\star v_\star^\top\right)\rangle\right| \lesssim \sqrt{\frac{\log T + \log n}{m}}\|\tilde{v}_t{}^\perp\|.$$

Inserting this inequality into (68) yields inequality (29).

It remains to show the inequality in (30). By applying the triangle inequality several times in combination with the RIP of $\mathcal{A}$ we obtain that

$$\begin{aligned}
&\left|\langle \mathcal{A}\left(u_\star\left(v_t^\perp\right)^\top\right), \mathcal{A}\left(u_{t+1/2}^\perp\left(v_t^\perp\right)^\top\right)\rangle\right| \\
&\le \delta\|u_{t+1/2}^\perp - \tilde{u}_{t+1/2}^\perp\| + 2\delta\|v_t^\perp - \tilde{v}_t{}^\perp\| \cdot \|u_{t+1/2}^\perp\| + \left|\langle \mathcal{A}\left(u_\star\left(\tilde{v}_t{}^\perp\right)^\top\right), \mathcal{A}\left(\tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\right)\rangle\right| \\
&\le \delta\|u_{t+1/2}^\perp - \tilde{u}_{t+1/2}^\perp\| + 4\delta\|v_t^\perp - \tilde{v}_t{}^\perp\| + \left|\langle \mathcal{A}\left(u_\star\left(\tilde{v}_t{}^\perp\right)^\top\right), \mathcal{A}\left(\tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\right)\rangle\right|,
\end{aligned} \tag{69}$$

where in the last inequality we used that $\|u_{t+1/2}^\perp\| \le 2$, which holds by Lemma 9 due to the RIP of $\mathcal{A}$. Next, we note that

$$\begin{aligned}
\left|\langle \mathcal{A}\left(u_\star\left(\tilde{v}_t{}^\perp\right)^\top\right), \mathcal{A}\left(\tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\right)\rangle\right| &= \frac{1}{m}\Big|\sum_{i=1}^m \langle A_i, u_\star\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F \langle A_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F\Big| \\
&= \frac{1}{m}\Big|\sum_{i=1}^m \langle O_i, u_\star\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F \langle D_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F\Big| \\
&= \frac{1}{m}\Big|\sum_{i=1}^m \langle O_i^\top e_1, \tilde{v}_t{}^\perp\rangle \langle D_i, \tilde{u}_{t+1/2}^\perp\left(\tilde{v}_t{}^\perp\right)^\top\rangle_F\Big|.
\end{aligned}$$

Hence, it follows from (27), the RIP of $\mathcal{A}$, and $\|\tilde{u}_{t+1/2}\| \le 2$ that

$$\begin{aligned}\left|\langle \mathcal{A}\left(u_\star \left(\tilde{v_t}^{\perp}\right)^\top\right), \mathcal{A}\left(\tilde{u}^{\perp}_{t+1/2}\left(\tilde{v_t}^{\perp}\right)^\top\right)\rangle\right| &\lesssim \sqrt{\frac{\log T + \log \eta}{m}} \cdot \|\mathcal{A}\left(\tilde{u}^{\perp}_{t+1/2}(\tilde{v_t}^{\perp})^\top\right)\| \\ &\lesssim \sqrt{\frac{\log T + \log \eta}{m}}.\end{aligned}$$

Combining this inequality with (69) yields (30).

B. Proofs of Lemmas in Phase 1

B.1. Proof of Lemma 9

It follows from the normal equations that

$$\begin{aligned}&u_{t+1/2} - \langle v_\star, v_t\rangle u_\star \\ =&\left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A})\left(u_{t+1/2}v_t^\top - u_\star v_\star^\top\right)\right] v_t \\ =&\left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A})\left(\left(u_{t+1/2} - \langle v_\star, v_t\rangle u_\star\right) v_t^\top\right)\right] v_t + \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A})\left(u_\star\left(\langle v_\star, v_t\rangle v_t^\top - v_\star^\top\right)\right)\right] v_t.\end{aligned}$$

In the following we will set for convenience that $\lambda_t = \langle v_\star, v_t\rangle$. Then we obtain by the previous calculation, the triangle inequality, and the Restricted Isometry Property that

$$\begin{aligned}&\|u_{t+1/2} - \lambda_t u_\star\| \\ \le&\left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A})\left(\left(u_{t+1/2} - \lambda_t u_\star\right) v_t^\top\right)\right] v_t\right\| + \left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A})\left(u_\star\left(\lambda_t v_t - v_\star\right)^\top\right)\right] v_t\right\| \\ \le&\delta\left(\|u_{t+1/2} - \lambda_t u_\star\| + \|\lambda_t v_t - v_\star\|\right),\end{aligned}$$

where in the last line we have used that $\|v_t\| = 1$. Rearranging terms yields that

$$\|u_{t+1/2} - \lambda_t u_\star\| \le \frac{\delta}{1-\delta}\left\|\lambda_t v_t - v_\star\right\|. \tag{70}$$

We compute that

$$\lambda_t v_t - v_\star = \lambda_t^2 v_\star + \lambda_t v_t^\perp - v_\star = \left(\lambda_t^2 - 1\right) v_\star + \lambda_t v_t^\perp.$$

Due to $1 - \lambda_t^2 = 1 - \langle v_\star, v_t\rangle^2 = \|v_t^\perp\|^2$ and $\lambda_t^2 = \|v_\star^{\|}\|^2$ this implies that

$$\begin{aligned}\|\lambda_t v_t - v_\star\|^2 &= \left(1 - \lambda_t^2\right)^2 \|v_\star\|^2 + \lambda_t^2 \|v_t^\perp\|^2 \\ &= \|v_t^\perp\|^4 + \|v_t^{\|}\|^2\|v_t^\perp\|^2 \\ &= \|v_t^\perp\|^2,\end{aligned}$$

where in the last line we used that $\|v_t^{\|}\|^2 + \|v_t^\perp\|^2 = \|v_t\|^2 = 1$. Together with (70) this shows (35). Since $\|u^\perp_{t+1/2}\| \le \left\|u_{t+1/2} - \langle v_\star, v_t\rangle u_\star\right\|$ this implies (36). In order to prove inequality (37) we note that

$$\begin{aligned}\|u_{t+1/2}\| &\le \|u_{t+1/2} - \langle v_\star, v_t\rangle u_\star\| + |\langle v_\star, v_t\rangle| \|u_\star\| \\ &\le \frac{\delta}{1-\delta}\|v_t^\perp\| + |\langle v_\star, v_t\rangle| \\ &\le \frac{\delta}{1-\delta} + |\langle v_\star, v_t\rangle| \\ &\le 2,\end{aligned}$$

where the third line follows from inequality (35) and from $\|u_\star\| = 1$. In the last line we used the assumption that $\delta \le \frac{1}{2}$ and $\|v_t\| = \|v_\star\| = 1$. This shows inequality (37).

B.2. Proof of Lemma 10

We will first show the following auxiliary inequality:

$$\|u^{\|}_{t+1/2} - \tilde{u}^{\|}_{t+1/2}\| \le (c_{2t} + C\delta\,(c_{2t}+1))\,\|v_t^{\|}\| + C\delta\|u^\perp_{t+1/2} - \tilde{u}^\perp_{t+1/2}\|, \tag{71}$$

where $C > 0$ is an absolute constant chosen large enough.

Proof of inequality (71): Recall that $u_{t+1/2}$ satisfies

$$u_{t+1/2} - \langle v_t, v_\star\rangle u_\star = [(\mathrm{Id} - \mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top - u_\star v_\star^\top)]v_t.$$

Then it follows that

$$u_{t+1/2} - \langle v_t, v_\star\rangle u_\star = [(\mathrm{Id} - \mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top - u_\star v_\star^\top)]v_t,$$

which is equivalently rewritten as

$$u^{\|}_{t+1/2} - \langle v_t, v_\star\rangle u_\star = \langle u_\star v_t^\top, (\mathrm{Id} - \mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top - u_\star v_\star^\top)\rangle u_\star. \tag{72}$$

Similarly $\tilde{u}_{t+1/2}$ also satisfies

$$\tilde{u}^{\|}_{t+1/2} - \langle \tilde{v}_t, v_\star\rangle u_\star = \langle u_\star \tilde{v}_t^\top, (\mathrm{Id} - \tilde{\mathcal{A}}^*\tilde{\mathcal{A}})(\tilde{u}_{t+1/2}\tilde{v}_t^\top - u_\star v_\star^\top)\rangle u_\star. \tag{73}$$

We obtain from (72) and (73) that

$$
\begin{aligned}
u_{t+1/2}^{\|}-\tilde{u}_{t+1/2}^{\|} =&\langle v_t-\tilde{v}_t, v_\star\rangle u_\star+\langle u_\star v_t^\top, (\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top-u_\star v_\star^\top)\rangle u_\star\\
&-\langle u_\star \tilde{v}_t^\top, (\mathrm{Id}-\tilde{\mathcal{A}}^*\tilde{\mathcal{A}})(\tilde{u}_{t+1/2}\tilde{v}_t^\top-u_\star v_\star^\top)\rangle_F u_\star\\
=&\langle v_t-\tilde{v}_t, v_\star\rangle u_\star+\langle u_\star\left(v_t-\tilde{v}_t\right)^\top, (\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top-u_\star v_\star^\top)\rangle_F u_\star\\
&+\langle u_\star\tilde{v}_t^\top, (\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top-u_\star v_\star^\top)-(\mathrm{Id}-\tilde{\mathcal{A}}^*\tilde{\mathcal{A}})(\tilde{u}_{t+1/2}\tilde{v}_t^\top-u_\star v_\star^\top)\rangle u_\star\\
=&\langle v_t-\tilde{v}_t, v_\star\rangle u_\star+\langle u_\star\left(v_t-\tilde{v}_t\right)^\top, (\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top-u_\star v_\star^\top)\rangle_F u_\star\\
&+\langle u_\star\tilde{v}_t^\top, (\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top)-(\mathrm{Id}-\tilde{\mathcal{A}}^*\tilde{\mathcal{A}})(\tilde{u}_{t+1/2}\tilde{v}_t^\top)\rangle_F u_\star\\
&+\langle u_\star\tilde{v}_t^\top, \left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(u_\star v_\star^\top\right)\rangle_F u_\star\\
=&\langle v_t-\tilde{v}_t, v_\star\rangle u_\star+\langle u_\star\left(v_t-\tilde{v}_t\right)^\top, (\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top-u_\star v_\star^\top)\rangle_F u_\star\\
&+\langle u_\star\tilde{v}_t^\top, (\mathrm{Id}-\mathcal{A}^*\mathcal{A})\left(u_{t+1/2}v_t^\top-\tilde{u}_{t+1/2}\tilde{v}_t^\top\right)\rangle_F u_\star\\
&+\langle u_\star\tilde{v}_t^\top, \left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t^\top\right)\rangle_F u_\star\\
&+\langle u_\star\tilde{v}_t^\top, \left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(u_\star v_\star^\top\right)\rangle_F u_\star.
\end{aligned}
$$

It follows from the triangle inequality, the restricted isometry property, and the Cauchy-Schwarz inequality that

$$
\begin{aligned}
\|u_{t+1/2}^{\|}-\tilde{u}_{t+1/2}^{\|}\| \leq&\|v_t^{\|}-\tilde{v}_t^{\|}\|+\delta\|v_t-\tilde{v}_t\|\cdot\|u_{t+1/2}v_t^\top-u_\star v_\star^\top\|_F\\
&+\delta\|\tilde{v}_t\|\cdot\|u_{t+1/2}v_t^\top-\tilde{u}_{t+1/2}\tilde{v}_t^\top\|_F+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t^\top\right)\right]\tilde{v}_t\right\|\\
&+\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(u_\star v_\star^\top\right)\right]\tilde{v}_t\|\\
\leq&\|v_t^{\|}-\tilde{v}_t^{\|}\|+\delta\|v_t-\tilde{v}_t\|\left(\|u_{t+1/2}\|\cdot\|v_t\|+\|u_\star\|\cdot\|v_\star\|\right)\\
&+\delta\|\tilde{v}_t\|\left(\|u_{t+1/2}-\tilde{u}_{t+1/2}\|\cdot\|v_t\|+\|\tilde{u}_{t+1/2}\|\cdot\|v_t-\tilde{v}_t\|\right)\\
&+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t^\top\right)\right]\tilde{v}_t\right\|\\
&+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(u_\star v_\star^\top\right)\right]\tilde{v}_t\right\|\\
\leq&\|v_t^{\|}-\tilde{v}_t^{\|}\|+5\delta\|v_t-\tilde{v}_t\|+2\delta\|u_{t+1/2}-\tilde{u}_{t+1/2}\|\\
&+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t^\top\right)\right]\tilde{v}_t\right\|\\
&+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(u_\star v_\star^\top\right)\right]\tilde{v}_t\right\|,
\end{aligned}
$$

where in the last inequality we have used the assumptions $\|u_{t+1/2}\|\leq 2$ and $\|\tilde{u}_{t+1/2}\|\leq 2$. Recall from Lemma 7 that

$$
\begin{aligned}
&\left\|\left[\left(\mathcal{A}^*\mathcal{A}-\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t^\top\right)\right]\tilde{v}_t\right\|\\
\lesssim&\sqrt{\frac{\log T}{m}}+\delta\|v_t^{\|}\|+\left(\delta+\sqrt{\frac{n_1}{m}}\right)\|\tilde{v}_t^{\|}\|+\delta\|\tilde{v}_t-v_t\|+\delta\|\tilde{u}_{t+1/2}-u_{t+1/2}\|.
\end{aligned}
\tag{74}
$$

From Lemma 6 it follows that

$$\left\| \left[\left(\tilde{\mathcal{A}}^* \tilde{\mathcal{A}} - \mathcal{A}^* \mathcal{A} \right) \left(u_\star v_\star^\top \right) \right] \tilde{v_t} \right\| \le C \left(\sqrt{\frac{\log T}{m}} + \sqrt{\frac{n_1}{m}} \| \tilde{v_t}^{\|} \| \right) + \delta \| v_t - \tilde{v_t} \|. \tag{75}$$

This implies that there is an absolute constant $\tilde{C}_1 > 0$ such that

$$\begin{aligned} &\| u_{t+1/2}^{\|} - \tilde{u}_{t+1/2}^{\|} \| \\ \le &\| v_t^{\|} - \tilde{v_t}^{\|} \| + \tilde{C}_1 \left(\sqrt{\frac{\log T}{m}} + \left(\sqrt{\frac{n_1}{m}} + \delta \right) \| \tilde{v_t}^{\|} \| + \delta \| v_t^{\|} \| + \delta \| v_t - \tilde{v_t} \| + \delta \| u_{t+1/2} - \tilde{u}_{t+1/2} \| \right). \end{aligned}$$

By using Assumption (32) and Condition ii) of Proposition 1, we obtain that

$$\begin{aligned} &\| u_{t+1/2}^{\|} - \tilde{u}_{t+1/2}^{\|} \| \\ \le &\| v_t^{\|} - \tilde{v_t}^{\|} \| + \tilde{C}_2 \left(\delta \| \tilde{v_t}^{\|} \| + \delta \| v_t^{\|} \| + \delta \| v_t - \tilde{v_t} \| + \delta \| u_{t+1/2} - \tilde{u}_{t+1/2} \| \right) \end{aligned}$$

with an absolute constant $\tilde{C}_2 > 0$ chosen large enough. By using the triangle inequality and Assumption (33) we obtain that

$$\| u_{t+1/2}^{\|} - \tilde{u}_{t+1/2}^{\|} \| \le \| v_t^{\|} - \tilde{v_t}^{\|} \| + \tilde{C}_3 \delta (1 + c_{2t}) \| v_t^{\|} \| + \tilde{C}_3 \delta \| u_{t+1/2} - \tilde{u}_{t+1/2} \|,$$

where $\tilde{C}_3 > 0$ is an absolute constant chosen large enough. By using the triangle inequality, by rearranging terms, and using the elementary inequality $1/(1-x) \le 1 + 2x$ for $0 < x < 1/2$ it follows that

$$\| u_{t+1/2}^{\|} - \tilde{u}_{t+1/2}^{\|} \| \le \left(1 + \tilde{C}_4 \delta \right) \| v_t^{\|} - \tilde{v_t}^{\|} \| + \tilde{C}_4 \delta (1 + c_{2t}) \| v_t^{\|} \| + \tilde{C}_4 \delta \| u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp} \|,$$

where $\tilde{C}_4 > 0$ is an absolute constant chosen large enough and we have used that $\delta > 0$ is chosen small enough. Using Assumption (33) we obtain that

$$\begin{aligned} \| u_{t+1/2}^{\|} - \tilde{u}_{t+1/2}^{\|} \| &\le c_{2t} \| v_t^{\|} \| + \tilde{C}_4 \delta (1 + c_{2t}) \| v_t^{\|} \| + \tilde{C}_4 \delta \| u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp} \| \\ &\le \left(\left(1 + 2 \tilde{C}_4 \delta \right) c_{2t} + \tilde{C}_4 \delta \right) \| v_t^{\|} \| + \tilde{C}_4 \delta \| u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp} \| \end{aligned}$$

This shows the auxiliary inequality (71).

Proof of inequality (42)**:** Having established the auxiliary inequality (71), we can in the next step prove inequality (42). It follows from the normal equations that

$$\begin{aligned} u_{t+1/2}^{\perp} &= P_{u_\star^\perp} [(\mathrm{Id} - \mathcal{A}^* \mathcal{A})(u_{t+1/2} v_t^\top - u_\star v_\star^\top)] v_t, \\ \tilde{u}_{t+1/2}^{\perp} &= P_{u_\star^\perp} [(\mathrm{Id} - \tilde{\mathcal{A}}^* \tilde{\mathcal{A}})(\tilde{u}_{t+1/2} \tilde{v_t}^\top - u_\star v_\star^\top)] \tilde{v_t}. \end{aligned}$$

Hence, we obtain that

$$\begin{aligned}
&\|u^{\perp}_{t+1/2}-\tilde{u}^{\perp}_{t+1/2}\|\\
\leq&\|[(\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top-u_\star v_\star^\top)]v_t-[(\mathrm{Id}-\tilde{\mathcal{A}}^*\tilde{\mathcal{A}})(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top-u_\star v_\star^\top)]\tilde{v}_t\|\\
\leq&\underbrace{\|[(\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top)]v_t-[(\mathrm{Id}-\tilde{\mathcal{A}}^*\tilde{\mathcal{A}})(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top)]\tilde{v}_t\|}_{=:(I)}\\
&+\underbrace{\|[(\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_\star v_\star^\top)]v_t-[(\mathrm{Id}-\tilde{\mathcal{A}}^*\tilde{\mathcal{A}})(u_\star v_\star^\top)]\tilde{v}_t\|}_{=:(II)}.
\end{aligned}$$

We estimate the first term by

$$\begin{aligned}
\|(I)\|\overset{(a)}{\leq}&\|(\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_{t+1/2}v_t^\top-\tilde{u}_{t+1/2}\tilde{v}_t{}^\top)v_t\|+\left\|\left[(\mathrm{Id}-\mathcal{A}^*\mathcal{A})(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top)\right](v_t-\tilde{v}_t)\right\|\\
&+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\right]\tilde{v}_t\right\|\\
\overset{(b)}{\leq}&\delta\|u_{t+1/2}v_t^\top-\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\|_F+\delta\|\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\|_F\|v_t-\tilde{v}_t\|\\
&+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\right]\tilde{v}_t\right\|\\
\overset{(c)}{\leq}&\delta\|u_{t+1/2}-\tilde{u}_{t+1/2}\|+3\delta\|v_t-\tilde{v}_t\|+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(\tilde{u}_{t+1/2}\tilde{v}_t{}^\top\right)\right]\tilde{v}_t\right\|\\
\overset{(d)}{\lesssim}&\sqrt{\frac{\log T}{m}}+\delta\|v_t^{\|}\|+\left(\delta+\sqrt{\frac{n_1}{m}}\right)\|\tilde{v}_t{}^{\|}\|+\delta\|\tilde{v}_t-v_t\|+\delta\|\tilde{u}_{t+1/2}-u_{t+1/2}\|.
\end{aligned}$$

In inequality (a) we used the triangle inequality and in inequality (b) we used the Restricted Isometry Property. In inequality (c) we used the triangle inequality as well as $\|u_{t+1/2}\|\leq 2$. Inequality (d) follows from inserting inequality (74). In the next step, we are going to estimate summand (II). For that, we observe

$$\begin{aligned}
\|(II)\|\leq&\|[(\mathrm{Id}-\mathcal{A}^*\mathcal{A})(u_\star v_\star^\top)]\,(v_t-\tilde{v}_t)\,\|+\left\|\left[\left(\tilde{\mathcal{A}}^*\tilde{\mathcal{A}}-\mathcal{A}^*\mathcal{A}\right)\left(u_\star v_\star^\top\right)\right]\tilde{v}_t\right\|\\
\leq&2\delta\|v_t-\tilde{v}_t\|+C\left(\sqrt{\frac{\log T}{m}}+\sqrt{\frac{n_1}{m}}\|\tilde{v}_t{}^{\|}\|\right),
\end{aligned}$$

where in the second inequality we have used inequality (75) and that $\mathcal{A}$ satisfies the Restricted Isometry Property. Hence, we have shown that

$$\begin{aligned}
\|u^{\perp}_{t+1/2}-\tilde{u}^{\perp}_{t+1/2}\|\leq&\|(I)\|+\|(II)\|\\
\lesssim&\sqrt{\frac{\log T}{m}}+\delta\|v_t^{\|}\|+\left(\delta+\sqrt{\frac{n_1}{m}}\right)\|\tilde{v}_t{}^{\|}\|+\delta\|\tilde{v}_t-v_t\|+\delta\|\tilde{u}_{t+1/2}-u_{t+1/2}\|\\
\leq&\sqrt{\frac{\log T}{m}}+\delta\|v_t^{\|}\|+\left(\delta+\sqrt{\frac{n_1}{m}}\right)\|\tilde{v}_t{}^{\|}\|+\delta\|\tilde{v}_t-v_t\|+\delta\|\tilde{u}^{\perp}_{t+1/2}-u^{\perp}_{t+1/2}\|\\
&+\delta\|\tilde{u}^{\|}_{t+1/2}-u^{\|}_{t+1/2}\|,
\end{aligned}$$

where for the last line we used the triangle inequality. Next, we obtain that

$$\begin{aligned}&\|u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp}\|\\ \lesssim&\delta\|v_t^{||}\| + \delta\|\tilde{v}_t{}^{||}\| + \delta\|\tilde{v}_t - v_t\| + \delta\|\tilde{u}_{t+1/2}^{\perp} - u_{t+1/2}^{\perp}\| + \delta\|\tilde{u}_{t+1/2}^{||} - u_{t+1/2}^{||}\|,\end{aligned}$$

where we have used Assumption (32) and Condition ii) of Proposition 1. By rearranging terms and using our assumption $\delta < 1/2$ we obtain that

$$\|u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp}\| \lesssim \delta\|v_t^{||}\| + \delta\|\tilde{v}_t{}^{||}\| + \delta\|\tilde{v}_t - v_t\| + \delta\|\tilde{u}_{t+1/2}^{||} - u_{t+1/2}^{||}\|.$$

By using the triangle inequality and Assumption (33) we obtain that

$$\|u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp}\| \lesssim \delta(1 + c_{2t})\|v_t^{||}\| + \delta\|\tilde{u}_{t+1/2}^{||} - u_{t+1/2}^{||}\|.$$

By inserting the auxiliary inequality (71) we obtain that

$$\|u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp}\| \lesssim\delta\left(1 + c_{2t}\right)\|v_t^{||}\| + \delta^2\|u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp}\|.$$

By rearranging terms we obtain that

$$\|u_{t+1/2}^{\perp} - \tilde{u}_{t+1/2}^{\perp}\| \lesssim \delta\left(1 + c_{2t}\right)\|v_t^{||}\|.$$

This shows the claimed inequality (42).

In order to finish the proof, it remains to prove inequality (41). For that, it suffices to note that this inequality follows from inserting inequality (42), which we have just shown, into the auxiliary inequality (71).

B.3. Proof of Lemma 11

For convenience, we set $\lambda_t = \langle v_t, v_\star\rangle$. We compute that

$$\begin{aligned}\|u_{t+1/2}^{||}\| &= |\langle u_{t+1/2}, u_\star\rangle|\\ &= |\langle u_{t+1/2} - \lambda_t u_\star, u_\star\rangle + \lambda_t\langle u_\star, u_\star\rangle|\\ &= |\langle u_{t+1/2} - \lambda_t u_\star, u_\star\rangle + \langle v_t, v_\star\rangle|.\end{aligned}$$

It follows from the triangle inequality and $|\langle v_t, v_\star\rangle| = \|v_t^{||}\|$ that

$$\|u_{t+1/2}^{||}\| - |\langle u_{t+1/2} - \lambda_t u_\star, u_\star\rangle| \le \|v_t^{||}\| \le \|u_{t+1/2}^{||}\| + |\langle u_{t+1/2} - \lambda_t u_\star, u_\star\rangle|. \qquad (76)$$

Hence, we need to bound $|\langle u_{t+1/2} - \lambda_t u_\star, u_\star \rangle|$ from above. For that purpose we compute that

$$\begin{aligned}
& u_{t+1/2} - \lambda_t u_\star \\
=& \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_t \\
=& \lambda_t \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star + \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_t^\perp \\
=& \lambda_t \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star + \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top \right) \right] v_t^\perp - \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_\star v_\star^\top \right) \right] v_t^\perp \\
=& \lambda_t \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star + \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top \right) \right] v_t^\perp + \left[(\mathcal{A}^*\mathcal{A}) \left(u_\star v_\star^\top \right) \right] v_t^\perp \\
=& \lambda_t \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star + \lambda_t \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_\star^\top \right) \right] v_t^\perp \\
& + \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} \left(v_t^\perp \right)^\top \right) \right] v_t^\perp + \left[(\mathcal{A}^*\mathcal{A}) \left(u_\star v_\star^\top \right) \right] v_t^\perp \\
=& \lambda_t \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star + \lambda_t \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_\star^\top \right) \right] v_t^\perp \\
& + \langle u_{t+1/2}, u_\star \rangle \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_\star \left(v_t^\perp \right)^\top \right) \right] v_t^\perp + \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2}^\perp \left(v_t^\perp \right)^\top \right) \right] v_t^\perp \\
& + \left[(\mathcal{A}^*\mathcal{A}) \left(u_\star v_\star^\top \right) \right] v_t^\perp .
\end{aligned}$$

It follows that

$$\begin{aligned}
& |\langle u_{t+1/2} - \lambda_t u_\star, u_\star \rangle| \\
\le& |\lambda_t| \cdot \left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star \right\| + |\lambda_t| \cdot \left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_\star^\top \right) \right] v_t^\perp \right\| \\
& + |\langle u_{t+1/2}, u_\star \rangle| \cdot \left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_\star \left(v_t^\perp \right)^\top \right) \right] v_t^\perp \right\| + \left| \langle u_\star, \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2}^\perp \left(v_t^\perp \right)^\top \right) \right] v_t^\perp \rangle \right| \\
& + \left| \langle \left[(\mathcal{A}^*\mathcal{A}) \left(u_\star v_\star^\top \right) \right] v_t^\perp, u_\star \rangle \right| \\
=& \|v_t^\parallel\| \cdot \left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star \right\| + \|v_t^\parallel\| \cdot \left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_\star^\top \right) \right] v_t^\perp \right\| \\
& + \|u_{t+1/2}^\parallel\| \cdot \left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_\star \left(v_t^\perp \right)^\top \right) \right] v_t^\perp \right\| + \left| \langle \mathcal{A} \left(u_{t+1/2}^\perp \left(v_t^\perp \right)^\top \right), \mathcal{A} \left(u_\star (v_t^\perp)^\top \right) \rangle \right| \\
& + \left| \langle \mathcal{A} \left(u_\star v_\star^\top \right), \mathcal{A} \left(u_\star \left(v_t^\perp \right)^\top \right) \rangle \right| .
\end{aligned}$$

By the RIP of $\mathcal{A}$ and the assumption $\|u_{t+1/2}\| \le 2$ we obtain that

$$\begin{aligned}
\left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_t^\top - u_\star v_\star^\top \right) \right] v_\star \right\| &\le \delta \| u_{t+1/2} v_t^\top - u_\star v_\star^\top \|_F \\
&\le \delta \left(\|u_{t+1/2}\| \cdot \|v_t\| + \|u_\star v_\star^\top\|_F \right) \\
&= \delta \left(\|u_{t+1/2}\| + 1 \right) \\
&\le 3\delta .
\end{aligned}$$

Furthermore, it follows from the Restricted Isometry Property and the assumption $\|u_{t+1/2}\| \le 2$ that

$$\left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_{t+1/2} v_\star^\top \right) \right] v_t^\perp \right\| \le \delta \|u_{t+1/2}\| \cdot \|v_\star\| \cdot \|v_t^\perp\| \le 2\delta$$

and

$$\left\| \left[(\mathrm{Id} - \mathcal{A}^*\mathcal{A}) \left(u_\star \left(v_t^\perp \right)^\top \right) \right] v_t^\perp \right\| \le \delta \|u_\star\| \cdot \|v_t^\perp\|^2 \le \delta.$$

We obtain that

$$\begin{aligned} &|\langle u_{t+1/2} - \lambda_t u_\star, u_\star \rangle| \\ &\le 5\delta \|v_t^\parallel\| + \delta \|u_{t+1/2}^\parallel\| + \left| \langle \mathcal{A}\left(u_{t+1/2}^\perp \left(v_t^\perp \right)^\top \right), \mathcal{A}\left(u_\star (v_t^\perp)^\top \right) \rangle \right| + \left| \langle \mathcal{A}\left(u_\star v_\star^\top \right), \mathcal{A}\left(u_\star \left(v_t^\perp \right)^\top \right) \rangle \right|. \end{aligned}$$

Recall from Lemma 8 that

$$\left| \langle \mathcal{A}\left(u_\star \left(v_t^\perp \right)^\top \right), \mathcal{A}\left(u_\star v_\star^\top \right) \rangle \right| \le \delta \|v_t^\perp - \tilde{v}_t{}^\perp\| + C\sqrt{\frac{\log T}{m}}$$

and

$$\left| \langle \mathcal{A}\left(u_\star \left(v_t^\perp \right)^\top \right), \mathcal{A}\left(u_{t+1/2}^\perp \left(v_t^\perp \right)^\top \right) \rangle \right| \le \delta \|u_{t+1/2}^\perp - \tilde{u}_{t+1/2}^\perp\| + 2\delta \|v_t^\perp - \tilde{v}_t{}^\perp\| + C\sqrt{\frac{\log T}{m}}.$$

Inserting these estimates into the above inequality we obtain that

$$\begin{aligned} &|\langle u_{t+1/2} - \lambda_t u_\star, u_\star \rangle| \\ &\le 5\delta \|v_t^\parallel\| + \delta \|u_{t+1/2}^\parallel\| + \delta \|u_{t+1/2}^\perp - \tilde{u}_{t+1/2}^\perp\| + 3\delta \|v_t^\perp - \tilde{v}_t{}^\perp\| + 2C\sqrt{\frac{\log T}{m}} \\ &\lesssim \delta \left(1 + c_{2t} \right) \|v_t^\parallel\| + \delta \|u_{t+1/2}^\parallel\|, \end{aligned}$$

where in the last line we used Assumptions (32), (33), Condition ii) of Proposition 1, and (42). By inserting this estimate into (76) and by rearranging terms we obtain inequality (43). This finishes the proof.

B.4. Proof of Lemma 13

Part 1 (Estimating $\|u^{\|}_{t+1} - \tilde{u}^{\|}_{t+1}\|$): First, we are going to estimate $\|u^{\|}_{t+1} - \tilde{u}^{\|}_{t+1}\|$. We compute that

$$\begin{aligned}
\|u^{\|}_{t+1} - \tilde{u}^{\|}_{t+1}\| &= \| \frac{u^{\|}_{t+1/2}}{\|u_{t+1/2}\|} - \frac{\tilde{u}^{\|}_{t+1/2}}{\|\tilde{u}_{t+1/2}\|} \| \\
&= \frac{\left\| \|\tilde{u}_{t+1/2}\| u^{\|}_{t+1/2} - \|u_{t+1/2}\| \tilde{u}^{\|}_{t+1/2} \right\|}{\|u_{t+1/2}\| \cdot \|\tilde{u}_{t+1/2}\|} \\
&\le \frac{\|u^{\|}_{t+1/2} - \tilde{u}^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|} + \frac{\left| \|\tilde{u}_{t+1/2}\| - \|u_{t+1/2}\| \right|}{\|u_{t+1/2}\|} \cdot \frac{\|\tilde{u}^{\|}_{t+1/2}\|}{\|\tilde{u}_{t+1/2}\|} \\
&= \frac{\|u^{\|}_{t+1/2} - \tilde{u}^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|} + \frac{\left| \|\tilde{u}_{t+1/2}\| - \|u_{t+1/2}\| \right|}{\|u_{t+1/2}\|} \cdot \|\tilde{u}^{\|}_{t+1}\| \\
&\le \underbrace{\frac{\|u^{\|}_{t+1/2} - \tilde{u}^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|}}_{=:(\S)} + \underbrace{\frac{\|\tilde{u}_{t+1/2} - u_{t+1/2}\|}{\|u_{t+1/2}\|} \cdot \|\tilde{u}^{\|}_{t+1}\|.}_{=:(\S\S)}
\end{aligned}$$

We estimate the two summands separately.

Estimation of (§)**:** We obtain that

$$\begin{aligned}
\frac{\|u^{\|}_{t+1/2} - \tilde{u}^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|} &\overset{(a)}{\le} (c_{2t} + C_1\delta(1+c_{2t})) \frac{\|v^{\|}_t\|}{\|u_{t+1/2}\|} \\
&\overset{(b)}{\le} (c_{2t} + C_1\delta(1+c_{2t}))\,(1 + C_2\delta(1+c_{2t}))\, \frac{\|u^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|} \\
&= (c_{2t} + C_1\delta(1+c_{2t}))\,(1 + C_2\delta(1+c_{2t}))\, \|u^{\|}_{t+1}\|
\end{aligned} \tag{77}$$

where in inequality (a) we have used Assumption (41) and in inequality (b) we have used Assumption (43).

Estimation of (§§)**:** By the triangle inequality we have

$$\frac{\|\tilde{u}_{t+1/2} - u_{t+1/2}\|}{\|u_{t+1/2}\|} \cdot \|\tilde{u}^{\|}_{t+1}\| \le \frac{\|\tilde{u}^{\|}_{t+1/2} - u^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|} \cdot \|\tilde{u}^{\|}_{t+1}\| + \frac{\|\tilde{u}^{\perp}_{t+1/2} - u^{\perp}_{t+1/2}\|}{\|u_{t+1/2}\|} \cdot \|\tilde{u}^{\|}_{t+1}\|. \tag{78}$$

Then we estimate the two summands in the right-hand side of (78) individually. It follows from (77) that the first summand is upper-bounded by

$$\frac{\|\tilde{u}^{\|}_{t+1/2} - u^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|} \cdot \|\tilde{u}^{\|}_{t+1}\| \le (c_{2t} + C_1\delta(1+c_{2t}))\,(1 + C_2\delta(1+c_{2t}))\, \|u^{\|}_{t+1}\| \cdot \|\tilde{u}^{\|}_{t+1}\|.$$

Moreover by Assumptions (42) and (43) the second summand is upper-bounded by

$$\begin{aligned}\frac{\|\tilde{u}_{t+1/2}^{\perp}-u_{t+1/2}^{\perp}\|}{\|u_{t+1/2}\|}\cdot\|\tilde{u}_{t+1}^{\|}\| &\overset{(a)}{\leq} C_1\delta\left(1+c_{2t}\right)\frac{\|v_t^{\|}\|}{\|u_{t+1/2}\|}\cdot\|\tilde{u}_{t+1}^{\|}\|\\ &\overset{(b)}{\leq} C_1\delta\left(1+c_{2t}\right)\left(1+C_2\delta(1+c_{2t})\right)\frac{\|u_{t+1/2}^{\|}\|}{\|u_{t+1/2}\|}\cdot\|\tilde{u}_{t+1}^{\|}\|\\ &= C_1\delta\left(1+c_{2t}\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\cdot\|\tilde{u}_{t+1}^{\|}\|.\end{aligned}$$

By combining the two estimates and inserting them into (78), we obtain that

$$\frac{\|\tilde{u}_{t+1/2}-u_{t+1/2}\|}{\|u_{t+1/2}\|}\cdot\|\tilde{u}_{t+1}^{\|}\| \leq \left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\cdot\|\tilde{u}_{t+1}^{\|}\|.$$

Combining the estimates: By combining the estimates for (§) and (§§) it follows that

$$\begin{aligned}&\|u_{t+1}^{\|}-\tilde{u}_{t+1}^{\|}\|\\ &\leq (\S)+(\S\S)\\ &\leq \left(c_{2t}+C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\\ &\quad+\left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\cdot\|\tilde{u}_{t+1}^{\|}\|\\ &\leq \left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\\ &\quad+\left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\left(\|\tilde{u}_{t+1}^{\|}\|+\|\tilde{u}_{t+1}^{\|}-u_{t+1}^{\|}\|\right),\end{aligned}$$

which is rearranged as

$$\begin{aligned}&\left(1-\left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\right)\|u_{t+1}^{\|}-\tilde{u}_{t+1}^{\|}\|\\ &\leq \left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\left(1+\|u_{t+1}^{\|}\|\right).\end{aligned} \tag{79}$$

Due to Lemma 12 we have $c_{2t}\lesssim 1$. Therefore one can choose c in (34) as a small absolute constant so that $\delta=\frac{c}{4\log n_2}$ satisfies

$$\left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|<\frac{1}{2}. \tag{80}$$

Then, since $\frac{1}{1-x}\leq 1+2x$ for $0<x<1/2$, it follows from (79) and (80) that

$$\begin{aligned}&\|u_{t+1}^{\|}-\tilde{u}_{t+1}^{\|}\|\\ &\leq \underbrace{\left(1+2\left(c_{2t}+2C_1\delta(1+c_{2t})\right)\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}^{\|}\|\right)}_{(i)}\\ &\quad\cdot\underbrace{\left(c_{2t}+2C_1\delta(1+c_{2t})\right)}_{(ii)}\underbrace{\left(1+C_2\delta(1+c_{2t})\right)}_{(iii)}\|u_{t+1}^{\|}\|\underbrace{\left(1+\|u_{t+1}^{\|}\|\right)}_{(iv)}\\ &\leq \left(1+\frac{C_3c}{\log n_2}\right)^3\left(c_{2t}+\frac{C_3c}{\log n_2}\right)\|u_{t+1}^{\|}\|\end{aligned} \tag{81}$$

for some absolute constant C_3, where the second inequality follows from the assumptions $\|u^{\|}_{t+1}\| < \frac{c}{\log n_2}$ and $\delta = \frac{c}{4\log n_2}$, and the fact that $c_{2t} \le C_0$ due to Lemma 12. Indeed, the above conditions imply

$$\begin{aligned}(i) &= 1 + 2\left(c_{2t} + 2C_1\delta(1+c_{2t})\right)\left(1 + C_2\delta(1+c_{2t})\right)\|u^{\|}_{t+1}\| \\ &\le 1 + \left(C_0 + \frac{2C_1(C_0+1)c}{4\log n_2}\right)\cdot\left(1 + \frac{C_2(C_0+1)c}{4\log n_2}\right)\cdot\frac{c}{\log n_2}.\end{aligned}$$

Then we need to choose C_3 so that

$$\left(C_0 + \frac{2C_1(C_0+1)c}{4\log n_2}\right)\cdot\left(1 + \frac{C_2(C_0+1)c}{4\log n_2}\right) \le C_3.$$

The constant C_3 also needs to satisfy

$$\begin{aligned}(ii) &= c_{2t} + 2C_1\delta(1+c_{2t}) \le c_{2t} + \frac{2C_1(C_0+1)c}{\log n_2} \le c_{2t} + \frac{C_3 c}{\log n_2}, \\ (iii) &= 1 + C_2\delta(1+c_{2t}) \le 1 + \frac{C_2(C_0+1)c}{\log n_2} \le 1 + \frac{C_3 c}{\log n_2},\end{aligned}$$

and

$$(iv) = 1 + \|u^{\|}_{t+1}\| \le 1 + \frac{c}{4\log n_2} \le 1 + \frac{C_3 c}{\log n_2}.$$

This is implied by

$$\max\left\{2C_1(C_0+1), C_2(C_0+1), \frac{1}{4}\right\} \le C_3.$$

Thus, there exists an absolute constant $C_3 > 0$ that satisfies the above conditions. Then one can choose an absolute constant $c > 0$ small enough so that the upper bound in (81) reduces to

$$\|u^{\|}_{t+1} - \tilde{u}^{\|}_{t+1}\| \le \underbrace{\left[\left(1 + \frac{1}{\log n_2}\right)c_{2t} + \frac{1}{\log n_2}\right]}_{=:c_{2t+1}}\|u^{\|}_{t+1}\|. \tag{82}$$

Thus we have shown the claimed bound for $\|u^{\|}_{t+1} - \tilde{u}^{\|}_{t+1}\|$.

Part 2 (Estimating $\|u^{\perp}_{t+1} - \tilde{u}^{\perp}_{t+1}\|$): Analogous as in the beginning of the proof, where we provided an estimate for $\|u^{\|}_{t+1} - \tilde{u}^{\|}_{t+1}\|$, we can show that

$$\|u^{\perp}_{t+1} - \tilde{u}^{\perp}_{t+1}\| \le \frac{\|u^{\perp}_{t+1/2} - \tilde{u}^{\perp}_{t+1/2}\|}{\|u_{t+1/2}\|} + \frac{\|\tilde{u}_{t+1/2} - u_{t+1/2}\|}{\|u_{t+1/2}\|}\cdot\|\tilde{u}^{\perp}_{t+1}\|.$$

By using the triangle inequality and $\|\tilde{u}^{\perp}_{t+1}\| \le 1$ it follows that

$$\|u^{\perp}_{t+1} - \tilde{u}^{\perp}_{t+1}\| \le \frac{2\|u^{\perp}_{t+1/2} - \tilde{u}^{\perp}_{t+1/2}\|}{\|u_{t+1/2}\|} + \frac{\|\tilde{u}^{\|}_{t+1/2} - u^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|}. \tag{83}$$

We are going to estimate the two summands individually. By Assumptions (42) and (43), the first summand is upper-bounded by

$$\begin{aligned}\frac{2\|u^{\perp}_{t+1/2}-\tilde{u}^{\perp}_{t+1/2}\|}{\|u_{t+1/2}\|} &\overset{(a)}{\leq} \frac{2C_1\delta(1+c_{2t})\|v^{\|}_t\|}{\|u_{t+1/2}\|}\\ &\overset{(b)}{\leq} 2C_1\delta(1+c_{2t})\left(1+C_2\delta(1+c_{2t})\right)\|u_{t+1}\|.\end{aligned}$$

Moreover, we use the estimate from the inequality chain (77) to obtain that

$$\frac{\|\tilde{u}^{\|}_{t+1/2}-u^{\|}_{t+1/2}\|}{\|u_{t+1/2}\|} \leq \left(c_{2t}+C_1\delta(1+c_{2t})\right)\left(1+C_2(1+c_{2t})\delta\right)\|u^{\|}_{t+1}\|.$$

Hence, by inserting these estimates into (83), we obtain that

$$\begin{aligned}\|u^{\perp}_{t+1}-\tilde{u}^{\perp}_{t+1}\| &\leq \left(c_{2t}+3C_1\delta(1+c_{2t})\right)\left(1+C_2(1+c_{2t})\delta\right)\|u^{\|}_{t+1}\|\\ &\leq \left(c_{2t}+\frac{C_4 c}{\log n_2}\right)\left(1+\frac{C_4 c}{\log n_2}\right)\|u^{\|}_{t+1}\|\\ &= \left[\left(1+\frac{C_4 c}{\log n_2}\right)c_{2t}+\frac{C_4 c}{\log n_2}\left(1+\frac{C_4 c}{\log n_2}\right)\right]\|u^{\|}_{t+1}\|\end{aligned}$$

for some absolute constant C_4, where the second inequality is dervied similarly to that of (81). Since $c_{2t}\lesssim 1$, by choosing c as a small enough absolute constant so that

$$\|u^{\perp}_{t+1}-\tilde{u}^{\perp}_{t+1}\| \leq \left[\left(1+\frac{1}{\log n_2}\right)c_{2t}+\frac{1}{\log n_2}\right]\|u^{\|}_{t+1}\| = c_{2t+1}\|u^{\|}_{t+1}\|, \qquad (84)$$

Then combining (82) and (84) provides (45). This finishes the proof.

B.5. Proof of Lemma 14

We observe that

$$\begin{aligned}\|u^{\|}_{t+1}\|^2 &\overset{(a)}{=} \frac{\|u^{\|}_{t+1/2}\|^2}{\|u_{t+1/2}\|^2}\\ &= \frac{\|u^{\|}_{t+1/2}\|^2}{\|u^{\|}_{t+1/2}\|^2+\|u^{\perp}_{t+1/2}\|^2}\\ &\overset{(b)}{\geq} \frac{\alpha\|v^{\|}_t\|^2}{\beta\|v^{\perp}_t\|^2+\alpha\|v^{\|}_t\|^2}\\ &\overset{(c)}{=} \frac{\alpha\|v^{\|}_t\|^2}{\beta+(\alpha-\beta)\|v^{\|}_t\|^2}.\end{aligned}$$

In equality (a) we used the definition of $u_{t+1/2}$. Inequality (b) from the inequalities (46) and (47). Equality (c) is due to $\|v_t\|=1$. This shows the first inequality in (48), from which the second inequality can be deduced immediately. In a similar manner we obtain that

$$\begin{aligned}
\|u_{t+1}^{\perp}\|^2 &= \frac{\|u_{t+1/2}^{\perp}\|^2}{\|u_{t+1/2}\|^2} \\
&= \frac{\|u_{t+1/2}^{\perp}\|^2}{\|u_{t+1/2}^{\perp}\|^2 + \|u_{t+1/2}^{\|}\|^2} \\
&\leq \frac{\beta\|v_t^{\perp}\|^2}{\beta\|v_t^{\perp}\|^2 + \alpha\|v_t^{\|}\|^2} \\
&\leq \frac{\beta}{\alpha\|v_t^{\|}\|^2} \cdot \|v_t^{\perp}\|^2,
\end{aligned}$$

which finishes the proof.