Beyond Sin-Squared Error: Linear Time Entrywise Uncertainty Quantification for Streaming PCA

Syamantak Kumar[1] **Shourya Pandey**[1] **Purnamrita Sarkar**[2]

[1]Department of Computer Science, University of Texas at Austin
[2]Department of Statistics and Data Sciences, University of Texas at Austin

Abstract

We propose a novel statistical inference framework for streaming principal component analysis (PCA) using Oja's algorithm, enabling the construction of confidence intervals for individual entries of the estimated eigenvector. Most existing works on streaming PCA focus on providing sharp sin-squared error guarantees. Recently, there has been some interest in uncertainty quantification for the sin-squared error. However, uncertainty quantification or sharp error guarantees for *entries of the estimated eigenvector* in the streaming setting remains largely unexplored. We derive a sharp Bernstein-type concentration bound for elements of the estimated vector matching the optimal error rate up to logarithmic factors. We also establish a Central Limit Theorem for a suitably centered and scaled subset of the entries. To efficiently estimate the coordinate-wise variance, we introduce a provably consistent subsampling algorithm that leverages the median-of-means approach, empirically achieving similar accuracy to multiplier bootstrap methods while being significantly more computationally efficient. Numerical experiments demonstrate its effectiveness in providing reliable uncertainty estimates with a fraction of the computational cost of existing methods.

1 INTRODUCTION

Principal Component Analysis (PCA) [Pearson, 1901, Ziegel, 2003] is a cornerstone for statistical data analysis and visualization. Given a dataset $\{X_i\}_{i=1}^n$, where each $X_i \in \mathbb{R}^d$ is independently drawn from a distribution $\mathcal{P}$ with mean zero and covariance matrix Σ, PCA computes the eigenvector v_1 of Σ that corresponds to the largest eigenvalue λ_1, and is the direction that explains the most variance in the data. It has been established [Wedin, 1972, Jain et al., 2016, Vershynin, 2012] that the leading eigenvector $\hat{v}$ of the empirical covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ is a nearly optimal estimator of v_1 under suitable assumptions on the data distribution.

While theoretically appealing, computing the empirical covariance matrix $\hat{\Sigma}$ explicitly requires $O(d^2)$ time and space, which is expensive in high-dimensional settings when both the sample size and the dimension are large. Oja's algorithm [Oja and Karhunen, 1985]— a streaming algorithm inspired by Hebbian learning [Hebb, 2005]— has emerged as an efficient and scalable algorithm for PCA. It maintains a running estimate of v_1 similar to a projected stochastic gradient descent (SGD) update

$$u_i \leftarrow u_{i-1} + \eta_n X_i (X_i^T u_{i-1}), \quad u_i \leftarrow \frac{u_i}{\|u_i\|_2} \tag{1}$$

for $i \in [n]$, where u_0 is a random unit vector and $\eta_n > 0$ is the learning rate. The algorithm is single-pass, runs in time $\mathcal{O}(nd)$, and takes only $\mathcal{O}(d)$ space. We call the output u_n of the above algorithm an *Oja vector* v_{oja}.

Oja's algorithm has fueled significant research in theoretical statistics, applied mathematics, and computer science [Jain et al., 2016, Allen-Zhu and Li, 2017, Chen et al., 2018, Yang et al., 2018, Henriksen and Ward, 2019, Price and Xun, 2024, Lunde et al., 2021, Monnez, 2022, Huang et al., 2021, Kumar and Sarkar, 2024a,b]. Despite the plethora of work on sharp rates for the sin-squared error $\sin^2(v_{\mathsf{oja}}, v_1) := 1 - (v_1^T v_{\mathsf{oja}})^2$, entrywise uncertainty estimation for streaming PCA has received only limited attention. Since the update rule in Oja's algorithm is similar to a broad class of important non-convex problems, uncertainty estimation for Oja's algorithm has potential implications for matrix sensing [Jain et al., 2013], matrix completion [Jain et al., 2013, Keshavan et al., 2010], subspace estimation [Balzano, 2022], and subspace tracking [Balzano et al., 2010]. A notable exception is Lunde et al. [2021], who show that $\sin^2(v_{\mathsf{oja}}, v_1) := 1 - (v_1^T v_{\mathsf{oja}})^2$ behaves asymptotically like a high-dimensional weighted chi-squared random variable. A main ingredient in their analysis is the

Hoeffding decomposition of the matrix product B_n. Their method takes $O(bnd)$ time and $O(bd)$ space, where b is the number of bootstrap replicas. While Lunde et al. [2021] do uncertainty estimation of the $\sin^2$ error, we are interested in coordinate-wise uncertainty estimation.

In contrast, in offline eigenvector analysis, there has been a surge of interest for *two-to-infinity* ($\ell_{2\to\infty}$) error bounds for empirical eigenvectors and singular vectors of random matrices [Eldridge et al., 2018, Mao et al., 2021, Abbe et al., 2020, Cape et al., 2019a, Abbe et al., 2022, Cape et al., 2019b]. However, none of these apply directly to the matrix product structure that arises from the Oja update in Eq (1). Recent advances on the concentration of matrix products [Huang et al., 2022, Kathuria et al., 2020] only provide operator norm or the ℓ_q moment of the Schatten norm of the deviation of a matrix product and do not provide non-trivial guarantees on the coordinates.

Our contributions:

In this paper, we obtain *finite sample and high probability deviation bounds* for elements of v_{oja}.

1. We show that the deviation of the elements of v_{oja} is governed by a suitably defined limiting covariance matrix $\mathbb{V}$. Furthermore, for a subset K of $[d]$ of interest, the distribution of the coordinate $v_{\mathsf{oja}}(k)$, when suitably centered and rescaled, is asymptotically normal with variance $\mathbb{V}_{kk}$.

2. We provide a sharp Bernstein-type concentration bound to show that *uniformly over entries of* v_{oja}, $\forall\, k \in [d]$,

$$|e_k^\top \underbrace{(v_{\mathsf{oja}} - (v_1^T v_{\mathsf{oja}})v_1)}_{:=r_{\mathsf{oja}}}| = \tilde{O}\left(\sqrt{\frac{\mathbb{V}_{kk}}{n}}\right). \tag{2}$$

where e_k denotes the kth standard basis vector. This is a surprising and sharp result because it can be used (see Lemma 8) to recover the optimal $\sin^2$ error up to logarithmic factors with high probability.

3. We provide an algorithm that couples a subsampling-based $O(nd)$ time and $O(d\log(d/\delta))$ space algorithm with Median of Means [Nemirovskij and Yudin, 1983] to estimate the marginal variances of the elements of $r_{\mathsf{oja}} := v_{\mathsf{oja}} - (v_1^T v_{\mathsf{oja}})v_1$. Theorem 2 provides high-probability error bounds of our variance estimator *uniformly* over $\forall k \in [d]$.

4. We present numerical experiments on synthetic and real-world data to show the empirical performance of our algorithm and also compare it to the multiplier bootstrap algorithm in Lunde et al. [2021] to show that our estimator achieves similar accuracy in significantly less time.

The paper is organized as follows: Section 1.1 discusses related work on streaming PCA, entrywise error bounds on eigenvectors, and statistical inference for Stochastic Gradient Descent. Section 2 provides our problem setup, assumptions, and necessary preliminaries. Section 3 provides our main results regarding entrywise concentration, CLT and our variance estimation algorithm, Algorithm 1. We provide proof sketches in Section 4 and experiments in Section 5.

1.1 RELATED WORK

Streaming PCA. A crucial measure of performance for Oja's algorithm is the $\sin^2$ error, which quantifies the discrepancy between the estimated direction and the principal eigenvector of Σ (the true population eigenvector, v_1) and the Oja vector, v_{oja}. Notably, several studies [Jain et al., 2016, Allen-Zhu and Li, 2017, Huang et al., 2021] have shown that Oja's algorithm attains the same error as its offline counterpart, which computes the leading eigenvector of the empirical covariance matrix directly. More concretely, it has been shown that for an appropriately defined variance parameter $\mathcal{V}$ (equation (3)),

$$\sin^2(v_1, v_{\mathsf{oja}}) := 1 - (v_1^T v_{\mathsf{oja}})^2 = O\left(\frac{\mathcal{V}}{n(\lambda_1 - \lambda_2)^2}\right).$$

ℓ_∞ error bounds. There is an extensive body of research on eigenvector perturbations of matrices. Most traditional bounds [Davis and Kahan, 1970, Wedin, 1972, Stewart and Sun, 1990] measure error using the ℓ_2 norm or other unitarily invariant norms. However, for machine learning and statistics applications, element-wise error bounds provide a better idea about the error in the estimated projection of *a feature* in a given direction. This area has recently gained traction for random matrices. Eldridge et al. [2018], Abbe et al. [2020], Cape et al. [2019a], Abbe et al. [2022] provide $\ell_{2\to\infty}$ bounds for eigenvectors and singular vectors of random matrices with low-rank structure. Cape et al. [2019a] show an $\ell_{2\to\infty}$ norm for the error of the singular vectors of a covariance matrix formed by n i.i.d. Gaussian vectors; as long as $\lambda_1 - \lambda_2 > 0$ and v_1 satisfies certain incoherence conditions, there exists a $w \in \{-1, 1\}$ such that with probability $1 - d^{-2}$, the top eigenvector $\hat{v}_1$ of the sample covariance matrix satisfies, up to logarithmic factors,

$$\|v_1 - w\hat{v}_1\|_\infty \lesssim \sqrt{\frac{\mathsf{Tr}\left(\Sigma\right)/\lambda_1}{n}}\left(\frac{\max_i \sqrt{\Sigma_{ii}}}{\sqrt{\lambda_1}} + \frac{\lambda_2}{\lambda_1}\right) + \frac{\mathsf{Tr}\left(\Sigma\right)/\lambda_1}{n}\left(\frac{1}{\sqrt{d}} + \sqrt{\frac{\lambda_2}{\lambda_1}}\right).$$

The guarantees of Cape et al. [2019a] are offline and provide a common upper bound on all coordinates. Our algorithm has error guarantees that scale with the variances of the coordinates.

Uncertainty estimation for SGD. For convex loss functions, the foundational work of Polyak and Juditsky [1992], Ruppert [1988], Bather [1989] in Stochastic Gradient Descent (SGD) demonstrates that averaged SGD iterates are asymptotically Gaussian. A significant body of research has focused on the convex setting. These include notable works

on covariance matrix estimation [Li et al., 2018, Su and Zhu, 2018, Fang et al., 2018, Chen et al., 2020, Lee et al., 2022, Zhu et al., 2023]. In comparison, work on uncertainty estimation for nonconvex loss functions is relatively few [Yu et al., 2021, Zhong et al., 2023]. Yu et al. [2021] establishes a Central Limit Theorem (CLT) under relaxations of strong convexity assumptions. Zhong et al. [2023] weakens the conditions but relies on online multiplier bootstrap methods to estimate the asymptotic covariance matrix. Existing methods for estimating and storing the full covariance matrix suffer from numerical instability or slow convergence rates (see Chee et al. [2023]). For convex functions and their relaxations, Zhu et al. [2024], Carter and Kuchibhotla [2025] present computationally efficient uncertainty estimation approaches that are related but different from ours.

In large-scale, high-dimensional problems, maintaining numerous bootstrap replicas is computationally expensive. Chee et al. [2023] introduce a scalable method for confidence intervals around SGD iterates, which are informative yet conservative under regularity conditions such as strong convexity at the optima. In their setting, for an appropriate initial learning rate, the covariance matrix can be approximated by a constant multiple of identity (see also Ljung et al. [1992]). In our setting, such an approximation requires knowledge of all eigenvalues and eigenvectors of Σ. The work most relevant to ours is by Lunde et al. [2021]. They provide asymptotic distributions for the sin-squared error of the Oja vector and present an online multiplier bootstrap algorithm to estimate the underlying distribution.

Resampling Methods and Bootstrapping. Nonparametric bootstrap [Efron, 1979, Hall, 1992, Efron and Tibshirani, 1993] is a resampling method where b resamples of a given size n dataset are drawn with replacement and treated as b independent samples drawn from the underlying distribution. Of these varieties of bootstraps, the one widely used in SGD inference is the online multiplier bootstrap, where multiple bootstrap resamples are updated in a streaming manner by sampling multiplier random variables to emulate the inherent uncertainty in the data [Ramprasad et al., 2023, Zhong et al., 2023, Lunde et al., 2021].

A major concern about the bootstrap is its computational bottleneck. Maintaining many bootstrap replicates is computationally prohibitive if the number of data points n and the dimension d are large. Some computationally cheaper alternatives to bootstrap are subsampling [Politis et al., 1999, Politis, 2023, Bertail et al., 1999, Levina and Priesemann, 2017, Chaudhuri et al., 2024, Chua et al., 2024] and m-out-of-n bootstrap [Bickel et al., 1997, Bickel and Sakov, 2008, Sakov, 1998, Andrews and Guggenberger, 2010] both of which rely on drawing $o(n)$ with-replacement samples. These methods are used in Kleiner et al. [2014] to create n with-replacement samples from smaller subsamples, but require multiple bootstrap replicates and are not directly applicable to the streaming setting.

2 PROBLEM SETUP AND PRELIMINARIES

Notation. Let $[n] = \{1, \ldots, n\}$ for all positive integers n. For a vector v, $\|v\| = \|v\|_2$ denotes its ℓ_2 norm. For a matrix A, $\|A\| = \|A\|_{\text{op}}$ is the operator norm, $\|A\|_{\text{F}}$ is the Frobenius norm, and $\|A\|_p$ is the Schatten p-norm of A, which is the ℓ_p norm of the vector of singular values of A. We define the *two-to-infinity* norm $\|A\|_{2\leftarrow\infty} := \sup_{\|x\|_2=1} \|Ax\|_\infty$. For a random matrix M and $p, q \geq 1$, we define the norm $|||M|||_{p,q} := \mathbb{E}[\|M\|_p^q]^{1/q}$. Let $I \in \mathbb{R}^{d\times d}$ be the identity matrix with i^{th} column e_i. Define the inner product of matrices as $\langle A, B\rangle = \mathsf{Tr}(A^T B)$. We use $\widetilde{O}$ and $\widetilde{\Omega}$ for bounds up to logarithmic factors and use $a \lesssim b$ to mean $a \leq Cb$ for some universal constant C. diag $(a_1, \ldots, a_d)$ denotes the diagonal matrix with entries $a_1, \ldots, a_d$. For a vector $v \in \mathbb{R}^d$ and $S \subseteq [d]$ with $|S| = k$, $v[S] \in \mathbb{R}^k$ is the "sub-vector" of v with its coordinates indexed by S.

Data. Let $\{X_i\}_{i\in[n]}$ be independent and identically distributed (i.i.d.) mean-zero vectors sampled from the distribution $\mathcal{P}$ over $\mathbb{R}^d$ with covariance matrix $\Sigma := \mathbb{E}\left[X_i X_i^T\right]$. Let $A_i := X_i X_i^\top$. Let $v_1, v_2, \ldots, v_d$ denote the eigenvectors of Σ with corresponding eigenvalues $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_d$. Let $V_\perp := [v_2, v_3, \ldots, v_d] \in \mathbb{R}^{d\times(d-1)}$.

We operate under the following assumptions unless otherwise specified.

Assumption 1. *For any $X_i \sim \mathcal{P}$, $A_i = X_i X_i^\top$, we assume the following moment bounds, where $\sqrt{\mathcal{V}} \leq \mathcal{M}_2 \leq \mathcal{M}_4$:*

$$\left\|\mathbb{E}\left[(A_i - \Sigma)^2\right]\right\|_{\text{op}} \leq \mathcal{V} \tag{3}$$

$$\mathbb{E}\left[\|A_i - \Sigma\|_{\text{op}}^2\right]^{\frac{1}{2}} \leq \mathcal{M}_2 \qquad \mathbb{E}\left[\|A_i - \Sigma\|_{\text{op}}^4\right]^{\frac{1}{4}} \leq \mathcal{M}_4. \tag{4}$$

Assumption 2. *There exists a universal constant $\kappa > 5$ such that $d = o\left(n^\kappa\right)$ and $\frac{n}{\log(n)} \geq 2\max\left\{\kappa, \frac{\kappa^2\mathcal{M}_2^4\log(d)}{(\lambda_1-\lambda_2)^4}\right\}$.*

Assumption 1 provides a suitable moment bound on the iterates A_i, and Assumption 2 shows that we can handle the dimension d growing polynomially with the sample size n, while requiring a mild base number of samples for convergence. We note that the constraint $\kappa > 5$ is arbitrary and our algorithm works as long as $d = \mathsf{poly}(n)$. These assumptions are commonly used in the streaming PCA literature (see for e.g. Jain et al. [2016]).

Oja's Algorithm with constant learning rate. With a constant learning rate, η_n, and initial vector, u_0, Oja's algorithm [Oja, 1982] (denoted as $\mathsf{Oja}\left(\{X_t\}_{t\in[n]}, \eta_n, u_0\right)$) performs the updates in Eq (1). Define $\forall t \in [n]$,

$$B_t := \prod_{i=0}^{t-1}\left(I + \eta_n X_{t-i} X_{t-i}^T\right); \qquad B_0 = I. \tag{5}$$

such that $u_t = B_t u_0 / \|B_t u_0\|_2$.

3 MAIN RESULTS

Recall the definition of Oja's algorithm with a constant learning rate, as defined in Section 2. For i.i.d. data $\mathcal{D}_n := \left\{X_i; X_i \in \mathbb{R}^d\right\}_{i\in[n]}$, the learning rate η_n defined in Lemma 9, and a random initial vector $u_0 := g/\|g\|$ where $g \sim \mathcal{N}(0, \mathbf{I}_d)$, define the *Oja vector*

$$v_{\mathsf{oja}}(\mathcal{D}_n) := \mathsf{Oja}(\mathcal{D}_n, \eta_n, u_0). \tag{6}$$

This is a random vector, with randomness over the data $\mathcal{D}_n$ as well as the initial vector u_0. While there are a myriad of works on the sin-squared error $1 - (v_1^T v_{\mathsf{oja}})^2$, there is, to our knowledge, no existing analysis on the concentration of the elements of the recovered vector around their population counterparts. One exception is [Kumar and Sarkar, 2024b], who showed that for sparse PCA, the elements of the Oja vector in the support of the true eigenvector are large, whereas those outside are small. However, these guarantees do not show concentration in our setting. We start our analysis with the Hoeffding decomposition of the matrix product (also see Lunde et al. [2021], van der Vaart [2000]). The Hoeffding decomposition is a powerful tool that allows one to write the *residual* of the Oja vector as

$$r_{\mathsf{oja}} := v_{\mathsf{oja}} - \left(v_1^\top v_{\mathsf{oja}}\right) v_1 = \Psi_{n,1} + \mathsf{Res}_n \tag{7}$$

where $\Psi_{n,1}$ is η_n times a sum of independent but non-identically distributed random vectors and the residual Res_n is negligible compared to $\Psi_{n,1}$ (see Lemma 2 for details).

First, we show that the covariance matrix $\mathbb{E}[\Psi_{n,1}\Psi_{n,1}^T]$ of the dominant term in the residual converges to $\mathbb{V}$ when suitably scaled. Later, in Proposition 1 we will show that the distribution of the entries of r_{oja} is asymptotically normal with covariance matrix $\mathbb{E}[\Psi_{n,1}\Psi_{n,1}^T]/(\eta_n(\lambda_1 - \lambda_2))$.

Lemma 1 (Asymptotic variance)**.** *Let*

$$\widetilde{M} := \mathbb{E}\left[V_\perp^\top (A_1 - \Sigma) v_1 v_1^\top (A_1 - \Sigma) V_\perp\right],$$
$$d_k := 1 - \left(\frac{\lambda_1 - \lambda_{k+1}}{1 + \eta_n \lambda_1}\right)\eta_n.$$

Then, the matrix $R^{(n)} \in \mathbb{R}^{(d-1)\times(d-1)}$ with entries

$$R_{k,l}^{(n)} := \frac{\widetilde{M}_{kl}}{(1+\eta_n\lambda_1)^2}\left(\frac{1-(d_k d_l)^n}{1-d_k d_l}\right),$$

satisfies $\mathbb{E}\left[\Psi_{n,1}\Psi_{n,1}^\top\right] = \eta_n^2 V_\perp R^{(n)} V_\perp^\top$.

Define the matrices $R_0 \in \mathbb{R}^{(d-1)\times(d-1)}$ and $\mathbb{V} \in \mathbb{R}^{d\times d}$ as

$$(R_0)_{k,l} := \frac{\widetilde{M}_{k\ell}}{2\lambda_1 - \lambda_{k+1} - \lambda_{\ell+1}}; \quad \mathbb{V} := \frac{1}{\lambda_1 - \lambda_2} V_\perp R_0 V_\perp^T. \tag{8}$$

then,

$$\left\|\frac{1}{\eta_n(\lambda_1 - \lambda_2)}\mathbb{E}[\Psi_{n,1}\Psi_{n,1}^T] - \mathbb{V}\right\|_F \lesssim \frac{\eta_n \lambda_1 \mathcal{M}_2^2}{(\lambda_1 - \lambda_2)^2}. \tag{9}$$

This shows that suitably scaled, $\mathbb{E}[\Psi_{n,1}\Psi_{n,1}^T]$ converges to the matrix $\mathbb{V}$. Note that the scaling factor $\eta_n(\lambda_1 - \lambda_2) = \frac{\alpha \log n}{n}$ is independent of model parameters for the choice of η_n defined in Lemma 9.

The next result establishes a Central Limit Theorem (CLT) for the subset of elements in the residual vector r_{oja} with sufficiently large limiting variance.

Proposition 1 (CLT for a suitable subset of entries)**.** *Let $\{X_i\}_{i=1}^n$ be independent mean-zero random vectors with covariance matrix Σ such that $\mathbb{E}\left[\exp(v^\top X_1)\right] \leq \exp\left(\frac{\sigma^2 v^\top \Sigma v}{2}\right)$ for all $v \in \mathbb{R}^d$ and $\sigma > 0$ is some constant.*

For all $i \in [n]$, let

$$H_i := \frac{\mathsf{sign}\left(v_1^\top u_0\right)}{(1+\eta_n\lambda_1)} V_\perp \Lambda_\perp^{n-i} V_\perp^\top (A_i - \Sigma) v_1,$$

Let $b > 0$ be a constant, and let $J \subseteq [d]$ be the set of coordinates with $\mathbb{V}_{jj} \geq b$. Let $p := |J|$.

Let $Y_i \in \mathbb{R}^p$ be independent mean-zero Gaussian vectors with covariance matrix

$$\mathbb{E}[Y_i Y_i^\top] = \frac{n\eta_n}{\lambda_1 - \lambda_2}\mathbb{E}[H_i[J]H_i[J]^\top],$$

and let $S_Y := \sum_{i=1}^n Y_i$.

Suppose the learning rate η_n, set according to Lemma 9, satisfies $\frac{\mathcal{M}_2^2 \lambda_1 \eta_n}{(\lambda_1-\lambda_2)^2} \lesssim b$. Then,

$$\sup_{A\in\mathcal{A}^{re}}\left|\mathbb{P}\left(\frac{r_{\mathsf{oja}}[J]}{\sqrt{(\lambda_1-\lambda_2)\eta_n}} \in A\right) - \mathbb{P}\left(\frac{S_Y}{\sqrt{n}} \in A\right)\right|$$
$$= \tilde{O}\left(\left(\frac{\mathcal{M}_4}{\lambda_1-\lambda_2}\right)^{1/3} n^{-1/6} + \left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^{1/2} n^{-1/8}\right),$$

where $\mathcal{A}^{re}$ is the collection of all hyperrectangles in $\mathbb{R}^p$, i.e, sets of the form $A = \{u \in \mathbb{R}^p : a_j \leq u_j \leq b_j \text{ for } j = 1, \ldots, p\}$ and each a_j and b_j belongs to $\mathbb{R} \cup \{-\infty, \infty\}$. Here, $\tilde{O}$ hides logarithmic factors in n, d, and polynomial factors in b and in model parameters $\lambda_1, \lambda_1 - \lambda_2, \mathcal{M}_2, \mathcal{M}_4$.

Remark 1. *Note that the first $n^{-1/6}$ term in the convergence rate arises from the high-dimensional CLT result by Chernozhukov et al. [2017a] applied to $\Psi_{n,1}$. The main bottleneck is the $n^{-1/8}$ term, resulting from the higher-order terms of the Hoeffding decomposition (Res_n in equation 7). We note that the second term may be tightened by using better concentration bounds. We point the reader to Proposition 2 in the Appendix for a complete statement and proof.*

Proposition 1 establishes a Gaussian approximation of suitably scaled $r_{\mathsf{oja}}[J]$, where J is a set of elements with large enough asymptotic variance. Our proof uses results from Chernozhukov et al. [2017b] on the Hájek projection (7) and bounds the effect of the remainder term by using Nazarov's Lemma [Nazarov, 2003] (Theorem 4). We use this to derive concentration bounds for all coordinates.

The lower bound on the variance is crucial and comes from Nazarov's inequality. It is also a condition of the results in Chernozhukov et al. [2017b]. A simple observation here is that when b_k is zero, i.e. $v_1(k) = 1$, then $\mathbb{V}_{kk} = 0$. Here, CLT may not hold since the Hájek projection is zero, and the perturbation arises from some of the smaller error terms in the error decomposition.

Theorem 1. *Let the learning rate η_n be set according to Lemma 9. Further, for $X_i \sim \mathcal{P}, A_i = X_i X_i^\top$, let $\|A_i - \Sigma\|_{\mathsf{op}} \leq \mathcal{M}$ almost surely. Then, with probability at least $3/4$, uniformly for all $k \in [d]$,*

$$\frac{\left|e_k^\top r_{\mathsf{oja}}\right|}{\sqrt{\eta_n(\lambda_1 - \lambda_2)}} \lesssim \sqrt{\mathbb{V}_{kk} \log(d)} + C b_k \sqrt{\frac{\log n}{n}},$$

where $b_k := \left\|e_k^\top V_\perp\right\|_2$, $\mathbb{V}$ is defined in Eq 8, and C is a constant that depends on $\lambda_1, \lambda_1 - \lambda_2, \mathcal{M}_2$, and $\mathcal{M}$.

Remark 2. *The limiting marginal variances $\mathbb{V}_{kk}$ also appear in the finite-sample bound for the elements of the residual vector. Estimating these variances enables us to quantify the uncertainty associated with each component of $\hat{v}_1$, even when the sample size is finite.*

In Appendix C, we provide a complete result with arbitrary failure probability δ in Lemma 28. The above guarantee can be boosted to a high probability one using geometric aggregation (see e.g. Alg. 3 in Kumar and Sarkar [2024b]).

3.1 UNCERTAINTY ESTIMATION

Proposition 1 shows that the asymptotic variance of elements of the residual $r_{\mathsf{oja}}(i)$ is governed by the variance of the entries $\mathbb{E}[(e_i^T \Psi_{n,1})^2]$ of $\Psi_{n,1}$. We cannot directly get to $\Psi_{n,1}$ since we only observe v_{oja}. If we could estimate r_{oja}, it would give us an idea of the error. However, we do not know v_1, and so cannot directly access r_{oja}. We alleviate this difficulty by using the following high-accuracy estimate of v_1 constructed using N samples,

$$\tilde{v} \leftarrow \mathsf{Oja}(\mathcal{D}_N, \eta_N, u_0), \tag{10}$$

where N satisfies the bounds of Theorem 2.

We now provide a subsampling-based approach (Alg. 1) to estimate $\mathbb{E}[(e_i^T \Psi_{n,1})^2]$ with high probability, allowing us to provide confidence intervals around the eigenvector elements. Algorithm 1 takes as input the data $\{X_i \in \mathbb{R}^d\}_{i \in [n]}$, a failure probability δ, and the proxy unit vector $\tilde{v}$. The n samples are split into m_1 batches with n/m_1 samples each. Then, the ℓth batch of n/m_1 samples is further split into m_2 batches of size $B := n/(m_1 m_2)$ each. Oja vectors $\{\hat{v}_{\ell,j}\}_{j \in [m_2]}$ are computed on each of these m_2 batches, and the variance of the kth coordinate is estimated as

$$\hat{\sigma}^2_{k,\ell} := \sum_{j \in [m_2]} \frac{\left(e_k^\top \left(\hat{v}_{\ell,j} - (\tilde{v}^\top \hat{v}_{\ell,j})\tilde{v}\right)\right)^2}{m_2}. \tag{11}$$

We will show that with a constant success probability, $\hat{\sigma}^2_{k,\ell}$ is close to the true variance of the corresponding coordinate. This is essentially the variance of a smaller dataset with scale η_B. To obtain a bound over all coordinates with an arbitrary failure probability, we take a median of the m_1 variances. For the final estimate of the diagonal elements $\mathbb{V}_{kk}$ of $\mathbb{V}$, the median is scaled by a factor $1/\eta_B(\lambda_1 - \lambda_2)$. In Theorem 2, we show that $\hat{\gamma}_k$ concentrates around $\mathbb{V}_{kk}$ (see (14)). For elements with large $\mathbb{V}_{kk}$, appropriate sample size N and batch size B, Theorem 2 also provides multiplicative error guarantees for the variance estimate (see (15)).

Remark 3. *We are using an estimate of $\mathbb{E}[(e_k^T \Psi_{n,1})^2]$ to provide the confidence interval around $\hat{v}_1(k)$. Algorithm 1 requires an estimate $\tilde{v}$ of v_1 for computing the estimates $\hat{\sigma}^2_{\ell,k}$ in Line 11, which is provided as an input to the algorithm and assumed to satisfy $\tilde{v} \leftarrow \mathsf{Oja}(\mathcal{D}_N, \eta_N, z/\|z\|_2)$ for $z \sim \mathcal{N}(0, I)$. For large N, this error of approximating v_1 by $\tilde{v}$ is small. In our experiments, we choose $N = n$ and obtain $\tilde{v}$ by running the algorithm on the entire data.*

Algorithm 1 $\mathsf{OjaVarEst}(\{X_i \in \mathbb{R}^d\}_{i \in [n]}, \delta, \tilde{v}, \lambda_1 - \lambda_2)$

1: **Input:** Data $\mathcal{D}_n := \{X_i \in \mathbb{R}^d\}_{i \in [n]}$, failure probability $\delta \in (0,1)$, unit vector $\tilde{v}$, eigengap $\lambda_1 - \lambda_2$
2: **Output:** Estimates $\{\hat{\gamma}_k\}_{k \in [d]}$ of $\{\mathbb{V}_{kk}\}_{k \in [d]}$
3: $m_1 \leftarrow 8\log(d/\delta)$, $m_2 \leftarrow \log n$, $B \leftarrow n/(m_1 m_2)$.
4: **for** $\ell \in [m_1]$ **do**
5: **for** $j \in [m_2]$ **do**
6: $\mathcal{D}_{\ell,j} \leftarrow \left\{X_{B(m_2(\ell-1)+(j-1))+t}\right\}_{t \in [B]}$
7: $g \leftarrow \mathcal{N}(0, I), \quad u \leftarrow g/\|g\|_2$
8: $\hat{v}_{\ell,j} \leftarrow \mathsf{Oja}(\mathcal{D}_{\ell,j}, \eta_B, u_0)$
9: **end for**
10: **for** $k \in [d]$ **do**
11: $\hat{\sigma}^2_{\ell,k} \leftarrow \frac{\sum_{j \in [m_2]} \left(e_k^\top \left(\hat{v}_{\ell,j} - (\tilde{v}^\top \hat{v}_{\ell,j})\tilde{v}\right)\right)^2}{m_2}$
12: **end for**
13: **end for**
14: **for** $k \in [d]$ **do**
15: $\hat{\gamma}_k \leftarrow \frac{1}{\eta_B(\lambda_1 - \lambda_2)} \mathsf{Median}\left(\left\{\hat{\sigma}^2_{\ell,k}\right\}_{\ell \in [m_1]}\right)$
16: **end for**
17: **return** $\{\hat{\gamma}_k\}_{k \in [d]}$

Theorem 2. *Let K be the set of indices in [d] that satisfy*

$$N = \tilde{\Omega}\left(B/c_k^2\right) \quad \textit{and} \tag{12}$$

$$B = \tilde{\Omega}\left(\left(\frac{b_k}{c_k}\right)^2 \left(\frac{\mathcal{M}_2}{\lambda_1 - \lambda_2}\right)^2\right) + \tilde{\Omega}\left(\left(\frac{b_k}{c_k}\right)^4 \left(\frac{\mathcal{M}_4}{\mathcal{M}_2}\right)^4 + \frac{\lambda_1}{c_k^2(\lambda_1 - \lambda_2)}\right), \tag{13}$$

where $b_k := \left\|e_k^\top V_\perp\right\|$, $c_k := \sqrt{\frac{\mathbb{E}\left[\left(e_k^\top \Psi_{B,1}\right)^2\right]}{\eta_B} \frac{\lambda_1 - \lambda_2}{\mathcal{M}_2^2}}$, and B, N are respectively the batch size and the number of samples used for the proxy estimate $\tilde{v}$ in Algorithm 1.

Then, with probability at least $1-\delta$, the output $\{\hat{\gamma}_k\}_{k\in[d]}$ of Algorithm 1 satisfies

$$|\hat{\gamma}_k - \mathbb{V}_{kk}| \lesssim \frac{\mathbb{V}_{kk}}{\sqrt{m}} + \tilde{O}\left(\frac{B}{N} + \frac{1}{B^{1/2}}\right) \quad \forall k \in [d], \text{ and} \tag{14}$$

$$|\hat{\gamma}_k - \mathbb{V}_{kk}| \lesssim \frac{\mathbb{V}_{kk}}{\sqrt{m}} \quad \forall k \in [K]. \tag{15}$$

Remark 4. *The output of Algorithm 1 rescales the median of the variances by the quantity $\eta_B(\lambda_1-\lambda_2) = \frac{\alpha \log B}{B}$. This is consistent with the entrywise concentration bounds in Theorem 1 (which shows that the error in the jth entry is $\sqrt{\eta_n(\lambda_1-\lambda_2)\mathbb{V}_{kk}}$, up to logarithmic terms) for a sufficiently large sample size and with Proposition 1 and Lemma 1 (which show that the limiting variance of suitable entries of r_{oja} is $\eta_n(\lambda_1-\lambda_2)\mathbb{V}_{kk}$).*

Remark 5. *Theorem 1 provides bounds about entries of the leading eigenvector. We believe our techniques can be generalized to provide uncertainty estimates for entries of top-k eigenvectors using deflation-based approaches (see e.g Jambulapati et al. [2024]).*

Equation (14) holds for all coordinates $k \in [d]$ and we show in the Appendix (see Remark 7) that for the choice of B and N in Theorem 2, the higher order terms are indeed $o\left(\frac{1}{\sqrt{m}}\right)$. Moreover, for any coordinate k for which equations (12) and (13) hold, the lower order terms of equation (14) are $O(\mathbb{V}_{kk}/\sqrt{m})$. This implies an $O(1/\sqrt{\log n})$-multiplicative guarantee on the error of $\hat{\gamma}_k$ like equation (15).

4 PROOF TECHNIQUES

Let $v_{\mathsf{oja}} \sim \mathsf{Oja}(\mathcal{D}_{\ell,j}, \eta_n, u_0)$ for uniform unit vector u_0 and $\tilde{v} \sim \mathsf{Oja}(\mathcal{D}_{\ell,j}, \eta_N, u_0)$. To estimator the uncertainty of the estimator, the residual vector $\tilde{r}_{\mathsf{oja}} := v_{\mathsf{oja}} - (\tilde{v}^\top v_{\mathsf{oja}})\tilde{v}$ is decomposed as the sum of five terms, as stated in Lemma 2. Proposition A.1 in Lunde et al. [2021] shows that B_n, defined in (5), can be written as

$$B_n = \sum_{k=0}^{n} T_{n,k}, \tag{16}$$

where

$$T_{n,k} := \sum_{S\subseteq[n],|S|=k} \prod_{i=1}^{n} M_{S,n+1-i}, \text{ and} \tag{17}$$

$$M_{S,i} := \begin{cases} \eta_n\left(X_i X_i^\top - \Sigma\right) & \text{if } i \in S, \\ I + \eta_n \Sigma & \text{if } i \notin S. \end{cases} \tag{18}$$

The term $T_{n,1}$ is called the Hájek projection of the random variable B_n on the random variables $X_1, \ldots, X_n$. $T_{n,1}$ is the best approximation to B_n among the estimators that can be written as the sum of independent random vectors and satisfy certain integrability conditions. Moreover,

- $T_{n,k}$ and $T_{n,j}$ are uncorrelated for all $k \neq j$, and
- the summands in $T_{n,k}$ are also pairwise uncorrelated.

We exploit this structure of the Hoeffding decomposition to decompose the residual vector $\tilde{r}_{\mathsf{oja}}$.

Lemma 2. *[Error Decomposition of v_{oja}] Let $v_{\mathsf{oja}}, \tilde{v}$ be defined as in (6) and (10) respectively. Then,*

$$v_{\mathsf{oja}} - (\tilde{v}^\top v_{\mathsf{oja}})\tilde{v} = \Psi_{n,0} + \Psi_{n,1} + \Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}, \tag{19}$$

where

$$\begin{aligned}
\Psi_{n,0} &:= (v_1^\top v_{\mathsf{oja}})v_1 - (\tilde{v}^\top v_{\mathsf{oja}})\tilde{v}, \\
\Psi_{n,1} &:= \frac{V_\perp V_\perp^\top T_{n,1} v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n}, \\
\Psi_{n,2} &:= \frac{V_\perp V_\perp^\top (\sum_{k\geq 2} T_{n,k}) v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n}, \\
\Psi_{n,3} &:= V_\perp V_\perp^\top B_n u_0 \left(\frac{1}{\|B_n u_0\|_2} - \frac{1}{|v_1^\top u_0|(1+\eta\lambda_1)^n}\right), \\
\Psi_{n,4} &:= \frac{V_\perp V_\perp^\top B_n V_\perp V_\perp^\top u_0}{|v_1^\top u_0|(1+\eta\lambda_1)^n}.
\end{aligned} \tag{20}$$

We bound the variance of each of these terms separately. The dominating term $\Psi_{n,1}$ corresponding to the Hájek projection $T_{n,1}$ has the largest variance. Recall from Lemma 1 that

$$\left|\mathbb{E}\left[\left(e_k^\top \Psi_{n,1}\right)^2\right] - \eta_n\lambda_1\mathbb{V}_{kk}\right| \leq \tilde{O}\left(\frac{1}{n^2}\right).$$

A finer analysis is needed for this term than the other residual terms in (20). To do this, we bound the variance of $\left(e_k^\top \Psi_{n,1}\right)^2$. Lemma 3 shows that $\sqrt{\mathrm{Var}((e_k^\top \Psi_{n,1})^2)}$ is a constant factor within $\mathbb{E}[(e_k^\top \Psi_{n,1})^2] = \tilde{O}(1/n)$ up to an additive error term $\tilde{O}(1/n^{3/2})$ which depends polynomially on model parameters.

Lemma 3 (Variance of the Hájek projection)**.** *Let $\Psi_{n,1}$ be defined as in Lemma 2. Then,*

$$\sqrt{\mathrm{Var}\left((e_k^\top \Psi_{n,1})^2\right)} \leq \sqrt{2}\mathbb{E}\left[\left(e_k^\top \Psi_{n,1}\right)^2\right] + \tilde{O}\left(\frac{1}{n^{3/2}}\right).$$

The three terms $\Psi_{n,2}, \Psi_{n,3}$, and $\Psi_{n,4}$ are lower order terms.

Lemma 4 (Bound on lower order terms)**.** *Let $\Psi_{n,2}$, $\Psi_{n,3}$, and $\Psi_{n,4}$ be defined as in Lemma 2. Then,*

$$\mathbb{E}\left[\left(e_k^\top \Psi_{n,2}\right)^2 + \left(e_k^\top \Psi_{n,3}\right)^2 + \left(e_k^\top \Psi_{n,4}\right)^2\right] = \tilde{O}\left(\frac{1}{n^2}\right).$$

The bound on the error term $e_k^\top \Psi_{n,2}$ stems from a more general analysis of the terms $T_{n,k}$ in the Hoeffding decomposition of B_n. Lemma 5 is shown by exploiting the Martingale structure of $T_{n,k}$ and using norm inequalities [Huang et al., 2022] to compare the operator norm with the $|||.|||_{p,q}$ norm.

Lemma 5. *Let $T_{n,k}$ be as defined in equation (17). Let for any $2 \leq q \leq 4\log d$, $\mathcal{M}_q$ be defined such that $\mathbb{E}\left[\|A_i - \Sigma\|^q\right]^{1/q} \leq \mathcal{M}_q$ and $\eta_n \mathcal{M}_q \sqrt{n \log d} \lesssim 1$. Then, for any $j \in [n]$, $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\left\| \sum_{k \geq j} T_{n,k} \right\| \leq \frac{3(1+\eta_n \lambda_1)^n \left(\eta_n \mathcal{M}_q \sqrt{4n \log d}\right)^j}{\delta^{\frac{1}{4\log d}}}$$

Proof sketch. Let $\mathcal{S}_{n,k}$ be the set of subsets of $[n]$ of size k.

$$T_{n,k} = (I + \eta_n \Sigma) T_{n-1,k} + \eta_n (A_n - \Sigma) T_{n-1,k-1}.$$

Proposition 4.3. of Huang et al. [2022] implies

$$\begin{aligned} \|\!|T_{n,k}|\!\|_{p,q}^2 &\leq \|\!|(I + \eta_n \Sigma) T_{n-1,k}|\!\|_{p,q}^2 \\ &\quad + (p-1) \|\!|\eta_n (A_n - \Sigma) T_{n-1,k-1}|\!\|_{p,q}^2. \end{aligned}$$

as long as $\mathbb{E}\left[\eta_n (A_n - \Sigma) T_{n-1,k-1} | (I + \eta_n \Sigma) T_{n-1,k}\right] = 0$, which is true due to $A_1, A_2, \ldots, A_n$ being mutually independent. Solving the recurrence shows the bound. □

The term $\Psi_{n,0}$ arises in the decomposition (20) because we use $\tilde{v}$ as a proxy to v_1 in Algorithm 1.

Lemma 6 (Variance of Approximating v_1)**.** *Let $\Psi_{n,0}$ be defined as in Lemma 2. Then, $\mathbb{E}\left[\left(e_k^\top \Psi_{n,0}\right)^2\right] = \tilde{O}\left(\frac{1}{N}\right)$, where $\tilde{v}$ (Eq 10) uses N samples.*

Theorem 2 follows by combining all these bounds. See Appendix B.2.6 for a complete argument.

5 EXPERIMENTS

In this section, we provide experiments on synthetic and real-world data to validate our theory. For all experiments, we estimate variance of the entries of r_{oja} (see Eq 7) by scaling the output of Algorithm 1 by $\eta_B(\lambda_1 - \lambda_2)$.

5.1 SYNTHETIC DATA EXPERIMENTS

We provide numerical experiments to compare Algorithm 1 (OjaVarEst) with the multiplier bootstrap based algorithm proposed in Lunde et al. [2021]. As discussed in Section 3.1, given a dataset $\mathcal{D}_n := \{X_i\}_{i \in [n]}$, we choose $\tilde{v}$ for OjaVarEst as $\tilde{v} := \mathsf{Oja}\left(\mathcal{D}_n, \eta_n, z / \|z\|_2\right)$ for $z \sim \mathcal{N}(0, I)$ and set $m_1 = 3$, $m_2 = \log(n)$, $N = n$. Given a variance estimate, $\hat{\sigma}^2_{\mathsf{OjaVarEst}}$, we construct a $(1-\alpha)$-confidence interval as $\tilde{v} \pm z_{\frac{\alpha}{2}} \hat{\sigma}_{\mathsf{OjaVarEst}}$.

For the bootstrap algorithm, using Algorithm 1 in the aforementioned paper, we use b bootstrap samples to generate estimates $v^{*(1)}, \cdots, v^{*(b)}$ and measure the empirical variance by computing the average squared residual with $\tilde{v}$. Again, given a variance estimate, $\hat{\sigma}^2_{\mathsf{BootstrapOja}}$, we construct a $(1-\alpha)$-confidence interval as $\tilde{v} \pm z_{\frac{\alpha}{2}} \hat{\sigma}_{\mathsf{BootstrapOja}}$.

We also use the data generation process proposed in Lunde et al. [2021] for our experiments. Specifically, we begin by generating independent samples $Z_{ij} \sim \mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$ for indices $i \in [n]$ and $j \in [d]$. Next, we define a positive semidefinite matrix K with entries $K_{ij} = \exp(-c\,|i-j|)$ using the constant $c = 0.01$. With this matrix, we construct a covariance matrix Σ via $\Sigma_{ij} = K(i,j)\,\sigma_i\,\sigma_j$, where the scaling factors are specified by $\sigma_i = 5\,i^{-\beta}$ for $\beta \in \{0.2, 1\}$. We finally transform the samples as $X_i = \Sigma^{1/2} Z_i$.

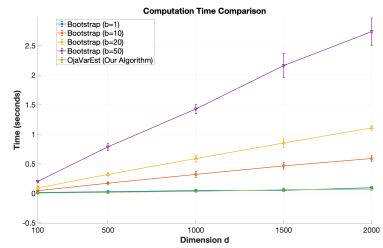


Figure 1: Time taken by the bootstrap methods and the OjaVarEst algorithm. Experiments verify that our proposed algorithm is as fast as bootstrap with $b = 1$.

The first experiment (see Figure 1) compares the computational performance of OjaVarEst with bootstrap to measure variance, varying the number of bootstrap samples, b, and recording performance for different values of d with a fixed $n = 5000$ and $\beta = 1$. We note that the performance of our algorithm is computationally at par with bootstrap when using only 1 bootstrap sample, and is substantially better if the number of bootstrap samples increase. This is to be expected since for our algorithm, only two passes over the entire dataset suffice, whereas for bootstrap, b bootstrap vectors are required to be maintained, which slows computation by a factor of b. Furthermore, it also requires b times as much space to maintain b different iterates, which may be costly in context of training large models.

The next experiment (Table 1) compares the quality of the variance estimates of our algorithm, $\hat{\sigma}^2_{\mathsf{OjaVarEst}}$ with that of bootstrap $\hat{\sigma}^2_{\mathsf{BootstrapOja}}$ for different number of bootstrap samples, b, and distributions, β. We record the average coverage rate, which is the proportion of times the confidence interval provided by the algorithm contains the coordinate of the true eigenvector, for a target coverage probability of 95% for the first two coordinates of the eigenvector. OjaVarEst performs similarly to Bootstrap with $b = 20$. However, as shown in Figure 1, the bootstrap method is 20 times slower. The time taken by bootstrap with $b = 1$ is similar to OjaVarEst but has a significantly worse average coverage rate.

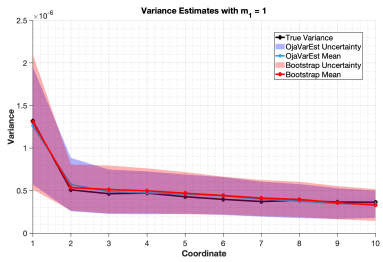
Our final experiment compares the Algorithm 1 with $m_1 = 3$ to using just the mean ($m_1 = 1$). Even with the choice $m_1 = 3$, the uncertainty in variance estimation is reduced.

5.2 REAL-WORLD DATA EXPERIMENTS

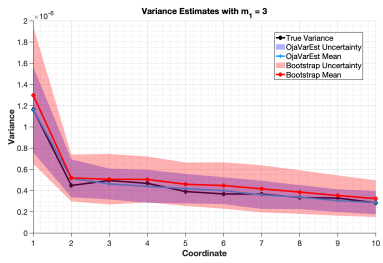
We provide experiments on two real-world datasets in this section. For each dataset, we show the 95% confidence

	Dist. 1 ($\beta = 1$), Coordinate 1				**Dist. 1** ($\beta = 1$), Coordinate 2			
(n, d)	OjaVarEst	BS ($b = 1$)	BS ($b = 10$)	BS ($b = 20$)	OjaVarEst	BS ($b = 1$)	BS ($b = 10$)	BS ($b = 20$)
2e3, 2e3	96.50%	65.00%	93.00%	95.00%	94.00%	69.50%	91.00%	91.50%
5e3, 2e3	95.50%	73.00%	91.50%	94.00%	95.50%	73.00%	89.00%	92.00%
1e4, 2e3	96.00%	69.00%	93.50%	94.50%	96.00%	71.50%	93.50%	96.00%
	Dist. 2 ($\beta = 0.02$), Coordinate 1				**Dist. 2** ($\beta = 2$), Coordinate 2			
(n, d)	OjaVarEst	BS ($b = 1$)	BS ($b = 10$)	BS ($b = 20$)	OjaVarEst	BS ($b = 1$)	BS ($b = 10$)	BS ($b = 20$)
2e3, 2e3	94.50%	74.00%	87.00%	93.50%	94.00%	75.00%	86.50%	92.00%
5e3, 2e3	96.00%	71.00%	87.50%	92.00%	96.50%	72.50%	87.00%	93.00%
1e4, 2e3	94.00%	65.00%	95.00%	94.00%	94.50%	66.50%	94.50%	93.50%

Table 1: Coverage statistics for our algorithm, OjaVarEst, and the Bootstrap(BS) estimator, with varying bootstrap samples ($b = 1, 10, 20$), data distributions ($\beta = 1, 0.02$) and sample sizes ($n = 2000, 5000, 10000$) with a fixed dimension $d = 2000$.

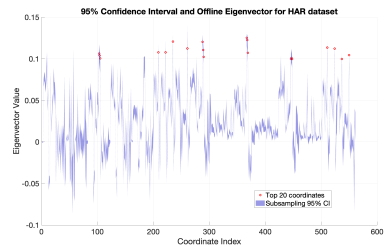


(a) Mean (with $m_1 = 1$)

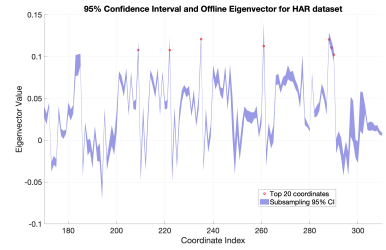


(b) Median (with $m_1 = 3$)

Figure 2: Comparison of Median and Mean in Algorithm 1 for $n = 5000$, $d = 2000$, $\beta = 1$, $b = 10$.



(a)



(b)

Figure 3: Uncertainty Estimation for HAR dataset ($n = 7352, d = 561$). The sin2 error of Oja's algorithm is equal to 0.057 for this dataset. (a) plot of the eigenvector with 95% confidence interval for all coordinates and (b) the same plot zoomed in on indices 170-310 for exposition.

intervals and plot the top 20 coordinates of the true offline eigenvector (red dot), used as a proxy for the ground truth.

Time series+missing data: The Human Activity Recognition (HAR) Dataset [Anguita et al., 2013] contains smartphone sensor readings from 30 subjects performing daily activities (walking, sitting, standing, etc.). Each data instance is a 2.56-second window of inertial sensor signals represented as a feature vector. Here, $n = 7352$ and $d = 561$. For each datum, we also replace 10% of features randomly by zero to simulate missing data. Even in this setting, which we do not analyze theoretically, most of the top 20 coordinates of the offline eigenvector are inside the 95% CI returned by our algorithm (see Figure 3).

Image data: We use the MNIST dataset [LeCun et al., 1998] of grayscale images of handwritten digits (0 through 9). Here, $n = 60,000, d = 784$, with each image normalized to a 28×28 pixel resolution. We see (Figure 4) that for the classes where Oja's algorithm converges (small $\sin^2$ error in Table 2), most of the top 20 coordinates are inside their confidence intervals (CIs). Notable exceptions are classes 3 and 4, where several of the top 20 coordinates are not contained inside the corresponding CIs. This is expected because our theory is applicable when Oja's algorithm converges.

Class	0	1	2	3	4	5	6	7	8	9
$\sin^2$ error	0.12	0.07	0.18	0.32	0.53	0.18	0.08	0.09	0.20	0.17

Table 2: $\sin^2$ of the angle between the offline eigenvector and the subsampling eigenvector output by our algorithm, computed separately after filtering the MNIST data for each class.

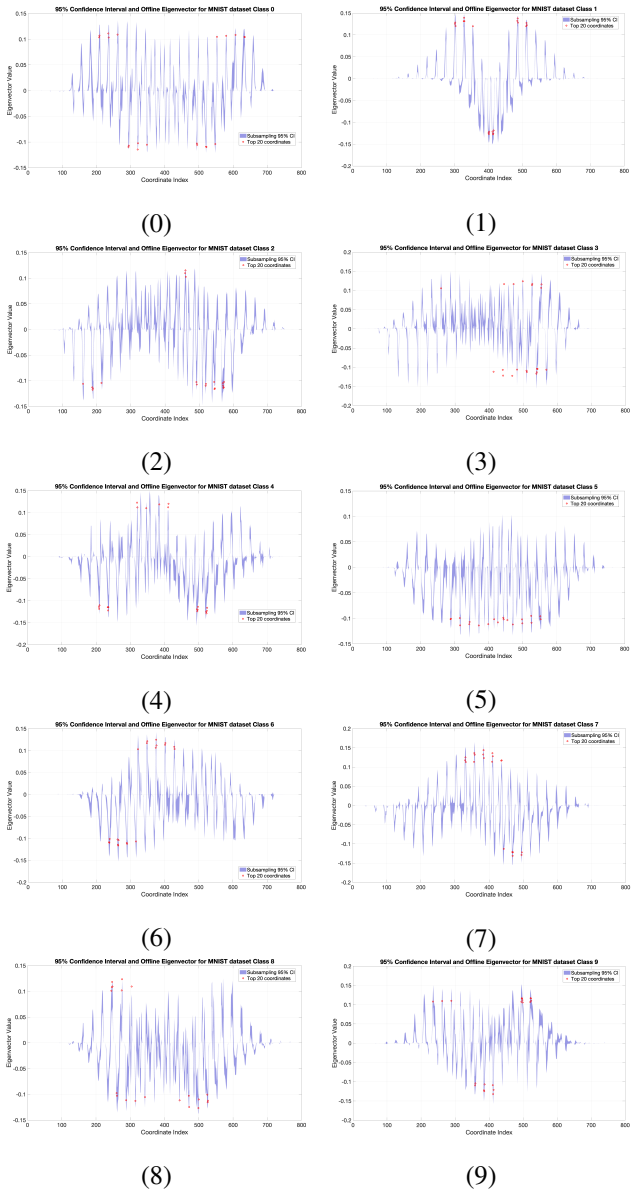


Figure 4: Uncertainty Estimation for MNIST dataset. The $\sin^2$ error of Oja's algorithm for each class is provided in Table 2.

6 CONCLUSION

In this work, we develop a novel statistical inference framework for streaming PCA using Oja's algorithm. We derive finite-sample and high-probability deviation bounds for the coordinates of the estimated eigenvector, establish a Bernstein-type concentration bound on the residual of the Oja vector, establish a Central Limit Theorem for suitable subsets of entries, and devise an efficient subsampling-based variance estimation algorithm. By leveraging the structure of the Oja updates, we provide entrywise confidence intervals, bypassing expensive resampling techniques such as bootstrapping. Our theoretical results are supported by extensive numerical experiments, indicating that our proposed estimator achieves accuracy similar to the multiplier bootstrap method while requiring significantly less time.

We believe that our subsampling algorithm can be adapted to any SGD problem where the covariance matrix of the estimator $\hat{\theta}_n$ scales as c_n times some scale-free matrix $\mathbb{V}$, where c_n is known. This structure aligns with subsampling and m-out-of-n bootstrap methods, where the variance estimated from a subsample of size m is scaled by m/n to approximate the variance of the full sample estimator. Our findings also highlight the potential for improved uncertainty quantification techniques in streaming non-convex optimization problems beyond PCA, since Oja-type updates can be found in many important non-convex optimization algorithms such as matrix sensing, matrix completion, and subspace estimation. Further directions include deflation-based methods to apply our method to variance estimation for top k eigenvectors.

ACKNOWLEDGMENTS

We gratefully acknowledge NSF grants 2217069, 2019844, and DMS 2109155. We are thankful to Soumendu Sundar Mukherjee and Arun Kuchibhotla for helpful discussions.

REFERENCES

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3): 1452–1474, 2020.

Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of PCA and spectral clustering. *The Annals of Statistics*, 50(4):2359 – 2385, 2022. doi: 10.1214/22-AOS2196. URL https://doi.org/10.1214/22-AOS2196.

Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on*

Foundations of Computer Science (FOCS), pages 487–492. IEEE, 2017.

Donald WK Andrews and Patrik Guggenberger. Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, 26(2):426–468, 2010.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Juan-Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 437–442, 2013.

Laura Balzano. On the equivalence of oja's algorithm and grouse. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7014–7030. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/balzano22a.html.

Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *2010 48th Annual allerton conference on communication, control, and computing*, pages 704–711. IEEE, 2010.

J.A. Bather. Stochastic approximation: a generalisation of the robbins-monro procedure. In *Proceedings of the Fourth Prague Symposium on Asymptotic Statistics (Prague, 1988)*, pages 13–27. Charles University, Prague, 1989.

Patrice Bertail, Dimitris N Politis, and Joseph P Romano. On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association*, 94(446):569–579, 1999.

Peter J Bickel and Anat Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, pages 967–985, 2008.

Peter J. Bickel, Friedrich Götze, and Willem van Zwet. Resampling fewer than n obersvations: Gains, losses, and remedies for losses. *Statistica Sinica*, pages 1–31, 1997.

Joshua Cape, Minh Tang, and Carey E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405 – 2439, 2019a. doi: 10.1214/18-AOS1752. URL https://doi.org/10.1214/18-AOS1752.

Joshua Cape, Minh Tang, and Carey E Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250, 2019b.

Selina Carter and Arun K Kuchibhotla. Statistical inference for online algorithms. *arXiv preprint arXiv:2505.17300*, 2025.

Kamalika Chaudhuri, Po-Ling Loh, Shourya Pandey, and Purnamrita Sarkar. On differentially private u statistics. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jerry Chee, Hwanwoo Kim, and Panos Toulis. "plus/minus the learning rate": Easy and scalable statistical inference with sgd. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2285–2309. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/chee23a.html.

Minshuo Chen, Lin Yang, Mengdi Wang, and Tuo Zhao. Dimensionality reduction for stationary time series via stochastic nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 02 2020. doi: 10.1214/18-AOS1801. URL https://doi.org/10.1214/18-AOS1801.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309 – 2352, 2017a. doi: 10.1214/16-AOP1113. URL https://doi.org/10.1214/16-AOP1113.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Detailed proof of nazarov's inequality, 2017b. URL https://arxiv.org/abs/1711.10696.

Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. Scalable dp-sgd: Shuffling vs. poisson subsampling. *Advances in Neural Information Processing Systems*, 37:70026–70047, 2024.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic learning theory*, pages 321–358. PMLR, 2018.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018. URL http://jmlr.org/papers/v19/17-370.html.

Peter Hall. *The Bootstrap and Edgeworth Expansion*. Springer, 1992.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

Amelia Henriksen and Rachel Ward. Adaoja: Adaptive learning rates for streaming pca, 2019. URL https://arxiv.org/abs/1905.12115.

De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja's algorithm, beyond rank-one updates. *CoRR*, abs/2102.03646, 2021. URL https://arxiv.org/abs/2102.03646.

De Huang, Jonathan Niles-Weed, Joel A Tropp, and Rachel Ward. Matrix concentration for products. *Foundations of Computational Mathematics*, 22(6):1767–1799, 2022.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 665–674, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488693. URL https://doi.org/10.1145/2488608.2488693.

Prateek Jain, Chi Jin, Sham Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming PCA: Matching matrix bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *Proceedings of The 29th Conference on Learning Theory (COLT)*, June 2016.

Arun Jambulapati, Syamantak Kumar, Jerry Li, Shourya Pandey, Ankit Pensia, and Kevin Tian. Black-box k-to-1-pca reductions: Theory and applications. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2564–2607. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/jambulapati24a.html.

Tarun Kathuria, Satyaki Mukherjee, and Nikhil Srivastava. On concentration inequalities for random matrix products. *arXiv preprint arXiv:2003.06319*, 2020.

Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56:2980 – 2998, 07 2010. doi: 10.1109/TIT.2010.2046205.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

Syamantak Kumar and Purnamrita Sarkar. Streaming pca for markovian data. *Advances in Neural Information Processing Systems*, 36, 2024a.

Syamantak Kumar and Purnamrita Sarkar. Oja's algorithm for streaming sparse pca. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

Yann LeCun, León Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7381–7389, 2022.

Anna Levina and Viola Priesemann. Subsampling scaling. *Nature communications*, 8(1):15140, 2017.

Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using sgd. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, pages 3571–3578. AAAI Press, 2018. URL http://dblp.uni-trier.de/db/conf/aaai/aaai2018.html#LiLKC18.

Lennart Ljung, Georg Pflug, and Harro Walk. *Stochastic approximation and optimization of random systems*. Birkhauser Verlag, CHE, 1992. ISBN 3764327332.

Robert Lunde, Purnamrita Sarkar, and Rachel Ward. Bootstrapping the error of oja's algorithm. *Advances in neural information processing systems*, 34:6240–6252, 2021.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating mixed memberships with sharp eigenvector deviations. *Journal of the American Statistical Association*, 116(536):1928–1940, 2021. doi: 10.1080/01621459.2020.1751645. URL https://doi.org/10.1080/01621459.2020.1751645.

Jean-Marie Monnez. Stochastic approximation of eigenvectors and eigenvalues of the q-symmetric expectation of a random matrix. *Communications in Statistics-Theory and Methods*, pages 1–15, 2022.

Fedor Nazarov. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a gaussian measure. In *Geometric Aspects of Functional Analysis (GAFA) 2001-2002*, volume 1807 of *Lecture Notes in Mathematics*, pages 169–187. Springer, 2003.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15: 267–273, 1982.

Erkki Oja and Juha 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.

Karl Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

Dimitris N Politis. Scalable subsampling: computation, aggregation and inference. *Biometrika*, 111 (1):347–354, 03 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad021. URL https://doi.org/10.1093/biomet/asad021.

Dimitris N. Politis, Joseph P. Romano, and Michael. Wolf. *Subsampling / by Dimitris N. Politis, Joseph P. Romano, Michael Wolf.* Springer Series in Statistics. Springer New York, New York, NY, 1st ed. 1999. edition, 1999. ISBN 1-4612-1554-4.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, July 1992. ISSN 0363-0129. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046.

Eric Price and Zhiyang Xun. Spectral guarantees for adversarial streaming pca. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1768–1785, 2024. doi: 10.1109/FOCS61266.2024.00108.

Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544): 2901–2914, 2023.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Anat Sakov. *Using the m out of n bootstrap in hypothesis testing*. University of California, Berkeley, 1998.

G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Weijie J. Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent, 2018.

Aad van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and GittaEditors Kutyniok, editors, *Compressed Sensing: Theory and Practice*, pages 210–268. Cambridge University Press, 2012. ISBN 9780511794308. doi: 10.1017/CBO9780511794308.006.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

Puyudi Yang, Cho-Jui Hsieh, and Jane-Ling Wang. History pca: A new algorithm for streaming pca. *arXiv preprint arXiv:1802.05447*, 2018.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. An analysis of constant step size sgd in the non-convex regime: asymptotic normality and bias. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Yanjie Zhong, Todd Kuffner, and Soumendra Lahiri. Online bootstrap inference with nonconvex stochastic gradient descent estimator. *arXiv preprint arXiv:2306.02205*, 2023.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541): 393–404, 2023.

Wanrong Zhu, Zhipeng Lou, Ziyang Wei, and Wei Biao Wu. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

Eric R Ziegel. Principal component analysis. *Technometrics*, 45(3):276–277, 2003.

The Appendix is organized as follows:

1. Section A provides some useful results used in subsequent analyses
2. Section C provides high probability Entrywise Error Bounds on the entries of v_{oja}
3. Section B has the Bias and Concentration calculation of our estimator designed in Algorithm 1
4. Section D provides a Central Limit Theorem for the entries of the Oja vector, v_{oja}, which ties the results developed in Section B to provide confidence intervals

A UTILITY RESULTS

Lemma 7. *For any integer $n \geq 2$, real $\epsilon \in (0,1)$, and reals $\{a_i\}_{i\in[n]}$,*

$$(1-\epsilon)a_1^2 - \frac{n-1}{\epsilon}\sum_{i=2}^{n} a_i^2 \leq \Big(\sum_{i=1}^{n} a_i\Big)^2 \leq (1+\epsilon)a_1^2 + \frac{2(n-1)}{\epsilon}\sum_{i=2}^{n} a_i^2.$$

Proof. We begin by writing

$$\Big(a_1 + \sum_{i=2}^{n} a_i\Big)^2 = a_1^2 + 2a_1\Big(\sum_{i=2}^{n} a_i\Big) + \Big(\sum_{i=2}^{n} a_i\Big)^2. \tag{21}$$

By Cauchy-Schwarz inequality,

$$0 \leq \Big(\sum_{i=2}^{n} a_i\Big)^2 \leq (n-1)\sum_{i=2}^{n} a_i^2. \tag{22}$$

The cross-term can be bounded using the inequality

$$-\epsilon x^2 - \frac{1}{\epsilon}\,y^2 \leq 2xy \leq \epsilon x^2 + \frac{1}{\epsilon}\,y^2$$

with $x = a_1$ and $y = \sum_{i=2}^{n} a_i$ to get

$$2a_1\Big(\sum_{i=2}^{n} a_i\Big) \geq -\epsilon a_1^2 - \frac{1}{\epsilon}\Big(\sum_{i=2}^{n} a_i\Big)^2 \geq -\epsilon a_1^2 - \frac{n-1}{\epsilon}\sum_{i=2}^{n} a_i^2,$$

and

$$2a_1\Big(\sum_{i=2}^{n} a_i\Big) \leq \epsilon a_1^2 + \frac{1}{\epsilon}\Big(\sum_{i=2}^{n} a_i\Big)^2 \leq \epsilon a_1^2 + \frac{n-1}{\epsilon}\sum_{i=2}^{n} a_i^2.$$

The proof follows by using the above inequalities in (21) followed by another application of (22). □

Lemma 8. *Let $\mathbb{V}$ be the asymptotic variance matrix defined in Lemma 1, and let v_{oja} be the Oja vector as defined in (6). If the coordinate-wise bound*

$$\left|e_i^\top\left(v_{\mathsf{oja}} - \left(v_1^\top v_{\mathsf{oja}}\right)v_1\right)\right| \lesssim C_{d,n}\sqrt{\frac{\mathbb{V}_{kk}}{n}}$$

holds for every $i \in [d]$, where $C_{d,n}^2$ hides logarithmic factors in d, n, then

$$\sin^2\left(v_{\mathsf{oja}}, v_1\right) = \sum_{i\in[d]}\left(e_i^\top\left(v_{\mathsf{oja}} - \left(v_1^\top v_{\mathsf{oja}}\right)v_1\right)\right)^2 \lesssim C_{d,n}^2\frac{\mathcal{V}}{\left(\lambda_1-\lambda_2\right)^2 n},$$

where $\mathcal{V}$ is the matrix variance statistic defined in Assumption 1.

Proof. By the definitions of $\mathbb{V}$ and R_0 as in Lemma 1,

$$\begin{aligned}
\sum_{i\in[d]} \left(e_i^\top \left(v_{\mathsf{oja}} - \left(v_1^\top v_{\mathsf{oja}}\right) v_1\right)\right)^2 \lesssim C_{d,n}^2 \frac{\mathsf{Tr}\left(\mathbb{V}\right)}{n} &\leq \left(\frac{C_{d,n}^2}{\lambda_1 - \lambda_2}\right) \frac{\mathsf{Tr}\left(R_0\right)}{n} = \left(\frac{C_{d,n}^2}{\lambda_1 - \lambda_2}\right) \frac{1}{n} \sum_{2\leq k\leq d} \frac{\widetilde{M}_{kk}}{2\left(\lambda_1 - \lambda_k\right)} \\
&\leq \frac{C_{d,n}^2}{\left(\lambda_1 - \lambda_2\right)^2} \frac{\mathsf{Tr}\left(\mathbb{E}\left[V_\perp \left(A - \Sigma\right) v_1 v_1^\top \left(A - \Sigma\right) V_\perp^\top\right]\right)}{n} \\
&= \frac{C_{d,n}^2}{\left(\lambda_1 - \lambda_2\right)^2} \frac{\mathbb{E}\left[\mathsf{Tr}\left(V_\perp \left(A - \Sigma\right) v_1 v_1^\top \left(A - \Sigma\right) V_\perp^\top\right)\right]}{n} \\
&= \frac{C_{d,n}^2}{\left(\lambda_1 - \lambda_2\right)^2} \frac{v_1^\top \mathbb{E}\left[\left(A - \Sigma\right) V_\perp V_\perp^\top \left(A - \Sigma\right)\right] v_1}{n} \\
&\leq \frac{C_{d,n}^2}{\left(\lambda_1 - \lambda_2\right)^2} \frac{v_1^\top \mathbb{E}\left[\left(A - \Sigma\right)^2\right] v_1}{n} \leq C_{d,n}^2 \frac{\mathcal{V}}{\left(\lambda_1 - \lambda_2\right)^2 n}.
\end{aligned}$$

□

Lemma 9 (Choice of learning rate)**.** *Let* $\eta_n := \frac{\alpha \log(n)}{n(\lambda_1 - \lambda_2)}$ *for* $\alpha > 1$*. Then, under Assumptions 1 and 2*

1. $nd \exp\left(-\eta_n n \left(\lambda_1 - \lambda_2\right)\right) = o\left(1\right)$.
2. $\max\left\{\eta_n, \frac{\log(d)}{\lambda_1 - \lambda_2}\right\} \frac{\mathcal{M}_2^4}{\lambda_1 - \lambda_2} \eta_n^2 = o\left(1\right)$.
3. $n\eta_n^2 (2\lambda_1^2 + \mathcal{M}_2^2) \leq 1$

Proof. The above conditions on η_n imply Corollary 1 in Lunde et al. [2021]. Let's start with the first condition. We have

$$nd \exp\left(-\eta_n n \left(\lambda_1 - \lambda_2\right)\right) \leq nd \exp\left(-\alpha \log\left(n\right)\right) = \frac{d}{n^{\alpha - 1}} = o(1), \text{ using the bound on } d$$

For the second condition, we first note that for $n \geq \alpha \log\left(n\right)$ provided by Assumption 2,

$$\eta_n \leq \frac{\log\left(d\right)}{\left(\lambda_1 - \lambda_2\right)}$$

Now for the second condition, we require,

$$\frac{\alpha^2 \mathcal{M}_2^4 \log^2\left(n\right) \log\left(d\right)}{n^2 \left(\lambda_1 - \lambda_2\right)^4} = o(1)$$

which is again ensured by the condition on n in Assumption 2. □

Lemma 10. *Let* t *be a positive integer,* $\delta \in (0, 1)$*, and let* I *be an interval in* $\mathcal{R}$*. Suppose* $a_1, a_2, \ldots, a_t$ *are independent random variables such that* $P\left(a_i \in I\right) \geq 3/4$*. Then, for* $t \geq 8 \log\left(1/\delta\right)$*,*

$$P\left(\mathsf{Median}\left(\left\{a_i\right\}_{i\in[t]}\right) \in I\right) \geq 1 - \delta.$$

Proof. Since I is an interval, the median does lies in I if at least half the a_i are in I. Let b_i be the indicator that $a_i \notin I$, and let $B = \sum_{i\in[t]} b_i$. Then, $b_1, b_2, \ldots, b_t$ are independent Bernoulli random variables each with mean at most $1/4$. By Hoeffding's inequality,

$$P\left(\mathsf{Median}\left(\left\{a_i\right\}_{i\in[t]}\right) \notin I\right) \leq P\left(B > t/2\right) \leq \exp\left(-2(t/2 - \mathbb{E}\left[B\right])^2/t\right) \leq \exp\left(-t/8\right) \leq \delta.$$

□

B ESTIMATOR CONCENTRATION

Lemma 2. *[Error Decomposition of v_{oja}] Let $v_{\mathsf{oja}}, \tilde{v}$ be defined as in (6) and (10) respectively. Then,*

$$v_{\mathsf{oja}} - (\tilde{v}^\top v_{\mathsf{oja}})\tilde{v} = \Psi_{n,0} + \Psi_{n,1} + \Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}, \tag{19}$$

where

$$\begin{aligned}
\Psi_{n,0} &:= (v_1^\top v_{\mathsf{oja}})v_1 - (\tilde{v}^\top v_{\mathsf{oja}})\tilde{v}, \\
\Psi_{n,1} &:= \frac{V_\perp V_\perp^\top T_{n,1} v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n}, \\
\Psi_{n,2} &:= \frac{V_\perp V_\perp^\top (\sum_{k\geq 2} T_{n,k}) v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n}, \\
\Psi_{n,3} &:= V_\perp V_\perp^\top B_n u_0 \left(\frac{1}{\|B_n u_0\|_2} - \frac{1}{\left|v_1^\top u_0\right| (1+\eta\lambda_1)^n} \right), \\
\Psi_{n,4} &:= \frac{V_\perp V_\perp^\top B_n V_\perp V_\perp^\top u_0}{\left|v_1^\top u_0\right| (1+\eta\lambda_1)^n}.
\end{aligned} \tag{20}$$

Proof. We have,

$$\begin{aligned}
v_{\mathsf{oja}} &= (v_1^\top v_{\mathsf{oja}})v_1 + V_\perp V_\perp^\top v_{\mathsf{oja}} \\
&= (v_1^\top v_{\mathsf{oja}})v_1 + \frac{V_\perp V_\perp^\top B_n u_0}{\|B_n u_0\|_2} \\
&= (v_1^\top v_{\mathsf{oja}})v_1 + \frac{V_\perp V_\perp^\top B_n u_0}{c_n} + \Psi_{n,3} \\
&= (v_1^\top v_{\mathsf{oja}})v_1 + \frac{V_\perp V_\perp^\top B_n v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n} + \Psi_{n,3} + \Psi_{n,4} \\
&= (v_1^\top v_{\mathsf{oja}})v_1 + \frac{V_\perp V_\perp^\top (B_n - \mathbb{E}[B_n]) v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n} + \Psi_{n,3} + \Psi_{n,4} \\
&= (v_1^\top v_{\mathsf{oja}})v_1 + \frac{V_\perp V_\perp^\top (\sum_{k\geq 1} T_{n,k}) v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n} + \Psi_{n,3} + \Psi_{n,4}, \text{ using Theorem A.1 Lunde et al. [2021]} \\
&= (\tilde{v}^\top v_{\mathsf{oja}})\tilde{v} + \Psi_{n,0} + \Psi_{n,1} + \Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}.
\end{aligned}$$

□

Lemma 11. *Let $\Psi_{n,1}$ be as defined in Lemma 2. Then,*

$$\Psi_{n,1} := \eta_n Y_n, \textit{ for } Y_n := \sum_{j=1}^n X_j^n \textit{ and } X_j^n := \frac{\mathsf{sign}\left(v_1^\top u_0\right)}{1+\eta_n\lambda_1} V_\perp \Lambda_\perp^{n-j} V_\perp^\top \left(A_j - \Sigma\right) v_1$$

where $\Lambda_\perp \in \mathbb{R}^{(d-1)\times(d-1)}$ is a diagonal matrix with entries $\Lambda_\perp(i,i) = \frac{1+\eta_n\lambda_{i+1}}{1+\eta_n\lambda_1}$.

Let $\{A_i\}_{i\in[n]}$ be symmetric independent matrices satisfying $\mathbb{E}[A_i] = \Sigma$, $\left\|\mathbb{E}\left[(A_i-\Sigma)^2\right]\right\|_2 \leq \mathcal{V}$ and $\|A_i - \Sigma\|_2 \leq \mathcal{M}$. Define,

$$\forall j \in [n], \quad X_j^n := V_\perp \Lambda_\perp^{n-j} V_\perp^\top \left(A_j - \Sigma\right) v_1, \text{ and } Y_n := \sum_{j\in[n]} X_j^n$$

B.1 ESTIMATOR BIAS

Proof of Lemma 1. Using the definitions of Y_n and X_j^n from Lemma 11, we have

$$\begin{aligned}
\frac{1}{\eta_n^2}\mathbb{E}\left[\Psi_{n,1}\Psi_{n,1}^\top\right] = \mathbb{E}\left[Y_n Y_n^\top\right] &= \sum_{j,k\in[i]} \mathbb{E}\left[X_j^n X_k^{n\top}\right] \\
&= \sum_{j\in[n]} \mathbb{E}\left[X_j^n X_j^{n\top}\right], \quad \text{since } A_j, A_k \text{ are independent for } j \neq k \\
&= \frac{1}{(1+\eta_n\lambda_1)^2} \sum_{j\in[n]} V_\perp \Lambda_\perp^{n-j} V_\perp^\top \mathbb{E}\left[(A_j-\Sigma)\, v_1 v_1^\top\, (A_j-\Sigma)\right] V_\perp \Lambda_\perp^{n-j} V_\perp^\top \\
&= \frac{1}{(1+\eta_n\lambda_1)^2} V_\perp \left(\sum_{j\in[n]} \Lambda_\perp^{n-j} \underbrace{V_\perp^\top \mathbb{E}\left[(A_j-\Sigma)\, v_1 v_1^\top\, (A_j-\Sigma)\right] V_\perp}_{:=\widetilde{M}} \Lambda_\perp^{n-j} \right) V_\perp^\top.
\end{aligned}$$

Recall $R^{(n)} := \frac{1}{(1+\eta_n\lambda_1)^2}\sum_{j\in[n]} \Lambda_\perp^{n-j}\widetilde{M}\Lambda_\perp^{n-j}$ and consider $(k,l)^{\text{th}}$ entry of $R^{(n)}$.

$$R_{kl}^{(n)} = \frac{1}{(1+\eta_n\lambda_1)^2} e_k^\top \sum_{j\in[n]} \Lambda_\perp^{n-j}\widetilde{M}\Lambda_\perp^{n-j} e_l = \frac{1}{(1+\eta_n\lambda_1)^2}\widetilde{M}_{kl}\sum_{j=1}^{n}(d_k d_l)^{n-j} = \frac{1}{(1+\eta_n\lambda_1)^2}\widetilde{M}_{kl}\left(\frac{1-(d_k d_l)^n}{1-d_k d_l}\right).$$

Let $R_0(k,l) = \widetilde{M}_{k\ell}/(2\lambda_1 - \lambda_{k+1} - \lambda_{\ell+1})$. Note that

$$\begin{aligned}
1 - d_k d_l &= \frac{\eta_n\left(2\lambda_1-\lambda_{k+1}-\lambda_{l+1}\right)}{1+\eta\lambda_1} - \frac{\eta_n^2\left(\lambda_1-\lambda_{k+1}\right)\left(\lambda_1-\lambda_{l+1}\right)}{\left(1+\eta\lambda_1\right)^2} \\
&= \frac{\eta_n\left(2\lambda_1-\lambda_{k+1}-\lambda_{l+1}\right)}{1+\eta\lambda_1}\left[1 - \frac{\eta_n\left(\lambda_1-\lambda_{k+1}\right)\left(\lambda_1-\lambda_{l+1}\right)}{\left(1+\eta\lambda_1\right)\left(\lambda_1-\lambda_{k+1}+\lambda_1-\lambda_{l+1}\right)}\right] \\
&\geq \frac{\eta_n\left(2\lambda_1-\lambda_{k+1}-\lambda_{l+1}\right)}{1+\eta\lambda_1}\left[1 - \frac{\eta_n\left(\lambda_1-\lambda_{k+1}\right)\left(\lambda_1-\lambda_{l+1}\right)}{\left(\lambda_1-\lambda_{k+1}+\lambda_1-\lambda_{l+1}\right)}\right] \\
&\geq \frac{\eta_n\left(2\lambda_1-\lambda_{k+1}-\lambda_{l+1}\right)}{1+\eta\lambda_1}\left[1-\eta_n\min\left\{\lambda_1-\lambda_{k+1},\lambda_1-\lambda_{l+1}\right\}\right] \\
&\geq \frac{\eta_n\left(2\lambda_1-\lambda_{k+1}-\lambda_{l+1}\right)}{1+\eta\lambda_1}\left[1-\eta_n\lambda_1\right] \\
&\geq \eta_n\left(2\lambda_1-\lambda_{k+1}-\lambda_{l+1}\right)\left(1-O\left(\eta_n\lambda_1\right)\right)
\end{aligned}$$

Then,

$$\begin{aligned}
R_{kl}^{(n)} - R_0(k,l)/\eta_n &= \frac{\widetilde{M}_{k\ell}}{\eta_n(2\lambda_1-\lambda_{k+1}-\lambda_{\ell+1})}\frac{(1+O(\eta_n\lambda_1))}{(1+\eta_n\lambda_1)^2} - \frac{\widetilde{M}_{k\ell}}{\eta_n(2\lambda_1-\lambda_{k+1}-\lambda_{\ell+1})} \\
&= \frac{\widetilde{M}_{k\ell}}{\eta_n(2\lambda_1-\lambda_{k+1}-\lambda_{\ell+1})}(1+O(\eta_n\lambda_1)) - \frac{\widetilde{M}_{k\ell}}{\eta_n(2\lambda_1-\lambda_{k+1}-\lambda_{\ell+1})} \\
&= \frac{\widetilde{M}_{k\ell}}{\eta_n(2\lambda_1-\lambda_{k+1}-\lambda_{\ell+1})}O(\eta_n\lambda_1)
\end{aligned}$$

So we have:

$$\frac{\eta_n R_{kl}^{(n)} - R_0(k,l)}{R_0(k,l)} = O(\eta_n\lambda_1)$$

Finally, we have:

$$\|\eta_n R^{(n)} - R_0\|_F \leq \frac{\eta_n\lambda_1}{\lambda_1-\lambda_2}\|\widetilde{M}\|_F/2$$

Note that

$$\|\widetilde{M}\|_F^2 \leq \mathbb{E}\left[\left\|(A_i-\Sigma)v_1 v_1^\top(A_i-\Sigma)\right\|\right] \leq \mathbb{E}\left[\|A_i-\Sigma\|^2\right] \leq \mathcal{M}_2^2.$$

□

B.2 ESTIMATOR CONCENTRATION

In this section, we estimate the bias of the variance estimate output by Algorithm 1. In the entirety of this section, we assume that the vector $\tilde{v}$ is "good", i.e $\sin^2(\tilde{v}, v_1) \lesssim \frac{\log(1/\delta)}{\delta^3}\frac{\eta_N \mathcal{M}_2^2}{(\lambda_1-\lambda_2)}$, which happens with probability at least $1-\delta$. Recall that $\tilde{v} \leftarrow \mathsf{Oja}(\mathcal{D}_N, \eta_N, u_0)$ is the high accuracy estimate of v_1. We present all results using a general n number of i.i.d. samples per split, which will later be replaced by $n/(m_1 m_2)$ as required by Algorithm 1. We denote $s_n := \frac{\log(1/\delta)}{\delta^3}\frac{\eta_n \mathcal{M}_2^2}{(\lambda_1-\lambda_2)}$ to be the upper bound on the $\sin^2$ error of the Oja vector due to Jain et al. [2016]. While our results henceforth are written using s_n and s_n is not guaranteed to be smaller than 1, it is straightforward to replace it by $\min\{s_n, 1\}$ since the $\sin^2$ error between any two vectors is always at most 1.

B.2.1 $\Psi_{n,0}$ Tail Bound

Lemma 12. *Let $\Psi_{n,0}$ be defined as in Lemma 2 for v_{oja} defined in (6). Let $\left\{\Psi_{n,0}^{(i)}\right\}_{i\in[m]}$ and $\left\{v_{\mathsf{oja}}^{(i)}\right\}_{i\in[m]}$ be m iid instances of $\Psi_{n,0}$ and v_{oja} respectively. Then, for any $k \in [d]$,*

$$P\left(\sum_{i\in[m]} \frac{\left(e_k^\top \Psi_{n,0}^{(i)}\right)^2}{m} \le \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^3}\frac{\eta_N \mathcal{M}_2^2}{(\lambda_1-\lambda_2)}\right) \ge 1-\delta.$$

Proof. For any $i \in [m]$,

$$\left|e_k^\top \Psi_{n,0}^{(i)}\right| = \left|e_k^\top \left(v_1 v_1^\top - \tilde{v}\tilde{v}^\top\right) v_{\mathsf{oja}}^{(i)}\right| \le \left\|v_1 v_1^\top - \tilde{v}\tilde{v}^\top\right\| = \sqrt{2}\left|\sin(\tilde{v}, v_1)\right|.$$

The result now follows from Corollary 1 of Lunde et al. [2021], which states that with probability at least $1-\delta$,

$$\sin^2(\tilde{v}, v_1) \le \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^3}\frac{\eta_N \mathcal{M}_2^2}{(\lambda_1-\lambda_2)}.$$

for some universal constant $C > 0$. □

B.2.2 $\Psi_{n,1}$ (Hajek Projection) Concentration

Lemma 13. *Let $\Psi_{n,1}$ be defined as in Lemma 2 for $u_0 = g/\|g\|_2$ with $g \sim \mathcal{N}(0, \mathbf{I}_d)$. Let $\left\{\Psi_{n,1}^{(i)}\right\}_{i\in[m]}$ and $\left\{g^{(i)}\right\}_{i\in[m]}$ be m i.i.d. instances of $\Psi_{n,1}$ and g respectively. Then, for any $\delta \in (0,1)$ and $k \in [d]$, with probability at least $1-\delta$,*

$$\left|\frac{\sum_{i\in[m]}\left(e_k^\top \Psi_{n,1}^{(i)}\right)^2}{m} - \mathbb{E}\left[\left(e_k^\top \Psi_{n,1}\right)^2\right]\right| \le \frac{\sqrt{2}\mathbb{E}\left[\left(e_k^\top \Psi_{n,1}\right)^2\right] + \eta_n^2 b_k^2 \mathcal{M}_4^2\sqrt{n}}{\sqrt{m\delta}}.$$

where $b_k := \left\|V_\perp^\top e_k\right\|_2$.

Proof. Recall the notations $X_j^n = V_\perp \Lambda_\perp^{n-j} V_\perp^\top (A_j - \Sigma) v_1$ and $Y_n = \sum_{j=1}^n X_j^n$ from Lemma 1. Since $V_\perp V_\perp^\top X_j^n = X_j^n$ and $T_{n,1} = \eta_n \sum_{i=1}^n X_j^n$, $e_k^\top \Psi_{n,1}$ can be written as

$$e_k^\top \Psi_{n,1} = \frac{e_k^\top V_\perp V_\perp^\top T_{n,1} v_1 \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)^n} = \frac{\eta_n \mathsf{sign}(v_1^\top u_0)}{(1+\eta_n\lambda_1)}\sum_{j=1}^n e_k^\top V_\perp V_\perp^\top X_j^n = \frac{\eta_n \mathsf{sign}(v_1^\top u_0)}{1+\eta_n\lambda_1} e_k^\top Y_n. \tag{23}$$

Next, we bound the variance of $(e_k^\top Y_n)^2$.

$$(e_k^\top Y_n)^2 = \sum_{j=1}^n \left(e_k^\top X_j^n\right)^2 + 2\sum_{j<j'}\left(e_k^\top X_j^n\right)\left(e_k^\top X_{j'}^n\right).$$

Most pairs of summands are uncorrelated.

- $\mathrm{Cov}\left((e_k^\top X_j^n)^2, (e_k^\top X_{j'}^n)^2\right) = 0$ for any distinct $j, j' \in [n]$.
- $\mathrm{Cov}\left((e_k^\top X_\ell^n)^2, (e_k^\top X_j^n)(e_k^\top X_{j'}^n)\right) = 0$ for any $\ell \in [n]$ and $1 \le j < j' \le n$.
- $\mathrm{Cov}\left((e_k^\top X_j^n)(e_k^\top X_{j'}^n), (e_k^\top X_\ell^n)(e_k^\top X_{\ell'}^n)\right) = 0$ for any $1 \le j < j' \le n$ and $1 \le \ell < \ell' \le n$ such that $(j, j') \neq (\ell, \ell')$.

It follows that

$$\mathrm{Var}\left((e_k^\top Y_n)^2\right) = \sum_{j=1}^n \mathrm{Var}\left((e_k^\top X_j^n)^2\right) + 4\sum_{j<j'} \mathrm{Var}\left((e_k^\top X_j^n)(e_k^\top X_{j'}^n)\right). \tag{24}$$

We bound both terms separately. By Lemma 1, the second term can be bounded as

$$\begin{aligned} 4\sum_{j<j'} \mathrm{Var}\left((e_k^\top X_j^n)(e_k^\top X_{j'}^n)\right) &= 4\sum_{i<j} \mathbb{E}\left[(e_k^\top X_j^n)^2\right] \mathbb{E}\left[(e_k^\top X_{j'}^n)^2\right] \\ &\le 2\sum_{j=1}^n \sum_{j'=1}^n \mathbb{E}\left[(e_k^\top X_j^n)^2\right] \mathbb{E}\left[(e_k^\top X_{j'}^n)^2\right] = 2\mathbb{E}\left[(e_k^\top Y_n)^2\right]^2. \end{aligned} \tag{25}$$

Next, we bound the first term of Equation (24). For any $j \in [n]$,

$$\left|e_k^\top X_j^n\right| = \left|e_k^\top V_\perp \Lambda_\perp^{n-j} V_\perp^\top (A_i - \Sigma) v_1\right| \le \left\|e_k^\top V_\perp\right\| \left\|\Lambda_\perp^{n-j}\right\| \left\|V_\perp^\top (A_j - \Sigma) v_1\right\| \le b_k \left\|A_j - \Sigma\right\|,$$

which implies

$$\sum_{j=1}^n \mathrm{Var}\left((e_k^\top X_j^n)^2\right) \le \sum_{j=1}^n \mathbb{E}\left[(e_k^\top X_j^n)^4\right] \le \sum_{j=1}^n \mathbb{E}\left[b_k^4 \left\|A_j - \Sigma\right\|^4\right] \le b_k^4 \mathcal{M}_4^4 n. \tag{26}$$

Combining equations (24), (25), and (26) and using equality (23),

$$\mathrm{Var}\left((e_k^\top \Psi_{n,1})^2\right) \le 2\mathbb{E}\left[\left((e_k^\top \Psi_{n,1})^2\right)\right]^2 + \frac{\eta_n^4}{(1+\eta_n\lambda_1)^4} b_k^4 \mathcal{M}_4^4 n.$$

By Chebyshev's inequality, for any $t > 0$,

$$\begin{aligned} P\left(\left|\frac{1}{m}\sum_{i=1}^m \left(e_k^\top \Psi_{n,1}^{(i)}\right)^2 - \mathbb{E}\left[\left(e_k^\top \Psi_{n,1}\right)^2\right]\right| \ge t\right) &\le \frac{\mathsf{Var}\left(\left(e_k^\top \Psi_{n,1}\right)^2\right)}{mt^2} \\ &\le \frac{2\mathbb{E}\left[\left((e_k^\top \Psi_{n,1})^2\right)\right]^2 + \frac{\eta_n^4}{(1+\eta_n\lambda_1)^4} b_k^4 \mathcal{M}_4^4 n}{mt^2}. \end{aligned}$$

The result follows by setting $t = \frac{\sqrt{2}\mathbb{E}\left[\left((e_k^\top \Psi_{n,1})^2\right)\right] + \eta_n^2 b_k^2 \mathcal{M}_4^2 \sqrt{n}}{\sqrt{m\delta}}$.

□

Remark 6. *Note that in Lemma 13, one can always provide a uniform bound on all elements using a Bernstein-type tail inequality rather than a Chebyshev bound. This is possible because we can use our concentration inequality in Lemma 26. However, there are two pitfalls of this approach; first, for failure probability δ, the errors of the lower order terms ($\Psi_{n,2}, \Psi_{n,3}, \Psi_{n,4}$) still depend polynomially on the $1/\delta$ (see Lemma 17, 19, 21), which limits the sample complexity of our estimator to have a* $\mathrm{poly}(1/\delta)$ *factor, and secondly, Lemma 26 requires a stronger a.s. upper bound on $A_i - \Sigma$ for $i \in [n]$. However, we can get both a uniform bound over all coordinates $k \in [d]$, and a $\log(1/\delta)$ dependence on the sample complexity, using our median of means based algorithm (Algorithm 1).*

B.2.3 $\Psi_{n,2}$ tail bound

We start by providing a tail bound on higher order terms in the Hoeffding decomposition of $B_n - \mathbb{E}[B_n]$, which may be of independent interest. Let $\mathcal{S}_{n,k} := \{\{i_1, \ldots, i_k\} : 1 \le i_1 < \cdots < i_k \le n\}$. Consider a general product of n matrices, where all but k of the matrices are constant, and k indexed by the subset S are mean zero independent random matrices.

With slight abuse of notation, let $M_{S,i}$ denote a constant matrix M_i with $\|M_i\| =: m_i$ when $i \notin S$ and W_i when $i \in S$, $EW_i = 0$, $W_i, i = 1, \ldots, n$ are mutually independent.

$$T_{n,k} := \sum_{S \in \mathcal{S}_{n,k}} \prod_{i=1}^{n} M_{S,n+1-i} \tag{27}$$

Let $T_{n,k}$ be a scaled version of the kth term in the Hoeffding projection of the matrix product $B_n := \prod_{i=1}^{n}(I + \eta_n A_i)$. Let $W_i = A_i - \Sigma$. We want a tail bound for $\sum_{k \geq 2} T_{n,k}$.

Lemma 14. *For $S \in \mathcal{S}_{n,k}$, denote a function $M_{S,i} := \eta_n(A_i - \Sigma)$ when $i \in S$ and $I + \eta_n \Sigma$ when $i \notin S$. Suppose $q \geq 2$ and $\mathcal{M}_q$ are such that $\mathbb{E}\left[\|A_i - \Sigma\|^q\right]^{1/q} \leq \mathcal{M}_q$. Then, for any $1 \leq j \leq n$ and any $p \geq q$,*

$$\left|\left|\left|\sum_{k \geq j} T_{n,j}\right|\right|\right|_{p,q} \leq 2d^{1/p}(1 + \eta_n \lambda_1)^n \left(\frac{\eta_n \mathcal{M}_q \sqrt{np}}{1 + \eta_n \lambda_1}\right)^j,$$

as long as $\frac{2\eta_n \mathcal{M}_q \sqrt{np}}{1+\eta_n \lambda_1} < 1$.

Proof. We start by deriving a recurrence relation for $T_{n,k}$ as follows:

$$\begin{aligned}
T_{n,k} &= \sum_{S \in \mathcal{S}_{n,k}} \prod_{i=1}^{n} M_{S,n+1-i} \\
&= \sum_{S \in \mathcal{S}_{n,k}, n \notin S} \prod_{i=1}^{n} M_{S,n+1-i} + \sum_{S \in \mathcal{S}_{n,k}, n \in S} \prod_{i=1}^{n} M_{S,n+1-i} \\
&= \sum_{S \in \mathcal{S}_{n-1,k}} (I + \eta_n \Sigma) \prod_{i=2}^{n} M_{S,n+1-i} + \sum_{S \in \mathcal{S}_{n-1,k-1}} \eta_n (A_n - \Sigma) \prod_{i=2}^{n} M_{S,n+1-i} M_{S,n+1-i} \\
&= (I + \eta_n \Sigma) \left(\sum_{S \in \mathcal{S}_{n-1,k}} \prod_{i=1}^{n-1} M_{S,n-i} \right) + \eta_n (A_n - \Sigma) \left(\sum_{S \in \mathcal{S}_{n-1,k-1}} \prod_{i=1}^{n-1} M_{S,n-i} \right) \\
&= (I + \eta_n \Sigma) T_{n-1,k} + \eta_n (A_n - \Sigma) T_{n-1,k-1}.
\end{aligned}$$

Next, we apply Proposition 4.3. of Huang et al. [2022] to bound $\|\!|T_{n,k}|\!\|_{p,q}$. To apply the proposition, we require $\mathbb{E}\left[\eta_n (A_n - \Sigma) T_{n-1,k-1} | (I + \eta_n \Sigma) T_{n-1,k}\right] = 0$. Indeed, by independence of $A_1, A_2, \ldots, A_n$,

$$\mathbb{E}\left[\eta_n (A_n - \Sigma) T_{n-1,k-1} | (I + \eta_n \Sigma) T_{n-1,k}\right] = \mathbb{E}\left[\eta_n (A_n - \Sigma)\right] \mathbb{E}\left[T_{n-1,k-1} | (I + \eta_n \Sigma) T_{n-1,k}\right] = 0.$$

Therefore, the proposition implies that

$$\|\!|T_{n,k}|\!\|_{p,q}^2 \leq \|\!|(I + \eta_n \Sigma) T_{n-1,k}|\!\|_{p,q}^2 + (p-1) \|\!|\eta_n (A_n - \Sigma) T_{n-1,k-1}|\!\|_{p,q}^2.$$

From Equation 4.1. and Equation 5.3. of Huang et al. [2022],

$$\begin{aligned}
\|\!|(I + \eta_n \Sigma) T_{n-1,k}|\!\|_{p,q} &\leq \|I + \eta_n \Sigma\|_{\text{op}} \|\!|T_{n-1,k}|\!\|_{p,q}, \text{ and} \\
\|\!|\eta_n (A_n - \Sigma) T_{n-1,k-1}|\!\|_{p,q} &\leq \eta_n \mathbb{E}\left[\|A_n - \Sigma\|^q\right]^{1/q} \|\!|T_{n-1,k-1}|\!\|_{p,q}.
\end{aligned}$$

Plugging these bounds into the recurrence yields

$$\|\!|T_{n,k}|\!\|_{p,q}^2 \leq (1 + \eta_n \lambda_1)^2 \|\!|T_{n,k-1}|\!\|_{p,q}^2 + \eta_n^2 \mathcal{M}_q^2 (p-1) \mathbb{E}\left[\|A_n - \Sigma\|^q\right]^{2/q} \|\!|T_{n-1,k-1}|\!\|_{p,q}^2.$$

Letting $f_{n,k} := \|\!|T_{n,k}|\!\|_{p,q}^2$, we have the following recurrence for all $n \geq k \geq 1$:

$$f_{n,k} \leq (1 + \eta_n \lambda_1)^2 f_{n-1,k} + \eta_n^2 \mathcal{M}_q^2 (p-1) f_{n-1,k-1}.$$

Defining $a_{n,k} := \frac{f_{n,k}}{(1+\eta_n\lambda_1)^{2(n-k)}(\eta_n^2\mathcal{M}_q^2(p-1))^k}$, we recover an inequality resembling Pascal's identity:

$$a_{n,k} \leq a_{n-1,k} + a_{n-1,k-1}.$$

Moreover, $a_{n,k} = 0$ for all $n < k$ and $a_{n,0} = (1+\eta_n\lambda_1)^{-2n}|||(I+\eta_n\Sigma)^n|||_{p,q}^2 \leq d^{2/p}$. Inducting on n and k shows

$$a_{n,k} \leq d^{2/p}\binom{n}{k}.$$

Translating this back to the bound on the norm of $T_{n,k}$, we conclude

$$|||T_{n,k}|||_{p,q} \leq \sqrt{(1+\eta_n\lambda_1)^{2(n-k)}\left(\eta_n^2\mathcal{M}_q^2(p-1)\right)^k d^{2/p}\binom{n}{k}} \leq d^{1/p}(1+\eta_n\lambda_1)^{n-k}\left(\eta_n\mathcal{M}_q\sqrt{np}\right)^k$$

Since norms are sub-additive and $\frac{\eta_n\mathcal{M}_q\sqrt{np}}{1+\eta_n\lambda_1} < \frac{1}{2}$,

$$\begin{aligned}
\left|\left|\left|\sum_{k\geq j} T_{n,k}\right|\right|\right|_{p,q} &\leq \sum_{k=j}^{n} d^{1/p}(1+\eta_n\lambda_1)^{n-k}\left(\eta_n\mathcal{M}_q\sqrt{np}\right)^k \\
&= d^{1/p}(1+\eta_n\lambda_1)^n \sum_{k=j}^{n}\left(\frac{\eta_n\mathcal{M}_q\sqrt{np}}{1+\eta_n\lambda_1}\right)^k \\
&\leq 2d^{1/p}(1+\eta_n\lambda_1)^n\left(\frac{\eta_n\mathcal{M}_q\sqrt{np}}{1+\eta_n\lambda_1}\right)^j.
\end{aligned}$$

□

Lemma 15. *For $S \in \mathcal{S}_{n,k}$, denote a function $M_{S,i} := \eta_n(A_i - \Sigma)$ when $i \in S$ and $I + \eta_n\Sigma$ when $i \notin S$. Then, for any $1 \leq j \leq n$, and $2 \leq q \leq 4\log d$,*

$$P\left(\left\|\sum_{k\geq j} T_{n,k}\right\| \geq \frac{3(1+\eta_n\lambda_1)^n\left(\eta_n\mathcal{M}_q\sqrt{4n\log d}\right)^j}{\delta^{\frac{1}{4\log d}}}\right) \leq \delta,$$

as long as $4\eta_n\mathcal{M}_q\sqrt{n\log d} < 1$.

Proof. Let $p = 4\log d$; note that the assumption $\frac{2\eta_n\mathcal{M}_q\sqrt{np}}{1+\eta_n\lambda_1} < 1$ holds. By Markov's inequality, Equation 4.2. of Huang et al. [2022], and Lemma 14,

$$\begin{aligned}
P\left(\left\|\sum_{k\geq j} T_{n,k}\right\| \geq (1+\eta_n\lambda_1)^n t\right) &\leq \inf_{p'\geq 2}\left((1+\eta_n\lambda_1)^n t\right)^{-p'}\mathbb{E}\left[\left\|\sum_{k\geq j} T_{n,k}\right\|^{p'}\right] \\
&\leq \inf_{p'\geq 2}\left((1+\eta_n\lambda_1)^n t\right)^{-p'}\mathbb{E}\left[\left|\left|\left|\sum_{k\geq j} T_{n,k}\right|\right|\right|_{p',q}^{p'}\right] \\
&\leq \left(\frac{2d^{1/p}\left(\frac{\eta_n\mathcal{M}_q\sqrt{np}}{1+\eta_n\lambda_1}\right)^j}{t}\right)^p \leq \left(\frac{3\left(\eta_n\mathcal{M}_q\sqrt{4n\log d}\right)^j}{t}\right)^{4\log d}.
\end{aligned}$$

for all $t > 0$. The lemma follows by setting $t = 3\left(\eta_n\mathcal{M}_q\sqrt{4n\log d}\right)^j\delta^{\frac{-1}{4\log d}}$.

□

Lemma 16. *Let $\Psi_{n,2}$ be as defined in Lemma 2 with $u_0 = g/\|g\|_2$. Then, for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\|\Psi_{n,2}\| \leq \frac{12\eta_n^2\mathcal{M}_2^2 n\log d}{\sqrt{\delta}}\right) \geq 1-\delta.$$

Proof. By Lemma 15, with probability at least $1-\delta$,

$$\left\|\sum_{k\geq 2} T_{n,k}\right\| \leq \frac{3(1+\eta_n\lambda_1)^n\left(\eta_n\mathcal{M}_2\sqrt{4n\log d}\right)^2}{\delta^{\frac{1}{4\log d}}} < \frac{12(1+\eta_n\lambda_1)^n\eta_n^2\mathcal{M}_2^2 n\log d}{\sqrt{\delta}}.$$

Conditioned on this event,

$$\begin{aligned}\|\Psi_{n,2}\| &= \frac{\left\|V_\perp V_\perp^\top(\sum_{k\geq 2}T_{n,k})v_1\mathsf{sign}(v_1^\top u_0)\right\|}{(1+\eta_n\lambda_1)^n} \leq \frac{\left\|V_\perp V_\perp^\top\right\|\left\|\sum_{k\geq 2}T_{n,k}\right\|\|v_1\|}{(1+\eta_n\lambda_1)^n}\\ &\leq \frac{\left\|\sum_{k\geq 2}T_{n,k}\right\|}{(1+\eta_n\lambda_1)^n} \leq \frac{12\eta_n^2\mathcal{M}_2^2 n\log d}{\sqrt{\delta}}.\end{aligned}$$

□

Lemma 17. *Let $\Psi_{n,2}$ be defined as in Lemma 2 for $u_0 = g/\|g\|_2$ with $g\sim\mathcal{N}(0,\mathbf{I}_d)$. Let $\left\{\Psi_{n,2}^{(i)}\right\}_{i\in[m]}$ and $\left\{g^{(i)}\right\}_{i\in[m]}$ be m i.i.d. instances of $\Psi_{n,2}$ and g respectively, and let $\delta\in(0,1)$. Then, with probability at least $1-\delta$,*

$$\frac{\sum_{i\in[m]}\left(e_k^\top\Psi_{n,2}^{(i)}\right)^2}{m} \leq \frac{144b_k^2\eta_n^4\mathcal{M}_2^4 n^2\log^2 d}{\delta},$$

for all $k\in[d]$, where $b_k := \left\|V_\perp^\top e_k\right\|_2$.

Proof. We have

$$\begin{aligned}\left|e_k^\top\Psi_{n,2}\right| &= \frac{\left|e_k^\top V_\perp V_\perp^\top(\sum_{k\geq 2}T_{n,k})v_1\mathsf{sign}(v_1^\top u_0)\right|}{(1+\eta_n\lambda_1)^n} = \frac{\left|e_k^\top V_\perp V_\perp^\top V_\perp V_\perp^\top(\sum_{k\geq 2}T_{n,k})v_1\mathsf{sign}(v_1^\top u_0)\right|}{(1+\eta_n\lambda_1)^n}\\ &= \left|e_k^\top V_\perp V_\perp^\top V_\perp\Psi_{n,2}\right| \leq \left\|e_k^\top V_\perp\right\|\|\Psi_{n,2}\| \leq \frac{b_k\left\|\sum_{k\geq 2}T_{n,k}\right\|}{(1+\eta_n\lambda_1)^n}.\end{aligned}$$

By Lemma 16, for each $i\in[m]$, with probability at least $1-\frac{\delta}{m}$,

$$\left|e_k^\top\Psi_{n,2}^{(i)}\right| \leq \frac{12b_k\eta_n^2\mathcal{M}_2^2 n\log d}{\sqrt{\delta/m}}.$$

By a union bound, the above holds for all $i\in[m]$ with probability at least $1-\delta$. Under this event,

$$\frac{\sum_{i\in[m]}\left(e_k^\top\Psi_{n,2}^{(i)}\right)^2}{m} \leq \frac{\sum_{i\in[m]}\left(\frac{12b_k\eta_n^2\mathcal{M}_2^2 n\log d}{\sqrt{\delta/m}}\right)^2}{m} = \frac{144b_k^2\eta_n^4\mathcal{M}_2^4 n^2\log^2 d}{\delta}.$$

□

B.2.4 $\Psi_{n,3}$ tail bound

Lemma 18. *Let $\Psi_{n,3}$ be as defined in Lemma 2 with $u_0 = g/\|g\|_2$. Let η_n be set according to Lemma 9. Fix $\delta\in(0,1)$. Then for any $\epsilon>0$ we have with probability at least $1-\delta$,*

$$\|\Psi_{n,3}\|_2 \lesssim \sqrt{s_n}\left(\frac{d\exp\left(-2\eta_n n(\lambda_1-\lambda_2)+\eta_n^2 n\left(\lambda_1^2+\mathcal{M}_2^2\right)\right)+\frac{\eta_n\mathcal{M}_2^2}{(\lambda_1-\lambda_2)}}{\delta^3(1-\delta)\log^{-1}(1/\delta)}\right)^{\frac{1}{2}} + \sqrt{s_n}\frac{\eta_n\sqrt{n}\mathcal{M}_2\log(d)}{\delta^{\frac{1}{2}}}.$$

where $s_n := \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^3}\frac{\eta_n\mathcal{M}_2^2}{(\lambda_1-\lambda_2)}$ for a universal constant $C>0$.

Proof. Let $c_n = (1+\eta_n\lambda_1)^n |u_0^T v_1|$. We first note that

$$\|\Psi_{n,3}\|_2 = \left\| V_\perp V_\perp^\top B_n u_0 \left(\frac{1}{\|B_n u_0\|_2} - \frac{1}{c_n} \right) \right\|_2 = \left\| \frac{V_\perp V_\perp^\top B_n u_0}{\|B_n u_0\|_2} \left(1 - \frac{\|B_n u_0\|_2}{c_n} \right) \right\|_2$$
$$\leq \left\| \frac{V_\perp V_\perp^\top B_n u_0}{\|B_n u_0\|_2} \right\|_2 \left| \frac{\|B_n u_0\|_2}{c_n} - 1 \right|. \tag{28}$$

We bound each of the two multiplicands separately. The first term corresponds to the $\sin$ error between v_{oja} and v_1:

$$\left\| \frac{V_\perp V_\perp^\top B_n u_0}{\|B_n u_0\|_2} \right\|_2^2 = 1 - \frac{\left(v_1^\top B_n u_0\right)^2}{\|B_n u_0\|_2^2} = \sin^2\left(v_{\mathsf{oja}}, v_1\right).$$

By Corollary 1 of Lunde et al. [2021],

$$\mathbb{P}\left(\left\| \frac{V_\perp V_\perp^\top B_n u_0}{\|B_n u_0\|_2} \right\|_2^2 > s_n \right) = \mathbb{P}\left(\sin^2\left(v_{\mathsf{oja}}, v_1\right) > s_n \right) \leq \delta. \tag{29}$$

It follows that for any $\epsilon > 0$,

$$\mathbb{P}\left(\|\Psi_{n,3}\|_2 > \epsilon\sqrt{s_n} \right) \leq \mathbb{P}\left(\left\| \frac{V_\perp V_\perp^\top B_n u_0}{\|B_n u_0\|_2} \right\|_2^2 > s_n \right) + \mathbb{P}\left(\left| \frac{\|B_n u_0\|_2}{c_n} - 1 \right| > \epsilon \right) \tag{30}$$
$$\leq \delta + \mathbb{P}\left(\left| \frac{\|B_n u_0\|_2}{c_n} - 1 \right| > \epsilon \right). \tag{31}$$

To bound the second term, we adapt the proof of Lemma B.2 in Lunde et al. [2021]. Letting $a_1 = \left|v_1^\top u_0\right|$,

$$\left| \frac{\|B_n u_0\|}{c_n} - 1 \right| \leq \left| \frac{\|B_n v_1 a_1\| - \|a_1 (I+\eta_n\Sigma)^n v_1\|}{c_n} \right| + \frac{\|B_n V_\perp V_\perp^T u_0\|}{c_n}$$
$$= \left| \frac{\|B_n v_1\| - \|(I+\eta_n\Sigma)^n v_1\|}{(1+\eta_n\lambda_1)^n} \right| + \frac{\|B_n V_\perp V_\perp^T u_0\|}{c_n}$$
$$\leq \frac{\|B_n - \mathbb{E}[B_n]\|_{\mathsf{op}}}{(1+\eta_n\lambda_1)^n} + \frac{\|B_n V_\perp V_\perp^T u_0\|}{c_n}. \tag{32}$$

For the first summand, using Eq 5.6 of Huang et al. [2022] with $q = 2$ and by Markov's inequality,

$$\mathbb{P}\left(\frac{\|B_n - \mathbb{E}\left[B_n\right]\|_{\mathsf{op}}}{(1+\eta_n\lambda_1/n)^n} > \frac{\epsilon}{2} \right) \leq \frac{\mathbb{E}\left[\|B_n - \mathbb{E}\left[B_n\right]\|_{\mathsf{op}}^2\right]}{(1+\eta_n\lambda_1)^n \epsilon^2} \leq \frac{C\eta_n^2 n \mathcal{M}_2^2 (1+\log d)^2}{\epsilon^2} \tag{33}$$

For the second summand of equation (32), define the event

$$\mathcal{G} = \left\{ \frac{\|B_n V_\perp V_\perp^T u_0\|^2}{|v_1^T u_0|^2} \leq \frac{C\log(1/\delta)}{\delta^2} \mathrm{trace}\left(V_\perp^T B_n^T B_n V_\perp\right) \right\}.$$

By Proposition B.6 of Lunde et al. [2021], $P(\mathcal{G}) \geq 1-\delta$ where $C > 0$ is some universal constant. Since $P(A|B)P(B) = P(A\cap B) \leq P(A)$, Markov's inequality together with Lemma 5.2 of Jain et al. [2016] with $\mathcal{V} \leq \mathcal{M}_2^2$ yields

$$\mathbb{P}\left(\frac{\|B_n V_\perp V_\perp^T u_0\|}{c_n} \geq \frac{\epsilon}{2} | \mathcal{G} \right) \tag{34}$$
$$\leq \frac{1}{1-\delta} \mathbb{P}\left(\mathrm{trace}(V_\perp B_n^T B_n V_\perp^T) \geq \frac{\epsilon^2}{4} \cdot \frac{\delta^2}{C\log(1/\delta)} \right)$$
$$\leq \frac{1}{1-\delta} C \frac{d\exp\left(-2\eta_n n\left(\lambda_1-\lambda_2\right) + \eta_n^2 n\left(\lambda_1^2+\mathcal{M}_2^2\right)\right) + \frac{\eta_n \mathcal{M}_2^2 \exp(n\eta_n^2(2\lambda_1^2+\mathcal{M}_2^2))}{2(\lambda_1-\lambda_2)}}{\epsilon^2\delta^2\log^{-1}(1/\delta)} \tag{35}$$
$$\leq \frac{1}{1-\delta} C \frac{d\exp\left(-2\eta_n n\left(\lambda_1-\lambda_2\right) + \eta_n^2 n\left(\lambda_1^2+\mathcal{M}_2^2\right)\right) + \frac{e\eta_n \mathcal{M}_2^2}{2(\lambda_1-\lambda_2)}}{\epsilon^2\delta^2\log^{-1}(1/\delta)}, \tag{36}$$

where the last bound follows from Lemma 9.

Finally, define the error ϵ as

$$\epsilon := \left(C \frac{d \exp\left(-2\eta_n n\left(\lambda_1 - \lambda_2\right) + \eta_n^2 n\left(\lambda_1^2 + \mathcal{M}_2^2\right)\right) + \frac{\eta_n \mathcal{M}_2^2}{(\lambda_1 - \lambda_2)}}{\delta^3 (1-\delta) \log^{-1}(1/\delta)} \right)^{\frac{1}{2}} + \frac{\eta_n \sqrt{n} \mathcal{M}_2 \log(d)}{\delta^{\frac{1}{2}}}. \tag{37}$$

Substituting ϵ in equations (36) and (33), and combining with equation (32),

$$\mathbb{P}\left(\left| \frac{\|B_n u_0\|}{c_n} - 1 \right| > \epsilon \right) \leq \mathbb{P}\left(\frac{\|B_n V_\perp V_\perp^T u_0\|}{c_n} > \frac{\epsilon}{2} \right) + \mathbb{P}\left(\frac{\|B_n - \mathbb{E}\left[B_n\right]\|_{\text{op}}}{(1+\eta_n \lambda_1/n)^n} > \frac{\epsilon}{2} \right) \tag{38}$$

$$\leq \mathbb{P}\left(\frac{\|B_n V_\perp V_\perp^T u_0\|}{c_n} > \frac{\epsilon}{2} | \mathcal{G} \right) + \mathbb{P}(\mathcal{G}^{\complement}) + \mathbb{P}\left(\frac{\|B_n - \mathbb{E}\left[B_n\right]\|_{\text{op}}}{(1+\eta_n \lambda_1/n)^n} > \frac{\epsilon}{2} \right) \leq 3\delta. \tag{39}$$

From equations (31) and (39), we conclude

$$\mathbb{P}\left(\|\Psi_{n,3}\|_2 > \epsilon \sqrt{s_n} \right) \leq 4\delta.$$

□

Lemma 19. *Let $\Psi_{n,3}$ be defined as in Lemma 2 for $u_0 = g/\|g\|_2$ with $g \sim \mathcal{N}(0, \mathbf{I}_d)$. Let η_n be set according to Lemma 9. Let $\left\{\Psi_{n,3}^{(i)}\right\}_{i\in[m]}$ and $\left\{g^{(i)}\right\}_{i\in[m]}$ be m i.i.d. instances of $\Psi_{n,3}$ and g respectively. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\frac{\sum_{i\in[m]} \left(e_k^\top \Psi_{n,3}^{(i)}\right)^2}{m} \lesssim s_n b_k^2 \left(m^3 \left(\frac{d \exp\left(-2\eta_n n\left(\lambda_1 - \lambda_2\right) + \eta_n^2 n\left(\lambda_1^2 + \mathcal{M}_2^2\right)\right) + \frac{\eta_n \mathcal{M}_2^2}{(\lambda_1 - \lambda_2)}}{\delta^3 (1-\delta/m) \log^{-1}(m/\delta)} \right) + m \frac{\eta_n^2 n \mathcal{M}_2^2 \log^2(d)}{\delta} \right).$$

for all $k \in [d]$, where $b_k := \left\|V_\perp^\top e_k\right\|_2$ and $s_n := \frac{C \log\left(\frac{1}{\delta}\right)}{\delta^3} \frac{\eta_n \mathcal{M}_2^2}{(\lambda_1 - \lambda_2)}$ for a universal constant $C > 0$.

Proof. Using Lemma 18, for any fixed $i \in [m]$, with probability at least $1-\delta$,

$$\left\|\Psi_{n,3}^{(i)}\right\|_2 \lesssim \sqrt{s_n} \left(\left(\frac{d \exp\left(-2\eta_n n\left(\lambda_1 - \lambda_2\right) + \eta_n^2 n\left(\lambda_1^2 + \mathcal{M}_2^2\right)\right) + \frac{\eta_n \mathcal{M}_2^2}{(\lambda_1 - \lambda_2)}}{\delta^3 (1-\delta) \log^{-1}(1/\delta)} \right)^{\frac{1}{2}} + \frac{\eta_n \sqrt{n} \mathcal{M}_2 \log(d)}{\delta^{\frac{1}{2}}} \right). \tag{40}$$

Furthermore, note that

$$\begin{aligned}
\left|e_k^\top \Psi_{n,3}^{(i)}\right|_2 &= \left| e_k^\top V_\perp V_\perp^\top B_n u_0 \left(\frac{1}{\|B_n u_0\|_2} - \frac{1}{c_n} \right) \right|_2 \\
&= \left| e_k^\top V_\perp V_\perp^\top V_\perp V_\perp^\top B_n u_0 \left(\frac{1}{\|B_n u_0\|_2} - \frac{1}{c_n} \right) \right|_2 \\
&\leq \left\|e_k^\top V_\perp V_\perp^\top\right\|_2 \left\| V_\perp V_\perp^\top B_n u_0 \left(\frac{1}{\|B_n u_0\|_2} - \frac{1}{c_n} \right) \right\|_2 \\
&= b_k \left\| V_\perp V_\perp^\top B_n u_0 \left(\frac{1}{\|B_n u_0\|_2} - \frac{1}{c_n} \right) \right\|_2 = b_k \left\|\Psi_{n,3}^{(i)}\right\|_2
\end{aligned} \tag{41}$$

The result then follows by a union bound over all $i \in [m]$ for the event in (40) and using (41).

□

B.2.5 $\Psi_{n,4}$ tail bound

Lemma 20. *Let $\Psi_{n,4}$ be defined as in Lemma 2 for $u_0 = g/\|g\|_2$ with $g \sim \mathcal{N}(0, \mathbf{I}_d)$. Let η_n be set according to Lemma 9. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\Psi_{n,4}\| \leq \frac{1}{\delta^{3/2}} \left(d \exp\left(-2\eta_n n (\lambda_1 - \lambda_2) + \eta_n^2 n \left(\lambda_1^2 + \mathcal{M}_2^2\right)\right) + \frac{e\eta_n^3 n \mathcal{M}_2^4 (1 + 2\log(d))}{2(\lambda_1 - \lambda_2) + \eta_n (\lambda_1^2 - \lambda_2^2 - \mathcal{M}_2^2)} \right)^{1/2}.$$

Proof. Recall that

$$\|\Psi_{n,4}\| = \frac{\left\|V_\perp V_\perp^\top B_n V_\perp V_\perp^\top u_0\right\|}{|v_1^\top u_0|(1+\eta_n\lambda_1)^n} = \frac{\left\|V_\perp V_\perp^\top B_n V_\perp V_\perp^\top g\right\|}{|v_1^\top g|(1+\eta_n\lambda_1)^n}.$$

To bound this quantity, we will bound its square instead. Using Markov's inequality, with probability at least $1-\delta/2$,

$$\begin{aligned}
\left\|V_\perp V_\perp^\top B_n V_\perp V_\perp^\top g\right\|^2 &\leq \frac{2}{\delta} E\left[\left\|V_\perp V_\perp^\top B_n V_\perp V_\perp^\top g\right\|^2\right] \\
&= \frac{2}{\delta} \mathsf{Tr}\left(E\left[\left(V_\perp V_\perp^\top B_n V_\perp V_\perp^\top g\right)\left(V_\perp V_\perp^\top B_n V_\perp V_\perp^\top g\right)^\top\right]\right) \\
&= \frac{2}{\delta} \mathbb{E}\left[\mathsf{Tr}\left(V_\perp^\top B_n V_\perp V_\perp^\top B_n^\top V_\perp\right)\right].
\end{aligned}$$

By Lemma B.3 of Lunde et al. [2021],

$$\frac{\mathbb{E}\left[\mathsf{Tr}\left(V_\perp^\top B_n V_\perp V_\perp^\top B_n^\top V_\perp\right)\right]}{(1+\eta_n\lambda_1)^{2n}} \leq d \exp\left(-2\eta_n n (\lambda_1 - \lambda_2) + \eta_n^2 n \left(\lambda_1^2 + \mathcal{M}_2^2\right)\right) + \frac{e\eta_n^3 n \mathcal{M}_2^4 (1 + 2\log(d))}{2(\lambda_1 - \lambda_2) + \eta_n (\lambda_1^2 - \lambda_2^2 - \mathcal{M}_2^2)}.$$

Also, with probability at least $1-\delta/2$, $|v_1^\top g| \geq \delta/2$ (see Proposition 7 from Lunde et al. [2021] for anticoncentration of gaussians). Combining the two bounds yields the result. □

Lemma 21. *Let $\Psi_{n,4}$ be defined as in Lemma 2 for $u_0 = g/\|g\|_2$ with $g \sim \mathcal{N}(0, \mathbf{I}_d)$. Let η_n be set according to Lemma 9. Let $\left\{\Psi_{n,4}^{(i)}\right\}_{i\in[m]}$ and $\left\{g^{(i)}\right\}_{i\in[m]}$ be m i.i.d. instances of $\Psi_{n,4}$ and g respectively. Fix $\delta \in (0,1)$. Then, conditioned on $\mathcal{E}$, with probability at least $1-\delta$,*

$$\begin{aligned}
&\frac{\sum_{i\in[m]}\left(e_k^\top \Psi_{n,4}^{(i)}\right)^2}{m} \\
&\quad \leq \frac{b_k^2 m^2}{\delta^3(1-\delta)} \left(d \exp\left(-2\eta_n n (\lambda_1 - \lambda_2) + \eta_n^2 n \left(\lambda_1^2 + \mathcal{M}_2^2\right)\right) + \frac{e\eta_n^3 n \mathcal{M}_2^4 (1 + 2\log(d))}{2(\lambda_1 - \lambda_2) + \eta_n (\lambda_1^2 - \lambda_2^2 - \mathcal{M}_2^2)} \right)
\end{aligned}$$

for all $k \in [d]$, where $b_k := \left\|V_\perp^\top e_k\right\|_2$.

Proof. Note that

$$\left(e_k^\top \Psi_{n,4}\right)^2 \leq \left\|V_\perp^\top e_k\right\|_2^2 \underbrace{\left(\frac{\left\|V_\perp^\top B_n V_\perp V_\perp^\top u_0\right\|_2}{\left|v_1^\top u_0\right|(1+\eta_n\lambda_1)^n}\right)^2}_{\Phi_n}$$

Let $\Phi_n^{(i)}$ correspond to the ith instance of the random variable Φ_n. Then, for any $k \in [d]$,

$$\frac{1}{m}\sum_{i\in[m]}\left(e_k^\top \Psi_{n,4}^{(i)}\right)^2 \leq \frac{\left\|V_\perp^\top e_k\right\|_2^2}{m} \sum_{i\in[m]} \Phi_n^{(i)}. \tag{42}$$

Define the event $\mathcal{E} := \left\{|v_1^\top g| \geq \frac{\delta}{m}\right\}$ and let $\mathcal{E}^{(i)}$, $i \in [m]$ be the ith instance of this event. First, observe that:

$$\begin{aligned}
\mathbb{E}[\Phi_n|\mathcal{E}] = \mathbb{E}\left[\left(\frac{\left\|V_\perp^\top B_n V_\perp V_\perp^\top u_0\right\|_2}{|v_1^\top u_0|\,(1+\eta_n\lambda_1)^n}\right)^2 \middle|\mathcal{E}\right] &= \mathbb{E}\left[\frac{V_\perp^\top B_n V_\perp V_\perp^\top g g^\top V_\perp V_\perp^\top B_n^\top V_\perp}{\left(v_1^\top g\right)^2 (1+\eta_n\lambda_1)^{2n}}\middle|\mathcal{E}\right] \\
&\leq \frac{m^2}{\delta^2(1+\eta_n\lambda_1)^{2n}}\mathbb{E}\left[V_\perp^\top B_n V_\perp V_\perp^\top g g^\top V_\perp V_\perp^\top B_n^\top V_\perp \middle|\mathcal{E}\right] \\
&\leq \frac{m^2}{\delta^2\mathbb{P}(\mathcal{E})}\frac{\mathbb{E}\left[\mathsf{Tr}\left(V_\perp^\top B_n V_\perp V_\perp^\top B_n^\top V_\perp\right)\right]}{(1+\eta_n\lambda_1)^{2n}} \qquad (43)
\end{aligned}$$

Now, using Markov's inequality conditioned on $\bigcap_{i\in[m]}\mathcal{E}^{(i)}$, we have with probability at least $1-\mathbb{P}(\bigcap_{i\in[m]}\mathcal{E}^{(i)})$,

$$\begin{aligned}
\frac{1}{m}\sum_{i\in[m]}\Phi_n^{(i)} &\leq \frac{1}{\delta}\mathbb{E}\left[\Phi_n^{(i)}\middle|\bigcap_{j\in[m]}\mathcal{E}^{(j)}\right] \\
\text{(By i.i.d. nature of the instances)} &= \frac{1}{\delta}\mathbb{E}\left[\Phi_n^{(i)}\middle|\mathcal{E}^{(i)}\right] = \frac{1}{\delta}\mathbb{E}[\Phi_n|\mathcal{E}] \\
&\leq \frac{m^2}{\delta^3\mathbb{P}(\mathcal{E})}\frac{\mathbb{E}\left[\mathsf{Tr}\left(V_\perp^\top B_n V_\perp V_\perp^\top B_n^\top V_\perp\right)\right]}{(1+\eta_n\lambda_1)^{2n}} \qquad (44)
\end{aligned}$$

The last step uses Eq 43. Using Lemma B.3 from Lunde et al. [2021], we have

$$\frac{\mathbb{E}\left[\mathsf{Tr}\left(V_\perp^\top B_n V_\perp V_\perp^\top B_n^\top V_\perp\right)\right]}{(1+\eta_n\lambda_1)^{2n}} \leq d\exp\left(-2\eta_n n\left(\lambda_1-\lambda_2\right)+\eta_n^2 n\left(\lambda_1^2+\mathcal{M}_2^2\right)\right)+\frac{e\eta_n^3 n\mathcal{M}_2^4\left(1+2\log\left(d\right)\right)}{2\left(\lambda_1-\lambda_2\right)+\eta_n\left(\lambda_1^2-\lambda_2^2-\mathcal{M}_2^2\right)} \qquad (45)$$

Finally, we note that using Proposition 7 from Lunde et al. [2021], we have

$$\forall i\in[m], \mathbb{P}\left(\mathcal{E}^{(i)}\right) \geq 1-\frac{\delta}{m} \implies \mathbb{P}\left(\left(\bigcap_{i\in[m]}\mathcal{E}^{(i)}\right)^{\complement}\right) \leq \sum_{i\in[m]}\mathbb{P}\left(\mathcal{E}_i^{\complement}\right) \leq \sum_{i\in[m]}\frac{\delta}{m} = \delta \qquad (46)$$

The result follows by substituting (45) in (44) and then using (42), along with the union-bound provided in (46). □

B.2.6 Total Variance Bound

We now put together the results from Lemmas 12, 13, 17, 19, and 21 to provide a high probability bound on the error of the variance estimator Algorithm 1.

Figure 5 summarizes how the variance estimation algorithm works. The algorithm first computes an Oja vector $\tilde{v}$ using N samples. Then, n samples are divided into m_1 batches, with each batch containing n/m_1 samples. These n samples need not be disjoint from the N samples used to compute the high-accuracy estimate $\tilde{v}$. Then, the ℓth batch of n/m_1 samples is split into $m = m_2$ batches of size $B := n/m_1m_2$ each. Oja vectors $\{\hat{v}_j\}_{j\in[m_2]}$ are computed on each of these m_2 batches, and

$$\hat{\sigma}_{k,\ell}^2 := \sum_{j\in[m_2]}\frac{\left(e_k^\top\left(\hat{v}_j-(\tilde{v}^\top\hat{v}_j)\tilde{v}\right)\right)^2}{m_2}. \qquad (47)$$

for all $k\in[d]$. The overall estimate for the variance of the kth coordinate is $\mathsf{Median}\left(\{\hat{\sigma}_{k,\ell}\}_{\ell\in[m_1]}\right)$. Since this variance scales with the inverse of the learning parameter η_B, we define the scale-free $\hat{\gamma}_k := \mathsf{Median}\left(\{\hat{\sigma}_{k,\ell}\}_{\ell\in[m_1]}\right)/(\eta_B\left(\lambda_1-\lambda_2\right))$. For each $k\in[d]$, define the quantities

$$b_k := \left\|e_k^\top V_\perp\right\|, \quad c_k := \sqrt{\frac{\mathbb{E}\left[\left(e_k^\top\Psi_{B,1}\right)^2\right]}{\eta_B}\frac{\lambda_1-\lambda_2}{\mathcal{M}_2^2}}.$$

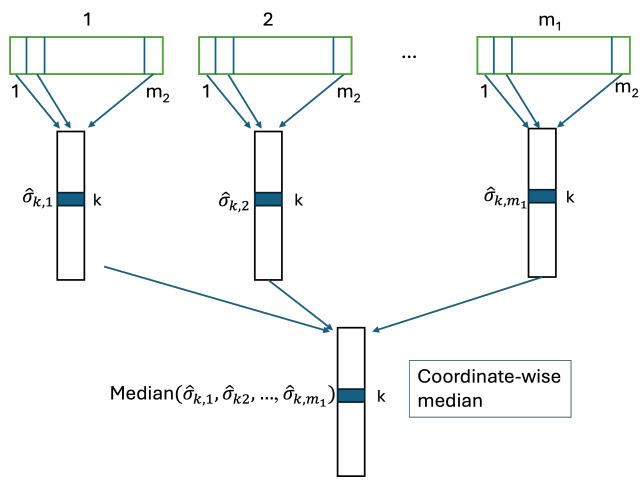


Figure 5: Schematic picture of Algorithm 1

Under this setting, we show that each $\hat{\sigma}_{k,\ell}^2$ approximates the true variance with at least $3/4$ probability. We assume that the learning rate η_B satisfies

$$\eta_B \leq \frac{1}{2\lambda_1} + \frac{\lambda_1 - \lambda_2}{2\mathcal{M}_2^2}. \tag{48}$$

It can be verified that this assumption is satisfied by the bounds on B provided in (56).

Lemma 22. *For any $\ell \in [m_1]$ and under assumption 48, with probability at least $3/4$,*

$$\begin{aligned}\left|\hat{\sigma}_{k,\ell}^2 - \eta_B\left(\lambda_1 - \lambda_2\right) e_k^\top \mathbb{V} e_k\right| \leq 8\left(\frac{1}{\sqrt{m}} + \frac{2}{m}\right)\eta_B\left(\lambda_1 - \lambda_2\right) e_k^\top \mathbb{V} e_k + O\left(\frac{b_k^2 \log^2 B}{B^{3/2} m^{1/2}}\left(\frac{\mathcal{M}_4}{\lambda_1 - \lambda_2}\right)^2 + \frac{\log N}{N}\left(\frac{\mathcal{M}_2}{\lambda_1 - \lambda_2}\right)^2\right) \\ + O\left(\frac{b_k^2 m^2 \log^2 d \log^4 B}{B^2}\left(\frac{\mathcal{M}_2}{\lambda_1 - \lambda_2}\right)^4 + \frac{\lambda_1 \mathcal{M}_2^2 \log^2 B}{B^2\left(\lambda_1 - \lambda_2\right)^3}\right).\end{aligned} \tag{49}$$

Proof. Drop the index ℓ for convenience of notation. Let $\delta_0 := 1/20$. By triangle inequality,

$$\left|\hat{\sigma}_k^2 - \eta_B\left(\lambda_1 - \lambda_2\right) e_k^\top \mathbb{V} e_k\right| \leq \left|\hat{\sigma}_{k,\ell}^2 - \mathbb{E}\left[\left(e_k^\top \Psi_{B,1}\right)^2\right]\right| + \left|\mathbb{E}\left[\left(e_k^\top \Psi_{B,1}\right)^2\right] - \eta_B\left(\lambda_1 - \lambda_2\right) e_k^\top \mathbb{V} e_k\right| \tag{50}$$

and by Lemma 1,

$$\left|\mathbb{E}\left[\left(e_k^\top \Psi_{B,1}\right)^2\right] - \eta_B\left(\lambda_1 - \lambda_2\right) e_k^\top \mathbb{V} e_k\right| \leq \frac{\eta_B^2 \mathcal{M}_2^2 \lambda_1}{\lambda_1 - \lambda_2} \lesssim \frac{\lambda_1 \mathcal{M}_2^2 \log^2 B}{B^2\left(\lambda_1 - \lambda_2\right)^3}. \tag{51}$$

By equation (19) and Lemma 7, for any $\epsilon \in (0, 1)$,

$$\begin{aligned}\left|\hat{\sigma}_k^2 - \mathbb{E}\left[\left(e_k^\top \Psi_{B,1}\right)^2\right]\right| \leq (1+\epsilon)\left|\frac{\sum_{j\in[m]}\left(e_k^\top \Psi_{B,1}^{(j)}\right)^2}{m} - \mathbb{E}\left[\left(e_k^\top \Psi_{B,1}\right)^2\right]\right| + \epsilon \mathbb{E}\left[\left(e_k^\top \Psi_{B,1}\right)^2\right] \\ + \underbrace{\frac{8}{\epsilon}\sum_{j\in[m]} \frac{\left(e_k^\top \Psi_{B,0}^{(j)}\right)^2 + \left(e_k^\top \Psi_{B,2}^{(j)}\right)^2 + \left(e_k^\top \Psi_{B,3}^{(j)}\right)^2 + \left(e_k^\top \Psi_{B,4}^{(j)}\right)^2}{m}}_{:= e_{\text{small}}}.\end{aligned} \tag{52}$$

Set $\epsilon = 2/\sqrt{m}$. By Lemmas 12, 17, 19, and 21, along with Lemma 9 to bound $nd\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) = o(1)$, we have

with probability at least $1-4\delta_0$

$$\frac{e_{\mathsf{small}}}{8/\epsilon} \lesssim \frac{\eta_N \mathcal{M}_2^2}{\lambda_1-\lambda_2} + b_k^2\eta_B^4\mathcal{M}_2^4 B^2\log^2 d + s_B b_k^2 m\eta_B^2 B\mathcal{M}_2^2\log^2 d + b_k^2 m^2 \frac{\eta_B^3 B\mathcal{M}_2^4\log d}{2(\lambda_1-\lambda_2)+\eta_B(\lambda_1^2-\lambda_2^2-\mathcal{M}_2^2)}$$
$$\lesssim \frac{\log N}{N}\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^2 + \frac{b_k^2 m\log^2 d\log^4 B}{B^2}\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^4 + \frac{b_k^2 m^2\log d\log^3 B}{B^2}\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^4. \tag{53}$$

where we used Assumption 48 to bound the last term. By Lemma 13, with probability $1-\delta_0$,

$$\left|\frac{\sum_{j\in[m]}\left(e_k^\top\Psi_{B,1}^{(j)}\right)^2}{m} - \mathbb{E}\left[\left(e_k^\top\Psi_{B,1}\right)^2\right]\right| \le \frac{\sqrt{2}\mathbb{E}\left[(e_k^\top\Psi_{B,1})^2\right]+\eta_B^2 b_k^2\mathcal{M}_4^2\sqrt{B}}{\sqrt{m\delta_0}}$$
$$\le 4\epsilon\mathbb{E}\left[(e_k^\top\Psi_{B,1})^2\right] + \frac{b_k^2\log^2 B}{B^{3/2}m^{1/2}}\left(\frac{\mathcal{M}_4}{\lambda_1-\lambda_2}\right)^2 \tag{54}$$

We now combine equations (51), (52), (53), and (54) in (50) to conclude that with probability at least $1-5\delta_0 = 3/4$,

$$\left|\hat{\sigma}_{k,\ell}^2 - \eta_B(\lambda_1-\lambda_2)e_k^\top\mathbb{V}e_k\right| \le (1+\epsilon)\left(4\epsilon\eta_B(\lambda_1-\lambda_2)e_k^\top\mathbb{V}e_k + \frac{b_k^2\log^2 B}{B^{3/2}m^{1/2}}\left(\frac{\mathcal{M}_4}{\lambda_1-\lambda_2}\right)^2\right) + (1+\epsilon)(1+4\epsilon)\frac{\eta_B^2\mathcal{M}_2^2\lambda_1}{\lambda_1-\lambda_2} + e_{\mathsf{small}},$$

which simplifies to the lemma statement.

□

Next, assume that the following relations hold:

$$N \gtrsim \frac{mB}{c_k^2\log B}\log\left(\frac{mB}{c_k^2\log B}\right). \tag{55}$$

$$B \gtrsim m^3\left(\frac{b_k}{c_k}\right)^2\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^2\log^3(B)\log^2(d). \tag{56}$$

$$B \gtrsim \max\left(m\left(\frac{b_k}{c_k}\right)^4\left(\frac{\mathcal{M}_4}{\mathcal{M}_2}\right)^4\log^2 B, \frac{m\lambda_1\log B}{c_k^2(\lambda_1-\lambda_2)}\right). \tag{57}$$

These assumptions on N and B *subsume* the assumption on the learning rate η_B in equation 48.

Using equation 51 and the relation

$$\frac{\mathbb{E}\left[\left(e_k^\top\Psi_{B,1}\right)^2\right]}{m} = \frac{\eta_B c_k^2}{m}\frac{\mathcal{M}_2^2}{\lambda_1-\lambda_2}. \tag{58}$$

and comparing it with each term in the smaller order error of Lemma 22 yields the following Lemma.

Lemma 23. *Under assumptions 55, 56, and 57, we have the following upper bound on the R.H.S of Eq 49 in Lemma 22.*

$$\frac{\log N}{N}\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^2 + \frac{b_k^2\log^2 B}{B^{3/2}m^{1/2}}\left(\frac{\mathcal{M}_4}{\lambda_1-\lambda_2}\right)^2 + \frac{b_k^2 m^2\log^2 d\log^4 B}{B^2}\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^4 + \frac{\lambda_1\mathcal{M}_2^2\log^2 B}{B^2(\lambda_1-\lambda_2)^3}$$
$$\le \frac{\eta_B(\lambda_1-\lambda_2)e_k^\top\mathbb{V}e_k}{m}. \tag{59}$$

It follows that a stronger multiplicative guarantee holds for any coordinate k that satisfies the above assumptions:

Lemma 24. *For any coordinate k that satisfies Lemma 22 and assumptions 55, 56, and 57,*

$$\left|\hat{\sigma}_k^2 - \eta_B(\lambda_1-\lambda_2)e_k^\top\mathbb{V}e_k\right| \le O\left(\frac{\eta_B(\lambda_1-\lambda_2)e_k^\top\mathbb{V}e_k}{\sqrt{m}}\right).$$

Given a per-coordinate guarantee that succeeds with probability $3/4$, we can boost the probability of success and give a uniform guarantee over all coordinates $k \in [d]$ using the median procedure described in Lemma 10.

Lemma 25. *Let $\{\hat{\gamma}_k\}_{k\in[d]}$ be the output of Algorithm 1. Under assumption (48), with probability $1-\delta$, for all $k \in [d]$,*

$$|\hat{\gamma}_k - \mathbb{V}_{kk}| \le 8\left(\frac{1}{\sqrt{m}} + \frac{2}{m}\right)\mathbb{V}_{kk} + O\left(\frac{b_k^2 \log B}{\sqrt{mB}}\left(\frac{\mathcal{M}_4}{\lambda_1-\lambda_2}\right)^2 + \frac{B\log N}{N\log B}\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^2\right)$$
$$+ O\left(\frac{b_k^2 m^2 \log^2 d \log^3 B}{B}\left(\frac{\mathcal{M}_2}{\lambda_1-\lambda_2}\right)^4 + \frac{\lambda_1 \mathcal{M}_2^2 \log B}{B\left(\lambda_1-\lambda_2\right)^3}\right).$$

Moreover, let K be the set of indices in [d] that satisfy assumptions (55), (56), and (57). Then, for all $k \in K$,

$$\left|\hat{\gamma}_k - e_k^\top \mathbb{V} e_k\right| = O\left(\frac{\mathbb{V}_{kk}}{\sqrt{m}}\right).$$

Proof. By Lemma 22, the bound for any $k \in [d]$, the bound of equation (49) holds with probability $3/4$. By Lemma 10 and the choice $m_1 = 8\log(d/\delta)$, the estimate $\hat{\gamma}_k$ satisfies the equation with probability at least $1-\delta/d$. The Lemma follows by a union bound over the indices in $[d]$. □

Remark 7. *The first term of the error of Lemma 25 is $O\left(\mathbb{V}_{kk}/\sqrt{m}\right)$, where $m = \log n$. We verify that the other terms are smaller asymptotically in n. Since $m = \log n$ and $m_2 = 8\log(20d)$ where $d = poly(n)$,*

$$B = \frac{n}{mm_1} = \Theta\left(\frac{n}{\log n \log d}\right).$$

Therefore, each summand with a $\sqrt{B}$ or B in the denominator of the error of Lemma 25 is $\tilde{O}(1/\sqrt{n})$. It suffices to show that $\frac{1}{\sqrt{m}}$ asymptotically dominates $\frac{B\log N}{N\log B}$. Note that $1 \le \log d \le 5\log n$, $B = \tilde{\Theta}(n)$ and $\log B = \Theta(\log n)$. Therefore,

$$\frac{B\log N}{N\log B} = \frac{\log B + \log m_1 + \log m}{m_1 m \log B} \le \frac{\log B + m_1 + m}{m_1 m \log B}$$
$$= \frac{1}{\log B \log n} + \frac{1}{8\log n \log(20d)} + \frac{1}{8\log(20d)\log B} = O\left(\frac{1}{\log n}\right) = o\left(\frac{1}{\sqrt{m}}\right).$$

C ENTRYWISE ERROR BOUNDS

Lemma 26. *Let the learning rate, η_n, be set according to Lemma 9. Further, for $X_i \sim \mathcal{P}$, $A_i = X_i X_i^\top$, let $\|A_i - \Sigma\|_{\text{op}} \le \mathcal{M}$ almost surely. Then, for $\delta \in (0,1)$, with probability at least $1-\delta$, we have for all $k \in [d]$,*

$$\left|e_k^\top \Psi_{n,1}\right| \lesssim \sqrt{\eta_n\left(e_k^\top V_\perp R_0 V_\perp^\top e_k\right)\log\left(\frac{d}{\delta}\right)} + \eta_n b_k\left(\mathcal{M}\log\left(\frac{d}{\delta}\right) + \mathcal{M}_2\sqrt{\frac{\lambda_1}{\lambda_1-\lambda_2}}\sqrt{\log\left(\frac{d}{\delta}\right)}\right)$$

where $\Psi_{n,1}$ is defined in Lemma 2, $b_k := \left\|V_\perp^\top e_k\right\|_2$, $\widetilde{M} := \mathbb{E}\left[V_\perp^\top\left(A_j-\Sigma\right)v_1 v_1^\top\left(A_j-\Sigma\right)^\top V_\perp\right]$ and $R_0 \in \mathbb{R}^{(d-1)\times(d-1)}$ with entires

$$R_0(k,l) := \frac{\widetilde{M}_{k\ell}}{2\lambda_1 - \lambda_{k+1} - \lambda_{\ell+1}}, \quad \forall k,l \in [d-1]$$

Proof. Using Lemma 11, we have

$$e_k^\top \Psi_{n,1} = \eta_n e_k^\top Y_n = \sum_{j=1}^n \eta_n e_k^\top X_j^n, \text{ where } X_j^n := V_\perp \Lambda_\perp^{n-j} V_\perp^\top (A_j - \Sigma) v_1$$

Let $\alpha_j := \eta_n e_k^\top X_j^n$. Then, note that $\mathbb{E}\left[\alpha_j\right] = 0$. Furthermore,

$$\mathbb{E}\left[\alpha_j^2\right] = \eta_n^2 e_k^\top V_\perp \Lambda_\perp^{n-j}\mathbb{E}\left[V_\perp^\top\left(A_j-\Sigma\right)v_1 v_1^\top\left(A_j-\Sigma\right)^\top V_\perp\right]\Lambda_\perp^{n-j}V_\perp^\top e_k = \eta_n^2 e_k^\top V_\perp \Lambda_\perp^{n-j}\widetilde{M}\Lambda_\perp^{n-j}V_\perp^\top e_k =: \sigma_{jk}^2,$$
$$|\alpha_j| = \left|\eta_n e_k^\top V_\perp \Lambda_\perp^{n-j} V_\perp^\top (A_j - \Sigma) v_1\right| \le \eta_n b_k \left\|\Lambda_\perp^{n-j}\right\|_{\text{op}} \mathcal{M} \le \eta_n b_k \mathcal{M}$$

Therefore, using the fact that α_j are independent of each other, along with Bernstein's inequality, (see e.g. Proposition 2.14 and the subsequent discussion in Wainwright [2019]), we have with probability at least $1-\delta$,

$$\left|e_k^\top \Psi_{n,1}\right| \leq \sqrt{\left(\sum_{j=1}^{n} \sigma_{jk}^2\right) \log\left(\frac{1}{\delta}\right)} + \eta_n \mathcal{M} b_k \log\left(\frac{1}{\delta}\right)$$

Furthermore, considering a union bound over $k \in [d]$, we have for all $k \in [d]$,

$$\left|e_k^\top \Psi_{n,1}\right| \leq \sqrt{\left(\sum_{j=1}^{n} \sigma_{jk}^2\right) \log\left(\frac{d}{\delta}\right)} + \eta_n \mathcal{M} \log\left(\frac{d}{\delta}\right)$$

Finally, using Lemma 1, we have

$$\begin{aligned}
\sum_{j=1}^{n} \sigma_{jk}^2 &= \eta_n^2 e_k^\top \left(\sum_{j=1}^{n} V_\perp \Lambda_\perp^{n-j} \widetilde{M} \Lambda_\perp^{n-j}\right) V_\perp^\top e_k \\
&= \eta_n^2 e_k^\top \mathbb{E}\left[Y_n {Y_n}^\top\right] e_k \\
&= \eta_n^2 e_k^\top V_\perp \left(R^{(n)}\right) V_\perp^\top e_k \\
&= \eta_n^2 e_k^\top V_\perp \left(\frac{R_0}{\eta_n} + \left(R^{(n)} - \frac{R_0}{\eta_n}\right)\right) V_\perp^\top e_k \\
&\leq \eta_n e_k^\top V_\perp R_0 V_\perp^\top e_k + \eta_n^2 b_k^2 \left\|R^{(n)} - \frac{R_0}{\eta_n}\right\|_F \\
&\leq \eta_n e_k^\top V_\perp R_0 V_\perp^\top e_k + \frac{\eta_n^2 b_k^2 \lambda_1 \mathcal{M}_2^2}{(\lambda_1 - \lambda_2)}
\end{aligned}$$

which completes our proof.

□

Lemma 27. *Let the learning rate, η_n, be set according to Lemma 9. Then, for $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\begin{aligned}
\left\|\Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}\right\|_2 \lesssim \frac{\eta_n^2 n \mathcal{M}_2^2 \log d}{\sqrt{\delta}} + \frac{\sqrt{s_n}\eta_n \sqrt{n} \mathcal{M}_2 \log(d)}{\sqrt{\delta}} \\
+ \frac{\log\left(\frac{1}{\delta}\right)}{\delta^3}\left(\sqrt{d}\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) + \frac{\sqrt{\eta_n^3 n}\mathcal{M}_2^2 \log(d)}{\sqrt{\lambda_1 - \lambda_2}}\right)
\end{aligned}$$

and for all $k \in [d]$,

$$\begin{aligned}
\left|e_k^\top\left(\Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}\right)\right| &\leq b_k \left\|\Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}\right\|_2 \\
&\lesssim \frac{b_k \eta_n^2 n \mathcal{M}_2^2 \log d}{\sqrt{\delta}} + \frac{b_k \sqrt{s_n}\eta_n \sqrt{n}\mathcal{M}_2 \log(d)}{\sqrt{\delta}} \\
&\quad + b_k \frac{\log\left(\frac{1}{\delta}\right)}{\delta^3}\left(\sqrt{d}\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) + \frac{\sqrt{\eta_n^3 n}\mathcal{M}_2^2 \log(d)}{\sqrt{\lambda_1 - \lambda_2}}\right)
\end{aligned}$$

where $\Psi_{n,2}, \Psi_{n,3}, \Psi_{n,4}$ are as defined in Lemma 2, $b_k := \left\|V_\perp^\top e_k\right\|_2$ and $s_n := \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^3}\frac{\eta_n \mathcal{M}_2^2}{(\lambda_1 - \lambda_2)}$ for a universal constant $C > 0$.

Proof. We have

$$\left\|\Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}\right\|_2 \leq \left|e_k^\top \Psi_{n,2}\right| + \left|e_k^\top \Psi_{n,3}\right| + \left|e_k^\top \Psi_{n,4}\right| \tag{60}$$

Using Lemma 16, we have for all $k \in [d]$, with probability at least $1 - \frac{\delta}{3}$,

$$\|\Psi_{n,2}\| \leq \frac{12\eta_n^2 \mathcal{M}_2^2 n \log d}{\sqrt{\delta/3}} \leq \frac{21\eta_n^2 \mathcal{M}_2^2 n \log d}{\sqrt{\delta}}. \tag{61}$$

Using Lemma 18 , along with the definition of η_n in Lemma 9, with probability at least $1 - \frac{\delta}{3}$,

$$\begin{aligned}
\|\Psi_{n,3}\|_2 &\lesssim \frac{\sqrt{s_n}\sqrt{\log\left(\frac{1}{\delta}\right)}}{\delta^{\frac{3}{2}}}\left(\sqrt{d}\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) + \frac{\sqrt{\eta_n}\mathcal{M}_2}{\sqrt{\lambda_1 - \lambda_2}}\right) + \sqrt{s_n}\frac{\eta_n\sqrt{n}\mathcal{M}_2 \log(d)}{\sqrt{\delta}} \\
&\lesssim \frac{\sqrt{\log\left(\frac{1}{\delta}\right)}}{\delta^{\frac{3}{2}}}\left(\sqrt{d}\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) + \sqrt{\frac{C\log(1/\delta)}{\delta^3}\frac{\eta_n\mathcal{M}_2^2}{\lambda_1 - \lambda_2}} \cdot \frac{\sqrt{\eta_n}\mathcal{M}_2 \log d}{\sqrt{\lambda_1 - \lambda_2}}\right) \\
&\lesssim \frac{\log\left(\frac{1}{\delta}\right)}{\delta^3}\left(\sqrt{d}\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) + \frac{\sqrt{\eta_n^3 n}\mathcal{M}_2^2 \log(d)}{\sqrt{\lambda_1 - \lambda_2}}\right),
\end{aligned} \tag{62}$$

where the second inequality used $s_n \leq 1$. Using Lemma 20, along with the definition of η_n in Lemma 9, with probability at least $1 - \frac{\delta}{3}$,

$$\|\Psi_{n,4}\|_2 \lesssim \frac{1}{\delta^{\frac{3}{2}}}\left(\sqrt{d}\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) + \frac{\sqrt{\eta_n^3 n}\mathcal{M}_2^2 \log(d)}{\sqrt{\lambda_1 - \lambda_2}}\right) \tag{63}$$

The first result follows by a union bound over (61), (62), (63) and substituting in (60). Finally, note that using Lemma 2, $\exists x_n, y_n, z_n \in \mathbb{R}^{d-1}$ such that of $\Psi_{n,2} = V_\perp V_\perp^\top x_n$, $\Psi_{n,3} = V_\perp V_\perp^\top x_n$, $\Psi_{n,4} = V_\perp V_\perp^\top x_n$. Therefore,

$$\begin{aligned}
\left|e_k^\top\left(\Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}\right)\right| &= \left|e_k^\top V_\perp V_\perp^\top\left(x_n + y_n + z_n\right)\right| \\
&= \left|e_k^\top V_\perp V_\perp^\top V_\perp V_\perp^\top\left(x_n + y_n + z_n\right)\right| \\
&\leq \left\|e_k^\top V_\perp V_\perp^\top\right\|_2 \left\|V_\perp V_\perp^\top\left(x_n + y_n + z_n\right)\right\|_2 \\
&= b_k \left\|\Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}\right\|_2
\end{aligned}$$

which completes the proof of the second result. □

Now we are ready to prove a detailed version of Theorem 1.

Lemma 28. *Let the learning rate, η_n, be set according to Lemma 9. Further, for $X_i \sim \mathcal{P}$, $A_i = X_i X_i^\top$, let $\|A_i - \Sigma\|_{\text{op}} \leq \mathcal{M}$ almost surely. Define $r_{\text{oja}} := v_{\text{oja}} - \left(v_1^\top v_{\text{oja}}\right) v_1$. Then, with probability at least $1 - \delta$, for all $k \in [d]$,*

$$\begin{aligned}
\left|e_k^\top r_{\text{oja}}\right| \lesssim &\sqrt{\eta_n\left(e_k^\top V_\perp R_0 V_\perp^\top e_k\right)\log\left(\frac{d}{\delta}\right)} + \eta_n b_k\left(\mathcal{M}\log\left(\frac{d}{\delta}\right) + \mathcal{M}_2\sqrt{\frac{\lambda_1}{\lambda_1 - \lambda_2}}\sqrt{\log\left(\frac{d}{\delta}\right)}\right) \\
&+ b_k\frac{\log\left(\frac{1}{\delta}\right)}{\delta^3}\left(\sqrt{d}\exp\left(-\eta_n n\left(\lambda_1 - \lambda_2\right)\right) + \frac{\sqrt{\eta_n^3 n}\mathcal{M}_2^2 \log(d)}{\sqrt{\lambda_1 - \lambda_2}}\right) \\
&+ \frac{b_k\eta_n^2 n\mathcal{M}_2^2 \log d}{\sqrt{\delta}} + \frac{b_k\sqrt{s_n}\eta_n\sqrt{n}\mathcal{M}_2 \log(d)}{\sqrt{\delta}}
\end{aligned}$$

where $b_k := \left\|V_\perp^\top e_k\right\|_2$, $s_n := \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^3}\frac{\eta_n\mathcal{M}_2^2}{(\lambda_1 - \lambda_2)}$, $\widetilde{M} := \mathbb{E}\left[V_\perp^\top\left(A_j - \Sigma\right)v_1 v_1^\top\left(A_j - \Sigma\right)^\top V_\perp\right]$ and $R_0 \in \mathbb{R}^{(d-1)\times(d-1)}$ with entires

$$R_0(k,l) = \frac{\widetilde{M}_{k\ell}}{2\lambda_1 - \lambda_{k+1} - \lambda_{\ell+1}}, k, l \in [d-1]$$

.

Proof. Using Lemma 2, we have

$$e_k^\top r_{\text{oja}} := e_k^\top \Psi_{n,1} + e_k^\top \Psi_{n,2} + e_k^\top \Psi_{n,3} + e_k^\top \Psi_{n,4}$$

Therefore,

$$\left|e_k^\top r_{\mathsf{oja}}\right| \leq \left|e_k^\top \Psi_{n,1}\right| + \left|e_k^\top \Psi_{n,2} + e_k^\top \Psi_{n,3} + e_k^\top \Psi_{n,4}\right|$$

The result then following by a union bound over the events defined in Lemma 26 and Lemma 27. □

D CENTRAL LIMIT THEOREM FOR ENTRIES OF THE OJA VECTOR

We consider the following setup from Chernozhukov et al. [2017a]. Let $\mathcal{A}^{\text{re}}$ denote the class of all hyperrectangles in $\mathbb{R}^p$. That is, $\mathcal{A}^{\text{re}}$ consists of all sets A of the form:

$$A = \{w \in \mathbb{R}^p : a_j \leq w_j \leq b_j \text{ for all } j = 1, \ldots, p\} \tag{64}$$

for some real values a_j and b_j satisfying $-\infty \leq a_j \leq b_j \leq \infty$ for each $j = 1, \ldots, p$.

Consider

$$S_n^X = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i.$$

where $X_i, i \in [n] \in \mathbb{R}^p$ are independent random vectors with $\mathbb{E}[X_{ij}] = 0$ and $\mathbb{E}[X_{ij}^2] < \infty$, for $i \in [n], j \in [p]$. Consider the following Gaussian approximation to S_n^X. Define the normalized sum for the Gaussian random vectors:

$$S_n^Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i,$$

where $Y_1, \ldots, Y_n$ be independent mean zero Gaussian random vectors in $\mathbb{R}^p$ such that each Y_i has the same covariance matrix as X_i. We are interested in bounding the quantity

$$\rho_n\left(\mathcal{A}^{\text{re}}\right) := \sup_{A \in \mathcal{A}^{\text{re}}} \left|\mathbb{P}\left(S_n^X \in A\right) - \mathbb{P}\left(S_n^Y \in A\right)\right|$$

Let $C_n \geq 1$ be a sequence of constants possibly growing to infinity as $n \to \infty$, and let $b, q > 0$ be some constants. Assume that X_i satisfy,

(M.1) $n^{-1} \sum_{i=1}^n \mathbb{E}[X_{ij}^2] \geq b$ *for all* $j = 1, \ldots, p$,

(M.2) $n^{-1} \sum_{i=1}^n \mathbb{E}[|X_{ij}|^{2+k}] \leq C_n^k$ *for all* $j = 1, \ldots, p$ *and* $k = 1, 2$.

Further, the authors consider examples where one of the following conditions also holds:

(E.1) $\mathbb{E}[\exp(|X_{ij}|/C_n)] \leq 2$ *for all* $i = 1, \ldots, n$ *and* $j = 1, \ldots, p$,

(E.2) $\mathbb{E}[(\max_{1 \leq j \leq p} |X_{ij}|/C_n)^q] \leq 2$ *for all* $i = 1, \ldots, n$.

Let

$$D_n^{(1)} = \left(\frac{C_n^2 \log^7(pn)}{n}\right)^{1/6}, \quad D_{n,q}^{(2)} = \left(\frac{C_n^2 \log^3(pn)}{n^{1-2/q}}\right)^{1/3}.$$

Now we present Proposition 2.1 [Chernozhukov et al., 2017a].

Theorem 3 (Proposition 2.1 [Chernozhukov et al., 2017a])**.** *Suppose that conditions (M.1) and (M.2) are satisfied. Then under (E.1), we have*

$$\rho_n(\mathcal{A}^{re}) \leq C D_n^{(1)},$$

where the constant C depends only on b; while under (E.2), we have

$$\rho_n(\mathcal{A}^{re}) \leq C\{D_n^{(1)} + D_{n,q}^{(2)}\},$$

where the constant C depends only on b and q.

Next, we will need the following result cited by Chernozhukov et al. [2017b].

Theorem 4 (Nazarov's inequality [Nazarov, 2003], Theorem 1 in [Chernozhukov et al., 2017a]). *Let $Y = (Y_1, \ldots, Y_p)^T$ be a centered Gaussian random vector in $\mathbb{R}^p$ such that*

$$\mathbb{E}[Y_j^2] \geq \sigma^2, \quad \text{for all } j = 1, \ldots, p,$$

for some constant $\sigma > 0$. Then, for every $y \in \mathbb{R}^p$ and $\delta > 0$,

$$\mathbb{P}(Y \leq y + \delta) - \mathbb{P}(Y \leq y) \leq \frac{\delta}{\sigma}(\sqrt{2 \log p} + 2).$$

Here, for vector $y \in \mathbb{R}^p$, $y + \delta$ denotes the vector constructed by adding δ to each entry of y.

Now we are ready to state our main result in Proposition 2,

Proposition 2 (CLT for a suitable subset of entries). *Suppose the learning rate η_n, set according to Lemma 9, satisfies $\frac{\mathcal{M}_2^2 \lambda_1 \eta_n}{(\lambda_1 - \lambda_2)^2} \leq \frac{C_0 b}{2}$ for some $b > 0$ and a small universal constant C_0. Let $\{X_i\}_{i=1}^n \in \mathbb{R}^d$ be i.i.d. mean-zero random vectors with covariance matrix Σ such that for all vectors $v \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[\exp\left(v^T X_1\right)\right] \leq \exp\left(\frac{\sigma^2 v^T \Sigma v}{2}\right).$$

Let $r_{\mathsf{oja}} := v_{\mathsf{oja}} - (v_1^\top v_{\mathsf{oja}}) v_1$. Consider the set $J := \{j : \mathbb{V}_{jj} \geq b\}$, and let $p := |J|$. Let $H_i := \frac{\mathsf{sign}(u_0^T v_1)}{1 + \eta_n \lambda_1} V_\perp \Lambda_\perp^{n-i} V_\perp^\top \left(X_i X_i^\top - \Sigma\right) v_1$. Let $Y_i \in \mathbb{R}^p$ be independent mean zero normal vectors such that

$$\mathbb{E}[Y_i Y_i^T] = \frac{n \eta_n}{\lambda_1 - \lambda_2} \mathbb{E}[H_i[J] H_i[J]^T].$$

Then,

$$\sup_{A \in \mathcal{A}_{re}} \left| P\left(\frac{r_{\mathsf{oja}}[J]}{\sqrt{(\lambda_1 - \lambda_2)\, \eta_n}} \in A\right) - P\left(\frac{\sum_i Y_i}{\sqrt{n}} \in A\right) \right| = \tilde{O}\left(\max\left(\left(\frac{\mathcal{M}_4}{\lambda_1 - \lambda_2}\right)^{1/3} n^{-1/6}, \left(\frac{\mathcal{M}_2}{\lambda_1 - \lambda_2}\right)^{1/2} n^{-1/8}\right)\right),$$

where $\tilde{O}$ hides logarithmic factors in n, p, and constants depending on b.

Proof of Proposition 2. Consider the error decomposition of the Oja vector in Lemma 2. We have $r_{\mathsf{oja}} = \Psi_{n,1} + \Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}$, where $\Psi_{n,1}, \Psi_{n,2}, \Psi_{n,3}, \Psi_{n,4}$ are defined in Equation (19). Let $R := \Psi_{n,2} + \Psi_{n,3} + \Psi_{n,4}$.

For any $\delta \in (0, 1)$, $\exists \epsilon > 0$ such that from Lemma 27 we have,

$$\mathbb{P}((\eta_n (\lambda_1 - \lambda_2))^{-1/2} \|R\|_2 \geq \epsilon) \leq \delta$$

we will specify ϵ as needed in the proof.

For all $i \in [n]$, let

$$U_i := \underbrace{\sqrt{n \eta_n / (\lambda_1 - \lambda_2)}}_{c_n} H_i \tag{65}$$

We show that $U_1, U_2, \ldots, U_n$ satisfy conditions (M.1) and (M.2) with suitable constants.

For (M.1), using equation (19),

$$\sum_{i=1}^n H_i = \Psi_{n,1}. \tag{66}$$

By Lemma 1 (equation (9)), there exists a universal constant C_0 such that

$$\left| e_j^\top \left(\frac{\eta_n}{\lambda_1 - \lambda_2} \sum_{i=1}^n \mathbb{E}\left[H_i H_i^\top\right] - \mathbb{V} \right) e_j \right| \leq \frac{\eta_n \lambda_1 \mathcal{M}_2^2}{C_0 (\lambda_1 - \lambda_2)^2} \leq \frac{b}{2} \leq \frac{\mathbb{V}_{jj}}{2}.$$

for all $j \in J$, where the last two inequalities follow by assumption and definition of J. This implies for all $j \in J$,

$$\frac{\eta_n}{\lambda_1 - \lambda_2} \sum_i^n \mathbb{E}\left[H_{ij}^2\right] \geq \mathbb{V}_{jj}/2 \geq b/2 \iff \frac{1}{n}\sum_i \mathbb{E}[U_{ij}^2] \geq \mathbb{V}_{jj}/2 \geq b/2$$

To show (M.2), by Lyapunov's inequality and Assumption 1:

$$\mathbb{E}\left[\|U_{ij}^{2+k}\|_2\right] = \mathbb{E}\left[c_n^{2+k}|H_{ij}|^{2+k}\right] \leq 2(c_n\mathcal{M}_4)^{2+k}$$

for $k \in \{1, 2\}$, where $C_n := 2c_n\mathcal{M}_4$.

We now check condition E.1. Now note that for any unit vector $u \in \mathbb{R}^d$, $u^T H_i$ is subexponential with parameter $\sigma^2\lambda_1$ (Proposition 2.7.1. of [Vershynin, 2018]). Hence, there exists a constant $C > 0$ such that

$$\mathbb{E}[\exp(|H_{ij}|/C\lambda_1\sigma^2)] \leq 2$$

Therefore,

$$\mathbb{E}[\exp(|U_{ij}|/C\lambda_1 c_n\sigma^2)] \leq 2.$$

Now we set $C_n := \max(2c_n\mathcal{M}_4, C\lambda_1 c_n\sigma^2)$.

Using Eq 66,

$$\frac{1}{\sqrt{(\lambda_1 - \lambda_2)\eta_n}}\Psi_{n,1}[J] = \sqrt{\eta_n/\left(\lambda_1 - \lambda_2\right)}\sum_i H_i[J] = \frac{1}{\sqrt{n}}\sum_i U_i[J],$$

the random variables $U_i[J], i \in [n]$ satisfy conditions (M.1), (M.2) and (E.1). By Theorem 3,

$$\rho(\mathcal{A}^{\text{re}}) \leq C\left(\frac{C_n^2\log^7(pn)}{n}\right)^{1/6}$$

Recall from the statement of the proposition that $Y_i, i \in [n]$ are mean zero independent Gaussian vectors in $\mathbb{R}^p$ with the same covariance structure as $U_i[J]$, i.e, $\mathbb{E}\left[Y_iY_i^\top\right] = \mathbb{E}\left[U_i[J]U_i[J]^\top\right]$.

Let S_W be the random variable $\sum_i W_i$ for any collection W of n random variables $W_1, W_2, \ldots, W_n$. Consider the vector $S_W[J]$ to be the projection of W on the set J, defined as $e_i^\top S_W[J] = e_i^\top S_W$ for $i \in J$.

Recall that

$$e_i^T r_{\mathsf{oja}} := e_i^\top\left(\sum_{j=1}^n \eta_n H_j + R\right).$$

Let $A := \{u \in \mathbb{R}^p | u_i \in [a_i, b_i], i \in J\}$. Let $A_\epsilon^+ := \{X | X_i \in [a_i - \epsilon, b_i + \epsilon], i \in [p]\}$ and $A_\epsilon^- := \{X | X_i \in [a_i + \epsilon, b_i - \epsilon], i \in J\}$.

Let $S_R[J] := \sum_{i\in J} e_i^T r_{\mathsf{oja}}$. Then, we have $S_R[J] = \eta_n S_H[J] + R[J]$.

We will use the following identity for vectors $G_1, G_2 \in \mathbb{R}^p$.

$$\mathbb{P}(G_1 \in A_\epsilon^-, \|G_2\| \leq \epsilon) \leq \mathbb{P}(G_1 + G_2 \in A, \|G_2\| \leq \epsilon) \leq P(G_1 \in A_\epsilon^+, \|G_2\| \leq \epsilon)$$

So,

$$\mathbb{P}(G_1 + G_2 \in A) \leq \mathbb{P}(G_1 \in A_\epsilon^+, \|V\| \leq \epsilon) + P(\|V\| \geq \epsilon)$$
$$\mathbb{P}(G_1 + G_2 \in A) \geq P(G_1 \in A_\epsilon^-, \|G_2\| \leq \epsilon)$$

Using $G_1 = S_U[J]/\sqrt{n}$ and $G_2 = (\eta_n(\lambda_1 - \lambda_2))^{-1/2} R$, we have:

$$\begin{aligned}
&\mathbb{P}(((\lambda_1 - \lambda_2)\eta_n)^{-1/2} r_{\mathsf{oja}}[J] \in A) - \mathbb{P}(S_Y/\sqrt{n} \in A) \\
&\leq \mathbb{P}(((\lambda_1 - \lambda_2)\eta_n)^{-1/2} r_{\mathsf{oja}}[J] \in A, (\eta_n(\lambda_1 - \lambda_2))^{-1/2}\|R\| \leq \epsilon) + \mathbb{P}((\eta_n(\lambda_1 - \lambda_2))^{-1/2}\|R\|_2 \geq \epsilon) \\
&\quad - \mathbb{P}(S_Y/\sqrt{n} \in A) \\
&\leq \mathbb{P}(S_U[J]/\sqrt{n} \in A_\epsilon^+) + \mathbb{P}((\eta_n(\lambda_1 - \lambda_2))^{-1/2}\|R\|_2 \geq \epsilon) - \mathbb{P}(S_Y/\sqrt{n} \in A) =: \gamma_A.
\end{aligned}$$

Note that γ_A can be written as

$$\gamma_A \leq |\mathbb{P}(S_U[J]/\sqrt{n} \in A_\epsilon^+) - \mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^+)| + |\mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^+) - \mathbb{P}(S_Y/\sqrt{n} \in A)| + \mathbb{P}((\eta_n(\lambda_1 - \lambda_2))^{-1/2}\|R\| \geq \epsilon).$$

Similarly,

$$\mathbb{P}(((\lambda_1 - \lambda_2)\eta_n)^{-1/2} r_{\mathsf{oja}}[J] \in A) - \mathbb{P}(S_Y/\sqrt{n} \in A) \geq \omega_A,$$

where

$$\begin{aligned}
\omega_A &:= \mathbb{P}(S_U[J]/\sqrt{n} \in A_\epsilon^-, (\eta_n(\lambda_1 - \lambda_2))^{-1/2}\|R\| \geq \epsilon) - \mathbb{P}(S_Y/\sqrt{n} \in A) \\
&\geq \mathbb{P}(S_U[J]/\sqrt{n} \in A_\epsilon^-) - \mathbb{P}((\eta_n(\lambda_1 - \lambda_2))^{-1/2}\|R\| \geq \epsilon) - \mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^-) + \mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^-) - \mathbb{P}(S_Y/\sqrt{n} \in A)
\end{aligned}$$

Therefore, we have by Theorem 3 that for some constant C' that depends only on b,

$$\sup_{A \in \mathcal{A}_{re}} |\gamma_A| \leq C' \left(\frac{C_n^2 \log^7(pn)}{n}\right)^{1/6} + \left|\mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^+) - \mathbb{P}(S_Y/\sqrt{n} \in A)\right| + \delta \tag{67}$$

Similarly,

$$\sup_{A \in \mathcal{A}_{re}} |\omega_A| \leq C' \left(\frac{C_n^2 \log^7(pn)}{n}\right)^{1/6} + \left|\mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^-) - \mathbb{P}(S_Y/\sqrt{n} \in A)\right| + \delta \tag{68}$$

For $\mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^+) - \mathbb{P}(S_Y/\sqrt{n} \in A)$, we will use Nazarov's inequality (Lemma 4):

$$\left|\mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^+) - \mathbb{P}(S_Y/\sqrt{n} \in A)\right| \leq \frac{\sqrt{2}\epsilon}{b^{1/2}}(\sqrt{2\log p} + 2) \tag{69}$$

For bounding the terms concerning A_ϵ^-, we need to be a little careful because if $b_i - a_i \leq 2\epsilon$, then A_ϵ^- has measure zero under the Gaussian distribution. If A_ϵ^- is nonempty, then we have the same bound as Eq 69. However, in case that is not true, note that there must be some $i \in [p]$ such that $b_i - a_i \leq 2\epsilon$. Hence

$$\begin{aligned}
\left|\mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^-) - \mathbb{P}(S_Y/\sqrt{n} \in A)\right| &= \mathbb{P}(S_Y/\sqrt{n} \in A) \\
&= \mathbb{P}(S_Y[i]/\sqrt{n} \in [a_i, b_i]) \\
&\leq \frac{2\epsilon}{\sqrt{\pi} b^{1/2}}
\end{aligned} \tag{70}$$

So overall,

$$\begin{aligned}
\left|\mathbb{P}(S_Y/\sqrt{n} \in A_\epsilon^-) - \mathbb{P}(S_Y/\sqrt{n} \in A)\right| &= \mathbb{P}(S_Y/\sqrt{n} \in A) \\
&= \mathbb{P}(S_Y[i]/\sqrt{n} \in [a_i, b_i]) \\
&\leq \max\left(\frac{2\epsilon}{\sqrt{\pi} b^{1/2}}, \frac{\sqrt{2}\epsilon}{b^{1/2}}(\sqrt{2\log p} + 2)\right)
\end{aligned} \tag{71}$$

Putting Eqs 67, 68, 69 and 71 together, we have, for some absolute constant C_1:

$$\begin{aligned}
&\sup_{A \in \mathcal{A}_{re}} |\mathbb{P}(((\lambda_1 - \lambda_2)\eta_n)^{-1/2} r_{\mathsf{oja}}[J] \in A) - \mathbb{P}(n^{-1/2} S_Y \in A)| \leq \max(\sup_{A \in \mathcal{A}_{re}} |\gamma_A|, \sup_{A \in \mathcal{A}_{re}} |\omega_A|) \\
&\lesssim \left(\frac{C_n^2 \log^7(pn)}{n}\right)^{1/6} + \frac{C_1 \epsilon}{b^{1/2}}\sqrt{\log p} + \delta
\end{aligned} \tag{72}$$

We invoke Lemma A.2.3 in Kumar and Sarkar [2024a] to see that: $\mathcal{M}_4 \leq \lambda_1 + \sigma^2 \text{trace}\left(\Sigma\right)$. Therefore, for some constant $C'' > 0$,

$$C_n = \max(2c_n\mathcal{M}_4, C\lambda_1 c_n \sigma^2) \leq C'' \sqrt{\frac{n\eta_n}{(\lambda_1 - \lambda_2)}}\mathcal{M}_4$$

From Lemma 27 and the assumption on the learning rate (Lemma 9),

$$\sqrt{\eta_n\left(\lambda_1 - \lambda_2\right)}\epsilon \lesssim \frac{\eta_n^2 n \mathcal{M}_2^2 \log d}{\sqrt{\delta}} + \frac{\sqrt{s_n}\eta_n\sqrt{n}\mathcal{M}_2 \log\left(d\right)}{\sqrt{\delta}} + \frac{\log\left(\frac{1}{\delta}\right)}{\delta^3}\left(\frac{\sqrt{\eta_n^3 n}\mathcal{M}_2^2 \log\left(d\right)}{\sqrt{\lambda_1 - \lambda_2}}\right) \tag{73}$$

Substituting the bound on ϵ from equation (73) into equation (72) and optimizing over δ yields

$$\delta = \tilde{O}\left(\left(\frac{\log p}{b}\right)^{1/8}\sqrt{\frac{\mathcal{M}_2}{\lambda_1 - \lambda_2}}n^{-1/8}\right). \tag{74}$$

Substituting the choice of δ from equation (74) in (72), we conclude

$$\begin{aligned}
&\sup_{A \in \mathcal{A}^{\text{re}}} \left|\mathbb{P}(((\lambda_1 - \lambda_2)\,\eta_n)^{-1/2} r_{\mathsf{oja}}[J] \in A) - \mathbb{P}(n^{-1/2} S_Y \in A)\right| \\
&\quad = \tilde{O}\left(\max\left(\left(\frac{\mathcal{M}_4}{\lambda_1 - \lambda_2}\right)^{1/3} n^{-1/6}, \left(\frac{\mathcal{M}_2}{\lambda_1 - \lambda_2}\right)^{1/2} n^{-1/8}\right)\right)
\end{aligned}$$

□