Streaming k-PCA: Efficient guarantees for Oja's algorithm, beyond rank-one updates

Preprint · February 2021

3 authors, including:

De Huang
Peking University
18 PUBLICATIONS **69** CITATIONS

SEE PROFILE

STREAMING k-PCA: EFFICIENT GUARANTEES FOR OJA'S ALGORITHM, BEYOND RANK-ONE UPDATES

DE HUANG, JONATHAN NILES-WEED, AND RACHEL WARD

Abstract. We analyze Oja's algorithm for streaming k-PCA, and prove that it achieves performance nearly matching that of an optimal offline algorithm. Given access to a sequence of i.i.d. $d \times d$ symmetric matrices, we show that Oja's algorithm can obtain an accurate approximation to the subspace of the top k eigenvectors of their expectation using a number of samples that scales polylogarithmically with d. Previously, such a result was only known in the case where the updates have rank one.

Our analysis is based on recently developed matrix concentration tools, which allow us to prove strong bounds on the tails of the random matrices which arise in the course of the algorithm's execution.

1. Introduction

Principal component analysis is one of the foundational algorithms of statistics and machine learning. From a practical perspective, perhaps no optimization problem is more widely used in data analysis [18]. From a theoretical perspective, it is one of the simplest examples of a non-convex optimization problem that can nevertheless be solved in polynomial time; as such, it has been an important proving ground for understanding the fundamental limits of efficient optimization [30].

In the basic setting, the practitioner has access to a sequence of independent symmetric random matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots$ with expectation $\boldsymbol{M} \in \mathbb{R}^{d\times d}$. The goal is to approximate the leading eigenspace of $\boldsymbol{M}$ or, more generally, to approximate the subspace spanned by its leading k eigenvectors. While it is natural to attempt to solve this problem by performing an eigen-decomposition of the empirical average $\bar{\boldsymbol{A}} = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{A}_i$, the amount of space required by this approach can be prohibitive when d is large. In particular, if the matrices $\boldsymbol{A}_i$ are sparse or low-rank, performing incremental updates with the matrices $\boldsymbol{A}_i$ may be significantly cheaper than storing all the iterates or their average. A tremendous amount of attention has therefore been paid to designing algorithms which can cheaply and provably estimate the subspace spanned by the top k eigenvectors of $\boldsymbol{M}$ using limited memory and a single pass over the data, a problem known as *streaming PCA* [17].

The simplest and most natural approach to this problem was proposed nearly 40 years ago by Oja [25, 26]:

(1) Randomly choose an initial guess $\boldsymbol{Z}_0 \in \mathbb{R}^{d\times k}$, and set $\boldsymbol{Q}_0 \leftarrow \mathsf{QR}[\boldsymbol{Z}_0]$
(2) For $t \geq 1$, set $\boldsymbol{Q}_t \leftarrow \mathsf{QR}[(\mathbf{I} + \eta_t \boldsymbol{A}_t)\boldsymbol{Q}_{t-1}]$.

Here, $\mathsf{QR}[\boldsymbol{Q}_t]$ returns an orthogonal $\mathbb{R}^{d\times k}$ matrix obtained by performing the Gram–Schmidt process to the columns of $\boldsymbol{Q}_t$. It is easy to see [1, Lemma 2.2] that the Gram–Schmidt step commutes with the multiplicative update, so that we can equivalently consider a version of the algorithm which performs a single orthonormalization at the end, and outputs

$$\boldsymbol{Q}_t = \mathsf{QR}[\boldsymbol{Z}_t]\,, \quad \boldsymbol{Z}_t = \boldsymbol{Y}_t \ldots \boldsymbol{Y}_1 \boldsymbol{Z}_0\,,$$

Date: 6 February 2021.

The authors gratefully acknowledge the funding for this work. DH was in part supported by NSF Grants DMS-1907977, DMS-1912654, and the Choi Family Postdoc Gift Fund. JNW was supported under NSF grant DMS-2015291. JNW and RW were supported in part by the Institute for Advanced Study, where some of this research was conducted. RW received support from AFOSR MURI Award N00014-17-S-F006 and NSF grant DMS-1952735 .

where $\boldsymbol{Y}_i := (\mathbf{I} + \eta_i \boldsymbol{A}_i)$.

Oja's algorithm can be viewed as a noisy version of the classic orthogonal iteration algorithm for computing invariant subspaces of a symmetric matrix [12, Section 7.3.2]; alternatively, it corresponds to projected stochastic gradient descent on the Stiefel manifold of matrices with orthonormal columns [9]. Despite its simplicity and practical effectiveness, Oja's algorithm has proven challenging to analyze because of its inherent non-convexity.

As a benchmark against which to compare Oja's algorithm, we may consider the performance of the simple offline algorithm which computes the leading k eigenvectors of $\bar{\boldsymbol{A}}$. We write $\boldsymbol{V} \in \mathbb{R}^{d\times k}$ for the orthogonal matrix whose columns are the leading k eigenvectors of $\boldsymbol{M}$ and $\hat{\boldsymbol{V}} \in \mathbb{R}^{d\times k}$ for the matrix containing the leading k eigenvectors of $\bar{\boldsymbol{A}}$, and measure the quality of $\hat{\boldsymbol{V}}$ by the following standard measure of distance between subspaces:

$$\operatorname{dist}(\hat{\boldsymbol{V}}, \boldsymbol{V}) := \|\boldsymbol{V}\boldsymbol{V}^* - \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^*\|$$

If $\|\boldsymbol{A}_i - \boldsymbol{M}\| \le M$ almost surely and the gap between the kth and $(k+1)$th eigenvalues is ρ_k, then the Matrix Bernstein inequality [31, Theorem 1.4] combined with Wedin's Theorem [33] implies that there exists a positive constant C such that

$$\operatorname{dist}(\hat{\boldsymbol{V}}, \boldsymbol{V}) \le C\frac{M}{\rho_k}\sqrt{\frac{\log(d/\delta)}{T}}. \tag{1.1}$$

with probability at least $1-\delta$.

The key question is whether Oja's algorithm is able to achieve similar performance. However, except in the special *rank-one* case where either $k = 1$ or $\operatorname{rank}(\boldsymbol{A}_i) = 1$ almost surely, no such bound is known.

1.1. **Our contribution.** We give the first results for Oja's algorithm nearly matching (1.1), for any $k \ge 1$ and updates of any rank. Our main result (Theorem 2.3) establishes that, after a burn-in period of $T_0 = \tilde{O}\left(\frac{kM^2}{\delta^2\rho_k^2}\right)$ steps, the output of Oja's algorithm satisfies

$$\operatorname{dist}(\boldsymbol{Q}_T, \boldsymbol{V}) \le C'\frac{M}{\rho_k}\sqrt{\frac{\log(kM/\delta\rho_k)}{T - T_0}}$$

with probability at least $1-\delta$ for a universal positive constant C'. Ours is the first work to show that Oja's algorithm can achieve a guarantee similar to (1.1) beyond the rank-one case.

The assumption that $k = 1$ or $\operatorname{rank}(\boldsymbol{A}_i) = 1$ is fundamental to the proof strategies used in prior works. To show that the error decays sufficiently quickly, prior work focuses on the quantity $\|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\|_2$, where the columns of $\boldsymbol{U}$ are the last $d-k$ eigenvectors of $\boldsymbol{M}$, which is an upper bound on $\operatorname{dist}(\boldsymbol{Q}_t, \boldsymbol{V})$. (See Lemma 2.6, below.) The key challenge is to control the inverse $(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}$. When $k = 1$, as in [17], this quantity is a scalar, so it can be pulled out of the norm and bounded separately. This is no longer possible when $k > 1$, but if $\operatorname{rank}(\boldsymbol{A}_i) = 1$, as in [1], then $\boldsymbol{V}^*\boldsymbol{Z}_t$ can be written as a rank-one perturbation of $\boldsymbol{V}^*\boldsymbol{Z}_{t-1}$. The Sherman–Morrison formula then implies that $\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}$ can be written as $\boldsymbol{U}^*\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*\boldsymbol{Z}_{t-1})^{-1}$ plus the sum of explicit, rank-one correction terms. However, if neither $k = 1$ nor $\operatorname{rank}(\boldsymbol{A}_i) = 1$, this approach quickly becomes infeasible, since the correction terms now involve a product of rank-k matrices whose norm is difficult to bound.

A more subtle difficulty implicit in prior work is that proofs must be carried out entirely in expected (squared) Frobenius norm. This requirement is necessitated by the fact that the Frobenius norm is Hilbertian, so it is possible to employ the crucial Pythagorean identity

$$\mathbb{E}\|\boldsymbol{Y}\|_2^2 = \|\mathbb{E}\boldsymbol{Y}\|_2^2 + \|\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y}\|_2^2 \tag{1.2}$$

for any random matrix $\boldsymbol{Y}$. It is this identity that makes it possible to control the evolution of $\mathbb{E}\|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\|_2^2$. However, as our proofs reveal, it is of significant utility to be able to recursively control the operator norm $\|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\|$ with high probability instead. Unfortunately, (1.2) is of no help in proving statements of this kind.

Our argument handles both challenges and represents a significant conceptual simplification over earlier proofs. Our crucial insight is that, rather than using the squared Frobenius norm, it is possible to prove a stronger recursion in a different norm, which implies high-probability bounds. Using techniques recently developed by [16] to prove concentration inequalities for products of random matrices, we show that conditioned on $\|\boldsymbol{U}^*\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*\boldsymbol{Z}_{t-1})^{-1}\|$ being well behaved, the probability that $\|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\|$ deviates significantly from its expectation is exponentially small.

In other words, good concentration properties for $\|\boldsymbol{U}^*\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*\boldsymbol{Z}_{t-1})^{-1}\|$ imply good concentration properties for the next iterate, $\|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\|$. These high-probability bounds significantly simplify the calculations, since they allow us to guarantee that the problematic error terms appearing in prior work are small.

If we knew that $\|\boldsymbol{U}^*\boldsymbol{Z}_0(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\| = O(1)$ with high probability, then the above induction argument would allow us to conclude that $\|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\| = O(1)$ for all t. Unfortunately, this is not the case: if $\boldsymbol{Z}_0$ is randomly initialized with i.i.d. Gaussian entries, then typically

$$\|\boldsymbol{U}^*\boldsymbol{Z}_0(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\| \asymp \sqrt{dk}\,.$$

We therefore adopt a two-phase approach: in the first, short phase, of length approximately $\log d$, we show that the operator norm decays from $O(\sqrt{dk})$ to $O(1)$, and in the second phase we use the above recursive argument to establish that the operator norm decays to zero at a $O(1/\sqrt{T})$ rate. To simplify the analysis of the first phase, we develop a coupling argument that allows us reduce without loss of generality to the case where the law $\mathrm{P}_{\boldsymbol{A}}$ of the random matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots$ has finite support and obtain almost-sure guarantees by a simple union bound. This weak control is enough to guarantee that $\|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\|$ decays exponentially fast, so that it is of constant order after approximately $\log d$ iterations.

1.2. **Prior work.** Obtaining non-asymptotic rates of convergence for Oja's algorithm and its variants has been an area of active recent interest [28, 29, 27, 21, 20, 2, 4, 13, 17, 23]. Apart from the results of [1] and [17], none of these works proves bounds matching (1.1).

A breakthrough in the project of obtaining optimal guarantees was due to [28], who gave an analysis of Oja's algorithm that works when provided with a warm start: he showed that, when $k = 1$ and $\mathrm{rank}(\boldsymbol{A}_i) = 1$ almost surely, Oja's algorithm converges in a number of steps logarithmic in d if it is initialized in a neighborhood of the optimum, but his result does not extend to random initialization and it is unclear how to find a warm start in practice. This restriction was lifted by [17], who were the first to show a global, efficient guarantee for Oja's algorithm when $k = 1$. Subsequently, [1] gave a global, efficient guarantee for Oja's algorithm in the $k > 1$ case, but under the restriction that $\mathrm{rank}(\boldsymbol{A}_i) = 1$ almost surely.

The idea of analyzing Oja's algorithm by developing concentration bounds for products of random matrices was suggested by [15], who also proved such non-asymptotic concentration bounds in a simplified setting. Those bounds were later improved by [16] who developed a different technique based on martingale inequalities for Schatten norms, following a strategy pursued by [19] and [24] for other Banach space norms. The concentration inequalities of [16] are not sharp enough to recover optimal rates for Oja's algorithm on their own; in this work, we use a similar proof techniques to establish tailor-made concentration results for the Oja setting.

1.3. **Organization of the remainder of the paper.** In Section 2, we give our main results and an overview of our techniques. Our main tool is a recursive inequality which proves a concentration result for the iterates of Oja's algorithm, which we state and prove in Section 3.

Our analysis of Oja's algorithm involves two distinct phases, which we analyze separately. Since the argument for the second phase is simpler, we present it first in Section 4, and present the slightly more complicated argument for the first phase in Section 5. We conclude in Section 6 with open questions and directions for future work. The appendices contain omitted proofs and supplementary results for each section.

1.4. **Notation.** We write $\lambda_1 \geq \cdots \geq \lambda_d$ for the eigenvalues of the symmetric matrix $\boldsymbol{M}$, and we write $\rho_k := \lambda_k - \lambda_{k+1}$ for the gap between the kth and $(k+1)$th eigenvalue. We write $\boldsymbol{V} \in \mathbb{R}^{d\times k}$ for the orthogonal matrix whose columns are the k leading eigenvectors of $\boldsymbol{M}$, and $\boldsymbol{U} \in \mathbb{R}^{d\times(d-k)}$ for the orthogonal matrix whose columns are the remaining eigenvectors. Given an orthogonal matrix $\boldsymbol{W} \in \mathbb{R}^{d\times k}$, we write [7]

$$\operatorname{dist}(\boldsymbol{W}, \boldsymbol{V}) = \|\boldsymbol{V}\boldsymbol{V}^* - \boldsymbol{W}\boldsymbol{W}^*\| = \|\boldsymbol{U}^*\boldsymbol{W}\|,$$

The symbol $\|\cdot\|$ denotes the spectral norm (i.e., ℓ_2 operator norm) of a matrix, which is equal to its maximum singular value. For $p \geq 1$, the symbol $\|\cdot\|_p$ denotes the Schatten p-norm, which is the ℓ_p norm of the singular values of its argument. We also define the L_p norm of a random matrix $\boldsymbol{X}$ as

$$\|\boldsymbol{X}\|_{p,p} := \left(\mathbb{E}\,\|\boldsymbol{X}\|_p^p\right)^{1/p}.$$

We employ standard asymptotic notation $a = O(b)$ to indicate that $a \leq Cb$ for a universal positive constant C, and write $a = \Theta(b)$ if $a = O(b)$ and $b = O(a)$. The notations $\tilde{O}(\cdot)$ and $\tilde{\Theta}(\cdot)$ suppress polylogarithmic factors in the problem parameters. When t is a positive integer, we write $[t] := \{1, \ldots, t\}$.

2. Techniques and main results

We focus throughout on the following setup:

Assumption 2.1. The matrices $\boldsymbol{A}_i$ are symmetric, independent, identically distributed samples from a distribution $\mathrm{P}_{\boldsymbol{A}}$, with expectation $\boldsymbol{M}$.

Note that while we require that each $\boldsymbol{A}_i$ is symmetric, we do not require that $\boldsymbol{A}_i \succeq \boldsymbol{0}$.

The requirement that $\boldsymbol{A}_i$ is symmetric is not as restrictive as it may seem, since we can replace $\boldsymbol{A}_i$ by its *Hermitian dilation*:

$$\mathscr{D}(\boldsymbol{A}_i) := \begin{pmatrix} \boldsymbol{0} & \boldsymbol{A}_i \\ \boldsymbol{A}_i^* & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{2d\times 2d}.$$

Estimating the leading eigenvectors of $\mathscr{D}(\boldsymbol{M})$ is equivalent to estimating the leading singular vectors of $\boldsymbol{M}$. Our results therefore extend to the non-symmetric streaming SVD problem as well. We refer the reader to [32] for more details about this standard reduction.

The second requirement establishes that the random errors are bounded in a suitable norm. We write $\mathcal{S}_{d,k}$ for the Stiefel manifold of $d \times k$ matrices with orthonormal columns.

Assumption 2.2. If $\boldsymbol{A} \sim \mathrm{P}_{\boldsymbol{A}}$, then $\sup_{\boldsymbol{P}\in\mathcal{S}_{d,k}} \|\boldsymbol{P}^*(\boldsymbol{A} - \boldsymbol{M})\|_2 \leq M$ almost surely.

Note that for any matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$,

$$\sup_{P\in\mathcal{S}_{d,k}} \|\boldsymbol{P}^*\boldsymbol{X}\|_2 = \left(\sum\nolimits_{i=1}^{k} \sigma_i(\boldsymbol{X})^2\right)^{1/2}, \quad 1 \leq k \leq d,$$

where $\sigma_1(\boldsymbol{X}) \geq \sigma_2(\boldsymbol{X}) \geq \cdots \geq \sigma_d(\boldsymbol{X})$ are the singular values of $\boldsymbol{X}$. This norm, sometimes known as the $(2,k)$ norm [22] or the Ky Fan 2-k norm [8], satisfies

$$\|\boldsymbol{X}\| \leq \sup_{P \in \mathcal{S}_{d,k}} \|\boldsymbol{P}^*\boldsymbol{X}\|_2 \leq \sqrt{k}\|\boldsymbol{X}\| \leq \|\boldsymbol{X}\|_2 .$$

This choice of norm generalizes the error assumptions in the literature. In the $k = 1$ case, it agrees with the operator norm, which is the condition used by [17]; and it weakens the requirement of [1] that $\|\boldsymbol{A}_i\|_2 \leq 1$ almost surely.

The following theorem summarizes our main results for Oja's algorithm.

Theorem 2.3 (Main, informal). *Adopt Assumptions 2.1 and 2.2. Let $\lambda_1 \geq \ldots \lambda_d$ be the eigenvalues of $\boldsymbol{M}$, and let $\rho_k = \lambda_k - \lambda_{k+1}$.*

For every $\delta \in (0,1)$, define learning rates

$$T_0 = \tilde{\Theta}\left(\frac{kM^2}{\delta^2\rho_k^2}\right), \quad \beta = \tilde{\Theta}\left(\frac{M^2}{\rho_k^2}\right), \quad \eta_t = \begin{cases} \tilde{\Theta}\left(\frac{1}{\rho_k T_0}\right), & t \leq T_0 \\ \Theta\left(\frac{1}{\rho_k(\beta+t-T_0)}\right), & t > T_0. \end{cases}$$

Let $\boldsymbol{V} \in \mathbb{R}^{d\times k}$ be the orthogonal matrix whose columns are the k leading eigenvectors of $\boldsymbol{M}$. Then for any $T > T_0$, the output $\boldsymbol{Q}_T$ of Oja's algorithm satisfies

$$\operatorname{dist}(\boldsymbol{Q}_T, \boldsymbol{V}) \leq C' \frac{M}{\rho_k}\sqrt{\frac{\log(Mk/\rho_k\delta)}{T - T_0}}$$

with probability at least $1-\delta$, where C' is a universal positive constant.

To prove Theorem 2.3, we adopt a two-phase analysis. Our first result shows that after T_0 iterations, the output of Oja's algorithm satisfies $\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1$ with high probability.

Theorem 2.4 (Phase I, informal). *Adopt the same setting as Theorem 2.3, and let $\boldsymbol{Z}_0 \in \mathbb{R}^{d\times k}$ have i.i.d. Gaussian entries. Let*

$$T_0 = \Theta\left(\frac{kM^2}{\delta^2\rho_k^2}\left(\log(dM/\delta\rho_k)\right)^4\right).$$

Then after T_0 iterations of Oja's algorithm with constant step size $\eta = \Theta\left(\frac{\log(d/\delta)}{\rho_k T_0}\right)$ and initialization $\boldsymbol{Z}_0$, the output $\boldsymbol{Q}_{T_0}$ satisfies

$$\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1$$

with probability at least $1-\delta$.

Our analysis of the second phase shows that, if Oja's algorithm is initialized with *any* matrix satisfying $\|\boldsymbol{U}^*\boldsymbol{Q}_0(\boldsymbol{V}^*\boldsymbol{Q}_0)^{-1}\| \leq 1$, then the output of Oja's algorithm decays at the rate $O(1/\sqrt{T})$.

Theorem 2.5 (Phase II, informal). *Adopt the same setting as Theorem 2.3, and suppose that $\boldsymbol{Z}_0 \in \mathbb{R}^{d\times k}$ satisfies $\|\boldsymbol{U}^*\boldsymbol{Z}_0(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\| \leq 1$. Then after T iterations of Oja's algorithm with step size $\eta_i = \frac{8}{(\beta+i)\rho_k}$ with $\beta = \Theta\left(\frac{M^2}{\rho_k^2}\log\left(\frac{Mk}{\rho_k\delta}\right)\right)$ and initialization $\boldsymbol{Q}_0$, the output $\boldsymbol{Q}_T$ satisfies*

$$\operatorname{dist}(\boldsymbol{Q}_T, \boldsymbol{V}) \leq 2\mathrm{e}\sqrt{\frac{\beta+1}{\beta+T}}$$

with probability at least $1-\delta$.

This error guarantee is completely dimension free, and depends only logarithmically on k and the failure probability δ.

Theorem 2.3 follows directly from Theorems 2.4 and 2.5. Theorem 2.4 guarantees that with probability $1-\delta$, the output of Phase I is a suitable initialization for Phase II, and, conditioned on this good event, Theorem 2.5 guarantees that the output of the second phase has error $O(\sqrt{\beta/T})$ with probability $1-\delta$. By concatenating the analysis of the two phases and using the union bound, we obtain that the resulting two-phase algorithm succeeds with probability at least $1-2\delta$, yielding Theorem 2.3.

In the remainder of this section, we describe the main technical tools we employ in our argument.

2.1. **A recursive expression.** To simplify the argument, we recall the following result of [1, Lemma 2.2]:

Lemma 2.6. *For all $t \geq 0$,*

$$\operatorname{dist}(\boldsymbol{Q}_t, \boldsymbol{V}) = \|\boldsymbol{U}^*\boldsymbol{Q}_t\| \leq \|\boldsymbol{U}^*\boldsymbol{Q}_t(\boldsymbol{V}^*\boldsymbol{Q}_t)^{-1}\| = \|\boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\| \,.$$

We therefore focus on bounding the norm of the matrix

$$\boldsymbol{W}_t := \boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1} \,. \tag{2.1}$$

Under the assumption that η_t is small, we might expect that we can write $\boldsymbol{W}_t$ as a sum of the dominant term

$$\boldsymbol{H}_t := \boldsymbol{U}^*(\mathbf{I}+\eta_t\boldsymbol{M})\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*(\mathbf{I}+\eta_t\boldsymbol{M})\boldsymbol{Z}_{t-1})^{-1}$$

plus lower order terms.

To argue that $\boldsymbol{W}_t$ is close to $\boldsymbol{H}_t$, we need to argue that the inverse $(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}$ does not blow up, which will be the case so long as the fluctuation term $\eta_t\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{Z}_{t-1}$ is smaller than the main term $\boldsymbol{V}^*(\mathbf{I}+\eta_t\boldsymbol{M})\boldsymbol{Z}_{t-1}$. In order to make this requirement precise, we write

$$\boldsymbol{\Delta}_t := \eta\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*(\mathbf{I}+\eta_t\boldsymbol{M})\boldsymbol{Z}_{t-1})^{-1} \,. \tag{2.2}$$

So long as this matrix has small norm, the inverse term will be well behaved. As we discuss in the following section, we will be able to guarantee that this is the case by conditioning on an appropriate good event.

The following lemma shows that, modulo a term involving $\boldsymbol{\Delta}_t$, we can indeed express $\boldsymbol{W}_t$ as $\boldsymbol{H}_t$ plus a small correction.

Lemma 2.7. *Let $\boldsymbol{W}_t$, $\boldsymbol{H}_t$, and $\boldsymbol{\Delta}_t$ be defined as in (2.1)–(2.2). Then we can write*

$$\boldsymbol{W}_t(\mathbf{I}-\boldsymbol{\Delta}_t^2) = \boldsymbol{H}_t + \boldsymbol{J}_{t,1} + \boldsymbol{J}_{t,2} \,, \tag{2.3}$$

for matrices $\boldsymbol{J}_{t,1}$ and $\boldsymbol{J}_{t,2}$ of norm $O(\eta_t)$ and $O(\eta_t^2)$, respectively.

Below, in Propositions A.1 and A.2, we use Lemma 2.7 to develop an explicit recursive bound on the norm of $\boldsymbol{W}_t$.

2.2. **Matrix concentration via smoothness.** In order to exploit the expression (2.3), we need concentration inequalities that allow us to conclude that $\boldsymbol{W}_t$ is near $\boldsymbol{H}_t$ with high probability. [16] recently developed new tools to control the norms of products of independent random matrices, in an attempt to extend the mature toolset for bounding *sums* of random matrices to the product setting. Their techniques are based on a simple but deep property of the Schatten p-norms known as *uniform smoothness*. The most elementary expression of this fact is the following inequality, which is the analogue of (1.2) for the L_p norm.

Proposition 2.8 ([16, Proposition 4.3])**.** *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be random matrices of the same size, with $\mathbb{E}[\boldsymbol{Y} : \boldsymbol{X}] = \mathbf{0}$. Then for any $p \geq 2$,*

$$\|\boldsymbol{X}+\boldsymbol{Y}+\boldsymbol{Z}\|_{p,p}^2 \leq \|\boldsymbol{X}\|_{p,p}^2 + (p-1)\|\boldsymbol{Y}\|_{p,p}^2 \,.$$

We will employ the following corollary of Proposition 2.8, which extends the inequality to non-centered random matrices.

Proposition 2.9. *Let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ be random matrices of the same size, with $\mathbb{E}[\boldsymbol{Y} : \boldsymbol{X}] = \mathbf{0}$. Then for any $p \geq 2$ and $\lambda > 0$,*

$$\|\boldsymbol{X} + \boldsymbol{Y} + \boldsymbol{Z}\|_{p,p}^2 \leq (1+\lambda)(\|\boldsymbol{X}\|_{p,p}^2 + (p-1)\|\boldsymbol{Y}\|_{p,p}^2 + \lambda^{-1}\|\boldsymbol{Z}\|_{p,q}^2)$$

The benefit of working in the L_p norm is that bounding this norm for p large yields good tail bounds on the operator norm, which are not available if the argument is carried out solely in expected Frobenius norm. We will rely heavily on this fact heavily in our argument.

2.3. **Conditioning on good events.** Obtaining control on $\boldsymbol{W}_t$ via (2.3) requires ensuring that the matrix $\mathbf{I} - \boldsymbol{\Delta}_t^2$ is invertible, with inverse of bounded norm. To accomplish this, we define a sequence of good events $\mathcal{G}_0 \supset \mathcal{G}_1 \supset \dots$, where each $\mathcal{G}_i$ is measurable with respect to the σ-algebra $\mathcal{F}_i := \sigma(\boldsymbol{Z}_0, \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_i)$. We write $\mathbb{1}_i$ for the indicator of the event $\mathcal{G}_i$, and we will define $\mathcal{G}_i$ in such a way that $(\mathbf{I} - \boldsymbol{\Delta}_t^2 \mathbb{1}_{t-1})$ is invertible almost surely.

During Phase II, the good events are defined by

$$\begin{aligned} \mathcal{G}_0 &:= \{\|\boldsymbol{W}_0\| \leq 1\} \\ \mathcal{G}_i &:= \{\|\boldsymbol{W}_i\| \leq \gamma\} \cap \mathcal{G}_{i-1}\,, \quad \forall i \geq 1 \end{aligned}$$

for some $\gamma \geq 1$ to be specified. Since Assumption 2.2 implies that $\|\boldsymbol{A}_i - \boldsymbol{M}\| \leq M$ almost surely, this definition guarantees that for all $i \geq 1$,

$$\|\boldsymbol{V}^*(\boldsymbol{A}_i - \boldsymbol{M})\boldsymbol{U}\boldsymbol{W}_{i-1}\mathbb{1}_{i-1}\| \leq M\gamma \quad \text{almost surely.} \tag{2.4}$$

As we show in Proposition A.1 below, if the step size is sufficiently small, then (2.4) implies that $\mathbf{I} - \boldsymbol{\Delta}_t^2$ is almost surely invertible on $\mathcal{G}_{t-1}$, which allows us to employ (2.3) to bound the norm of $\boldsymbol{W}_t\mathbb{1}_{t-1}$.

During Phase I, we condition on a slightly more complicated set of events, which we describe explicitly in Section 5. However, these events are constructed so that (2.4) still holds for all $i \geq 1$.

Our matrix concentration results described in Section 2.2 allow us to show that, during both Phase I and Phase II, $\|\boldsymbol{W}_t\mathbb{1}_{t-1}\|$ is small with high probability, for all $t \geq 1$. Using this fact, we show that, conditioned on $\mathcal{G}_{t-1}$, the probability that $\mathcal{G}_t$ holds is also large. Bounding the failure probability at each step, we are able to conclude that, conditioned on the initialization event $\mathcal{G}_0$, the good events $\mathcal{G}_t$ hold for all $t \geq 1$ with high probability.

3. Main recursive bound

In this section, we state our main recursive bound, which we use in both Phase I and Phase II. A proof appears in Section B.

Theorem 3.1. *Let t be a positive integer, and for all $i \in [t]$, let $\varepsilon_i = 2\eta_i M(1+\gamma)$. Let $\mathbb{1}_1, \dots, \mathbb{1}_t$ be the indicator functions of a sequence of good events satisfying (2.4) for all $i \in [t]$.*

Assume that for all $i \in [t]$,

$$\varepsilon_i \leq \frac{1}{2}\,, \qquad \eta_i\|\boldsymbol{M}\| \leq \frac{1}{2}\,, \qquad \mathrm{e}^{-\eta_i\rho_k/4} \leq \frac{\varepsilon_i}{\varepsilon_{i-1}}\,, \tag{3.1}$$

with the convention that the last requirement is vacuous when $i = 1$. Then for any $p \geq 2$,

$$\|\boldsymbol{W}_t\mathbb{1}_t\|_{p,p}^2 \leq \|\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \mathrm{e}^{-s_t\rho_k}\|\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + C_1 p\varepsilon_t^2 \sum_{i=0}^{t-1} \|\boldsymbol{W}_i\mathbb{1}_i\|_{p,p}^2 + C_2 p k^{2/p}\varepsilon_t^2 t\,, \tag{3.2}$$

where $s_t = \sum_{i=1}^t \eta_i$, $C_1 = 21$, and $C_2 = 5$. Moreover, if in addition for all $i \in [t]$,

$$p\varepsilon_i^2 \leq \frac{\eta_i\rho_k}{50}\,, \tag{3.3}$$

then

$$\|\boldsymbol{W}_t \mathbb{1}_t\|_{p,p}^2 \le \|\boldsymbol{W}_t \mathbb{1}_{t-1}\|_{p,p}^2 \le \mathrm{e}^{-s_t \rho_k/2} \|\boldsymbol{W}_0 \mathbb{1}_0\|_{p,p}^2 + C_2 p k^{2/p} \varepsilon_t^2 t\,.$$

Theorem 3.1 shows that, up to small error, $\|\boldsymbol{W}_t \mathbb{1}_{t-1}\|_{p,p}^2$ decays exponentially fast. We will use this fact to prove high probability bounds on $\|\boldsymbol{W}_t \mathbb{1}_{t-1}\|$, which then imply bounds on $\|\boldsymbol{W}_t\|$.

4. Phase II

In this section, we use Theorem 3.1 to prove a formal version of Theorem 2.5.

For this phase, recall that we define the good events $\mathcal{G}_i$ by

$$\mathcal{G}_0 = \{\|\boldsymbol{W}_0\| \le 1\}\,, \qquad \mathcal{G}_i = \{\|\boldsymbol{W}_i\| \le \gamma\} \cap \mathcal{G}_{i-1}\,, \quad \forall i \ge 1\,. \tag{4.1}$$

For Phase II, we set $\gamma = \sqrt{2}\mathrm{e}$.

We first show that, with a specific step-size schedule, we obtain good bounds on the norm of the last iterate.

Proposition 4.1. *Define the good events as in (4.1). Set $\eta_i = \frac{\alpha}{(\beta+i)\rho_k}$, for positive quantities α and β, and define the normalized gap*

$$\bar{\rho}_k = \min\left\{\frac{\rho_k}{M}, \frac{\rho_k}{\|\boldsymbol{M}\|}, 1\right\}\,. \tag{4.2}$$

If

$$\alpha \ge 8\,, \quad \beta \ge \frac{4(1+\sqrt{2}\mathrm{e})\alpha}{\bar{\rho}_k}\,, \tag{4.3}$$

then for any $t \ge 1$,

$$\|\boldsymbol{W}_t \mathbb{1}_t\|_{p,p}^2 \le k^{2/p}\left(\frac{\beta+1}{\beta+t}\right)^{\alpha} + p k^{2/p} \cdot \left(\frac{C_3 \alpha}{\bar{\rho}_k}\right)^2 \cdot \frac{t}{(\beta+t)^2}\,, \tag{4.4}$$

where C_3 is a numerical constant less than 175.

Proof. Since the good events defined in (4.1) satisfy (2.4), we can apply Theorem 3.1. In the appendix, we show (Lemma C.1) that (4.3) implies that the assumptions in (3.1) hold. Theorem 3.1 then yields

$$\begin{aligned}\|\boldsymbol{W}_t \mathbb{1}_t\|_{p,p}^2 &\le \mathrm{e}^{-s_t\rho_k}\|\boldsymbol{W}_0 \mathbb{1}_0\|_{p,p}^2 + C_1 p \varepsilon_t^2 \sum_{i=1}^{t-1} \|\boldsymbol{W}_i \mathbb{1}_i\|_{p,p}^2 + C_2 p k^{2/p} \varepsilon_t^2 t \\ &\le \mathrm{e}^{-s_t\rho_k} k^{2/p} + (C_1\gamma^2 + C_2) p k^{2/p} \varepsilon_t^2 t\,,\end{aligned}$$

since (4.1) implies $\|\boldsymbol{W}_0 \mathbb{1}_0\|_{p,p}^2 \le k^{2/p}$ and $\|\boldsymbol{W}_i \mathbb{1}_i\|_{p,p}^2 \le \gamma^2 k^{2/p}$ for all $i \ge 1$.

The definition of η_i implies

$$\rho_k s_t = \alpha \sum_{i=1}^{t} \frac{1}{\beta+i} \ge \alpha \log\left(\frac{\beta+t}{\beta+1}\right).$$

We obtain

$$\|\boldsymbol{W}_t \mathbb{1}_t\|_{p,p}^2 \le k^{2/p}\left(\frac{\beta+1}{\beta+t}\right)^{\alpha} + p k^{2/p} \cdot \left(\frac{C_3 \alpha}{\bar{\rho}_k}\right)^2 \cdot \frac{t}{(\beta+t)^2}\,,$$

where

$$C_3 = (C_1\gamma^2 + C_2)^{1/2} C_\varepsilon < 175\,,$$

as desired. □

Finally, we remove the conditioning and prove the full version of Theorem 2.5.

Theorem 4.2. *Assume $\|\boldsymbol{W}_0\| \le 1$, and adopt the step size $\eta_i = \frac{\alpha}{(\beta+i)\rho_k}$, with*

$$\alpha \ge 8\,, \quad \beta \ge 2\left(\frac{C_3\alpha}{\bar{\rho}_k}\right)^2 \log\left(\frac{C_3\alpha}{\bar{\rho}_k}\cdot 2k/\delta\right)\,,$$

where $\bar{\rho}_k$ is as in (4.2) and C_3 is as in (4.4). Then

$$\|\boldsymbol{W}_T\| \le 2\mathrm{e}\sqrt{\frac{\beta+1}{\beta+T}}$$

with probability at least $1-\delta$.

Proof. For any $s \ge 0$, it holds $\mathbb{P}\left\{\|\boldsymbol{W}_T\| \ge s\right\} \le \mathbb{P}\left\{\|\boldsymbol{W}_T\mathbb{1}_T\| \ge s\right\} + \mathbb{P}\left\{\mathcal{G}_T^C\right\}$. First, we have

$$\mathbb{P}\left\{\mathcal{G}_T^C\right\} \le \mathbb{P}\left\{\mathcal{G}_0^C\right\} + \sum\nolimits_{j=1}^{T}\mathbb{P}\left\{\mathcal{G}_j^C \cap \mathcal{G}_{j-1}\right\}\,.$$

Since we have assumed that the initialization satisfies $\|\boldsymbol{W}_0\| \le 1$, the event $\mathcal{G}_0$ holds with probability 1, so it suffices to bound the second term. By Markov's inequality, we have

$$\mathbb{P}\left\{\mathcal{G}_j^C \cap \mathcal{G}_{j-1}\right\} = \mathbb{P}\left\{\|\boldsymbol{W}_j\mathbb{1}_{j-1}\| \ge \gamma\right\} \le \inf_{p\ge 2}\gamma^{-p}\|\boldsymbol{W}_j\mathbb{1}_{j-1}\|_{p,p}^p\,.$$

For fixed $j \ge 1$, we choose $p = (\beta+j)\cdot\frac{\bar{\rho}_k^2}{C_3^2\alpha^2}$. It follows from (4.4) that,

$$\begin{aligned}
\gamma^{-p}\|\boldsymbol{W}_j\mathbb{1}_{j-1}\|_{p,p}^p &\le \left(\frac{1}{\gamma^2}k^{2/p}\left(\frac{\beta+1}{\beta+j}\right)^{\alpha} + \frac{1}{\gamma^2}pk^{2/p}\cdot\frac{C_3^2\alpha^2}{\bar{\rho}_k^2}\cdot\frac{j}{(\beta+j)^2}\right)^{p/2}\\
&\le k\left(\frac{1}{2\mathrm{e}^2} + \frac{1}{2\mathrm{e}^2}\frac{j}{\beta+j}\right)^{p/2}\\
&\le k\mathrm{e}^{-p} = k\exp\left(-(\beta+j)\cdot\frac{\bar{\rho}_k^2}{C_3^2\alpha^2}\right)\,.
\end{aligned}$$

Therefore, for any $T \ge 1$,

$$\sum\nolimits_{j=1}^{T}\mathbb{P}\left\{\mathcal{G}_j^C|\mathcal{G}_{j-1}\right\} \le k\sum\nolimits_{j=1}^{T}\exp\left(-(\beta+j)\cdot\frac{\bar{\rho}_k^2}{C_3^2\alpha^2}\right) \le k\frac{C_3^2\alpha^2}{\bar{\rho}_k^2}\mathrm{e}^{-\beta\cdot\frac{\bar{\rho}_k^2}{C_3^2\alpha^2}}.$$

This quantity is smaller than $\delta/2$ if

$$\beta \ge 2\frac{C_3^2\alpha^2}{\bar{\rho}_k^2}\log\left(\frac{C_3\alpha M}{\bar{\rho}_k}\cdot 2k/\delta\right)\,.$$

It remains to bound $\mathbb{P}\left\{\|\boldsymbol{W}_T\mathbb{1}_T\| \ge s\right\}$. A simple argument (Lemma C.2) based on (4.4) shows that this probability is at least $\delta/2$ for

$$s = 2\mathrm{e}\sqrt{\frac{\beta+1}{\beta+T}}\,.$$

The claim follows. □

5. Phase I

In this section, we describe the slightly more delicate proof of the formal version of Theorem 2.4. As in Section 4, we will employ Theorem 3.1. However, we will also need to develop an auxiliary recurrence to bound the growth of an additional matrix sequence.

Before we analyze Phase I, we first show that we can reduce to the case that that $\mathrm{P}_{\boldsymbol{A}}$ has finite support. We prove the following result in Appendix E.

Proposition 5.1. *Fix $\rho > 0$. Suppose that there exists a choice of constant step size η and $T_0 \geq \frac{9M}{\rho\delta}\log(d/\delta)$ such that for any* finitely-supported *distribution with support size at most T_0^3 satisfying Assumptions 2.1 and 2.2 and with $\rho_k \geq \rho/2$, we have*

$$\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq \frac{1}{6} \tag{5.1}$$

with probability at least $1-\delta/3$.

Then for this same η and T_0 it in fact holds that for any *distribution satisfying Assumptions 2.1 and 2.2 and with $\rho_k \geq \rho$, we have*

$$\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1$$

with probability at least $1-\delta$.

Proposition 5.1 implies that it suffices to prove the error guarantee (5.1) in the special case when $\mathrm{P}_{\boldsymbol{A}}$ has finite support of cardinality at most T_0^3.

Let us fix a time horizon T_0 and assume in what follows that $m := |\mathrm{supp}(\mathrm{P}_{\boldsymbol{A}})| \leq T_0^3$.We begin by defining the good events for Phase I. We adopt a constant step size η, to be specified. Denote

$$\mathscr{E} := \{M^{-1}(\boldsymbol{A}-\boldsymbol{M})\boldsymbol{U}\boldsymbol{U}^* : \boldsymbol{A} \in \mathrm{supp}(\mathrm{P}_{\boldsymbol{A}})\}.$$

For $i \geq 1$, we will set

$$\mathscr{G}_i = \{\max_{\boldsymbol{E}\in\mathscr{E}} \|\boldsymbol{V}^*\boldsymbol{E}\boldsymbol{U}\boldsymbol{W}_i\| \leq \gamma\} \cap \mathscr{G}_{i-1}\,.$$

Note that this choice satisfies (2.4) for all $i > 1$.

To define the initial good event $\mathscr{G}_0$, we need to define a larger set of matrices to condition on. For all $r, \ell \geq 1$, set

$$\begin{aligned}\mathscr{E}_{r,\ell} := \{\boldsymbol{V}^*\boldsymbol{F}_1\cdots\boldsymbol{F}_r\boldsymbol{U} : &\boldsymbol{F}_i \in \mathscr{E} \text{ for at most } \ell \text{ distinct indices } i \in [r],\\ &\text{and } \boldsymbol{F}_i = (1+\eta\lambda_{k+1})^{-1}(\mathbf{I}+\eta\boldsymbol{M})\boldsymbol{U}\boldsymbol{U}^* \text{ otherwise}\}\end{aligned}$$

The set $\mathscr{E}_{r,\ell}$ has cardinality less than $(r(m+1))^\ell$, and $\|\boldsymbol{E}\|_2 \leq 1$ for any $\boldsymbol{E} \in \mathscr{E}_{r,\ell}$, and any $r, \ell \geq 1$. We have defined $\mathscr{E}_{r,\ell}$ so that control over $\max_{\boldsymbol{E}\in\mathscr{E}_{r+1,\ell+1}} \|\boldsymbol{E}\boldsymbol{W}_{t-1}\|$ gives control over $\max_{\boldsymbol{E}\in\mathscr{E}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_t\|$.

Finally, we define

$$\mathscr{G}_0 := \bigcap_{r,\ell=1}^{T_0+1} \left\{\max_{\boldsymbol{E}\in\mathscr{E}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_0\|_2 \leq \frac{\sqrt{\ell}\gamma}{\sqrt{2}\mathrm{e}}\right\} \cap \{\|\boldsymbol{W}_0\|_2 \leq \sqrt{d}\gamma\}\,. \tag{5.2}$$

Since $\boldsymbol{V}^*(\boldsymbol{A}_1-\boldsymbol{M})\boldsymbol{U} \in \mathscr{E}_{1,1}$ almost surely, this choice satisfies (2.4) for $i = 1$.

Our strategy will be similar to the one used in Section 4. However, in order to show that the good events $\mathscr{G}_i$ hold with high probability, we will also need a second recurrence that allows us to control the norm of matrices of the form $\boldsymbol{E}\boldsymbol{W}_t$, for $\boldsymbol{E} \in \mathscr{E}_{r,\ell}$. The details appear in Section D.

6. Conclusion

This work gives the first nearly optimal analysis of Oja's algorithm for streaming PCA beyond the rank one case. Our analysis is conceptually simple: we show that the spectral norm of the matrix $\boldsymbol{W}_t$ concentrates well around its expectation, once we condition on $\boldsymbol{W}_{t-1}$ having the same behavior. And our concentration results are strong enough that we can pay to union bound over the entire course of the algorithm, to show that $\boldsymbol{W}_t$ is well behaved for all $t \geq 1$.

The matrix concentration techniques we have applied here could be useful in analyzing other PCA-like algorithms, or, more generally, other stochastic algorithms for simple non-convex optimization

problems. An interesting question is whether these techniques can prove *gap-free* rates for Oja's algorithm outside the rank-one setting. This would extend the results of [1] to the general case.

Finally, we stress that the algorithm we have described here requires *a priori* knowledge of the problem parameters (including the gap ρ_k) to set the step sizes, which is a serious limitation in practice. Recently, [14] developed a data-driven procedure to adaptively select the optimal step sizes. Obtaining theoretical guarantees for this or similar algorithms is an important open problem.

Acknowledgement

We thank Joel Tropp and Amelia Henriksen for valuable discussions which greatly improved this manuscript.

Appendix A. Additional results for Section 3

The following proposition develops the expansion described in Lemma 2.7 and gives explicit bounds on the norms of the error matrices $\boldsymbol{J}_{t,1}$ and $\boldsymbol{J}_{t,2}$.

We recall the following definitions

$$\begin{aligned}\boldsymbol{W}_t &= \boldsymbol{U}^*\boldsymbol{Z}_t(\boldsymbol{V}^*\boldsymbol{Z}_t)^{-1}\\ \boldsymbol{H}_t &= \boldsymbol{U}^*(\mathbf{I}+\eta\boldsymbol{M})\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*(\mathbf{I}+\eta\boldsymbol{M})\boldsymbol{Z}_{t-1})^{-1}\\ \boldsymbol{\Delta}_t &= \eta_t\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*(\mathbf{I}+\eta_t\boldsymbol{M})\boldsymbol{Z}_{t-1})^{-1}\end{aligned}$$

Proposition A.1. *Let $t \geq 1$. Assume that η_t is small enough that $\boldsymbol{M} \succeq -\frac{1}{2\eta_t}\mathbf{I}$, and assume that (2.4) holds for $i = t$. Let*

$$\begin{aligned}E_t &= (k^{1/p} + 2\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p})\\ \varepsilon_t &= 2\eta_t M(1+\gamma)\,.\end{aligned}$$

Then $\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\| \leq \varepsilon_t$ almost surely, and

$$\boldsymbol{W}_t(\mathbf{I}-\boldsymbol{\Delta}_t^2) = \boldsymbol{H}_t + \boldsymbol{J}_{t,1} + \boldsymbol{J}_{t,2}$$

for $\boldsymbol{J}_{t,1}$ and $\boldsymbol{J}_{t,2}$ satisfying

$$\begin{aligned}\|\boldsymbol{J}_{t,1}\mathbb{1}_{t-1}\|_{p,p} &\leq E_t\varepsilon_t\\ \|\boldsymbol{J}_{t,2}\mathbb{1}_{t-1}\|_{p,p} &\leq E_t\varepsilon_t^2\,,\end{aligned}$$

and $\mathbb{E}[\boldsymbol{J}_{t,1} : \mathcal{F}_{t-1}] = \mathbf{0}$.

Proof. We employ the notation of the proof of Lemma 2.7. (See Appendix G.) First, we show the bound on $\boldsymbol{\Delta}_t$. Since $\eta_t\boldsymbol{M} \succeq -\frac{1}{2}\mathbf{I}$, we have $\|\boldsymbol{V}^*(\mathbf{I}+\eta_t\boldsymbol{M})^{-1}\boldsymbol{V}\| \leq 2$. Moreover, since $\|\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{U}\boldsymbol{W}_{t-1}\| \leq M\gamma$ almost surely, we have that

$$\begin{aligned}\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\| &\leq 2\|\eta_t\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})(\boldsymbol{U}\boldsymbol{U}^*+\boldsymbol{V}\boldsymbol{V}^*)\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*\boldsymbol{Z}_{t-1})^{-1}\mathbb{1}_{t-1}\|\\ &\leq 2\eta_t\|\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{Z}_{t-1}(\boldsymbol{V}^*\boldsymbol{Z}_{t-1})^{-1}\mathbb{1}_{t-1}\| + 2\eta_t\|\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{V}\|\\ &= 2\eta_t\|\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{U}\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\| + 2\eta_t\|\boldsymbol{V}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{V}\|\\ &\leq 2\eta_t M(1+\gamma) =: \varepsilon_t\,.\end{aligned}$$

We can bound $\|\widehat{\boldsymbol{\Delta}}_t\mathbb{1}_{t-1}\|_{p,p}$ by a similar argument. First, note that Assumption 2.2 implies that $\|\boldsymbol{A}_t-\boldsymbol{M}\| \leq M$ almost surely. Hence

$$\|\widehat{\boldsymbol{\Delta}}_t\mathbb{1}_{t-1}\|_{p,p} \leq 2\eta_t\|\boldsymbol{U}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{U}\boldsymbol{U}^*Z_{t-1}(\boldsymbol{V}^*\boldsymbol{Z}_{t-1})^{-1}\mathbb{1}_{t-1}\|_{p,p} + 2\eta_t\|\boldsymbol{U}^*(\boldsymbol{A}_t-\boldsymbol{M})\boldsymbol{V}\mathbb{1}_{t-1}\|_{p,p}$$

$$
\begin{aligned}
&= 2\eta_t \|\boldsymbol{U}^*(\boldsymbol{A}_t - \boldsymbol{M})\boldsymbol{U}\| \, \|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + 2\eta_t \|\boldsymbol{U}^*(\boldsymbol{A}_t - \boldsymbol{M})\boldsymbol{V}\mathbb{1}_{t-1}\|_{p,p} \\
&\leq (\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + k^{1/p}) 2\eta_t M \\
&\leq (\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + k^{1/p})\varepsilon_t \,,
\end{aligned}
$$

Finally, we have

$$
\|\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p} \leq \frac{1+\eta_t\lambda_{k+1}}{1+\eta_t\lambda_k}\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} \leq \|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} \,.
$$

We now employ Lemma 2.7. The term $\boldsymbol{J}_{t,1}$ satisfies

$$
\mathbb{E}[\boldsymbol{J}_{t,1}\mathbb{1}_{t-1}|\mathcal{F}_{t-1}] = \boldsymbol{0} \,,
$$

and we have

$$
\begin{aligned}
\|\boldsymbol{J}_{t,1}\mathbb{1}_{t-1}\|_{p,p} &\leq \|\widehat{\boldsymbol{\Delta}}_t\mathbb{1}_{t-1}\|_{p,p} + \|\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\| \\
&\leq (\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + k^{1/p})\varepsilon_t + \|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}\varepsilon_t \\
&\leq E_t\varepsilon_t \,.
\end{aligned}
$$

Finally,

$$
\|\boldsymbol{J}_{t,2}\|_{p,p} \leq \|\widehat{\boldsymbol{\Delta}}_t\mathbb{1}_{t-1}\|_{p,p}\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\| \leq (\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + k^{1/p})\varepsilon_t^2 \leq E_t\varepsilon_t^2 \,.
$$

□

Combining Proposition A.1 with Proposition 2.9 immediately yields a recursive bound.

Proposition A.2. *Adopt the setting of Proposition A.1. If $\varepsilon_t \leq 1/2$, then*

$$
\|\boldsymbol{W}_t\mathbb{1}_t\|_{p,p}^2 \leq \|\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq K_{1,t}\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + K_{2,t} \,, \tag{A.1}
$$

where

$$
\begin{aligned}
K_{1,t} &= (1+5\varepsilon_t^2)\left\{\left(\frac{1+\eta_t\lambda_k}{1+\eta_t\lambda_{k+1}}\right)^2 + 8p\varepsilon_t^2\right\} \\
K_{2,t} &= 5pk^{2/p}\varepsilon_t^2 \,.
\end{aligned}
$$

Proof. Reusing the notation of Proposition A.1, we have

$$
\boldsymbol{W}_t\mathbb{1}_{t-1}(\mathbf{I} - \boldsymbol{\Delta}_t^2) = \boldsymbol{H}_t\mathbb{1}_{t-1} + \boldsymbol{J}_{t,1}\mathbb{1}_{t-1} + \boldsymbol{J}_{t,2}\mathbb{1}_{t-1} \,,
$$

where $\mathbb{E}[\boldsymbol{J}_{t,1}\mathbb{1}_{t-1} : \mathcal{F}_{t-1}] = \boldsymbol{0}$. Since $\boldsymbol{H}_t\mathbb{1}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable, Proposition 2.9 therefore yields for any $\lambda > 0$

$$
\|\boldsymbol{W}_t\mathbb{1}_{t-1}(\mathbf{I} - \boldsymbol{\Delta}_t^2)\|_{p,p}^2 \leq (1+\lambda)(\|\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}^2 + (p-1)E_t^2\varepsilon_t^2 + \lambda^{-1}E_t^2\varepsilon_t^4) \,.
$$

Choosing $\lambda = \varepsilon_t^2$, we obtain

$$
\|\boldsymbol{W}_t\mathbb{1}_{t-1}(\mathbf{I} - \boldsymbol{\Delta}_t^2)\|_{p,p}^2 \leq (1+\varepsilon_t^2)(\|\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}^2 + pE_t^2\varepsilon_t^2) \,.
$$

Finally, under the assumption that $\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\| \leq \varepsilon_t \leq \frac{1}{2}$ almost surely, on the event $\mathcal{G}_{t-1}$ the matrix $\mathbf{I} - \boldsymbol{\Delta}_t^2$ is invertible and satisfies

$$
\|(\mathbf{I} - \boldsymbol{\Delta}_t^2)^{-1}\mathbb{1}_{t-1}\| \leq (1 - \|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\|^2)^{-1} \leq (1-\varepsilon_t^2)^{-1}
$$

Hence

$$
\|\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \|\boldsymbol{W}_t\mathbb{1}_{t-1}(\mathbf{I} - \boldsymbol{\Delta}_t^2)\|_{p,p}^2\|(\mathbf{I} - \boldsymbol{\Delta}_t)^{-1}\mathbb{1}_{t-1}\| \leq \frac{1+\varepsilon_t^2}{(1-\varepsilon_t^2)^2}(\|\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}^2 + pE_t^2\varepsilon_t^2) \,.
$$

Since $\frac{1+\varepsilon_t^2}{(1-\varepsilon_t^2)^2} \le 1 + 5\varepsilon_t^2$ for all $\varepsilon_t \le \frac{1}{2}$ and

$$(1+5\varepsilon_t^2)E_t^2 \le (1+5\varepsilon_t^2)(2k^{2/p} + 8\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2)$$

and $2(1+5\varepsilon_t^2) \le 5$ for all $\varepsilon_t \le \frac{1}{2}$, this proves the claim. □

Appendix B. Proof of Theorem 3.1

We will unroll the one-step recurrence of Proposition A.2. We first bound $K_{1,i}$. We have

$$K_{1,i} \le \left(\frac{1+\eta_i\lambda_k}{1+\eta_i\lambda_{k+1}}\right)^2 + (5+8p)\varepsilon_i^2 + 40p\varepsilon_i^4 \le \left(\frac{1+\eta_i\lambda_k}{1+\eta_i\lambda_{k+1}}\right)^2 + (5+18p)\varepsilon_i^2\,,$$

where the second inequality follows from the first assumption in (3.1). The second assumption in (3.1) implies that $0 \le 1+\eta_i\lambda_k \le 2$, so

$$\left(\frac{1+\eta_i\lambda_{k+1}}{1+\eta_i\lambda_k}\right)^2 = \left(1 - \frac{\eta_i\rho_k}{1+\eta_i\lambda_k}\right)^2 \le \left(1 - \frac{1}{2}\eta_i\rho_k\right)^2 \le \mathrm{e}^{-\eta_i\rho_k}\,.$$

Since $5+18p \le 21p$ for all $p \ge 2$, we obtain

$$K_{1,i} \le \mathrm{e}^{-\eta_i\rho_k} + C_1 p\varepsilon_i^2\,.$$

We now proceed to prove the first claim by induction. When $t=1$, we use (A.1) to obtain

$$\begin{aligned}\|\boldsymbol{W}_1\mathbb{1}_1\|_{p,p}^2 \le \|\boldsymbol{W}_1\mathbb{1}_0\|_{p,p}^2 &\le K_{1,1}\|\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + K_{2,1}\\ &\le \mathrm{e}^{-\eta_1\rho_k}\|\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + C_1 p\varepsilon_1^2\|\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + C_2 pk^{2/p}\varepsilon_1^2\,,\end{aligned}$$

which is the desired bound.

Proceeding by induction, for $t > 1$ we have

$$\begin{aligned}\|\boldsymbol{W}_t\mathbb{1}_t\|_{p,p}^2 &\le \|\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2\\ &\le K_{1,t}\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + K_{2,t}\\ &\le \mathrm{e}^{-\eta_t\rho_k}\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + C_1 p\varepsilon_t^2\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + K_{2,t}\\ &\le \mathrm{e}^{-\eta_t\rho_k}\left(\mathrm{e}^{-s_{t-1}\rho_k}\|\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + C_1 p\varepsilon_{t-1}^2\sum_{i=0}^{t-2}\|\boldsymbol{W}_i\mathbb{1}_i\|_{p,p}^2 + C_2 pk^{2/p}\varepsilon_{t-1}^2(t-1)\right)\\ &\quad + C_1 p\varepsilon_t^2\|\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + C_2 pk^{2/p}\varepsilon_t^2\\ &\le \mathrm{e}^{-s_t\rho_k}\|\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + C_1 p\varepsilon_t^2\sum_{i=0}^{t-1}\|\boldsymbol{W}_i\mathbb{1}_i\|_{p,p}^2 + C_2 pk^{2/p}\varepsilon_t^2 t\,,\end{aligned}$$

where in the final inequality we have used that $\mathrm{e}^{-\eta_t\rho_k}\varepsilon_{t-1}^2 \le \varepsilon_t^2$ by the third assumption of (3.1). This proves the first bound.

For the second bound, we proceed in a similar way, but with a sharper bound on $K_{1,i}$. The second assumption of (3.1) again implies

$$\left(\frac{1+\eta_i\lambda_{k+1}}{1+\eta_i\lambda_k}\right)^2 = \left(1 - \frac{\eta_i\rho_k}{1+\eta_i\lambda_k}\right)^2 \le 1 - \eta_i\rho_k + \frac{1}{4}(\eta_i\rho_k)^2 \le 1 - \frac{3}{4}\eta_i\rho_k\,,$$

and therefore

$$\begin{aligned}K_{1,i} &\le (1+5\varepsilon_i^2)\left(1 - \frac{3}{4}\eta_i\rho_k + 8p\varepsilon_i^2\right)\\ &\le \exp\left(-\frac{3}{4}\eta_i\rho_k + (5+8p)\varepsilon_i^2\right)\end{aligned}$$

$$\leq \mathrm{e}^{-\eta_i \rho_k/2},$$

where the final step uses Assumption (3.3) and the fact that $5 + 8p \leq \frac{25}{2}p$ for all $p \geq 2$.

When $t = 1$, we therefore have

$$\begin{aligned}\|\boldsymbol{W}_1 \mathbb{1}_1\|_{p,p}^2 \leq \|\boldsymbol{W}_1 \mathbb{1}_0\|_{p,p}^2 &\leq K_{1,1}\|\boldsymbol{W}_0 \mathbb{1}_0\|_{p,p}^2 + K_{2,1} \\ &\leq \mathrm{e}^{-\eta_1 \rho_k/2}\|\boldsymbol{W}_0 \mathbb{1}_0\|_{p,p}^2 + C_2 p k^{2/p} \varepsilon_1^2,\end{aligned}$$

as desired, and for $t > 1$ the induction hypothesis yields

$$\begin{aligned}\|\boldsymbol{W}_t \mathbb{1}_t\|_{p,p}^2 &\leq \|\boldsymbol{W}_t \mathbb{1}_{t-1}\|_{p,p}^2 \\ &\leq K_{1,t}\|\boldsymbol{W}_{t-1} \mathbb{1}_{t-1}\|_{p,p}^2 + K_{2,t} \\ &\leq \mathrm{e}^{-\eta_t \rho_k/2}\left(\mathrm{e}^{-s_{t-1}\rho_k/2}\|\boldsymbol{W}_0 \mathbb{1}_0\|_{p,p}^2 + C_2 p k^{2/p} \varepsilon_{t-1}^2 (t-1)\right) \\ &\leq \mathrm{e}^{-s_t \rho_k/2}\|\boldsymbol{W}_0 \mathbb{1}_0\|_{p,p}^2 + C_2 p k^{2/p} \varepsilon_t^2 t,\end{aligned}$$

where the final inequality again uses the third assumption in (3.1). This proves the second bound. □

Appendix C. Additional results for Section 4

Lemma C.1. *Under the conditions of Proposition 4.1, the assumptions of (3.1) hold.*

Proof. First assumption. We have

$$\varepsilon_i = 2\eta_i M(1+\gamma) = 2(1+\sqrt{2}\mathrm{e})\frac{\alpha M}{(\beta+i)\rho_k} \leq C_\varepsilon \frac{\alpha}{\beta \bar{\rho}_k},$$

where $C_\varepsilon = 2(1+\sqrt{2}\mathrm{e})$. So the first assumption is fulfilled as long as

$$\beta/\alpha \geq 2C_\varepsilon/\bar{\rho}_k. \tag{C.1a}$$

Second assumption. As above, we have

$$\eta_i \|\boldsymbol{M}\| \leq \frac{\alpha\|\boldsymbol{M}\|}{\beta\rho_k} \leq \frac{\alpha}{\beta\bar{\rho}_k},$$

so the assumption is fulfilled if (C.1a) holds.

Third assumption. It suffices to show that

$$\frac{\varepsilon_{i-1}}{\varepsilon_i} \leq 1 + \frac{\eta_i \rho_k}{4} \qquad \forall i \geq 2,$$

which is equivalent to

$$\frac{1}{\beta+i-1} \leq \frac{\alpha/4}{\beta+i} \qquad \forall i \geq 2.$$

This holds as long as

$$\alpha \geq 8. \tag{C.1b}$$

We obtain that all three assumptions hold under (C.1a) and (C.1b), as claimed.

□

Lemma C.2. *In the setting of Theorem D.5, if* $s = 2\mathrm{e}\sqrt{\frac{\beta+1}{\beta+T}}$, *then*

$$\mathbb{P}\{\|\boldsymbol{W}_T\| \geq s\} \leq \delta/2.$$

Proof. We have

$$\mathbb{P}\{\|\boldsymbol{W}_T \mathbb{1}_T\| \geq s\} \leq \inf_{p \geq 2} s^{-p}\|\boldsymbol{W}_T \mathbb{1}_T\|_{p,p}^p.$$

In particular, we choose

$$s^2 = \mathrm{e}^2\left(\frac{\beta+1}{\beta+T}\right)^{\alpha} + \mathrm{e}^2\frac{C_3^2\alpha^2}{\bar{\rho}_k^2}\frac{T}{(\beta+T)^2}\log(2k/\delta), \quad \text{and} \quad p = \log(2k/\delta)\,.$$

It then follows from (4.4) that

$$\mathbb{P}\left\{\|\boldsymbol{W}_T\mathbb{1}_T\| \geq s\right\} \leq s^{-p}\|\boldsymbol{W}_T\mathbb{1}_T\|_p^p \leq k\left(\frac{1}{s^2}\left(\frac{\beta+1}{\beta+T}\right)^{\alpha} + \frac{1}{s^2}p\frac{C_3^2\alpha^2}{\bar{\rho}_k^2}\frac{T}{(\beta+T)^2}\right)^{p/2} = k\mathrm{e}^{-p} = \delta/2.$$

Combining the above bounds, we obtain that

$$\|\boldsymbol{W}_T\| \leq s \leq \mathrm{e}\left(\frac{\beta+1}{\beta+T}\right)^{\alpha/2} + \mathrm{e}\frac{C_3\alpha M}{\rho_k}\sqrt{\frac{\log(2k/\delta)}{T}}\,,$$

with probability at least $1-\delta$. Since both terms are smaller than $\mathrm{e}\sqrt{\frac{\beta+1}{\beta+T}}$, the claim follows. □

Appendix D. Additional results for Section 5

Our main tool will be the following slight variation on Proposition A.1.

Proposition D.1. *Let $t \geq 1$. Assume that η_t is small enough that $\boldsymbol{M} \geq -\frac{1}{2\eta_t}\mathbf{I}$, and assume that (2.4) holds for $i = t$. Consider an arbitrary deterministic matrix $\boldsymbol{E} \in \mathscr{C}_{r,\ell}$.*

Let

$$\bar{E}_t = 1 + 2\max_{\boldsymbol{E}'' \in \mathscr{C}_{r+1,\ell+1}} \|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}$$
$$\varepsilon = 2\eta M(1+\gamma)\,.$$

Then $\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\| \leq \varepsilon$ almost surely, and

$$\boldsymbol{E}\boldsymbol{W}_t(\mathbf{I} - \boldsymbol{\Delta}_t^2) = \boldsymbol{E}\boldsymbol{H}_t + \boldsymbol{E}\boldsymbol{J}_{t,1} + \boldsymbol{E}\boldsymbol{J}_{t,2}$$

for $\boldsymbol{E}\boldsymbol{J}_{t,1}$ and $\boldsymbol{E}\boldsymbol{J}_{t,2}$ satisfying

$$\|\boldsymbol{E}\boldsymbol{J}_{t,1}\mathbb{1}_{t-1}\|_{p,p} \leq \bar{E}_t\varepsilon$$
$$\|\boldsymbol{E}\boldsymbol{J}_{t,2}\mathbb{1}_{t-1}\|_{p,p} \leq \bar{E}_t\varepsilon^2\,,$$

and $\mathbb{E}[\boldsymbol{E}\boldsymbol{J}_{t,1} : \mathscr{F}_{t-1}] = \mathbf{0}$.

Proof. The proof is a slight modification on the proof of Proposition A.1. By construction,

$$\|\boldsymbol{E}\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \left(\frac{1+\eta\lambda_k}{1+\eta\lambda_{k+1}}\right)^2\|\boldsymbol{E}'\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2\,,$$

where $\boldsymbol{E}' = \frac{1}{1+\eta\lambda_{k+1}}\boldsymbol{E}\boldsymbol{U}^*(\mathbf{I}+\eta\boldsymbol{\Sigma})\boldsymbol{U} \in \mathscr{C}_{r+1,\ell} \subseteq \mathscr{C}_{r+1,\ell+1}$.

Similarly, we have

$$\begin{aligned}\|\boldsymbol{E}\widehat{\boldsymbol{\Delta}}_t\mathbb{1}_{t-1}\|_{p,p} &\leq 2\eta\|\boldsymbol{E}\boldsymbol{U}^*(\boldsymbol{A}_t - \boldsymbol{M})\boldsymbol{U}\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + 2\eta\|\boldsymbol{E}\boldsymbol{U}^*(\boldsymbol{A}_t - \boldsymbol{M})\boldsymbol{V}\|_{p,p}\\ &\leq 2\eta M(\|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + \|\boldsymbol{E}\|_{p,p})\\ &\leq \varepsilon(\|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + 1)\end{aligned}$$

where $\boldsymbol{E}'' = \frac{1}{M}\boldsymbol{E}\boldsymbol{U}^*(\boldsymbol{A}_t - \boldsymbol{M})\boldsymbol{U} \in \mathscr{C}_{r+1,\ell+1}$, and we have used $\|\boldsymbol{E}\|_p \leq \|\boldsymbol{E}\|_2 \leq 1$.

We therefore obtain

$$\|\boldsymbol{E}\boldsymbol{J}_{t,1}\mathbb{1}_{t-1}\|_{p,p} \leq \|\boldsymbol{E}\widehat{\boldsymbol{\Delta}}_t\mathbb{1}_{t-1}\|_{p,p} + \|\boldsymbol{E}\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\|$$

$$\leq \left(\|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + \|\boldsymbol{E}'\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p} + 1\right)\varepsilon$$
$$\leq \bar{E}_t\varepsilon\,,$$

and

$$\|\boldsymbol{E}\boldsymbol{J}_{t,2}\mathbb{1}_{t-1}\|_{p,p} \leq \|\boldsymbol{E}\widehat{\boldsymbol{\Delta}}_t\mathbb{1}_{t-1}\|_{p,p}\|\boldsymbol{\Delta}_t\mathbb{1}_{t-1}\| \leq (\|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}+1)\varepsilon^2 \leq \bar{E}_t\varepsilon^2\,.$$

□

The following two results are the appropriate analogues of Proposition A.2 and Theorem 3.1.

Proposition D.2. *Adopt the setting of Proposition D.1. If $\varepsilon \leq 1/2$, then*

$$\max_{\boldsymbol{E}\in\mathscr{E}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \bar{K}_1 \max_{\boldsymbol{E}'\in\mathscr{E}_{r+1,\ell}} \|\boldsymbol{E}'\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + \bar{K}_2 \max_{\boldsymbol{E}''\in\mathscr{E}_{r+1,\ell+1}} \|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + \bar{K}_2\,, \tag{D.1}$$

where

$$\bar{K}_1 = (1+5\varepsilon^2)\left(\frac{1+\eta\lambda_k}{1+\eta\lambda_{k+1}}\right)^2$$
$$\bar{K}_2 = (1+5\varepsilon^2)8p\varepsilon^2$$

Proof. As in the proof of Proposition A.2, we have for any $\boldsymbol{E}\in\mathscr{E}_{r,\ell}$,

$$\|\boldsymbol{E}\|_{p,p}^2 \leq (1+5\varepsilon^2)(\|\boldsymbol{E}\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}^2 + p\bar{E}_t\varepsilon^2)\,.$$

As in the proof of Proposition D.1, we can write

$$\|\boldsymbol{E}\boldsymbol{H}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \left(\frac{1+\eta\lambda_k}{1+\eta\lambda_{k+1}}\right)^2 \|\boldsymbol{E}'\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2$$

where $\boldsymbol{E}' = \frac{1}{1+\eta\lambda_{k+1}}\boldsymbol{E}\boldsymbol{U}^*(\mathbf{I}+\eta\boldsymbol{\Sigma})\boldsymbol{U} \in \mathscr{E}_{r+1,\ell}$. Since

$$\bar{E}_t^2 \leq 8 \max_{\boldsymbol{E}''\in\mathscr{E}_{r+1,\ell+1}} \|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + 8\,,$$

taking the maximum over all $\boldsymbol{E}\in\mathscr{E}_{r,\ell}$ and $\boldsymbol{E}'\in\mathscr{E}_{r+1,\ell}$ yields the claim. □

Theorem D.3. *Let $t \leq T_0$ be a positive integer, and assume the following requirements hold for some $p \geq 2$:*

$$\varepsilon \leq \frac{1}{2}\,, \tag{D.2a}$$
$$\eta\|\boldsymbol{M}\| \leq \frac{1}{2}\,, \tag{D.2b}$$
$$p\varepsilon^2 \leq \frac{\eta\rho_k}{50} \tag{D.2c}$$
$$\gamma \geq 2\,. \tag{D.2d}$$

Then for any $r, \ell \in [T_0 - t + 1]$ and $p \geq 2$,

$$\max_{\boldsymbol{E}\in\mathscr{E}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_t\mathbb{1}_t\|_{p,p}^2 \leq \max_{\boldsymbol{E}\in\mathscr{E}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \frac{\ell\gamma^2}{2\mathrm{e}^2}\mathrm{e}^{-t\eta\rho_k/2} + C_4 p\gamma^2\varepsilon^2 t\,.$$

where $C_4 = 6$.

Proof. First, as in the proof of Theorem 3.1, Assumptions (D.2b) and (D.2c) imply

$$\bar{K}_1 + \bar{K}_2 = (1+5\varepsilon^2)\left\{\left(\frac{1+\eta\lambda_k}{1+\eta\lambda_{k+1}}\right)^2 + 8p\varepsilon^2\right\}$$
$$\leq \mathrm{e}^{-\eta\rho_k/2}\,.$$

In particular, $\bar{K}_1 + \bar{K}_2 \le 1$. Assumption (D.2a) likewise implies that $\bar{K}_2 \le 18$.

We now turn to the proof of the main claim, which we prove by induction on t. For convenience, we introduce the notation $\gamma_e = \gamma/\sqrt{2}e$. When $t = 1$ and $r, \ell \le T_0$, (D.1) implies

$$\begin{aligned}
\max_{\boldsymbol{E}\in\mathscr{C}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_1\mathbb{1}_1\|_{p,p}^2 &\le \max_{\boldsymbol{E}\in\mathscr{C}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_1\mathbb{1}_0\|_{p,p}^2 \\
&\le \bar{K}_1 \max_{\boldsymbol{E}'\in\mathscr{C}_{r+1,\ell}} \|\boldsymbol{E}'\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + \bar{K}_2 \max_{\boldsymbol{E}''\in\mathscr{C}_{r+1,\ell+1}} \|\boldsymbol{E}''\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2\varepsilon^2 + \bar{K}_2 \\
&\le \bar{K}_1\ell\gamma_e^2 + \bar{K}_2(\ell+1)\gamma_e^2 + \bar{K}_2 \\
&\le \ell\gamma_e^2(\bar{K}_1+\bar{K}_2) + (1+\gamma_e^2)\bar{K}_2 \\
&\le \ell\gamma_e^2 e^{-\eta\rho_k/2} + \frac{\gamma^2}{3}\bar{K}_2
\end{aligned}$$

where we have used the definition of $\mathscr{C}_0$ and where the last step uses (D.2d). Proceeding by induction, we have

$$\begin{aligned}
\max_{\boldsymbol{E}\in\mathscr{C}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_t\mathbb{1}_t\|_{p,p}^2 &\le \max_{\boldsymbol{E}\in\mathscr{C}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \\
&\le \bar{K}_1 \max_{\boldsymbol{E}'\in\mathscr{C}_{r+1,\ell}} \|\boldsymbol{E}'\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + \bar{K}_2 \max_{\boldsymbol{E}''\in\mathscr{C}_{r+1,\ell+1}} \|\boldsymbol{E}''\boldsymbol{W}_{t-1}\mathbb{1}_{t-1}\|_{p,p}^2 + \bar{K}_2 \\
&\le \bar{K}_1(\ell\gamma_e^2 e^{-(t-1)\eta\rho_k/2} + (t-1)\gamma^2\bar{K}_2) \\
&\quad + \bar{K}_2((\ell+1)\gamma_e^2 e^{-(t-1)\eta\rho_k/2} + (t-1)\gamma^2\bar{K}_2) + \bar{K}_2 \\
&\le \ell\gamma_e^2(\bar{K}_1+\bar{K}_2)e^{-(t-1)\eta\rho_k/2} + (t-1)(\bar{K}_1+\bar{K}_2)\gamma^2\bar{K}_2 + (1+\gamma_e^2)\bar{K}_2 \\
&= \ell\gamma_e^2 e^{-t\eta\rho_k/2} + \frac{\gamma^2}{3}\bar{K}_2 t\,,
\end{aligned}$$

as claimed. □

Proposition D.4. *Fix $s \in (0,1)$, $2 \le \gamma \le C_\gamma \frac{d}{\delta^2}$, and $p \ge 2$, where $C_\gamma = 144\gamma$ is the constant in Lemma H.4. Given $\rho > 0$, define the normalized gap*

$$\bar{\rho} = \min\left\{\frac{M}{\rho}, \frac{\|\boldsymbol{M}\|}{\rho}, 1\right\},$$

and adopt the step size

$$\eta = \frac{C_\eta \log(ed/s\delta)}{\rho T_0}.$$

If $\rho_k \ge \rho/2$ and

$$T_0 \ge p \cdot \frac{C_T\gamma^2\log(ed/s\delta)^2}{s^2\bar{\rho}^2}$$

where

$$C_\eta \ge 8 + 4\log 2C_\gamma\,, \qquad C_T \ge 600e^2C_\eta^2\,,$$

then

$$\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0-1}\|_{p,p}^2 \le \frac{s^2}{2e^2}\left(1 + k^{2/p}\right)$$

and

$$\max_{\boldsymbol{E}\in\mathscr{C}_{1,1}} \|\boldsymbol{E}\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p} \le \frac{\gamma}{e}$$

for all $1 \le t \le T_0$.

Proof. We will apply Theorems 3.1 and D.3. First, note that (D.2d) holds by assumption. We now turn to the other conditions.
Assumption (D.2a): Since $\gamma \geq 2$, we have

$$\varepsilon = 2\eta M(1+\gamma) \leq \frac{3C_\eta \gamma \log(ed/s\delta)M}{\rho T_0}\,.$$

The assumption therefore holds as long as

$$C_T \geq 3C_\eta\,. \tag{D.3}$$

Assumption (D.2b): As above, we have

$$\eta\|\boldsymbol{M}\| \leq \frac{2C_\eta \log(ed/s\delta)\|\boldsymbol{M}\|}{\rho T_0}\,,$$

and the requirement (D.3) implies that this quantity is also smaller than $1/2$.
Assumption (D.2c): Since $\eta\rho_k = \frac{C_\eta \log(ed/s\gamma)}{T_0} \geq \frac{1}{T_0}$ and $36e^2 > 50$, it suffices to prove the stronger claim

$$p\varepsilon^2 \leq \frac{s^2}{36e^2 T_0}\,. \tag{D.4}$$

This is satisfied so long as

$$p \cdot \frac{16C_\eta^2\gamma^2 \log^2(ed/s\delta)M^2}{\rho^2 T_0^2} \leq \frac{s^2}{36e^2 T_0}\,.$$

which will hold if

$$C_T \geq 600e^2 C_\eta^2\,. \tag{D.5}$$

This requirement is stronger than (D.3), so Assumptions (D.2a)–(D.2c) hold under the sole condition (D.5).

We now turn to the two claimed bounds. First, we instantiate Theorem 3.1 with the choice $\eta_i = \eta$ for $1 \leq i \leq T_0$. The third assumption of (3.1) is trivially satisfied when when η_i is constant, since in that case $\varepsilon_i = \varepsilon_{i-1}$ for all $i \geq 1$. The remaining assumptions correspond directly to Assumptions (D.2a), (D.2b), and (D.2c). The assumptions of Theorem 3.1 are therefore satisfied, so we obtain,

$$\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0-1}\|_{p,p}^2 \leq e^{-T_0\eta\rho_k/2}\|\boldsymbol{W}_0\mathbb{1}_0\|_{p,p}^2 + 5pk^{2/p}\varepsilon^2 T_0\,.$$

The definition of $\mathcal{G}_0$ in (5.2) and the fact that $\rho_k \geq \rho/2$ implies that the first term is at most

$$e^{-T_0\eta\rho_k/2}d\gamma^2 = (ed/s\delta)^{-C_\eta/4}d\gamma^2\,,$$

and this will be less than $\frac{s^2}{2e^2}$ if

$$C_\eta \geq 8 + 4\log(2C_\gamma)\,.$$

Since (D.4) holds, the second term satisfies

$$5pk^{2/p}\varepsilon^2 T_0 \leq \frac{5s^2}{36e^2}k^{2/p} < \frac{s^2}{2e^2}k^{2/p}\,.$$

We obtain

$$\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0-1}\|_{p,p}^2 \leq \frac{s^2}{2e^2}\left(1 + k^{2/p}\right)\,,$$

as claimed.

For the second claim, we rely on Theorem D.3. Assumptions (D.2a)–(D.2d) having already been verified, we obtain for all $1 \leq t \leq T_0$,

$$\max_{\boldsymbol{E}\in\mathcal{E}_{1,1}} \|\boldsymbol{E}\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \frac{\gamma^2}{2e^2}e^{-t\eta\rho_k/2} + 18p\gamma^2\varepsilon^2 t\,.$$

Since $\rho_k \geq 0$, the first term is at most $\frac{\gamma^2}{2e^2}$, and the second term is also at most $\frac{\gamma^2}{2e^2}$ by (D.4). We obtain that

$$\max_{\boldsymbol{E}\in\mathscr{E}_{1,1}} \|\boldsymbol{E}\boldsymbol{W}_t \mathbb{1}_{t-1}\|_{p,p}^2 \leq \frac{\gamma^2}{e^2},$$

as claimed.

□

With Proposition D.4 in hand, we can prove a full version of Theorem 2.4.

Theorem D.5. *Fix a $\rho > 0$ and assume $|\mathrm{supp}(\mathrm{P}_A)| = m$. Let*

$$\bar{\rho} = \max\left\{\frac{\rho}{M}, \frac{\rho}{\|\boldsymbol{M}\|}, 1\right\},$$

and set $s = 1/6$.

Adopt the step size

$$\eta = \frac{C_\eta \log(ed/\delta s)}{\rho T_0}$$

where

$$T_0 \geq \frac{C_T k(\log 12ed/\delta\bar{\rho}s)^4}{s^2\delta^2\bar{\rho}^2}.$$

and

$$C_\eta \geq 8 + 2\log 144C_\gamma, \qquad C_T \geq (12000e^2 C_\eta^2 C_\gamma^2)^{5/4}.$$

If $m \leq T_0^3$ and $\rho_k \geq \rho/2$, then

$$\|\boldsymbol{W}_{T_0}\| \leq 1/6$$

with probability at least $1 - \delta/3$.

Proof. We first show that we can assume that $\log T_0 \leq 5\log(C_T d/\delta\bar{\rho}s)$. Indeed, if $T_0 > \left(\frac{C_T d}{\delta\bar{\rho}s}\right)^5$, a crude argument similar to the one employed in the analysis of Phase II yields the claim. We give the full details in Appendix F. In what follows, we therefore assume

$$\log T_0 \leq 5\log(C_T d/\delta\bar{\rho}s). \tag{D.6}$$

Set

$$\gamma = 144C_\gamma \min\left\{\frac{\sqrt{21k\log(C_T d/\delta\bar{\rho}s)}}{\delta}, \frac{d}{\delta^2}\right\},$$

where C_γ is as in Lemma H.4.

Recall that our goal is to show $\|\boldsymbol{W}_{T_0}\| \leq s$ with probability at least $1 - \delta/3$. The failure probability can be bounded as

$$\mathbb{P}\left\{\|\boldsymbol{W}_{T_0}\| \geq s\right\} \leq \mathbb{P}\left\{\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0}\| \geq s\right\} + \mathbb{P}\left\{\mathscr{G}_{T_0}^C\right\} \leq \inf_{p\geq 2} s^{-p}\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0}\|_{p,p}^p + \mathbb{P}\left\{\mathscr{G}_{T_0}^C\right\}.$$

If we choose $p = \log(6k/\delta)$, then since $\log(C_T) \leq C_T^{1/5}\log(12)$ for any value of C_T, we have

$$\begin{aligned}
T_0 &\geq \frac{C_T k(\log(12ed/\delta\bar{\rho}s))^4}{s^2\delta^2\bar{\rho}^2} \\
&\geq \log(6k/\delta)\cdot C_T^{4/5}\frac{k\log(C_T d/\delta\bar{\rho}s)}{\delta^2}\cdot\frac{\log(ed/s\delta)^2}{s^2\bar{\rho}^2} \\
&\geq p\frac{600e^2C_\eta^2\gamma^2\log(ed/s\delta)^2}{s^2\bar{\rho}^2},
\end{aligned}$$

as long as

$$C_T \geq (12000e^2 C_\eta^2 (144 C_\gamma)^2)^{5/4},$$

which verifies the assumption of Proposition D.4.

We obtain

$$\|\boldsymbol{W}_{T_0} \mathbb{1}_{T_0}\|_{p,p}^2 \leq \frac{s^2}{2e^2}(1 + k^{2/p}) \leq k^{2/p} \frac{s^2}{e^2}.$$

We therefore have

$$s^{-p} \|\boldsymbol{W}_{T_0} \mathbb{1}_{T_0}\|_{p,p}^p \leq e^{-\log(6k/\delta)} \leq \delta/6.$$

It remains to bound $\mathbb{P}\left\{\mathscr{G}_{T_0}^C\right\}$. Clearly

$$\mathbb{P}\left\{\mathscr{G}_{T_0}^C\right\} \leq \mathbb{P}\left\{\mathscr{G}_0^C\right\} + \sum_{j=1}^{T_0} \mathbb{P}\left\{\mathscr{G}_j^C \cap \mathscr{G}_{j-1}\right\}.$$

Since $m \leq T_0^3$ and we have assumed $\log T_0 \leq 5 \log(C_T d/\delta\bar{\rho}s)$, we have

$$\log(emT_0/\delta) \leq 4 \log(T_0) + \log(e/\delta) \leq 20 \log(C_T d/\delta\bar{\rho}s) + \log(e/\delta) \leq 21 \log(C_T d/\delta\bar{\rho}s),$$

so Lemma H.4 guarantees that $\mathscr{G}_0$ holds with probability at least $1 - \delta/12$.

For the second term, we have

$$\mathbb{P}\left\{\mathscr{G}_j^C \cap \mathscr{G}_{j-1}\right\} = \mathbb{P}\left\{\max_{\boldsymbol{E} \in \mathscr{C}_{1,1}} \|\boldsymbol{E}\boldsymbol{W}_j \mathbb{1}_{j-1}\| \geq \gamma\right\} \leq \sum_{\boldsymbol{E} \in \mathscr{C}_{1,1}} \mathbb{P}\left\{\|\boldsymbol{E}\boldsymbol{W}_j \mathbb{1}_{j-1}\| \geq \gamma\right\}.$$

Choose $p = 21 \log(C_T d/\delta\bar{\rho}s)$. The same argument as above yields

$$T_0 \geq p \cdot C_T^{3/5} \frac{k \log(C_T d/\delta\bar{\rho}s)}{\delta^2} \cdot \frac{\log^2(ed/s\delta)}{s^2\bar{\rho}^2},$$

and this will be larger than the lower bound required on T_0 that was assumed in Proposition D.4 as long as

$$C_T \geq (12000e^3 C_\eta^2 (144 C_\gamma)^2)^{5/3}$$

Proposition D.4 therefore yields

$$\mathbb{P}\left\{\|\boldsymbol{E}\boldsymbol{W}_j \mathbb{1}_{j-1}\| \geq \gamma\right\} \leq \gamma^{-p} \|\boldsymbol{E}\boldsymbol{W}_j \mathbb{1}_{j-1}\|_{p,p}^p \leq e^{-p} = e^{-21 \log(C_T d/\delta\bar{\rho}s)} \quad \text{for all } \boldsymbol{E} \in \mathscr{C},$$

and thus

$$\mathbb{P}\left\{\mathscr{G}_j^C | \mathscr{G}_{j-1}\right\} \leq \sum_{\boldsymbol{E} \in \mathscr{C}_{1,1}} \mathbb{P}\left\{\|\boldsymbol{E}\boldsymbol{W}_j \mathbb{1}_{j-1}\| \geq \gamma\right\} \leq m e^{-21 \log(C_T d/\delta\bar{\rho}s)}.$$

This yields

$$\sum_{j=1}^{T_0} \mathbb{P}\left\{\mathscr{G}_j^C | \mathscr{G}_{j-1}\right\} \leq m T_0 e^{-21 \log(C_T d/\delta\bar{\rho}s)} \leq e^{-21 \log(C_T d/\delta\bar{\rho}s) + 4 \log T_0} \leq \delta/12,$$

where the last step uses (D.6). Finally, choosing $s = 1/6$, we obtain

$$\mathbb{P}\left\{\|\boldsymbol{W}_{T_0}\| \geq 1/6\right\} \leq \delta/3,$$

as claimed. □

Appendix E. A reduction to finite support

Let Ω be the space of $d \times d$ symmetric matrices. We argue that it suffices to assume that P_A has finite support of cardinality at most T_0^3 in Phase I. We prove this by comparing the product measure $P_A^{\otimes T_0}$ with another distribution P_m on $\Omega^{\otimes T_0}$. We specify this distribution by the following procedure: drawing a T_0-tuple $(A_1, \ldots, A_{T_0})$ from the distribution P_m is accomplished by

(1) Drawing m independent samples $\hat{\boldsymbol{A}}_1, \ldots, \hat{\boldsymbol{A}}_m$ from P_A.

(2) Drawing $\boldsymbol{A}_1, \dots, \boldsymbol{A}_{T_0}$ independently from the discrete distribution

$$P_{\hat{A}} = \frac{1}{m} \sum_{i=1}^{m} \delta_{\hat{\boldsymbol{A}}_i}.$$

That is, drawing $\boldsymbol{A}_1, \dots, \boldsymbol{A}_{T_0}$ independently and uniformly from the set $\{\hat{\boldsymbol{A}}_i\}_{i=1}^m$ **with** replacement.

We will rely on the fact that the two distributions, $P_A^{\otimes T_0}$ and P_m, are close in total variation distance when m is large. To see this, we first recognize that drawing $(A_1, \dots, A_{T_0})$ from $P_A^{\otimes T_0}$ is equivalent to the following:

(1) Draw m independent samples $\hat{\boldsymbol{A}}_1, \dots, \hat{\boldsymbol{A}}_m$ from P_A.
(2) Draw $\boldsymbol{A}_1, \dots, \boldsymbol{A}_{T_0}$ sequentially and uniformly from the set $\{\hat{\boldsymbol{A}}_i\}_{i=1}^m$ **without** replacement. Denote by $P_{\hat{A}}^{(T_0)}$ the distribution of this sampling.

It is a standard result [11] that, given any $\{\hat{A}_i\}_{i=1}^m$,

$$d_{\mathrm{TV}}\left(P_{\hat{A}}^{\otimes T_0}, P_{\hat{A}}^{(T_0)}\right) \le \frac{1}{2}\frac{T_0^2}{m}.$$

We thus have the following:

Proposition E.1. *For any $\delta \in (0,1)$, it holds that*

$$d_{TV}\left(P_m, P_A^{\otimes T_0}\right) \le \delta$$

for all $m \ge T_0^2/2\delta$.

Proof. For any set $S \subset \Omega^{\otimes T_0}$, we have

$$\begin{aligned}
\left|P_m(S) - P_A^{\otimes T_0}(S)\right| &= \left|\mathbb{E}_{\hat{A}_i \sim P_A, 1\le i\le m}\left[P_{\hat{A}}^{\otimes T_0}(S) - P_{\hat{A}}^{(T_0)}(S)\right]\right| \\
&\le \mathbb{E}_{\hat{A}_i \sim P_A, 1\le i\le m}\left|P_{\hat{A}}^{\otimes T_0}(S) - P_{\hat{A}}^{(T_0)}(S)\right| \\
&\le \mathbb{E}_{\hat{A}_i \sim P_A, 1\le i\le m} d_{\mathrm{TV}}\left(P_{\hat{A}}^{\otimes T_0}, P_{\hat{A}}^{(T_0)}\right) \\
&\le \frac{1}{2}\frac{T_0^2}{m} \le \delta.
\end{aligned}$$

The claim follows from taking the maximum of $|P_m(S) - P_A^{\otimes T_0}(S)|$ over all subsets of $\Omega^{\otimes T_0}$. □

Given any $\hat{\boldsymbol{A}}_1, \dots, \hat{\boldsymbol{A}}_m$, define the empirical average

$$\hat{\boldsymbol{M}}_m := \mathbb{E}_{A \sim P_{\hat{A}}} \boldsymbol{A} = \frac{1}{m} \sum_{i=1}^{m} \hat{\boldsymbol{A}}_i.$$

Denote by $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_d$ the eigenvalues of $\hat{\boldsymbol{M}}_m$, and write $\hat{\rho}_k = \hat{\lambda}_k - \hat{\lambda}_{k+1}$. Let $\hat{V} \in \mathbb{R}^{d\times k}$ be the orthogonal matrix whose columns are the leading k eigenvectors of $\hat{\boldsymbol{M}}_m$, and let $\hat{\boldsymbol{U}} \in \mathbb{R}^{d\times(d-k)}$ be the orthogonal matrix consisting of the remaining eigenvectors. Standard results of matrix concentration implies that $\hat{\boldsymbol{M}}_m$ is close to $\boldsymbol{M}$. In particular, we have the following:

Proposition E.2. *Suppose that $m \ge \frac{35M^2}{\rho_k^2}\log(2d/\delta)$. Let $\hat{\boldsymbol{A}}_1, \dots, \hat{\boldsymbol{A}}_m$ be drawn independently from P_A. Then it holds with probability at least $1-\delta$ that*

$$\|\hat{\boldsymbol{M}}_m - \boldsymbol{M}\| \le \rho_k/4,$$

and, in particular,

$$\hat{\rho}_k \ge \rho_k/2 \quad \text{and} \quad \|\boldsymbol{U}^*\hat{\boldsymbol{V}}\| \le 1/3.$$

Proof. By assumption 2, we have that $\|\hat{\boldsymbol{M}}_m - \boldsymbol{M}\| \le M$ almost surely. Then the matrix Bernstein inequality [31, Theorem 1.4] implies that, for any $t \ge 0$,

$$\mathbb{P}\left\{\|\hat{\boldsymbol{M}}_m - \boldsymbol{M}\| \ge t\right\} \le 2d \exp\left(\frac{-mt^2/2}{M^2 + Mt/3}\right).$$

Substituting $t = \rho_k/4$ yields the first claim. Using the perturbation theory of eigenvalues of symmetric matrices, we have

$$\hat{\lambda}_k \ge \lambda_k - \|\hat{\boldsymbol{M}}_m - \boldsymbol{M}\| \quad \text{and} \quad \hat{\lambda}_{k+1} \le \lambda_{k+1} - \|\hat{\boldsymbol{M}}_m - \boldsymbol{M}\|.$$

Therefore, conditioned on the first claim, it holds that

$$\hat{\rho}_k \ge \rho_k - 2\|\hat{\boldsymbol{M}}_m - \boldsymbol{M}\| \ge \frac{\rho_k}{2}.$$

Furthermore, it follows from Wedin's inequality [33] that

$$\|\boldsymbol{U}^*\hat{\boldsymbol{V}}\| \le \frac{\|\hat{\boldsymbol{M}}_m - \boldsymbol{M}\|}{\hat{\lambda}_k - \lambda_{k+1}} \le \frac{1}{3}.$$

□

Proposition E.3. *Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be orthogonal matrices such that $\boldsymbol{U}\boldsymbol{U}^* + \boldsymbol{V}\boldsymbol{V}^* = \mathbf{I}$, and let $\hat{\boldsymbol{U}}$ and $\hat{\boldsymbol{V}}$ be matrices of the same size satisfying the same requirement. Suppose $\|\boldsymbol{U}^*\hat{\boldsymbol{V}}\| \le 1/2$ and $\|\hat{\boldsymbol{U}}^*\boldsymbol{S}(\hat{\boldsymbol{V}}^*\boldsymbol{S})^{-1}\| \le \gamma \le 1$. Then*

$$\|\boldsymbol{U}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| \le \frac{2+4\gamma}{3-2\gamma}.$$

Proof. A direct calculation yields

$$\begin{aligned}
\|\boldsymbol{U}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| &= \|\boldsymbol{U}^*(\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^* + \hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^*)\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| \\
&\le \|\hat{\boldsymbol{U}}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| + \|\boldsymbol{U}^*\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| \\
&\le \|\hat{\boldsymbol{U}}^*\boldsymbol{S}(\hat{\boldsymbol{V}}^*\boldsymbol{S})^{-1}\hat{\boldsymbol{V}}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| + \frac{1}{2}\|\hat{\boldsymbol{V}}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| \\
&\le (\gamma + \frac{1}{2})\|\hat{\boldsymbol{V}}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\|.
\end{aligned}$$

We also have

$$\|\hat{\boldsymbol{V}}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| \le \|\hat{\boldsymbol{V}}^*\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| + \|\hat{\boldsymbol{V}}^*\boldsymbol{V}\boldsymbol{V}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| \le \frac{1}{2}\|\boldsymbol{U}^*\boldsymbol{S}(\boldsymbol{V}^*\boldsymbol{S})^{-1}\| + 1.$$

Sequencing the two displays above and rearrange the inequality yields the claim. □

Now let T_0 be given as in Theorem D.5 and choose $m = T_0^2/2\delta$. As long as $T_0 \ge \frac{9M}{\rho_k\delta}\log(d/\delta)$, we have

$$\frac{35M^2}{\rho_k^2}\log(2d/\delta) \le m \le T_0^3.$$

It then follows from Proposition E.2 that, when drawing $\hat{\boldsymbol{A}}_1, \ldots, \hat{\boldsymbol{A}}_m$ independently from P_A, the event

$$\mathcal{G} := \{\hat{\rho}_k \ge \rho_k/2 \text{ and } \|\boldsymbol{U}^*\hat{\boldsymbol{V}}\| \le 1/2\} \tag{E.1}$$

happens with probability at least $1-\delta$. Conditioned on $\mathcal{G}$, we consider running T_0 steps of Oja's algorithm, with $A_1, \ldots, A_{T_0}$ drawn i.i.d from $P_{\hat{A}}$. Note that the discrete distribution $P_{\hat{A}}$ also satisfies Assumption 1 and Assumption 2 (with M replaced by $2M$). Our main theorem thus guarantees that with appropriately chosen step size, the output $Q_{T_0} = Q_{T_0}(A_1, \ldots, A_{T_0})$ of this algorithm after T_0 steps satisfies

$$\|\hat{\boldsymbol{U}}^*\boldsymbol{Q}_{T_0}(\hat{\boldsymbol{V}}^*\boldsymbol{Q}_{T_0})^{-1}\| \le \frac{1}{6}$$

with probability $1-\delta$. Combining (E.1) and Proposition E.3, we obtain that with probability at least $(1-\delta)^2 \geq 1-2\delta$, the output of the algorithm satisfies

$$\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1,$$

that is,

$$P_m\left(\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1\right) \geq 1-2\delta.$$

Finally, we obtain from Proposition E.1 that

$$\begin{aligned} &P_m\left(\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1\right) \\ &\geq P_A^{\otimes T_0}\left(\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1\right) - d_{\mathrm{TV}}\left(P_m, P_A^{\otimes T_0}\right) \\ &\geq 1-3\delta. \end{aligned}$$

In other words, with the same choice of T_0, the output of T_0 steps of Oja's algorithm with $A_1, \ldots, A_{T_0}$ drawn i.i.d from the original distribution P_A satisfies

$$\|\boldsymbol{U}^*\boldsymbol{Q}_{T_0}(\boldsymbol{V}^*\boldsymbol{Q}_{T_0})^{-1}\| \leq 1$$

with probability at least $1-3\delta$.

Appendix F. Phase I succeeds if T_0 is large

In this section, we prove Theorem D.5 when $T_0 > \frac{C_T^5 d^5}{\delta^5\bar{\rho}^5 s^5}$. Note that this value of T_0 is far larger than the optimal choice (which is of order $\tilde{\Theta}(k/\delta^2\bar{\rho}^2 s^2)$), which makes the theorem much easier to prove. Indeed, if T_0 is this large, we can prove Theorem D.5 directly by using the same conditioning argument as in Phase II.

Proposition F.1. *Assume η and T_0 satisfy the requirements of Theorem D.5, and assume $\rho \geq \rho_k/2$. If $T_0 \geq \frac{C_T^5 d^5}{\delta^5\bar{\rho}^5 s^5}$, then*

$$\|\boldsymbol{W}_{T_0}\| \leq s$$

with probability at least $1-\delta/3$.

Proof. Set $\gamma = \frac{144 C_\gamma d}{\delta^2}$ where C_γ is defined in Lemma H.4 and define the good events

$$\begin{aligned} \mathcal{G}_0 &:= \{\|\boldsymbol{W}_0\| \leq \gamma/(\sqrt{2}\mathrm{e})\} \\ \mathcal{G}_i &:= \{\|\boldsymbol{W}_0\| \leq \gamma\} \cap \mathcal{G}_{i-1}\,, \quad \forall i \geq 1\,. \end{aligned}$$

In order to apply Theorem 3.1, we verify (3.1)
First assumption. We have

$$\varepsilon = 2\eta M(1+\gamma) \leq \frac{3C_\eta \log(\mathrm{e}d/\delta s) M\gamma}{\rho T_0}\,,$$

and this quantity is smaller than $1/2$ so long as

$$C_T^5 \geq 864 C_\eta C_\gamma\,. \tag{F.1}$$

Second assumption. We again have

$$\eta\|\boldsymbol{M}\| = \frac{C_\eta \log(\mathrm{e}d/\delta s)\|\boldsymbol{M}\|}{\rho T_0}\,,$$

and (F.1) guarantees that this quantity is smaller than $1/2$ as well.
Third assumption. Since $\varepsilon_i = \varepsilon$ for all i and $\eta\rho \geq 0$, this requirement trivially holds.

Our goal is to bound

$$\mathbb{P}\left\{\|\boldsymbol{W}_{T_0}\| \geq s\right\} \leq \mathbb{P}\left\{\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0}\| \geq s\right\} + \mathbb{P}\left\{\mathcal{G}_0^C\right\} + \sum_{j=1}^{T_0} \mathbb{P}\left\{\mathcal{G}_j^C \cap \mathcal{G}_{j-1}\right\}.$$

Having verified (3.1), we can employ (3.2), obtaining

$$\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0}\|_{p,p}^2 \leq \mathrm{e}^{-T_0\eta\rho_k}k^{2/p}\gamma^2/2\mathrm{e}^2 + (C_1\gamma^2 + C_2)pk^{2/p}\varepsilon^2 T_0\,.$$

For the first term, the fact that $\rho_k \geq \rho/2$ implies that

$$\mathrm{e}^{-T_0\eta\rho_k}\frac{\gamma^2}{2\mathrm{e}^2} = (\delta s/\mathrm{e}d)^{C_\eta/2}\frac{\gamma^2}{2\mathrm{e}^2}\,,$$

and this is smaller than $\frac{s^2}{2\mathrm{e}^2}$ as long as

$$C_\eta \geq 8 + 2\log(144C_\gamma)\,.$$

Letting C_3 be as in Proposition 4.1 and choosing $p = \log(6k/d\delta)$, we also have

$$p(C_1\gamma^2 + C_2)\varepsilon^2 T_0 \leq p\frac{144^2C_3^2C_\eta^2\log^2(\mathrm{e}d/\delta s)M^2\gamma^2}{\rho^2 T_0} \leq \frac{144^2C_3^2C_\eta^2C_\gamma^2\log^3(6d/\delta s)}{C_T^5}\cdot\frac{\delta s}{d}$$

Since $\log^3(6d/\delta s) \leq 9\frac{d}{\delta s}$ for all positive d, δ, and s, this quantity will be less than $\frac{s^2}{2\mathrm{e}^2}$ so long as

$$C_T^5 \geq 2(432\mathrm{e}C_3C_\eta C_\gamma)^2\,, \tag{F.2}$$

and this requirement subsumes (F.1).

We therefore obtain, for $p = \log(6k/\delta)$,

$$\mathbb{P}\left\{\|\boldsymbol{W}_0\mathbb{1}_{T_0}\| \geq s\right\} \leq s^{-p}\|\boldsymbol{W}_0\mathbb{1}_{T_0}\|_{p,p}^p \leq k\mathrm{e}^{-p} \leq \delta/6\,,$$

In a similar way, (3.2) yields for all $t \in [T_0]$,

$$\gamma^{-2}\|\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^2 \leq \frac{k^{2/p}}{2\mathrm{e}^2} + (C_1\gamma^2 + C_2)pk^{2/p}\varepsilon^2 T_0\,.$$

If we choose $p = \log(12kT_0/\delta)$, then we have

$$p(C_1\gamma^2 + C_2)\varepsilon^2 T_0 \leq p\frac{C_3^2C_\eta^2\log^2(\mathrm{e}d/\delta s)M^2\gamma^2}{\rho^2 T_0} \leq \frac{2144^2C_3^2C_\eta^2C_\gamma^2\log^3(T_0)}{C_T^4T_0^{1/5}}\,,$$

and since $\log^3(T_0) \leq 169T_0^{1/5}$ for all T_0, we have that this quantity will be at most $\frac{1}{2\mathrm{e}^2}$ if

$$C_T^5 \geq (3744\mathrm{e}C_3C_\eta C_\gamma)^{5/2}\,,$$

and this requirement subsumes (F.2), and it holds under the assumptions of Theorem D.5.

By Lemma H.4, the event $\mathcal{G}_0$ holds with probability at least $1 - \delta/12$.

Finally, we have for any $j \in [T_0]$,

$$\mathbb{P}\left\{\mathcal{G}_j^C \cap \mathcal{G}_{j-1}\right\} \leq \mathbb{P}\left\{\|\boldsymbol{W}_j\mathbb{1}_{j-1}\| \geq \gamma\right\} \leq \inf_{p\geq 2}\gamma^{-p}\|\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^p\,,$$

and choosing $p = \log(12kT_0/\delta)$ we have

$$\gamma^{-p}\|\boldsymbol{W}_t\mathbb{1}_{t-1}\|_{p,p}^p \leq k\mathrm{e}^{-p} \leq \frac{12}{\delta T_0}\,,$$

and summing these probabilities for $j \in [T_0]$, yields that

$$\mathbb{P}\left\{\|\boldsymbol{W}_{T_0}\| \geq s\right\} \leq \mathbb{P}\left\{\|\boldsymbol{W}_{T_0}\mathbb{1}_{T_0}\| \geq s\right\} + \mathbb{P}\left\{\mathcal{G}_0^C\right\} + \sum_{j=1}^{T_0}\mathbb{P}\left\{\mathcal{G}_j^C \cap \mathcal{G}_{j-1}\right\} \leq \frac{1}{6} + \frac{1}{12} + \frac{1}{12} = \frac{1}{3}\,,$$

as claimed. □

Appendix G. Omitted proofs

G.1. **Proof of Lemma 2.7.** We will show that

$$W_t(\mathbf{I} - \Delta_t^2) = H_t + J_{t,1} + J_{t,2}\,,$$

where

$$H_t = U^*(\mathbf{I} + \eta_t M)Z_{t-1}(V^*(\mathbf{I} + \eta_t M)Z_{t-1})^{-1}, \quad J_{t,1} = \widehat{\Delta}_t - H_t\Delta_t, \quad \text{and} \quad J_{t,2} = -\widehat{\Delta}_t\Delta_t$$

and where we write

$$\widehat{\Delta}_t = \eta_t U^*(A_t - M)Z_{t-1}(V^*(\mathbf{I} + \eta_t M)Z_{t-1})^{-1}\,.$$

By the definition of Z_t, we have

$$W_t = U^*Z_t(V^*Z_t)^{-1} = U^*Y_tZ_{t-1}(V^*Y_tZ_{t-1})^{-1}.$$

We have

$$\begin{aligned} V^*Y_tZ_{t-1} &= V^*(\mathbf{I} + \eta_t M)Z_{t-1} + \eta_t V^*(A_t - M)Z_{t-1} \\ &= \left(\mathbf{I} + \eta_t V^*(A_t - M)Z_{t-1}(V^*(\mathbf{I} + \eta_t M)Z_{t-1})^{-1}\right) V^*(\mathbf{I} + \eta_t M)Z_{t-1} \\ &= (\mathbf{I} + \Delta_t)V^*(\mathbf{I} + \eta_t M)Z_{t-1}, \end{aligned}$$

which implies

$$\begin{aligned} (V^*Y_tZ_{t-1})^{-1}(\mathbf{I} - \Delta_t^2) &= (V^*(\mathbf{I} + \eta_t M)Z_{t-1})^{-1}(\mathbf{I} + \Delta_t)^{-1}(\mathbf{I} + \Delta_t)(\mathbf{I} - \Delta_t) \\ &= (V^*(\mathbf{I} + \eta_t M)Z_{t-1})^{-1}(\mathbf{I} - \Delta_t)\,. \end{aligned}$$

We also have

$$\begin{aligned} U^*Y_tZ_{t-1} &= U^*(\mathbf{I} + \eta_t M)Z_{t-1} + \eta_t U^*(A_t - M)Z_{t-1} \\ &= U^*(\mathbf{I} + \eta_t M)Z_{t-1} + \widehat{\Delta}_t(V^*(\mathbf{I} + \eta_t M)Z_{t-1}). \end{aligned}$$

Therefore

$$\begin{aligned} W_t(\mathbf{I} - \Delta_t^2) &= U^*Y_tZ_{t-1}(V^*Y_tZ_{t-1})^{-1} \\ &= U^*(\mathbf{I} + \eta_t M)Z_{t-1}(V^*(\mathbf{I} + \eta_t M)Z_{t-1})^{-1} \\ &\quad + \widehat{\Delta}_t - U^*(\mathbf{I} + \eta_t M)Z_{t-1}(V^*(\mathbf{I} + \eta_t M)Z_{t-1})^{-1}\Delta_t \\ &\quad - \widehat{\Delta}_t\Delta_t\,. \end{aligned}$$

That is

$$W_t(\mathbf{I} - \widehat{\Delta}_t^2) = H_t + J_{t,1} + J_{t,2}\,.$$

Since Δ_t and $\widehat{\Delta}_t$ are both $O(\eta_t)$, the claim follows. □

G.2. **Proof of Proposition 2.9.** By the triangle inequality, we have

$$\|X + Y + Z\|_{p,p} \le \|X + Y\|_{p,p} + \|Z\|_{p,p}\,,$$

which implies

$$\begin{aligned} \|X + Y + Z\|_{p,p}^2 &\le (\|X + Y\|_{p,p} + \|Z\|_{p,p})^2 \\ &\le (1 + \lambda)(\|X + Y\|_{p,p}^2 + \lambda^{-1}\|Z\|_{p,p}^2)\,, \end{aligned}$$

where in the second step we have applied the elementary inequality

$$(a + b)^2 \le (1 + \lambda)(a^2 + \lambda^{-1}b^2)\,,$$

valid for all real numbers a and b and $\lambda > 0$. Applying Proposition 2.8 to $\|X + Y\|_{p,p}^2$ then yields the claim. □

Appendix H. Additional Lemmas

Lemma H.1. *For any deterministic matrices $\boldsymbol{A}, \boldsymbol{B}$ and any standard Gaussian matrix $\boldsymbol{Z}$ of suitable sizes, it holds that*

$$\mathbb{P}\left\{\|\boldsymbol{AZB}\|_2 \geq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_2(1+t)\right\} \leq \mathrm{e}^{-t^2/2}.$$

Proof. Let $f(\boldsymbol{X}) := \|\boldsymbol{AXB}\|_2$, then

$$|f(\boldsymbol{X}_1) - f(\boldsymbol{X}_2)| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\| \cdot \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_2.$$

By Gaussian concentration, we have

$$\mathbb{P}\left\{f(\boldsymbol{Z}) \geq \mathbb{E}f(\boldsymbol{Z}) + \|\boldsymbol{A}\|\|\boldsymbol{B}\|t\right\} \leq \mathrm{e}^{-t^2/2}.$$

Moreover, we have

$$\mathbb{E}f(\boldsymbol{Z}) \leq (\mathbb{E}\|\boldsymbol{AZB}\|_2^2)^{1/2} = \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_2.$$

It thus follows that

$$\mathbb{P}\left\{f(\boldsymbol{Z}) \geq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_2(1+t)\right\} \leq \mathbb{P}\left\{f(\boldsymbol{Z}) \geq \mathbb{E}f(\boldsymbol{Z}) + \|\boldsymbol{A}\|\|\boldsymbol{B}\|t\right\} \leq \mathrm{e}^{-t^2/2},$$

which is the stated result. □

Lemma H.2 ([6, Theorem II.13])**.** *Let $\boldsymbol{Q} \in \mathbb{R}^{d\times k}$ be a standard Gaussian matrix. Then*

$$\mathbb{P}\left\{\|\boldsymbol{Q}\| \geq \sqrt{d} + \sqrt{k} + t\right\} \leq 2 \cdot \mathrm{e}^{-t^2/2}.$$

Lemma H.3 ([1, Lemma i.A.3])**.** *Let $\boldsymbol{Q} \in \mathbb{R}^{k\times k}$ be a standard Gaussian matrix. Then for every $\delta \in (0,1)$,*

$$\mathbb{P}\left\{\|\boldsymbol{Q}^{-1}\|_2 \geq \frac{6\sqrt{k}}{\delta}\right\} \leq \delta.$$

The next lemma bounds the probability of $\mathcal{G}_0$ from below.

Lemma H.4. *Let $\mathcal{G}_0$ be the event defined in (5.2). There exists a positive constant $C_\gamma = 144\mathrm{e}$ such that for any $\delta \in (0,1)$, if $\gamma \geq C_\gamma \min\{\sqrt{k\log(emT_0/\delta)}/\delta, d/\delta^2\}$, then $\mathcal{G}_0$ holds with probability at least $1-\delta$.*

Proof. We have $\boldsymbol{W}_0 = \boldsymbol{U}^*\boldsymbol{Z}_0(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}$, where $\boldsymbol{Z}_0$ is a matrix with i.i.d. Gaussian entries. Since $\boldsymbol{U}$ and $\boldsymbol{V}$ have orthonormal columns and are themselves orthogonal, the two matrices $\boldsymbol{V}^*\boldsymbol{Z}_0$ and $\boldsymbol{U}^*\boldsymbol{Z}_0$ are independent matrices with i.i.d. Gaussian entries. Using Lemma H.1 and conditioning on $\boldsymbol{V}^*\boldsymbol{Z}_0$, we have that with probability at least $1 - \delta/3(T_0+1)^2$,

$$\max_{\boldsymbol{E}\in\mathcal{E}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{U}^*\boldsymbol{Z}_0(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\|_2 \leq \|(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\|_2 \cdot 2\sqrt{8\ell\log(emT_0/\delta)}, \tag{H.1}$$

where we have taken a union bound over the fewer than $((m+1)(T_0+1))^\ell$ elements of $\mathcal{E}_{r,\ell}$. Taking a uniform bound again over all $r, \ell \in [T_0+1]$ yields that, with probability at least $1-\delta/3$, the event (H.1) holds for all $r, \ell \in [T_0+1]$. By Lemma H.3, we also have that that $\|(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\|_2 \leq 18\sqrt{k}/\delta$ with probability at least $1-\delta/3$. Furthermore, Lemma H.2 implies that $\|\boldsymbol{U}^*\boldsymbol{Z}_0\| \leq 2\sqrt{2d\log(3/\delta)}$ with probability at least $1-\delta/3$. Combining these bounds, we obtain that with probability at least $1-\delta/$,

$$\max_{\boldsymbol{E}\in\mathcal{E}_{r,\ell}} \|\boldsymbol{E}\boldsymbol{U}^*\boldsymbol{Z}_0(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\|_2 \leq 36\sqrt{8\ell\log(emT_0/\delta)},$$

which is less than $\frac{\sqrt{\ell}\gamma}{\sqrt{2}\mathrm{e}}$ as long as $C_\gamma \geq 144\mathrm{e}$, and under this same assumption

$$\|\boldsymbol{W}_0\|_2 \leq \|\boldsymbol{U}^*\boldsymbol{Z}_0\|\|(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\|_2 \leq 36\sqrt{2d\log(3/\delta)} \leq \sqrt{d}\gamma$$

as well.

So $\mathcal{G}_0$ holds with probability at least $1-\delta$ if $\gamma \geq C_\gamma \sqrt{k \log(emT_0/\delta)}/\delta$ for $C_\gamma \geq 144e$.

On the other hand, We have $\mathbb{E}\|\boldsymbol{U}^*\boldsymbol{Z}_0\| \leq 2\sqrt{d}$, so that $\|\boldsymbol{U}^*\boldsymbol{Z}_0\| \leq 4\sqrt{d}/\delta$ with probability at least $1-\delta/2$, and Lemma H.3 implies that $\|\boldsymbol{V}^*\boldsymbol{Z}_0\|_2 \leq 12\sqrt{k}/\delta$ with probability at least $1-\delta/2$, so with probability at least $1-\delta$ we have

$$\|\boldsymbol{W}_0\|_2 \leq \|\boldsymbol{U}^*\boldsymbol{Z}_0\|\|(\boldsymbol{V}^*\boldsymbol{Z}_0)^{-1}\|_2 \leq 48\sqrt{dk}/\delta^2 < 50d/\delta^2\,.$$

as claimed. On this event, we also have $\|\boldsymbol{E}\boldsymbol{W}_0\|_2 \leq \|\boldsymbol{W}_0\|_2 \leq 50d/\delta^2$. Therefore, if $\gamma \geq 50\sqrt{2}ed/\delta^2$, then $\mathcal{G}_0$ holds.

So $\mathcal{G}_0$ holds with probability at least $1-\delta$ if $\gamma \geq C_\gamma d/\delta^2$ for $C_\gamma \geq 50\sqrt{2}e$. Therefore, taking $C_\gamma = 144e$ satisfies both requirements and proves the claim. □

References

[1] Z. Allen-Zhu and Y. Li. First efficient convergence for streaming k-PCA: a global, gap-free, and near-optimal rate. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 487–492. IEEE Computer Soc., Los Alamitos, CA, 2017.

[2] M. Balcan, S. S. Du, Y. Wang, and A. W. Yu. An improved gap-dependency analysis of the noisy power method. In Feldman et al. [10], pages 284–309.

[3] M. Balcan and K. Q. Weinberger, editors. *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2016.

[4] A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental PCA. In Burges et al. [5], pages 3174–3182.

[5] C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2013.

[6] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.

[7] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[8] X. V. Doan and S. Vavasis. Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm. *SIAM J. Optim.*, 26(1):274–312, 2016.

[9] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[10] V. Feldman, A. Rakhlin, and O. Shamir, editors. *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2016.

[11] D. Freedman. A remark on the difference between sampling with and without replacement. *J. Amer. Statist. Assoc.*, 72(359):681, 1977.

[12] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[13] M. Hardt and E. Price. The noisy power method: A meta algorithm with applications. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2861–2869, 2014.

[14] A. Henriksen and R. Ward. Adaoja: Adaptive learning rates for streaming pca. 05 2019, 1905.12115.

[15] A. Henriksen and R. Ward. Concentration inequalities for random matrix products. *Linear Algebra Appl.*, 594:81–94, 2020.

[16] D. Huang, J. Niles-Weed, J. A. Tropp, and R. Ward. Matrix concentration for products. 03 2020, 2003.05437.

[17] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford. Streaming PCA: matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In Feldman et al. [10], pages 1147–1164.

[18] I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002.

[19] A. Juditsky and A. S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. 09 2008, 0809.0813.

[20] C. Li, H. Lin, and C. Lu. Rivalry of two families of algorithms for memory-restricted streaming PCA. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*

2016, Cadiz, Spain, May 9-11, 2016, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 473–481. JMLR.org, 2016.

[21] C. J. Li, M. Wang, H. Liu, and T. Zhang. Near-optimal stochastic approximation for online principal component estimation. *Math. Program.*, 167(1, Ser. B):75–97, 2018.

[22] C.-K. Li and N.-K. Tsing. Some isometries of rectangular complex matrices. *Linear and Multilinear Algebra*, 23(1):47–53, 1988.

[23] I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming PCA. In Burges et al. [5], pages 2886–2894.

[24] A. Naor. On the banach-space-valued azuma inequality and small-set isoperimetry of alon–roichman graphs. *Combinatorics, Probability and Computing*, 21(4):623–634, 2012.

[25] E. Oja. A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, 15(3):267–273, 1982.

[26] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69–84, 1985.

[27] C. D. Sa, C. Ré, and K. Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2332–2341. JMLR.org, 2015.

[28] O. Shamir. Convergence of stochastic gradient descent for PCA. In Balcan and Weinberger [3], pages 257–265.

[29] O. Shamir. Fast stochastic algorithms for SVD and PCA: convergence properties and convexity. In Balcan and Weinberger [3], pages 248–256.

[30] M. Simchowitz, A. El Alaoui, and B. Recht. Tight query complexity lower bounds for PCA via finite sample deformed Wigner law. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1249–1259. ACM, New York, 2018.

[31] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[32] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[33] P.-A. Wedin. Perturbation bounds in connection with singular value decomposition. *Nordisk Tidskrift for Informationsbehandling*, 12:99–111, 1972.