Bootstrapping the error of Oja's algorithm

Robert Lunde
University of Michigan
rlunde@umich.edu

Purnamrita Sarkar
University of Texas at Austin
purna.sarkar@austin.utexas.edu

Rachel Ward
University of Texas at Austin
rward@math.utexas.edu

Abstract

We consider the problem of quantifying uncertainty for the estimation error of the leading eigenvector from Oja's algorithm for streaming principal component analysis, where the data are generated IID from some unknown distribution. By combining classical tools from the U-statistics literature with recent results on high-dimensional central limit theorems for quadratic forms of random vectors and concentration of matrix products, we establish a weighted χ^2 approximation result for the $\sin^2$ error between the population eigenvector and the output of Oja's algorithm. Since estimating the covariance matrix associated with the approximating distribution requires knowledge of unknown model parameters, we propose a multiplier bootstrap algorithm that may be updated in an online manner. We establish conditions under which the bootstrap distribution is close to the corresponding sampling distribution with high probability, thereby establishing the bootstrap as a consistent inferential method in an appropriate asymptotic regime.

1 Introduction

Since its discovery over a century ago [13], principal component analysis (PCA) has been a cornerstone of data analysis. In many applications, dimension reduction is paramount and PCA offers an optimal low-rank approximation of the original data. PCA is also highly interpretable as it projects the dataset onto the directions that capture the most variance known as principal components.

Important applications of PCA include image and document analysis, where the largest few principal components may be used to compress a large dimensional dataset to a manageable size without incurring much loss; for a discussion of some other applications of PCA, see for example, [28]. In these settings, the original dimensionality, which could be the number of pixels in an image or the vocabulary size after removing stop-words, is in the tens of thousands. An offline computation of the principal components would require the computation of eigenvectors of the sample covariance matrix. However, in high-dimensional settings, storing the covariance matrix and subsequent eigen-analysis can be challenging. Streaming PCA methods have gained significant traction owing to their ability to iteratively update the principal components by considering one data-point at a time.

One of the most widely used algorithms for streaming PCA is Oja's algorithm, proposed in the seminal work of [41]. Oja's algorithm involves the following update rule:

$$w_{t+1} - w_t = \eta(w_t^T X_t)X_t; \qquad w_{t+1}^T w_{t+1} = 1, \tag{1}$$

where $X_t \in \mathbb{R}^d$ is the t^{th} data point and w_t is the current estimate for the leading eigenvector of $\Sigma = \mathbb{E}XX^T$ after t data-points have been seen. The parameter η can be thought of as a learning rate, which can either be fixed or varied as a function of t. In this paper we fix the learning rate, similar to [26].

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Contribution: In the present work, we consider the problem of uncertainty quantification for the estimation error of the leading eigenvector from Oja's algorithm, which is one of the most commonly used streaming PCA algorithms. Our contributions may be summarized as follows:

1. We derive a high-dimensional weighted χ^2 approximation to the $\sin^2$ error for the leading eigenvector of Oja's algorithm. We recover the optimal convergence rate $O(1/n)$ while allowing d to grow at a sub-exponential rate under suitable structural assumptions on the covariance matrix, matching state-of-the-art theoretical results for consistency of Oja's algorithm. Our result provides a distributional characterization of the $\sin^2$ error for Oja's algorithm for the first time in the literature. The approximation holds for a wide range of step sizes.
2. Since the weighted χ^2 approximation depends on unknown parameters, we propose an online bootstrap algorithm and establish conditions under which the bootstrap is consistent. Our bootstrap procedure allows the approximation of important quantities such as the quantiles of the error associated with Oja's algorithm for the first time.

Prior analysis of Oja's algorithm. While Oja's algorithm was invented in 1982 it was not until recently that the theoretical workings of Oja's algorithm have been understood. A number of papers in recent years have focused on proving guarantees of convergence of the iterative update in (1) toward the principal eigenvector of the (unknown) covariance matrix $\mathbb{E}XX^T$, which can be recast as stochastic gradient descent (SGD) on the quadratic objective function

$$\min_{\substack{w \\ w^T w = 1}} -\mathrm{trace}(w^T \Sigma w), \qquad \Sigma = \mathbb{E}XX^T, \tag{2}$$

projected onto the non-convex unit sphere. We assume that the data-points are mean zero. Despite being non-convex and thus falling outside the framework for which theory for stochastic gradient descent convergence is firmly established, the output of Oja's algorithm be viewed as a product of random matrices and shares similar structure to other important classes of non-convex problems, such as matrix completion [27, 29], matrix sensing [27], and subspace tracking [4]. Thus, studying this optimization problem serves as a natural first step toward understanding the behavior of SGD in more general non-convex settings.

Let v_1 denote the principal eigenvector of Σ, and let $\hat{v}_1 = w_n$ be the solution to the stochastic iterative method applying Eq 1. Finally, let $\lambda_1 > \lambda_2$ be the first and second principal eigenvalues of Σ. Sharp rates of convergence for Oja's updates were established in [25]. Under boundedness assumptions on $\|X_i X_i^T - \Sigma\|$, they show that with constant probability, the square of the sine of the angle between v_1 and w satisfies:

$$1 - (v_1^T \hat{v}_1)^2 = O\left(\frac{1}{n}\right) \tag{3}$$

where the O hides a constant which depends in the optimal way on the eigengap between the top two eigenvalues, and independent of n or d, improving on previous error bounds for Oja's algorithm [46, 18, 3, 47, 38, 2] which showed convergence rates that deteriorate with the ambient dimension d, and thus did not fully explain the efficiency of Oja's update. This sharp rate is remarkable, as it matches the error of the principal eigenvector of the sample covariance matrix, which is the batch or offline version of PCA. Other notable work include [31, 33] for unbounded X_i, analysis of Oja's algorithm for computing top k principal components [1, 24].

The bootstrap. The bootstrap, proposed by [9], is one of the most widely used methods for uncertainty quantification in machine learning and statistics and accordingly has a vast literature. We refer the reader to [17, 49] for expositions on the classical theory of the bootstrap for IID data. Recently, since the groundbreaking work of [7, 8], the bootstrap has seen a renewed surge of interest in the context of high-dimensional data where d can be potentially exponentially larger than n. Of particular relevance to the present work are high-dimensional central limit theorems (CLTs) for quadratic forms, which have been studied by [43, 51, 15]. In particular, our CLT for the estimation error of Oja's algorithm invokes a modest adaptation of [51] to independent but non-identically distributed random variables. In machine learning, bootstrap methods have been used to estimate the uncertainty of randomized algorithms such as bagging and random forests [35], sketching for large scale singular value decomposition (SVD) [36], randomized matrix multiplication [37], and randomized least squares [34].

A standard notion of bootstrap consistency is that, conditioned on the data, the distribution of the suitably centered and scaled bootstrap functional approaches the true distribution with high probability in some norm on probability measures, typically the Kolmogorov distance, which is the supremum of the absolute pointwise difference between two CDFs. Bootstrap consistency is often established by deriving a Gaussian approximation for the sampling distribution and showing that the bootstrap distribution is close to the corresponding Gaussian approximation with high probability.

It may seem that if one knows that the approximating distribution of a statistic is Gaussian, this defeats the purpose of bootstrap. However, *for most statistics, the parameters of the normal approximation depend on unknown model parameters, and have to be estimated if one intends to use the normal approximation*. Furthermore, the CLT only gives a *first-order* correct approximation of the target distribution, i.e. with $O(1/\sqrt{n})$ error. In contrast, the bootstrap of a suitably centered and scaled statistic has been shown to be higher order correct for many functionals [16, 17, 19].

Quantifying uncertainty for SGD. Behind the recent success of neural networks in a wide range of sub-fields of machine learning, the workhorse algorithm has become Stochastic gradient descent (SGD) [42, 40, 44]. For establishing consistency of bootstrap, one requires to establish asymptotic normality [11, 42, 45, 39]. There has also been many works on uncertainty estimation of SGD [6, 32, 12, 48]. However, *all these works are for convex, and predominantly strongly convex loss functions.* Only recently, [52] has established asymptotic normality for nonconvex loss functions under dissipativity conditions and appropriate growth conditions on the gradient, which are weaker conditions than strong convexity but not significantly so.

Now, in Section 2 we present notation and do setup, present our main theoretical results in Section 3, followed by simulations in Section 4.

2 Preliminaries

We consider a row-wise IID triangular array, where the random vectors $\{X_i\}$ in the n^{th} row take values in $\mathbb{R}^{d_n}$, with $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = \Sigma_n$. Note that the triangular array allows $\{X_1, \ldots, X_n\}$ to come from a different distribution for each n and the setting where d is fixed and n grows is a special case. For readability, we drop the subscript n from Σ_n. We use $\|\cdot\|$ to denote the Euclidean norm for vectors and the operator norm for matrices and $\|\cdot\|_F$ to denote the Frobenius norm.

Expanding out the recursive definition in Eq 1, we see that Oja's iteration can be expressed as $w_{t+1} = (I_d + \eta X_t X_t^T) w_t$. Thus, after n iterations the vector can be written as a matrix-vector product, where the matrix is a product of n independent matrices. Expanding out the recursive definition, we get:

$$B_n := \prod_{i=1}^{n} (I_d + \eta X_i X_i^T) \qquad \hat{v}_1 = \frac{B_n u_0}{\|B_n u_0\|}, \tag{4}$$

where I_d is a $d \times d$ identity matrix. where u_0 is a random unit vector in d dimensions. In the scalar case, when $\eta = 1/n$, for large n, the numerator of Eq 4 behaves like $\exp(\sum_i X_i^2/n)$, which in turn converges to $\exp(E[X_1^2])$. For matrices, one hopes that, by independence, a result of the same flavor will hold. And in fact if it does hold, then for $\eta = \frac{\log n}{n}$, the numerator in Eq 4 will concentrate around $\exp(\log n \Sigma)$. The spectrum of this matrix is dominated by the principal eigenvector, i.e. the ratio of the first eigenvalue to the second one is $\exp(\log n(\lambda_1 - \lambda_2))$, where λ_i is the i^{th} eigenvalue of the covariance matrix Σ. This makes it clear that *Oja's algorithm is essentially a matrix vector product of this matrix exponential (suitably scaled) and a random unit vector.*

However, the intuition from the scalar case is nontrivial to generalize to matrices due to non-commutativity. Limits of products of random matrices have been studied in mathematics in the context of ergodic theory on Markov chains (see [14, 30, 5, 10] etc.). However, until recent results of [23], which extended and improved results in [21], there has not been much work on quantifying the exact *rate of convergence*, or finite-sample large deviation bounds for how a random matrix product deviates from its expectation.

We reparametrize η as η_n/n, where η_n is chosen carefully to obtain a suitable error rate. Note that this is not a scheme where we decrease η over time as in [20], but hold it as a constant which is a function of the total number of data-points.

2.1 The Hoeffding decomposition

The Hoeffding decomposition, attributed to [22], is a key technical tool for studying the asymptotic properties of U-statistics. However, the idea generalizes far beyond U-statistics; see Supplement Section A for further discussion. In the present work, we use Hoeffding decompositions for matrix and vector-valued functions of independent random variables taking values in $\mathbb{R}^d$ to facilitate analysis for B_n.

A concept closely related to the Hoeffding decomposition is the more well-known Hájek projection, which gives the best approximation (in an L_2 sense) of a general function of n independent random variables by a function of the form $\sum_i g_i(X_i)$, where g_i are measurable functions satisfying a square integrability condition. The Hájek projection facilitates distributional approximations for complicated statistics since this linear projection is typically more amenable to analysis. However, establishing a central limit theorem requires showing the negligibility of a remainder term, which can be large if the projection is not accurate enough.

The Hájek projection may be viewed as the first-order term in the Hoeffding decomposition, a general way of representing functions of independent random variables. The Hoeffding decomposition consists of a sum of projections onto a linear space, quadratic space, cubic space, and so on. Each new space is chosen to be orthogonal to the previous space. Thus, the Hoeffding decomposition can be thought of as a sum of terms of increasing levels of complexity. Even if the remainder of the Hájek projection turns out to be small, the Hoeffding decomposition can be easier to work with due to the orthogonality of the projections.

The Hoeffding decomposition for the matrix product. Let $Y_i = X_i X_i^T - \Sigma$ and let $S \subseteq \{1, \dots n\}$. By Corollary A.1 of the Supplement Section A, the Hoeffding Decomposition for B_n is given by:

$$B_n = \sum_{k=0}^{n} T_k, \qquad T_k = \sum_{|S|=k} H^{(S)}. \tag{5}$$

where $H^{(S)} = \prod_{i=1}^n A_i^{(S)}$ and $A_i^{(S)}$ is given by: $A_i^{(S)} = \begin{cases} \frac{\eta_n}{n} Y_i & \text{if } i \in S \\ I + \frac{\eta_n}{n}\Sigma & \text{otherwise} \end{cases}$.

The above expansion has favorable properties that facilitate second-moment calculations. In fact, as a consequence of the orthogonality property of Hoeffding projections, we have that

$$\mathbb{E}\left[\|B_n\|_F^2\right] = \sum_{k=0}^{n} \sum_{|S|=k} \mathbb{E}\left[\|\prod_{i=1}^{n} H_i^{(S)}\|_F^2\right]$$

$$\mathbb{E}\left[\|B_n x\|^2\right] = \sum_{k=0}^{n} \sum_{|S|=k} \mathbb{E}\left[\|\prod_{i=1}^{n} H_i^{(S)} x\|^2\right]$$

where the second statement holds for any $x \in \mathbb{R}^d$; see Proposition A.2 in Supplement Section A.

2.2 Online bootstrap for streaming PCA

To approximate the sampling distribution, we consider a Gaussian multiplier bootstrap procedure. As observed by [7], a Gaussian multiplier random variable eliminates the need to establish a Gaussian approximation for the bootstrap since conditional on the data, it is already Gaussian. It is not hard to see that this is a natural candidate for the online setting; the multiplier bootstrap has been used for bootstrapping the stochastic gradient descent estimator in [12].

We present our bootstrap in Algorithm 1. In our procedure, we update $m+1$ vectors at every iteration. The first one is $\hat{v}$, which will result in the final Oja estimate of the first principal component. The other vectors $\{v^{*(j)}, j = 1, \dots m\}$ are obtained by perturbing the basic Oja update (Eq 1).

The W_i's are the multiplier random variables, which are scaled mean zero scaled Gaussians with variance $1/2$. The update of the $v^{*(j)}$ is novel because it preserves the mean and the variance of the original Oja estimator while not requiring access to the full sample covariance matrix. Consequently, we can make our updates online and attain both a point estimate and a confidence interval for the principal eigenvector, *while increasing the computation and storage by only a factor of* m.

Algorithm 1: Bootstrap for Oja's algorithm

Input: Datapoints $X_1, \ldots, X_n$, stepsize η, number of bootstrap replicates m
Output: Oja's solution $\hat{v}_1$ and m bootstrapped versions of it $v_1^{*(1)}, \ldots, v_1^{*(m)}$
Draw $g \sim N(0, I_d)$
Create unit vector $u_0 \leftarrow g/\|g\|$
Initialize $\hat{v}_1, v_1^{*(1)}, \ldots, v_1^{*(m)} \leftarrow u_0$
for *t=2,..., n* **do**
  Update $\hat{v}_1 \leftarrow \hat{v}_1 + \eta(X_t^T \hat{v}_1)\hat{v}_1$
  Normalize $\hat{v}_1$ to have unit norm;
  for *i=1:m* **do**
    Draw $W_i \sim N(0, 1/2)$;
    Let $h^{(i)} \leftarrow (X_t^T v_1^{*(i)}) X_t$;
    Let $g^{(i)} \leftarrow (X_{t-1}^T v_1^{*(i)}) X_{t-1}$;
    Update $v_1^{*(i)} \leftarrow v_1^{*(i)} + \eta\left(h^{(i)} + W_i(h^{(i)} - g^{(i)})\right)$;
    Normalize $v_1^{*(i)}$ to have unit norm;
  end
end

3 Main results

In this section we present our main contributions: a CLT for the error of Oja's algorithm and consistency of an online multiplier bootstrap for error.

3.1 Central limit theorem for the error of Oja's algorithm

We start by stating a CLT for the error of Oja's algorithm. To state this theorem, we will need to introduce some notation.

Let $\hat{v}_1$ denote the Oja vector and $V_\perp$ the $d \times d-1$ matrix with $2, \ldots, d$ eigenvectors of Σ on its columns. Note that $V_\perp$ is not uniquely defined, but $V_\perp V_\perp^T = I - v_1 v_1^T$ is if the leading eigenvalue is distinct and consequently, norms of the form $\|V_\perp^T x\|$ for $x \in \mathbb{R}^d$ are well-defined. Let $\lambda_1 \geq \cdots \geq \lambda_d$ denote the eigenvalues of Σ and $\Lambda_\perp$ be a diagonal matrix with $\Lambda_\perp(i,i) = (1 + \eta_n \lambda_{i+1}/n)/(1 + \eta_n \lambda_1/n)$, $i = 1, \ldots, d-1$. Also let

$$\mathbb{M} := \mathbb{E}\left[V_\perp^T (X_1^T v_1)^2 X_1 X_1^T V_\perp\right] \tag{6}$$

Now we define

$$\begin{aligned}\bar{\mathbb{V}}_n &= \frac{\eta_n}{n} \sum_i \mathbb{E}[V_\perp \Lambda_\perp^{i-1} V_\perp^T (X_i X_i^T - \Sigma) v_1 v_1^T (X_i X_i^T - \Sigma) V_\perp \Lambda_\perp^{i-1} V_\perp^T] \\ &= \frac{\eta_n}{n} V_\perp \left(\sum_i \Lambda_\perp^{i-1} \mathbb{M} \Lambda_\perp^{i-1}\right) V_\perp^T \end{aligned} \tag{7}$$

We have the following result:

Theorem 1. *Suppose that u_0 is drawn from the uniform distribution on $\mathcal{S}^{d-1}$, $\lambda_1 = O(1)$. Choose $\eta_n \to \infty$ such that $nd \cdot \exp(-\eta_n(\lambda_1 - \lambda_2)) \to 0$, $\frac{(\eta_n \vee \log d)\, \eta_n^2 (M_d^2 \vee 1)}{n} \to 0$, where $M_d = \mathbb{E}[\left\|X_i X_i^T - \Sigma\right\|^2]$. Further, let $\widetilde{Z}_n$ be a mean 0 Gaussian matrix such that Var$(\widetilde{Z}_n) =$ Var$((X_1 X_1^T - \Sigma)v_1)$ and suppose that:*

$$\|\mathbb{M}\|_F \geq c > 0 \tag{8}$$

$$\frac{\mathbb{E}\left[\left\|V_\perp^T \widetilde{Z}_n\right\|^6\right] \vee \mathbb{E}\left[\left\|V_\perp^T (X_1 X_1^T - \Sigma) v_1\right\|^6\right]}{\|\mathbb{M}\|_F^3} = o(n) \tag{9}$$

Then, for a sequence of Gaussian distributions $\{Z_n\}_{n\geq 1}$ with mean 0 and covariance matrix $\bar{\mathbb{V}}_n$ (see Eq 7), the following holds:

$$\sup_{t\in\mathbb{R}} \left| P\left(n/\eta_n \ \sin^2(\hat{v}_1, v_1) \leq t\right) - P(Z_n^T Z_n \leq t)\right| \to 0 \tag{10}$$

Theorem 1 is very general. We allow the dimension to grow with the number of observations, which is typical in the high-dimensional bootstrap literature. Note that the case of fixed d and growing n is also a special case of this setup.

We want to point out that while previous literature obtained sharp bounds on the $\sin^2$ error $1-(v_1^T\hat{v}_1)^2$, we go a step further. *We establish an approximating distribution for $n/\eta_n(1-(v_1^T\hat{v}_1)^2)$.*

Remark 1 (Condition on norm)**.** *For simplicity, we assume $\lambda_1 = O(1)$, which can be easily relaxed to grow slowly with n. We do not assume that the $\|X_iX_i^T - \Sigma\|_2$ is bounded almost surely. However, the norm of $X_iX_i^T - \Sigma$ comes into play implicitly via the assumption in Eq 9. Consider the case where X_i are drawn from some multivariate Gaussian distribution. We use this to build intuition about the assumptions in Eq 8 and 9. In this case, $X_1^TV_\perp$ is a Gaussian of independent entries and thus $\mathbb{E}\left[\left\|V_\perp(X_1X_1^T-\Sigma)v_1\right\|^6\right] = \mathbb{E}\|X_1^Tv_1\|^6\mathbb{E}\left(\sum_{j>1}(X_j^Tv_j)^2\right)^3$. Note that $\sum_{j>1}((X_j^Tv_j)^2-\lambda_j)$ is a sub-exponential random variable with parameters $(c_1\sum_{j>1}\lambda_j, c_2)$. Furthermore, $\|\mathbb{M}\|_F^2 = \lambda_1\sum_{i>1}\lambda_i$. Thus Eq 9 reduces to checking if*

$$\frac{\lambda_1^{3/2}(\sum_{j>1}\lambda_j)^3}{(\sum_i \lambda_i^2)^{3/2}} = o(n)$$

Remark 2 (Coordinates with summable sub-Gaussian parameters)**.** *Eq 9 imposes a growth condition on the moments of both the data and a Gaussian analog. One setting for which both growth rates are in fact bounded is if the coordinates of X are sub-Gaussian and the sub-Gaussian parameters satisfy $\sum_{i=1}^d \nu_i < C < \infty$ following similar arguments to Proposition 1.*

Remark 3 (Constant vs Adaptive Learning Rate)**.** *Adaptive learning rates are also commonly studied in the literature on Oja's algorithm and have the advantage that they require no prior knowledge of the sample size. It should be noted that our results hold for a wide range of learning rates, ranging from $\log(nd) \ll \eta_n \ll n^{1/3}$, so our results will still apply so long as in the initial guess of the sample size is not off by orders of magnitude. We leave a detailed study of the adaptive learning rate setting to future work.*

As a corollary of our main theorem, we obtain the following error bound on the $\sin^2$ error.

Corollary 1. *Under the conditions in Theorem 1, we have*

$$\sin^2(\hat{v}_1, v_1) = O_P\left(\frac{\eta_n M_d}{n(\lambda_1-\lambda_2)}\right)$$

Remark 4 (Comparison with previous work)**.** *As a byproduct of our analysis, we recover the sharpest convergence rates for Oja's algorithm in the literature. If we set $\eta_n = c_1 \log nd/(\lambda_1-\lambda_2)$, for large enough c_1, the dominating term in the error is $O_P\left(\dfrac{M_d\log nd}{n(\lambda_1-\lambda_2)^2}\right)$ under mild conditions on d. This matches the bound in [25].*

Remark 5 (Rate of convergence in Kolmogorov distance)**.** *To simplify the theorem statement, we have stated Theorem 1 without giving an explicit rate of convergence in the Kolmogorov distance. Convergence rates depend on the rate of decay of the remainder terms, which are worked out in Supplement Section B.3, and the magnitude of the quantity in Eq 9. The contribution of the latter quantity to the rate is worked out in the IID case in [51].*

Remark 6 (Lower bound on norm)**.** *While our rate matches the sharp bounds in literature and our assumptions on norm upper bounds are similar or weaker than previous work, we do assume a lower bound on the Frobenius norm of the covariance matrix as in Eq 8. Note that if indeed all X_i's were a scalar multiple of v_1, then the $\bar{\mathbb{V}}_n$ matrix in Eq 7 will be zero. This will lead to a perfect point estimate, but there will not be any variability from the data and hence there will be no non-degenerate approximation. The lower bound on the norm is not resulting from loose analysis. Similar lower bounds on the variance are imposed in the high-dimensional CLT literature [7, 8].*

Now we provide a proof sketch of Theorem 1 below.

Proof sketch for Theorem 1. We provide the main steps in our derivation. The detailed calculations can be found in Supplement Section B.

1. We start by expressing the $\sin^2$ error as a quadratic form:

$$\begin{aligned}\sin^2(v_1, \hat{v}_1) = 1 - \frac{u_0^T B_n^T v_1 v_1^T B_n u_0}{u_0^T B_n^T B_n u_0} &= \frac{u_0^T B_n^T (I - v_1 v_1^T) B_n u_0}{u_0^T B_n^T B_n u_0} \\ &= \frac{(V_\perp V_\perp^T B_n u_0)^T (V_\perp V_\perp^T B_n u_0)}{\|B_n u_0\|^2}\end{aligned} \tag{11}$$

   where in the last line we used the fact that $V_\perp V_\perp^T$ is idempotent. Our proof strategy for the central limit theorem involves further approximating Eq 11 with an inner product of the Hájek projection (first-order) term in Eq 5.

2. Our second step is to show that $\|B_n u_0\|$ concentrates around its expectation $(1+\eta_n\lambda_1/n)^n |v_1^T u_0|$.

3. Next we establish that $\frac{\|V_\perp V_\perp B_n V_\perp V_\perp^T u_0\|_2}{\|B_n u_0\|}$ is $O_P\left(\sqrt{d} \cdot \exp\{-\eta_n(\lambda_1 - \lambda_2)\} + \sqrt{\frac{\eta_n^3 M_d^2 \log d}{n^2}}\right)$. This is achieved by using a similar recursive argument as in [25], but with the crucial observation that the residual or common difference term is of a lower order because it can be replaced by a matrix product minus its expectation.

4. Now we go back to the expansion in Eq 5.

$$(v_1^T u_0) V_\perp V_\perp^T B_n v_1 = (v_1^T u_0) \sum_k V_\perp V_\perp^T T_k v_1$$

   Since $T_0 = (I + \eta_n/n\Sigma)^n$, $V_\perp V_\perp^T T_0 v_1$ is the zero vector. Now we examine the $(v_1^T u_0) V_\perp V_\perp^T (B_n - T_1) v_1$ term. Here we use the structure of the higher order terms T_k. In particular, we use the fact that it is a matrix product interlaced with k $X_i X_i^T - \Sigma$ matrices. For example, for $k = 2$ we have

$$T_2 = \frac{\eta_n^2}{n^2} \sum_{i<j} \left(I + \frac{\eta_n}{n}\Sigma\right)^{i-1} Y_i \left(I + \frac{\eta_n}{n}\Sigma\right)^{j-i-1} Y_j \left(I + \frac{\eta_n}{n}\Sigma\right)^{n-j}$$

   We show that the norm of $(v_1^T u_0) V_\perp V_\perp^T (B_n - T_1) v_1$, normalized by the denominator, is $O(\eta_n^2 M_d^2/n^2)$. The fact that the summands of T_k are uncorrelated and T_k and T_ℓ are uncorrelated for $k \neq \ell$ makes this possible.

5. Finally, we are left with $V_\perp V_\perp^T T_1 v_1 (v_1^T u_0)$. Note that this is of the following form:

$$\frac{\eta_n}{n} \frac{(v_1^T u_0) V_\perp V_\perp^T T_1 v_1}{|v_1^T u_0|(1 + \lambda_1 \eta_n/n)^n} = \frac{\eta_n \text{sgn}(v_1^T u_0)}{n} \sum_{i=1}^n V_\perp \Lambda_\perp^{i-1} V_\perp^T (X_i X_i^T - \Sigma) v_1$$

   It is not hard to see that this is a sum of independent random vectors with covariance matrix $\eta_n/n\bar{\mathbb{V}}_n$ (see Eq 7).

6. We adapt a result of distributional convergence of squared norm of sums of IID random vectors in [51] to squared norm of sums of independent random vectors. Under the assumptions 9 and 8, the conditions of distributional convergence are satisfied.

7. Finally, all the error terms are combined along with an anti-concentration argument for χ^2 to establish the final result. The full proof and accompanying lemmas are in Section B of the Supplement.

□

3.2 Bootstrap consistency

Using the weighted χ^2 approximation for inference requires estimating the eigenvalues of Σ and other population quantities; however, accurate estimates may not be available in a streaming setting. Instead, we propose a streaming bootstrap procedure that mimics the properties of the original Oja algorithm. While a similar structure leads to error terms that are similar to the CLT, the analysis of the bootstrap presents its own technical challenges. In what follows let P^* denote the bootstrap measure, which is conditioned on the data, and let $\mathbb{E}^*[\cdot]$ denote the corresponding expectation operator.

A common strategy for establishing consistency of the Gaussian multiplier bootstrap is to invoke a Gaussian comparison lemma. Since the multipliers are themselves Gaussian and the data is treated as fixed, the idea is that one can use specialized results for comparing the distributions of two Gaussians (bootstrapped Z_n^* and approximating Z_n from the CLT) that only depend on how close the covariance matrices $\mathbb{E}^*[Z_n^* Z_n^{*T}]$ and $\mathbb{E}[Z_n Z_n^T]$ are in an appropriate metric. Using a Gaussian comparison lemma for quadratic forms (see Supplement Section C.3), we have the following result for the bootstrapped $\sin^2$ error:

Lemma 1. *[Bounding the difference between the bootstrap covariance and true covariance] Let:*

$$Z_n^* = sgn(v_1^T u_0)\sqrt{\frac{\eta_n}{n}}\sum_i W_i V_\perp \Lambda_\perp^{i-1} V_\perp^T (X_i X_i^T - X_{i-1}X_{i-1}^T) v_1. \tag{12}$$

Recall the definition of $\bar{\mathbb{V}}_n$ from Eq 7. We have,

$$|\text{trace}(\mathbb{E}^*[Z_n^* Z_n^{*T}] - \bar{\mathbb{V}}_n)|, \|\mathbb{E}^*[Z_n^* Z_n^{*T}] - \bar{\mathbb{V}}_n\|_F = O_P\left(\sqrt{\frac{\mathbb{E}\|X_1X_1^T - \Sigma\|^4}{n(\lambda_1 - \lambda_2)}}\right)$$

With this lemma in hand, we are ready to state our bootstrap result.

Theorem 2 (Bootstrap Consistency)**.** *Suppose that the conditions of Theorem 1 are satisfied. Furthermore, let α_n be a sequence such that $P(\mathcal{A}_n^c) \to 0$, where $\mathcal{A}_n$ is defined as $\mathcal{A}_n = \{\max_{i \le i \le n} \|X_i\|^2 \le \alpha_n\}$. Further suppose that $\frac{M_d \log^2 d\ \eta_n^2}{n} \to 0$, $\frac{(\alpha_n^3 \vee M_d \log d)\ \alpha_n \eta_n^3}{n} \to 0$, $\frac{\alpha_n M_d \eta_n^2}{n(\lambda_1 - \lambda_2)} \to 0$, and $\frac{\mathbb{E}[\|X_1X_1^T - \Sigma\|^4]}{n(\lambda_1-\lambda_2)} \to 0$. Then,*

$$\sup_{t\in\mathbb{R}} \left|P^*(n/\eta_n \sin^2(v_1^*, \hat{v}_1) \le t) - P(n/\eta_n \sin^2(\hat{v}_1, v_1) \le t)\right| \xrightarrow{P} 0$$

Proof sketch of Theorem 2. The proof follows a similar route to Theorem 2. We provide a detailed analysis in Supplementary Section. We use a bootstrap version of the Hoeffding decomposition conditioned on the data, stated in Supplement Section. In step one we have B_n^* replace B_n, where B_n^* is given by:

$$B_n^* = \prod_{i=1}^n \left(I + \eta_n/n(X_iX_i^T + W_i(X_iX_i^T - X_{i-1}X_{i-1}^T)\right)$$

We work out Step 1 using concentration of matrix products [23]. For steps 2-3, we see that T_k^* has the same structure as T_k with the difference that $(I + \eta_n\Sigma/n)^i$ is replaced by its sample counterpart which is a product of i independent matrices of the form $I + \eta_n/nX_jX_j^T$. Concentration of these terms in operator norm are established with results from [23]. Finally for step 4, we see that the main term that approximates the bootstrap residual $\widehat{V}_\perp \widehat{V}_\perp^T B_n^* u_0$ is given by $\sqrt{\eta_n/n}Z_n^*$, where Z_n^* is given in Eq 12. Conditioned on the data, this is already Normally distributed since the multiplier random variables W_i are themselves Gaussian. We then invoke the Gaussian comparison result Lemma 1 to obtain convergence to the weighted χ^2 approximation. $\square$

We now make a couple of points regarding our analysis. It should be noted that the terms in the product are weakly dependent, which is different from the CLT and would seem to complicate concentration arguments used to establish bootstrap consistency. However, the dependence is not strong and second-moment methods may be used. We also operate on a good set in which the norms of the the updates are not too large, which is far less restrictive than assuming an almost sure bound on the norm.

In theorem above, we have stated the good set $\mathcal{A}_n$ in an abstract manner, but one may wonder how stringent the condition is in various problem settings. Below, we describe a general setup with sub-Gaussian entries of X_i in which α_n grows as $\log n$; under milder forms of various decay, all we need is for α_n to grow slowly with n. Here $\|\cdot\|_{\psi_1}$ is the sub-Exponential Orlicz norm and $\|\cdot\|_{\psi_2}$ is the sub-Gaussian Orlicz norm (see, for example [50]).

Proposition 1 (The effect of variance decay on the norm)**.** *For each $1 \leq j \leq p$, suppose that X_{1j} satisfies $\|X_{1j}\|_{\psi_2} \leq \nu_j$ $\sum_{j=1}^{p} \nu_j \leq C_1 < \infty$. Then, for some universal constant $C_2 > 0$, $\left\|\sum_{j=1}^{p}(X_{1j}^2 - \mathbb{E}[X_{1j}^2])\right\|_{\psi_1} < C_2$, and for some $c_1, c_2 > 0$,*

$$P\left(\max_{1\leq i\leq n} \|X_i\|^2 > c_1 \log n\right) \leq \frac{c_2}{n}$$

We now present experimental validation of our bootstrap procedure below.

4 Experimental validation of the online multiplier bootstrap

We draw $Z_{ij} \overset{IID}{\sim} \text{Uniform}(-\sqrt{3}, \sqrt{3})$, for $i = 1, \ldots, n$ and $j = 1, \ldots d$. Consider a PSD matrix $K_{ij} = \exp(-|i - j|c)$ with $c = 0.01$. We create a covariance matrix such that $\Sigma_{ij} = K(i, j)\sigma_i\sigma_j$. We consider $\sigma_i = 5i^{-\beta}$ for $\beta = 0.2$ and $\beta = 1$. Now we transform the data to introduce dependence by letting $X_i = \Sigma^{1/2} Z_i$. By construction, we have that $\mathbb{E}[X_i X_i^T] = \Sigma$ for all $1 \leq i \leq n$. Our goal is to simply demonstrate that the bootstrap distribution of $\sin^2$ errors closely match that of the sampling distribution. To this effect, we fix u_0 and draw 500 datasets and run streaming PCA on each and then construct an empirical CDF (F) from the $\sin^2$ error with the true v_1. This is the point of comparison for the bootstrap distribution (F^*), for which we fix a dataset X. We then invoke algorithm 1 to obtain 500 bootstrap replicates $\hat{v}_1^*$ as well as the Oja vector for the dataset $\hat{v}_1$. The bootstrap distribution is the empirical CDF of $1 - (\hat{v}_1^T \hat{v}_1^*)^2$. We use $\eta_n = \log n$. In Figure 1, we see that for $\beta = 0.2$ (see (A) and (B)), where the variance decay is slow and therefore the error bounds of the residual terms are expected to be large, the quality of approximation is poorer compared to (C) and (D), where $\beta = 1$. However, even for $\beta = 0.2$, increasing n improves performance. Also note that, for (A) and (B) the variance decay does not satisfy our theorem's conditions and thus, the normalized error does not behave like a $O_P(1)$ random variable. However, for (C) and (D) the variance decay satisfies the conditions and in this case the normalized error is $O_P(1)$, which happens to be in the [0,1] range for this example.

5 Discussion

Modern tools in non-asymptotic random matrix theory have given rise to recent breakthroughs in establishing pointwise convergence rates for stochastic iterative methods in optimizing certain nonconvex objectives, including the classic Oja's algorithm for online principal component analysis. By synthesizing modern random matrix theory tools with classic results from the U-statistics literature and recently developed high-dimensional central limit theorems, we extend the error analysis of Oja's algorithm from pointwise convergence rates to *distributional* convergence and moreover establish an efficient online bootstrap method for Oja's algorithm to quantify the error on the fly. Our results are a first step toward incorporating uncertainty estimation into the general framework of stochastic optimization algorithms, but we acknowledge the present limitations of our analysis: new tools will be needed to extend the current analysis to estimating higher-dimensional principal subspaces, and additional tools will be needed to account for non-independent matrix products which appear beyond the setting of online PCA.

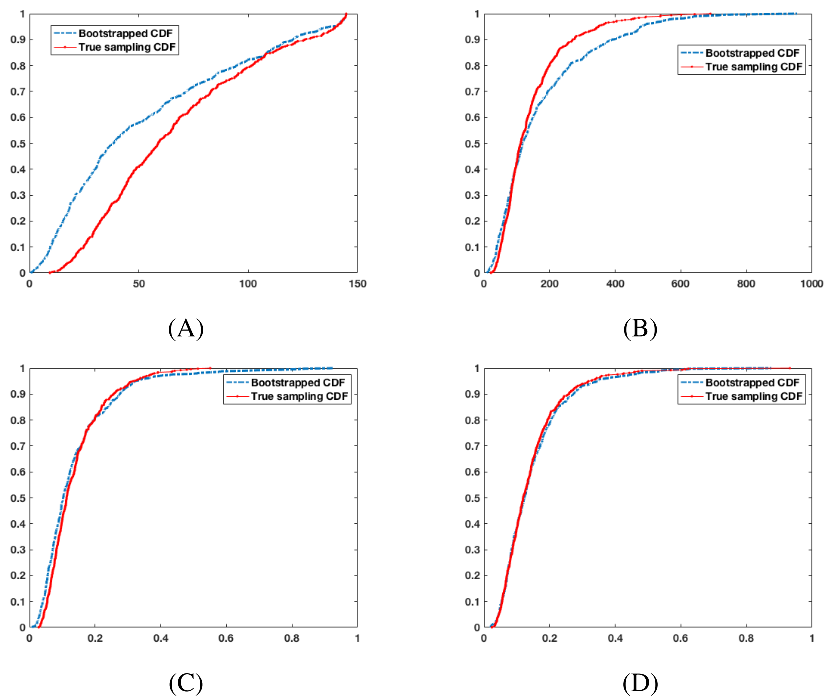

Figure 1: Bootstrapped and sampling CDF for $n = 1000, d = 500$ in (A) and (C) and for $n = 10,000, d = 500$ in (B) and (D). (A) and (B) use $\beta = 0.2$ whereas (C) and (D) use $\beta = 1$.

Acknowledgment

P.S. and R.L. are supported in part by NSF 2019844 and NSF HDR-1934932. R.W. is supported in part by AFOSR MURI FA9550-19-1-0005, NSF DMS 1952735, NSF HDR-1934932, and NSF 2019844.

References

[1] Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson. Much Faster Algorithms for Matrix Scaling. In *Proceedings of the 58th Symposium on Foundations of Computer Science*, FOCS '17, 2017.

[2] M.-F. Balcan, S. S. Du, Y. Wang, and A. W. Yu. An improved gap-dependency analysis of the noisy power method. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 284–309, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[3] A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental PCA. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3174–3182. Curran Associates, Inc., 2013.

[4] L. Balzano, R. Nowak, and B. Recht. Online Identification and Tracking of Subspaces from Highly Incomplete Information. *arXiv e-prints*, page arXiv:1006.4046, June 2010.

[5] Y. Benoist and J.-F. Quint. *Random Walks on Reductive Groups*. Springer International Publishing, June 2016.

[6] X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 02 2020.

[7] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786 – 2819, 2013.

[8] V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352, 07 2017.

[9] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.*, 1(1):54–75, 02 1986.

[10] J. Emme and P. Hubert. Limit laws for random matrix products. *arXiv e-prints*, page arXiv:1712.03698, Dec. 2017.

[11] V. Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Statist.*, 39(4):1327–1332, 08 1968.

[12] Y. Fang, J. Xu, and L. Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *J. Mach. Learn. Res.*, 19(1):3053–3073, Jan. 2018.

[13] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[14] A. Furman. Chapter 12 random walks on groups and random transformations. volume 1 of *Handbook of Dynamical Systems*, pages 931 – 1014. Elsevier Science, 2002.

[15] A. Giessing and J. Fan. Bootstrapping ℓ_p-statistics in high dimensions, 2020.

[16] F. Götze and H. R. Künsch. Second-order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.*, 24(5):1914–1933, 10 1996.

[17] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York, 1992.

[18] M. Hardt and E. Price. The noisy power method: A meta algorithm with applications. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2861–2869. Curran Associates, Inc., 2014.

[19] R. Helmers. On the edgeworth expansion and the bootstrap approximation for a studentized u-statistic. *Ann. Statist.*, 19(1):470–484, 03 1991.

[20] A. Henriksen and R. Ward. AdaOja: Adaptive Learning Rates for Streaming PCA. *arXiv e-prints*, page arXiv:1905.12115, May 2019.

[21] A. Henriksen and R. Ward. Concentration inequalities for random matrix products. *arXiv e-prints*, page arXiv:1907.05833, July 2019.

[22] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 09 1948.

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

[24] D. Huang, J. Niles-Weed, and R. Ward. Streaming k-PCA: Efficient guarantees for Oja's algorithm, beyond rank-one updates, 2021.

[25] P. Jain, C. Jin, S. Kakade, P. Netrapalli, and A. 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.

[26] P. Jain, D. Nagaraj, and P. Netrapalli. SGD without Replacement: Sharper Rates for General Smooth Convex Functions. *arXiv e-prints*, page arXiv:1903.01463, Mar. 2019.

[27] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 665–674, New York, NY, USA, 2013. Association for Computing Machinery.

[28] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A.*, 2016.

[29] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56:2980 – 2998, 07 2010.

[30] F. Ledrappier. *Some asymptotic properties of random walks on free groups*, pages 117–152. 06 2001.

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

[32] T. Li, L. Liu, A. Kyrillidis, and C. Caramanis. Statistical inference using SGD. *arXiv e-prints*, page arXiv:1705.07477, May 2017.

[33] X. Liang. On the optimality of the Oja's algorithm for online PCA, 2021.

[34] M. Lopes, S. Wang, and M. Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. volume 80 of *Proceedings of Machine Learning Research*, pages 3217–3226, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[35] M. E. Lopes. Estimating the algorithmic variance of randomized ensembles via the bootstrap. *Ann. Statist.*, 47(2):1088–1112, 04 2019.

[36] M. E. Lopes, N. B. Erichson, and M. W. Mahoney. Error Estimation for Sketched SVD via the Bootstrap. In *Proceedings of the International Conference of Machine Learning*, 2020.

[37] M. E. Lopes, S. Wang, and M. W. Mahoney. A bootstrap method for error estimation in randomized matrix multiplication. *Journal of Machine Learning Research*, 20(39):1–40, 2019.

[38] I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming PCA. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2886–2894, Red Hook, NY, USA, 2013. Curran Associates Inc.

[39] E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 451–459. Curran Associates, Inc., 2011.

[40] A. Nemirovski, A. B. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimization*, 19(4):1574–1609, 2009.

[41] E. Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, Nov. 1982.

[42] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, July 1992.

[43] D. Pouzo. Bootstrap consistency for quadratic forms of sample averages with increasing dimension. *Electronic Journal of Statistics*, 9(2):3046 – 3097, 2015.

[44] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.

[45] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. 1988.

[46] C. D. Sa, K. Olukotun, and C. Ré. Global convergence of stochastic gradient descent for some nonconvex matrix problems. *CoRR*, abs/1411.1134, 2014.

[47] O. Shamir. Convergence of stochastic gradient descent for pca. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 257–265. JMLR.org, 2016.

[48] W. J. Su and Y. Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent, 2018.

[49] A. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[50] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, Cambridge, UK, 2018.

[51] M. Xu, D. Zhang, and W. B. Wu. Pearson's chi-squared statistics: approximation theory and beyond. *Biometrika*, 106(3):716–723, 04 2019.

[52] L. Yu, K. Balasubramanian, S. Volgushev, and M. A. Erdogdu. An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias, 2020.

Supplementary material for "Bootstrapping the error of Oja's algorithm"

Robert Lunde
University of Michigan
rlunde@umich.edu

Purnamrita Sarkar
University of Texas at Austin
purna.sarkar@austin.utexas.edu

Rachel Ward
University of Texas at Austin
rward@math.utexas.edu

Supplementary Material

In this document we provide the detailed proofs of results presented in the main manuscript. In Section A, we provide a proof for the Hoeffding expansion of the matrix product in Eq 5 of the main document. We also provide the Hoeffding decomposition for the bootstrap in Proposition A.4. In Section B we provide all results needed for a complete proof of Theorem 1. In Sections B.1, B.2, and B.3 we provide the proof of Theorem 1, the adaptation of high dimensional CLT of [8] to our setting and all supporting lemmas, respectively.

In Section C we provide all details of the proof of the Bootstrap consistency, i.e. Theorem 2. To be specific, Section C.1 has the proof of Theorem 2; Section C.2 has the proof of Lemma 1, Section C.3 has the statement and proof of the Gaussian comparison lemma, and Section C.4 has all the supporting lemmas. Finally, in Section D, we provide a proof of Proposition 1.

A On the Hoeffding decomposition

We discuss Hoeffding decompositions for a function f of n independent random variables $X_1, \dots X_n$, where the random variables take values in an arbitrary space and the function takes values[1] in $\mathbb{R}^{d\times d}$ or $\mathbb{R}^d$. The following exposition largely follows [6].

With Hoeffding decompositions, we project $T(X_1, \dots, X_n)$ onto spaces of increasing complexity that are orthogonal to each other. In our setup, orthogonality means $\langle f, g\rangle_{L^2} = 0$ where $\langle f, g\rangle_{L^2} = \int \langle f, g\rangle dP$. Here, $\langle f, g\rangle = \mathrm{Trace}(f^T g)$ in the matrix case and $\langle f, g\rangle = f^T g$ in the vector case. The first-order projection, also known as a Hájek projection, involves projecting our function onto a space of functions of the form

$$g^{(i)}(X_i)$$

where $g^{(i)}$ satisfies $E[g^{(i)}] = 0$. We will let $H^{(i)}(X_i)$ denote the corresponding projection. Since the functions $g^{(i)}, g^{(j)}$ are mutually orthogonal for $i \neq j$, the sum of the projections is equivalent to the projection onto the space spanned by functions of the form:

$$\sum_{i=1}^{n} g^{(i)}(X_i)$$

The higher-order spaces have the form:

$$g^{(S)}(X_i : i \in S)$$

[1]The math generalizes to Hilbert spaces due to the Hilbert projection theorem but we specialize to these cases for concreteness.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

where $S \subseteq \{1, \ldots, n\}$ and the functions satisfy $\mathbb{E}[g^{(S)} \mid X_i : i \in R] = 0$ for any $R \subset S$, including $R = \emptyset$, which implies $\mathbb{E}[g^{(S)}] = 0$. If $R \not\subset S$ and $S \not\subset R$, $\langle g^{(S)}, g^{(R)} \rangle_{L^2} = 0$ since, by conditional independence given $\{X_i : i \in R \cap S\}$:

$$\mathbb{E}[\mathbb{E}[\langle g^{(S)}, g^{(R)} \rangle \mid X_i : i \in R \cap S]\,] = \mathbb{E}\left[\left\langle \mathbb{E}[g^{(S)} \mid X_i : i \in R \cap S],\ \mathbb{E}[g^{(R)} \mid X_i : i \in R \cap S]\ \right\rangle\right] = 0 \tag{S.1}$$

Combining these projections leads to the following representation, known as the Hoeffding decomposition:

$$T(X_1, \ldots, X_n) = \sum_{k=0}^{n} \sum_{|S|=k} H^{(S)}(X_i : i \in S)$$

While the following proposition is stated for real-valued functions in [6][Lemma 11.11], it turns out that the proof there generalizes to our setting without difficulty due to machinery for projections in Hilbert spaces.

Proposition A.1 (Hoeffding projections)**.** *Let $X_1, \ldots, X_n$ be arbitrary random variables and let suppose $\langle T, T \rangle_{L_2} < \infty$. Then the projection on the the space of functions of the form $g^{(S)}(X_i : i \in S)$ with $\mathbb{E}[g^{(S)} \mid X_i : i \in R] = 0$ for any $R \subset S$ has the form:*

$$H^{(S)}(T) = \sum_{R \subseteq S} (-1)^{|S|-|R|}\, \mathbb{E}\left[T \mid X_i : i \in R\right]$$

For completeness, we provide a proof of the proposition below.

Proof. We begin by verifying that the space of all random matrices (vectors) satisfying $\|A\|_{L^2} < \infty$ forms a Hilbert Space. First, it is clear that $\langle \cdot, \cdot \rangle_{L^2}$ is indeed an inner product. Linearity follows from linearity of the inner product $\langle \cdot, \cdot \rangle$ and linearity of expectations and conjugate symmetry follows from this property holding pointwise in Ω for $\langle \cdot, \cdot \rangle$. Positive definiteness again follows from the fact that this property holds pointwise in Ω; then a standard contradiction argument yields that if $\langle x, x \rangle_{L^2} = 0$, but x is not equal to 0 almost surely, there exists some M such that for some $\delta > 0$, $P(\|x\| > \frac{1}{M}) \geq \delta$ and hence $\int \langle x, x \rangle dP \geq \delta / M > 0$, a contradiction.

One can again adapt standard arguments for completeness of L_2 spaces to our setting; namely, show that Cauchy sequences converging in L_2 implies convergence almost everywhere, and then invoke completeness of the Hilbert space over matrices/vectors along with integral convergence theorems; see for example, the proof of Theorem 1.2, page 159 in [5].

Now to verify that this function is indeed the projection, we invoke the Hilbert Projection Theorem; see for example, Lemma 4.1 of [5]. To use this theorem, we need to check that the space spanned by functions of the form $g^{(S)}$ satisfying the condition $\mathbb{E}[g^{(S)} \mid X_i : i \in R] = 0$ for any $R \subset S$ is a closed subspace. Linearity of the space follows from the fact that the sum of such functions satisfies the constraint; therefore it is a subspace. To check closure, let $\|f\|^2 = \langle f, f \rangle$ and consider some (convergent) sequence in this subspace $(g_\alpha^{(S)})_{\alpha \geq 1}$ where $g_\alpha^{(S)} \to g^{(S)}$ and observe that, for any $R \subset S$:

$$\begin{aligned}
\mathbb{E}[\|g_\alpha^{(S)} - g^{(S)}\|^2] &= \mathbb{E}[\,\mathbb{E}[\|g_\alpha^{(S)} - g^{(S)}\|^2 \mid X_i : i \in R]\,] \\
&\geq \mathbb{E}\left[\|\mathbb{E}[g_\alpha^{(S)} - g^{(S)} \mid X_i : i \in R]\|^2\right] \\
&\geq \mathbb{E}\left[\|\mathbb{E}[g^{(S)} \mid X_i : i \in R]\|^2\right]
\end{aligned}$$

where above we used the fact that $\mathbb{E}[g_\alpha^{(S)} \mid X_i : i \in R] = 0$ for all α by assumption. Since the LHS converges to 0, it follows that $\mathbb{E}[g^{(S)} \,|X_i : i \in R]$ must be equal to 0 almost surely. Since the limit satisfies $\mathbb{E}[g^{(S)} \,|X_i : i \in R] = 0$ for all $R \subset S$, it belongs in the space, proving closure.

Now, we show that the stated expression is indeed the Hoeffding projection. First, to show that belongs in this space, we have, following analogous reasoning to [6], for any $C \subset A$,

$$
\begin{aligned}
\mathbb{E}[H^{(A)}(T) \mid X_i : i \in C] &= \sum_{B \subseteq A} (-1)^{|A|-|B|} \mathbb{E}[T \mid X_i : i \in B \cap C] \\
&= \sum_{D \subseteq C} \sum_{j=0}^{|A|-|C|} (-1)^{|A|-(|D|+j)} \binom{|A|-|C|}{j} \mathbb{E}[T \mid X_i : i \in D] \\
&= \sum_{D \subseteq C} (-1)^{|C|-|D|}\, \mathbb{E}[T \mid X_i : i \in D]\, (1-1)^{|A|-|C|} = 0
\end{aligned}
$$

where the last line follows from the Binomial Theorem. Now as a consequence of the Hilbert Projection Theorem, it suffices to show that $H^{(A)}(T)$ satisfies the property:

$$
\langle T - H^{(A)}(T), g^{(A)} \rangle_{L^2} = 0
$$

for any $g^{(A)}$ in the space. In the matrix case, we have

$$
\begin{aligned}
\langle T - H^{(A)}(T), g^{(A)} \rangle_{L^2} =& \sum_{j=1}^{d} \sum_{k=1}^{d} \mathbb{E}[(T_{jk} - \mathbb{E}[T_{jk} \mid X_i : i \in A]) \cdot g_{jk}^{(A)}] \\
&+ \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{B \subset A} \mathbb{E}\left[(-1)^{|A|-|B|} \mathbb{E}[T_{jk} \mid X_i : i \in B] \cdot \mathbb{E}[g_{jk}^{(A)} \mid X_i : i \in B]\right]
\end{aligned}
$$

The first term above is 0 since conditional expectations may be viewed as an orthogonal projection in the Hilbert Space with inner product $\int fg \, dP$ into the closed subspace of $\sigma(X_i : i \in A)$-measurable functions. The second term is zero since $\mathbb{E}[g_{jk}^{(A)} \mid X_i : i \in B] = 0$ for any $B \subset A$. The vector case is analogous.

Since this property holds, it must be the unique (up to measure 0 sets) minimizer and projection. $\square$

Now an immediate corollary for our setting follows.

Proposition A.2 (Orthogonality of Hoeffding projections)**.** *Let:*

$$
B_n = \sum_{k=0}^{n} \sum_{|S|=k} H^{(S)}
$$

where $A^{(S)}$ is the Hoeffding projection corresponding to the set $S \subseteq \{1, \ldots, n\}$. Then,

$$
\begin{aligned}
\mathbb{E}\left[\|B_n\|_F^2\right] &= \sum_{k=0}^{n} \sum_{|S|=k} \mathbb{E}\left[\|A^{(S)}\|_F^2\right] \\
\mathbb{E}\left[\|B_n x\|^2\right] &= \sum_{k=0}^{n} \sum_{|S|=k} \mathbb{E}\left[\|A^{(S)} x\|^2\right]
\end{aligned}
$$

where the last inequality holds for all $x \in \mathbb{R}^d$.

Proof. Letting $g^{(S)} = H^{(S)}$ and $g^{(R)} = H^{(R)}$ in Eq S.1, we have that $\langle H^{(S)}, H^{(R)} \rangle_{L^2} = 0$ for all $R \neq S$ and the result follows. $\square$

It remains to be shown that Hoeffding decomposition has the form stated in Eq 5. Deriving all projections in the Hoeffding decomposition for a general function is typically non-trivial, but the product structure facilitates our proof below. Before establishing the Hoeffding decomposition, following for example, [1] observe that the following inverse relation holds:

Proposition A.3 (Conditional expectation and Hoeffding projections)**.**

$$
\mathbb{E}\left[T \mid X_i : i \in S\right] = \sum_{R \subseteq S} H^{(R)}(T)
$$

Proof. Observe that:

$$\mathbb{E}[T\ |X_i : i \in S] = \sum_{k=0}^{n} \sum_{|R|=k} \mathbb{E}[H^{(R)}(T) \mid X_i : i \in S]$$

Since the conditional expectation is zero for $R \not\subseteq S$ and for $R \subseteq S$, the Hoeffding projection is fixed, the result follows. $\square$

Now we are ready to establish the form of the Hoeffding projection for any $S \subseteq \{1, \ldots, n\}$. We in fact prove a slightly stronger statement, which makes the induction argument more natural. In what follows let $S[i]$ denote the ith element in S. We will also use $H^{(S)}$ instead of $H^{(S)}(T)$ when it is clear from the context.

Theorem A.1 (Hoeffding projections for Oja's algorithm). *Define:*

$$T_{-j} = \prod_{i=j+1}^{n} \left(I + \frac{\eta_n}{n} X_i X_i^T\right), \quad T = T_{-0} = \prod_{i=1}^{n} \left(I + \frac{\eta_n}{n} X_i X_i^T\right),$$

Then for any $S \subseteq \{1, \ldots, n\}$ and for all $0 \leq j < S[1]$, we have the Hoeffding projection of T_{-j} onto $\{X_i : i \in S\}$ may be expressed as:

$$H_{-j}^{(S)} = \prod_{i=j+1}^{n} A_i^{(S)}, \;\; H^{(S)} = H_{-0}^{(S)} \tag{S.2}$$

where:

$$A_i^{(S)} = \begin{cases} \frac{\eta_n}{n}(X_i X_i^T - \Sigma) & i \in S \\ I + \frac{\eta_n}{n}\Sigma & i \notin S \end{cases}$$

Proof. We will conduct (strong) induction on $k = |R|$, where $R \subseteq S$. We will start with the base case $k = 1$; $k = 0$ is simply the expectation. For the base case $|R| = 1$, a direct calculation is possible, since:

$$H_{-j}^{(R)} = \mathbb{E}[T_{-j} \mid X_i : i \in R] - \mathbb{E}[T_{-j}],$$

which has the stated form. Now, we will suppose that the inductive hypothesis holds. In what follows, let $S[1] = k$ and define the conditional expectation for any set S as:

$$\mathbb{E}\left[T_{-j} \mid X_i : i \in S\right] = \prod_{i=j+1}^{n} E_i^{(S)},$$

where:

$$E_i^{(S)} = \begin{cases} I + \frac{\eta_n}{n} X_i X_i^T & i \in S \\ I + \frac{\eta_n}{n}\Sigma & i \notin S \end{cases}$$

We will now add and subtract a product where an entry corresponding to $S[1]$ in $\mathbb{E}[T_{-j} \mid X_i : i \in S]$ is replaced by $(I + \frac{\eta_n}{n}\Sigma)$. Doing, so we have

$$\begin{aligned}\mathbb{E}[T_{-j} \mid X_i : i \in S] = \mathbb{E}\left[T_{-j} \mid X_i : i \in S\right] - (I + \frac{\eta_n}{n}\Sigma)^{k-j} \times \prod_{i=k+1}^{n} E_i^{(S)} \\ + (I + \frac{\eta_n}{n}\Sigma)^{k-j} \times \prod_{i=k+1}^{n} E_i^{(S)}\end{aligned}$$

We recognize the second summand as $\mathbb{E}[T_{-j} \mid X_i : i \in S_{-k}]$, where $S_{-k} = \{i \in S, i \neq k\}$. Now for the first summand, taking the difference we have the term

$$\begin{aligned}&(I + \frac{\eta_n}{n}\Sigma)^{k-j-1} \times \frac{\eta_n}{n}(X_k X_k^T - \Sigma) \times \prod_{i=k+1}^{n} E_i^{(S)} \\ &= (I + \frac{\eta_n}{n}\Sigma)^{k-j-1} \times \frac{\eta_n}{n}(X_k X_k^T - \Sigma) \times \mathbb{E}\left[T_{-k} \mid X_i : i \in S_{-k}\right]\end{aligned}$$

By Proposition A.3, we may represent a conditional expectation as:

$$\mathbb{E}\left[T_{-k} \mid X_i : i \in S_{-k}\right] = \sum_{R \subseteq S_{-k}} H_{-k}^{(R)} \tag{S.3}$$

Furthermore, by the inductive hypothesis, each $H_{-k}^{(R)}$ takes the form in Eq S.2. Now, combining the two parts, we have

$$\begin{aligned}
\mathbb{E}[T_{-j} \mid X_i : i \in S] = & \sum_{R \subseteq S_{-k}} (I + \frac{\eta_n}{n}\Sigma)^{k-j-1} \times \frac{\eta_n}{n}(X_k X_k^T - \Sigma) \times H_{-k}^{(R)} \\
& + \sum_{R \subseteq S_{-k}} (I + \frac{\eta_n}{n}\Sigma)^{k-j} \times H_{-k}^{(R)} \\
& = \prod_{i=j+1}^{n} A_i^{(S)} + \sum_{R \subset S} H_{-j}^{(R)}
\end{aligned}$$

For the last step, notice that with the exception of $R = S_{-k}$ in the first sum, each product in the sum corresponds to a Hoeffding projection of some set of size less than k by the inductive hypothesis. The first term must be the Hoeffding projection onto S (with $S[1] = k > j$) by the same argument as Eq S.3, i.e.

$$H_{-j}^{(S)} = \prod_{i=j+1}^{n} A_i^{(S)},$$

proving the desired result. □

Now, since the Hoeffding decomposition is a sum of Hoeffding projections by definition, we have the following corollary.

Corollary A.1 (Hoeffding decomposition for Oja's algorithm)**.**

$$B_n = \sum_{k=0}^{n} \sum_{|S|=k} H^{(S)}$$

where $A^{(S)}$ is given by $H^{(S)}$ in Eq S.2.

It turns out that the bootstrap Hoeffding decomposition can be proved using the same strategy in Theorem A.1, where $X_1, \ldots, X_n$ is treated as fixed in the bootstrap measure. We state the result below.

Proposition A.4 (Hoeffding decomposition for the bootstrap)**.**

$$B_n^* = \sum_{k=0}^{n} \sum_{|S|=k} \alpha^{(S)}$$

where $\alpha^{(S)} = \prod_{i=1}^{n} \alpha_i^{(S)}$ and $\alpha_i^{(S)}$ is given by:

$$\alpha_i^{(S)} = \begin{cases} \frac{\eta_n}{n} W_i \cdot (X_i X_i^T - X_{i-1} X_{i-1}^T) & \textit{if } i \in S \\ I + \frac{\eta_n}{n} X_i X_i^T & \textit{otherwise} \end{cases}$$

B Central limit theorem for Oja's algorithm

B.1 Proof of Theorem 1

Proof of Theorem 1. Our strategy will be to approximate $\sin^2$ distance for estimated eigenvector with a quadratic form, and invoke a high-dimensional central limit theorem result. The remainder terms will be bounded using an anti-concentration result for weighted χ^2 random variables due to [8].

Observe that $\sin^2(\hat{v}_1, v_1)$ has the representation:

$$1 - \left(v_1^T \frac{B_n u_0}{\|B_n u_0\|}\right)^2 = \frac{u_0^T B_n^T (I - v_1 v_1^T) B_n u_0}{\|B_n u_0\|^2}$$

Let $V_\perp V_\perp^T = I - v_1 v_1^T$. Clearly, $V_\perp V_\perp^T$ is idempotent and is a projection matrix, implying that it is also symmetric. Therefore,

$$\frac{n}{\eta_n} \cdot \sin^2(u_n, v_1) = \frac{(\sqrt{n/\eta_n} V_\perp V_\perp^T B_n u_0)^T (\sqrt{n/\eta_n} V_\perp V_\perp^T B_n u_0)}{\|B_n u_0\|^2} \tag{S.4}$$

Let $a_1 = (v_1^T u_0)$ denote the scalar projection of u_0 so that $u_0 = a_1 v_1 + w$, where w is in the orthogonal complement of v_1.

Our first reduction of (S.4) is to approximate the denominator with a more convenient quantity. By Lemma B.2, we have that (S.4) may be written as

$$\frac{(\sqrt{n}/\eta_n \cdot V_\perp V_\perp^T B_n u_0)^T (\sqrt{n}/\eta_n \cdot V_\perp V_\perp^T B_n u_0)}{a_1^2 (1 + \frac{\eta_n}{n}\lambda_1)^{2n}} \cdot R_1$$

where

$$R_1 = \frac{\|B_n u_0\|^2}{a_1^2 (1 + \frac{\eta_n}{n}\lambda_1)^{2n}} = 1 - O_P\left(\sqrt{d} \exp\left(-\frac{\eta_n}{2}(\lambda_1 - \lambda_2)\right) + \sqrt{\frac{\eta_n^2 M_d \log d}{n}}\right)$$

While the aforementioned Lemma is stated for $\frac{\|B_n u_0\|}{|a_1|(1+\frac{\eta_n}{n}\lambda_1)^n}$, the relationship holds for the squared quantity since with high probability for n large enough, $|\frac{\|B_n u_0\|}{|a_1|(1+\frac{\eta_n}{n}\lambda_1)^n}| \leq 2$ and $|x^2 - 1^2| \leq 3|x-1|$ for all $-2 \leq x \leq 2$.

We will further approximate the quantity $\sqrt{n}/\eta_n \cdot V_\perp V_\perp^T B_n u_0$. First we will bound the contribution of $V_\perp V_\perp^T B_n V_\perp V_\perp^T$. By Lemma B.3 we have that:

$$R_2 := \sqrt{\frac{n}{\eta_n}} \cdot \frac{V_\perp V_\perp^T B_n V_\perp V_\perp^T u_0}{|a_1|(1 + \frac{\eta_n \lambda_1}{n})^n} = O_P\left(\sqrt{\frac{nd}{\eta_n}} \cdot \exp\{-\eta_n(\lambda_1 - \lambda_2)\} + \sqrt{\frac{\eta_n^2 M_d^2 \log d}{n}}\right)$$

Now it remains to bound the term $V_\perp V_\perp^T B_n v_1 (v_1^T u_0)$. First, by Corollary A.1, B_n can be decomposed as:

$$B_n = \sum_{k=0}^{n} T_k$$

where for $S \subseteq \{1, \ldots, n\}$, T_k is defined as:

$$T_k = \sum_{|S|=k} A^{(S)} \tag{S.5}$$

with $A^{(S)}$ taking the form in Eq S.2.

Since v_1 is orthogonal to $V_\perp$:

$$\sqrt{\frac{n}{\eta_n}} \cdot \frac{V_\perp V_\perp^T T_0 \, v_1 a_1}{|a_1|(1 + \eta_n/n\lambda_1)^n} = \sqrt{\frac{n}{\eta_n}} \cdot \mathrm{sign}(a_1)(I - v_1 v_1^T) v_1 = 0.$$

Furthermore, by Lemma B.4, since $\frac{\eta_n^3 M_d^2}{n} \to 0$ by assumption,

$$R_3 := \sqrt{\frac{n}{\eta_n}} \cdot \frac{V_\perp V_\perp^T (B_n - T_1) v_1 a_1}{|a_1|(1 + \eta_n/n\lambda_1)^n} = O_P\left(\sqrt{\frac{\eta_n^3 M_d^2}{n}}\right) \tag{S.6}$$

Now our term of interest is given by:

$$\frac{(\sqrt{n/\eta_n} \cdot V_\perp V_\perp^T T_1 v_1)^T (\sqrt{n/\eta_n} \cdot V_\perp V_\perp^T T_1 v_1)}{(1 + \frac{\eta_n}{n}\lambda_1)^{2n}} \tag{S.7}$$

Now, observe that $(I + \frac{\eta_n}{n}\Sigma)$ and $v_1 v_1^T$ share a common eigenspace and therefore commute. Therefore, the terms in the product to the left of T_1 may be written as:

$$\frac{V_\perp V_\perp^T (I + \frac{\eta_n}{n}\Sigma)^{i-1}}{(1 + \frac{\eta_n}{n}\lambda_1)^{i-1}} = \sum_{j=2}^{d} \left(\frac{1 + \frac{\eta_n}{n}\lambda_j}{1 + \frac{\eta_n}{n}\lambda_1} \right)^{i-1} v_j v_j^T := D_{i-1}, \quad \text{say.} \tag{S.8}$$

Hence,

$$\begin{aligned}
\sqrt{\frac{n}{\eta_n}} \cdot \frac{V_\perp V_\perp^T T_1 v_1}{(1 + \frac{\eta_n}{n}\lambda_1)^n} &= \sqrt{\frac{\eta_n}{n}} \sum_{i=1}^{n} \left(1 + \frac{\eta_n}{n}\lambda_1\right)^{-1} D_{i-1}(X_i X_i^T - \Sigma) v_1 \\
&= S_n = \sqrt{n}\left(1 + \frac{\eta_n}{n}\lambda_1\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} U_i, \quad \text{say,}
\end{aligned}$$

where

$$U_i = D_{i-1}(X_i X_i^T - \Sigma) v_1. \tag{S.9}$$

Observe that S_n is a sum of independent but non-identically distributed random variables with mean 0. Therefore, if the conditions of Proposition B.5 are satisfied, we may approximate $S_n^T S_n$ with $Z_n^T Z_n$, where $\mathbb{E}[Z_n] = 0$, $\mathrm{Var}(Z_n) = \mathrm{Var}(S_n)$. Below define $\tilde{Z}_i$ to be a Gaussian vector with $\mathrm{Var}(\tilde{Z}_i) = \mathrm{Var}((X_i X_i^T - \Sigma)v_1)$. Now define $Z_i = D_{i-1}\tilde{Z}_i$. We now verify these conditions.

First, we derive a lower bound on $\left\|\bar{\mathbb{V}}_n\right\|_F$ that will be used in all of the following bounds. Observe that $\left\|\bar{\mathbb{V}}_n\right\|_F = \frac{\eta_n}{n}\left\|\sum_i \Lambda_\perp^{i-1}\mathbb{M}\Lambda_\perp^{i-1}\right\|_F$ and the $k l$th entry of $\sum_i \Lambda_\perp^{i-1}\mathbb{M}\Lambda_\perp^{i-1}$ is lower bounded by:

$$\begin{aligned}
&\frac{\eta_n}{n}\sum_{i\geq 1}\left(\frac{1+\eta_n\lambda_{k+1}/n}{1+\eta_n\lambda_1/n}\right)^{i-1}\left(\frac{1+\eta_n\lambda_{\ell+1}/n}{1+\eta_n\lambda_1/n}\right)^{i-1}\mathbb{M}(k,\ell) \\
&\geq \frac{1-\exp(-2\eta_n(\lambda_1-\lambda_2))\left(1-\frac{\eta_n^2\lambda_1^2}{n}\right)^{-2}}{2\lambda_1-(\lambda_{k+1}+\lambda_{k+1})+\frac{\eta_n}{n}(\lambda_1^2-\lambda_k\lambda_l)}\mathbb{M}(k,\ell) \\
&\geq \frac{1-\exp(-2\eta_n(\lambda_1-\lambda_2))\left(1-\frac{\eta_n^2\lambda_1^2}{n}\right)^{-2}}{2\lambda_1+\frac{\eta_n}{n}\lambda_1^2}\mathbb{M}(k,\ell) \\
&\geq \frac{c}{\lambda_1}\,\mathbb{M}(k,\ell)
\end{aligned} \tag{S.10}$$

for some $c > 0$ and n large enough since $\exp(-\eta_n(\lambda_1 - \lambda_2)) \to 0$.

For the first term of L_q, $q = 3$ we have

$$\begin{aligned}
L_{3,1}^U &\leq \frac{1}{\sqrt{n}}\max_i \frac{\mathbb{E}(U_i^T\bar{\mathbb{V}}_n U_i)^{3/2}}{\|\bar{\mathbb{V}}_n\|_F^3} \\
&\leq \frac{M_d^{3/2}}{\sqrt{n}}\frac{\mathbb{E}\|V_\perp^T(X_iX_i^T-\Sigma)v_1\|^3}{\|\bar{\mathbb{V}}_n\|_F^3} \qquad \text{Since } \|\bar{\mathbb{V}}_n\| \leq M_d\eta_n \text{ from Eq 7} \\
&\leq C\frac{M_d^{3/2}\eta_n^3\lambda_1^3}{\sqrt{n}}\mathbb{E}\left(\frac{\|V_\perp^T X_1X_1^Tv_1\|}{\|\mathbb{M}\|_F}\right)^3
\end{aligned}$$

Similarly, for the Gaussian analog, we have that:

$$
\begin{aligned}
L_{3,1}^{Z} &\leq \frac{1}{\sqrt{n}} \max_{i} \frac{\mathbb{E}(Z_i^T \bar{\mathbb{V}}_n Z_i)^{3/2}}{\|\bar{\mathbb{V}}_n\|_F^3} \\
&\leq \frac{M_d^{3/2}\eta_n^{3/2}}{\sqrt{n}} \max_{i} \frac{\mathbb{E}\|Z_i\|^3}{\|\bar{\mathbb{V}}_n\|_F^3} \\
&\leq \frac{M_d^{3/2}\eta_n^{3/2}}{\sqrt{n}} \frac{\mathbb{E}\|\tilde{Z}_i\|^3}{\|\bar{\mathbb{V}}_n\|_F^3} \\
&\leq C \frac{M_d^{3/2}\eta_n^{3}\lambda_1^3}{\sqrt{n}} \mathbb{E}\left(\frac{\|\tilde{Z}_1\|}{\|\mathbb{M}\|_F}\right)^3
\end{aligned}
$$

For the second term, using the definition of U_i in Eq S.9 we have:

$$
\begin{aligned}
L_{3,2}^{U} &\leq \frac{1}{n} \max_{i<j} \frac{E|U_i^T U_j|^3}{\|\bar{\mathbb{V}}_n\|_F^3} \\
&= \frac{1}{n} \max_{i<j} \frac{E|v_1^T (X_i X_i^T - \Sigma) D_{i+j-2} (X_j X_j^T - \Sigma) v_1|^3}{\|\bar{\mathbb{V}}_n\|_F^3} \\
&\leq \frac{1}{n} \frac{\left(\mathbb{E}\|V_\perp^T (X_i X_i^T - \Sigma) v_1\|^3\right)^2}{\|\bar{\mathbb{V}}_n\|_F^3} \leq \frac{\eta_n^3 \lambda_1^3}{n} \frac{\left(\mathbb{E}\|V_\perp^T (X_i X_i^T) v_1\|^3\right)^2}{\|\mathbb{M}\|_F^3}
\end{aligned}
$$

For K_3, we have:

$$
\begin{aligned}
K_3^3 &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left|\frac{U_i^T U_i - E(U_i^T U_i)}{f}\right|^3 \\
&\leq \max_{i} \frac{\mathbb{E}(U_i^T U_i)^3 + (E U_i^T U_i)^3}{f^3} \leq 2 \max_{i} \frac{\mathbb{E}(U_i^T U_i)^3}{\|\bar{\mathbb{V}}_n\|_F^3} \\
&\leq 2\eta_n^3 \lambda_1^3 \frac{\mathbb{E}\|V_\perp^T (X_i X_i^T - \Sigma) v_1\|^6}{\|\mathbb{M}\|_F^3}
\end{aligned}
$$

Finally, for J_1 we have:

$$
\begin{aligned}
J_n &= \frac{\sum_{i=1}^n \mathrm{Var}(U_i^T U_i)}{(nf)^2} \leq \frac{\sum_{i=1}^n \mathbb{E}(U_i^T U_i)^2}{n^2 f^2} \\
&\leq \frac{\eta_n^2 \lambda_1^2}{n} \frac{\mathbb{E}[\|V_\perp (X_1 X_1^T - \Sigma) v_1\|^4]}{\|\mathbb{M}\|_F^2}
\end{aligned}
$$

The first makes $L_{3,2}$, K_3^3/n and J_n go to zero. The two conditions also imply $\frac{\mathbb{E}[\|V_\perp (X_1 X_1^T - \Sigma) v_1\|^3]}{\|\mathbb{M}\|_F^3} = o(\sqrt{n})$, which implies $L_{3,1} \to 0$.

Finally, we collect remainder terms and show that their contribution to the inner product is negligible using anti-concentration. Observe that,

$$
\begin{aligned}
&\sup_{t\in\mathbb{R}} \left|P(n/\eta_n \sin^2(w,v) \leq t) - P(Z_n^T Z_n \leq t)\right| \\
= &\sup_{t\in\mathbb{R}} \left|P\left(R_1 \cdot \frac{(S_n + R_2 + R_3)^T (S_n + R_2 + R_3)}{f} \leq t\right) - P\left(\frac{Z_n^T Z_n}{f} \leq t\right)\right|
\end{aligned} \tag{S.11}
$$

Now will will lower bound the above quantity. Observe that

$$\begin{aligned} & P\left(R_1 \cdot \frac{(S_n+R_2+R_3)^T(S_n+R_2+R_3)}{f} \le t\right) \\ & \ge P\left(R_1 \cdot \frac{S_n^T S_n}{f}\left(1+\frac{2\|R_2\|+2\|R_3\|_2}{\sqrt{S_n^T S_n}}\right)+\frac{R_1 \cdot \|R_2+R_3\|^2}{f} \le t\right) \\ & = P\left(R' \cdot \frac{S_n^T S_n}{f}+\widetilde{R} \le t\right), \text{ say.} \end{aligned} \tag{S.12}$$

Now, for $\delta_n = o(\sqrt{f})$, we have that:

$$P\left(S_n^T S_n \le \delta_n^2\right) \le \sup_{t\in\mathbb{R}} \left|P(S_n^T S_n \le t) - P(Z_n^T Z_n \le t)\right| + P(Z_n^T Z_n \le \delta_n^2) \to 0 \tag{S.13}$$

Note that $\delta_n = o(1)$ suffices since f is bounded away from zero under Eq 8 as shown in Eq S.10.

Now, choose ϵ_n satisfying $\epsilon_n = o(1)$ $\epsilon_n = \omega\left(\sqrt{\frac{\eta_n^3 M_d^2 \log d}{n}}\right)$, define the set:

$$\mathcal{G} = \left\{|R'-1| \le \epsilon_n,\ \left|\widetilde{R}\right| \le \epsilon_n\right\}$$

so that $P(\mathcal{G}^c) \to 0$ with the choice of δ_n in Eq. S.13. By using the fact that, for any two sets A and B, $1 \ge P(A) + P(B) - P(A \cap B)$ and hence $P(A \cap B) \ge P(A) - P(B^c)$, we have that:

$$\begin{aligned} & P\left(R' \cdot \frac{S_n^T S_n}{f}+\widetilde{R} \le t\right) \\ & = P\left(R' \cdot S_n^T S_n/f + \widetilde{R} \le t \ \cap\ \mathcal{G}\right) + P\left(R' \cdot S_n^T S_n/f + \widetilde{R} \le t \ \cap\ \mathcal{G}^c\right) \\ & \ge P\left(\frac{S_n^T S_n}{f} \le \frac{t}{1+\epsilon_n} - \epsilon_n\right) - P(\mathcal{G}^c) \end{aligned} \tag{S.14}$$

Therefore,

$$\begin{aligned} & P\left(\frac{n/\eta_n \sin^2(w,v)}{f} \le t\right) - P\left(\frac{Z_n^T Z_n}{f} \le t\right) \\ & \ge P\left(\frac{S_n^T S_n}{f} \le \frac{t}{1+\epsilon_n} - \epsilon_n\right) - P\left(\frac{Z_n^T Z_n}{f} \le \frac{t}{1+\epsilon_n} - \epsilon_n\right) \\ & \quad + P\left(\frac{Z_n^T Z_n}{f} \le \frac{t}{1+\epsilon_n} - \epsilon_n\right) - P\left(\frac{Z_n^T Z_n}{f} \le t\right) - P(\mathcal{G}^c) = I + II - III \end{aligned} \tag{S.15}$$

Now, we may upper bound $III \to 0$ arising from our choice of δ_n, and II goes to 0 if the conditions of Proposition B.5 are satisfied, and $I \to 0$ due to Proposition B.7.

Now for the upper bound, since $\|R_i\|_2 \ge 0$, observe that we may bound Eq S.11 with:

$$\begin{aligned} & P\left(R_1 \cdot \frac{(S_n+R_2+R_3)^T(S_n+R_2+R_3)}{f} \le t\right) \\ & \le P\left(R_1 \cdot \frac{S_n^T S_n}{f}\left(1-\frac{2\|R_2\|+2\|R_3\|}{\sqrt{S_n^T S_n}}\right)-\frac{R_1 \cdot \|R_2\|\,\|R_3\|}{f} \le t\right) \end{aligned}$$

We may now lower bound the negative terms and arrive at an identical expression to the lower bound. The result follows. □

With the central limit theorem in hand, we are now ready to give the proof for Corollary 1.

Proof of Corollary 1. Observe that the approximating distribution $Z_n^T Z_n$ has expectation $\mathrm{trace}(\bar{\mathbb{V}}_n)$ and variance $f = \left\|\bar{\mathbb{V}}_n\right\|_F$. Therefore, for any $M > 0$, it follows that:

$$P\left(\frac{n/\eta_n \sin^2(\hat{v}_1, v_1) - \mathrm{trace}(\bar{\mathbb{V}}_n)}{f} > M\right)$$
$$\leq \sup_{t\in\mathbb{R}} \left|P\left(n/\eta_n \sin^2(\hat{v}_1, v_1) > t\right) - P\left(Z_n^T Z_n > t\right)\right| + P\left(\frac{Z_n^T Z_n - \mathrm{trace}(\bar{\mathbb{V}}_n)}{f} > M\right)$$

The first term goes to zero under the conditions of Theorem 1. Chebychev's inequality implies that there exists $M > 0$ such that the latter probability can be made smaller than $\epsilon/2$ for any $\epsilon > 0$. Hence,

$$\frac{n/\eta_n \sin^2(\hat{v}_1, v_1) - \mathrm{trace}(\bar{\mathbb{V}}_n)}{f} = O_P(1).$$

Therefore, under the conditions in Theorem 1,

$$\sin^2(\hat{v}_1, v_1) = \frac{\eta_n}{n}\left[\mathrm{trace}(\bar{\mathbb{V}}_n) + O_P\left(\left\|\bar{\mathbb{V}}_n\right\|_F\right)\right]$$

We now derive bounds for $\mathrm{trace}(\bar{\mathbb{V}}_n)$ and $\left\|\bar{\mathbb{V}}_n\right\|_F$. Let $\Lambda_\perp$ be a diagonal matrix with $\Lambda_\perp(i,i) = (1+\eta_n\lambda_{i+1}/n)/(1+\eta_n\lambda_1/n)$, $i = 1, \ldots, d-1$. Recall that:

$$\mathbb{M} := \mathbb{E}\left[V_\perp^T (X_1^T v_1)^2 X_1 X_1^T V_\perp\right]. \tag{S.16}$$

$$\bar{\mathbb{V}}_n = \frac{\eta_n}{n} V_\perp \left(\sum_i \Lambda_\perp^{i-1}\mathbb{M}\Lambda_\perp^{i-1}\right) V_\perp^T$$

So now observe that,

$$\left\|\bar{\mathbb{V}}_n\right\|_F = \frac{\eta_n}{n}\left\|\sum_i \Lambda_\perp^{i-1}\mathbb{M}\Lambda_\perp^{i-1}\right\|_F$$
$$\mathrm{trace}(\bar{\mathbb{V}}_n) = \frac{\eta_n}{n}\mathrm{trace}\left(\sum_i \Lambda_\perp^{i-1}\mathbb{M}\Lambda_\perp^{i-1}\right)$$

A direct calculation shows that the k, ℓ^{th} entry of the sum $\sum_i \Lambda_\perp^{i-1}\mathbb{M}\Lambda_\perp^{i-1}$ is:

$$\begin{aligned}
&\sum_{i\geq 1}\left(\frac{1+\eta_n\lambda_{k+1}/n}{1+\eta_n\lambda_1/n}\right)^{i-1}\left(\frac{1+\eta_n\lambda_{\ell+1}/n}{1+\eta_n\lambda_1/n}\right)^{i-1}\mathbb{M}(k,\ell)\\
&\leq \frac{n\mathbb{M}(k,\ell)}{\eta_n}\frac{(1+\frac{\lambda_1\eta_n}{n})^2}{2\lambda_1 - (\lambda_{k+1}+\lambda_{k+1}) + \frac{\eta_n}{n}(\lambda_1^2 - \lambda_k\lambda_l)}\\
&\leq \frac{n}{\eta_n}\frac{C\mathbb{M}(k,\ell)}{\lambda_1-\lambda_2}
\end{aligned}\tag{S.17}$$

for some $0 < C < \infty$.

Therefore, by Eq 7, we have

$$\mathrm{trace}(\bar{\mathbb{V}}_n) \leq C\frac{\mathrm{trace}(\mathbb{M})}{\lambda_1-\lambda_2} \leq C\frac{M_d}{\lambda_1-\lambda_2}$$
$$\|\bar{\mathbb{V}}_n\|_F \leq \frac{C\|\mathbb{M}\|_F}{\lambda_1-\lambda_2} \leq C'\frac{M_d}{\lambda_1-\lambda_2}$$

The last step is true since:

$$\begin{aligned}
\mathrm{trace}(\mathbb{M}) &= \mathrm{trace}(\mathbb{E}\left[V_\perp^T (X_1^T v_1)^2 X_1 X_1^T V_\perp\right])\\
&= \mathrm{trace}(\mathbb{E}\left[V_\perp^T (X_1 X_1^T - \Sigma) v_1 v_1^T (X_1 X_1^T - \Sigma) V_\perp\right])\\
&= \mathbb{E}\left(\mathrm{trace}\left[V_\perp^T (X_1 X_1^T - \Sigma) v_1 v_1^T (X_1 X_1^T - \Sigma) V_\perp\right]\right)\\
&= \mathbb{E}\|V_\perp^T (X_1 X_1^T - \Sigma) v_1\|^2 \leq M_d
\end{aligned}$$

Similarly,

$$\begin{aligned}\|\mathbb{M}\|_F &= \left\|\mathbb{E}\left[V_\perp^T (X_1^T v_1)^2 X_1 X_1^T V_\perp\right]\right\|_F \\ &= \left\|\mathbb{E}\left[V_\perp^T (X_1 X_1^T - \Sigma) v_1 v_1^T (X_1 X_1^T - \Sigma) V_\perp\right]\right\|_F \\ &\leq \mathbb{E}\|X_1 X_1^T - \Sigma\|_{op}^2 = M_d\end{aligned}$$

where in the last line we used the fact that $\left\|xx^T\right\|_{op} = \left\|xx^T\right\|_F$ for $x \in \mathbb{R}^d$ since xx^T is rank 1. □

B.2 Adaptation of high-dimensional central limit theorem

Let $U_1, \dots, U_n$, be independent random vectors in $\mathbb{R}^p$ such that $E(U_i) = 0$ and $\mathrm{Var}(U_i) = \mathbb{V}_i$. Define a Gaussian analog of Y_i, denoted Z_i, which satisfies $E(Z_i) = 0$ and $\mathrm{Var}(Z_i) = \mathbb{V}_i$. Furthermore, let $\bar{\mathbb{V}}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{V}_i$, $g_i = \mathrm{Var}(U_i^T U_i)$, $f_1 = \mathrm{trace}(\bar{\mathbb{V}}_n)$, and $f = \left\|\bar{\mathbb{V}}_n\right\|_F$. For $0 < \delta \leq 1$, $q = 2 + \delta$, and $\beta \geq 2$ define the following quantities:

$$\begin{aligned}L_q^U &= \frac{1}{n}\sum_{i=1}^n \frac{E(U_i^T \bar{\mathbb{V}}_n U_i)^{q/2}}{n^{\delta/2} f^q} + \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \frac{E(|U_i^T U_j|^q)}{n^\delta f^q} \\ L_q^Z &= \frac{1}{n}\sum_{i=1}^n \frac{E(Z_i^T \bar{\mathbb{V}}_n Z_i)^{q/2}}{n^{\delta/2} f^q} \\ K_\beta^\beta &= \frac{1}{n}\sum_{i=1}^n E\left|\frac{U_i^T U_i - E(U_i^T U_i)}{f}\right|^\beta \\ J_n &= \frac{\sum_{i=1}^n g_i}{(nf)^2}\end{aligned}$$

The following proposition is an adaptation of [8], which is stated for IID random variables, to independent but non-identically distributed random variables. While the changes are minor, we provide a proof below detailing the adaptation for completeness.

Proposition B.5. *Suppose that $L_q^U \to 0$, $L_q^Z \to 0$, $J_n \to 0$, $n^{1-\beta} K_\beta^\beta \to 0$. Then,*

$$\sup_{t \in \mathbb{R}} \left| P\left(n \bar{U}_n^T \bar{U}_n \leq t\right) - P\left(n \bar{Z}_n^T \bar{Z}_n \leq t\right)\right| \to 0$$

Proof. Since a Lindeberg argument is easier with diagonals removed, we will show that the removal of these terms is negligible. Observe that:

$$\begin{aligned}&\sup_{t \in \mathbb{R}} \left|P(n\bar{U}_n^T \bar{U}_n \leq t) - P(n \bar{Z}_n^T \bar{Z}_n \leq t)\right| \\ &\leq \sup_{t' \in \mathbb{R}} \left| P\left(\frac{n\bar{U}_n^T \bar{U}_n - f_1}{f} \leq t'\right) - P\left(\frac{\sum_{i \neq j} U_i^T U_j}{nf} \leq t'\right)\right| \\ &+ \sup_{t' \in \mathbb{R}} \left| P\left(\frac{\sum_{i \neq j} U_i^T U_j}{nf} \leq t'\right) - P\left(\frac{\sum_{i \neq j} Z_i^T Z_j}{nf} \leq t'\right)\right| \\ &+ \sup_{t' \in \mathbb{R}} \left| P\left(\frac{\sum_{i \neq j} Z_i^T Z_j}{nf} \leq t'\right) - P\left(\frac{n\bar{Z}_n^T \bar{Z}_n - f_1}{f} \leq t'\right)\right| \\ &= I + II + III, \text{ say.}\end{aligned}$$

We will start by bounding III. First note that $\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i \sim \mathcal{N}(0, \bar{\mathbb{V}}_n)$. Let $\bar{\mathbb{V}}_n = Q^T D Q$ denote the eigendecomposition, with diagonal entries of D given by $\lambda_1 \geq \ldots \geq \lambda_d$ and let $g \sim \mathcal{N}(0, \mathrm{I}_d)$. It follows that:

$$\begin{aligned}n\bar{Z}_n^T \bar{Z}_n &\stackrel{d}{=} (QD^{1/2}Q^T g)^T (QD^{1/2} Q^T g) \\ &\stackrel{d}{=} g^T D g\end{aligned}$$

Notice that $V := g^T D g \sim \sum_{r=1}^d \lambda_r \eta_r$, where $\eta_1, \ldots, \eta_d \sim \chi^2(1)$. Now define $R_n^Z = \frac{\frac{1}{n}\sum_{i=1}^n Z_i^T Z_i - f_1}{f}$. Notice that:

$$\begin{aligned}
&P\left(\frac{n\bar{Z}_n^T \bar{Z}_n - f_1}{f} \leq t\right) - P\left(\frac{\sum_{i\neq j} Z_i^T Z_j}{f} \leq t\right) \\
&= P\left(\frac{n\bar{Z}_n^T \bar{Z}_n - f_1}{f} \leq t\right) - P\left(\frac{n\bar{Z}_n^T \bar{Z}_n - f_1}{f} - R_n^Z \leq t\right) \\
&\leq P(t' \leq V \leq t' + h_n) + P(|R_n^Z| > h_n)
\end{aligned} \tag{S.18}$$

Under the conditions $J_n \to 0$, $n^{1-\beta} K_\beta^\beta \to 0$, Nagaev's inequality implies that one may choose $h_n \to 0$ such that $P(|R_n^Z| > h_n) \to 0$. The desired anti-concentration for the first term in the previous display follows from Lemma S2 of [8]. We may also derive the lower bound $P(t' \leq V \leq t' + h_n) - P(|R_n^Z| > h_n)$ in a similar manner.

To adapt II, consider the smoothed indicator function:

$$g_{\psi,t}(x) = \left[1 - \min\{1, \max(x-t, 0)\}^4\right]^4.$$

This function satisfies:

$$\begin{aligned}
&\max_{x,t}\{|g'_{\psi,t}(x)| + |g''_{\psi,t}(x)| + |g'''_{\psi,t}(x)|\} < \infty \\
&\mathbb{1}_{x \leq t} \leq g_{\psi,t} \leq \mathbb{1}_{x \leq t+\psi^{-1}}.
\end{aligned}$$

Therefore, we may bound the approximation error with smoothed indicator function by again using anti-concentration of the weighted χ^2. In what follows, let:

$$S_n^U = \frac{1}{nf}\sum_{i\neq j} U_i^T U_j, \quad S_n^Z = \frac{1}{nf}\sum_{i\neq j} Z_i^T Z_j$$

We have that:

$$\begin{aligned}
&P(S_n^U \leq t) - P(S_n^Z \leq t) \\
&\leq P(S_n^U \leq t) - P(S_n^Z \leq t + \psi^{-1}) + P(S_n^Z \leq t + \psi^{-1}) - P(S_n^Z \leq t) \\
&\leq Eg_{\psi,t}(S_n^U) - Eg_{\psi,t}(S_n^Z) + III + P(t \leq V \leq t + \psi^{-1}).
\end{aligned}$$

An analogous argument establishes a lower bound of $g_{\psi,t}(S_n^U) - Eg_{\psi,t}(S_n^Z) - III - P(t - \psi^{-1} \leq V \leq t)$. Choosing $\psi_n \to \infty$, the last term goes to zero. A Lindeberg telescoping sum argument leads to the following bound for the leading term:

$$\left|Eg_{\psi,t}(S_n^U) - Eg_{\psi,t}(S_n^Z)\right| \leq \sum_{i=1}^n c_q(E|\Delta_i|^q + E|\Gamma_i|^q),$$

where:

$$H_i = \sum_{j=1}^{i=1} U_i + \sum_{j=i+1}^n Z_i, \quad \Delta_i = \frac{U_i^T H_i}{nf}, \quad \Gamma_i = \frac{Z_i^T H_i}{nf}.$$

We may use analogous reasoning to bound these terms. Let $\xi \sim N(0,1)$. Conditioning on $U_1 = u_i$, by Rosenthal's inequality:

$$\begin{aligned}
\mathbb{E}\left[|\Delta_i|^q \mid U_i\right] &\leq \sum_{j=1}^{i-1} \frac{\mathbb{E}[|U_j^T u_i|^q]}{n^q f^q} + \sum_{j=i+1}^n \frac{\mathbb{E}[|Z_j^T u_i|^q]}{n^q f^q} + n^{q/2}\frac{\left(u_i^T \bar{\mathbb{V}}_n u_i\right)^{q/2}}{n^q f^q} \\
&\leq \sum_{j=1}^{i-1} \frac{\mathbb{E}[|U_j^T u_i|^q]}{n^q f^q} + \sum_{j=i+1}^n \|\xi\|_q^q \frac{\left(u_i^T \mathbb{V}_j u_i\right)^{q/2}}{n^q f^q} + \frac{\left(u_i^T \bar{\mathbb{V}}_n u_i\right)^{q/2}}{n^{q/2} f^q}
\end{aligned} \tag{S.19}$$

Taking expectations, it follows that:

$$\sum_{i=1}^n \mathbb{E}\left[|\Delta_i|^q\right] \lesssim \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \frac{\mathbb{E}\left[|U_i^T U_j|^q\right]}{n^\delta f^q} + \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}\left|U_i^T \bar{\mathbb{V}}_n U_i\right|^{q/2}}{n^{\delta/2} f^q}$$

Now, for Γ_i, we may use Rosenthal's inequality so that:

$$\sum_{i=1}^{n} \mathbb{E}\left[\left|\Gamma_i\right|^q\right] \leq \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}\left|U_i^T \bar{\mathbb{V}}_n U_i\right|^{q/2}}{n^\delta \delta f^q} + \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}\left[\left|Z_i^T \bar{\mathbb{V}}_n Z_i\right|^{q/2}\right]}{n^\delta \delta f^q} + \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}\left(Z_i^T \bar{\mathbb{V}}_n Z_i]\right)^{q/2}}{n^{q/2} f^q}$$

While omitted in the original proof, in the IID case, the latter terms may be bounded by using an eigendecomposition along with properties of the Gaussian. However, since the Z_i do not have variance matrix $\mathbb{V}_n$, we instead oppose the additional condition for L_q^Z. By the assumptions made in theorem, it follows that $II \to 0$.

Finally, for I, we have that:

$$\begin{aligned}
&P\left(\frac{n\bar{U}_n^T \bar{U}_n - f_1}{f} \leq t\right) - P\left(\frac{\sum_{i\neq j} U_i^T U_j}{nf} \leq t\right) \\
&\leq P(S_n^X \leq t + h_n) - P(S_n^U \leq t + h_n) + P(|R_n^X| > h_n) \\
&\quad + P(t \leq V \leq t + 2h_n) + P(|S_n^Z| > h_n)
\end{aligned}$$

Using bounds from II and III along with anti-concentration properties, we may conclude that $I \to 0$.

□

B.3 Supporting lemmas for CLT

In several of our lemmas, we use the following technique from [4] that facilitates analysis for initializations from a uniform distribution on $\mathcal{S}^{d-1}$ particularly when d is large.

Proposition B.6 (Trace trick)**.** *Suppose that u is drawn from a uniform distribution on $\mathcal{S}^{d-1}$. Then, for any $A \in \mathbb{R}^{d\times d}$ and $v \in \mathbb{R}^d$ satisfying $\|v\| = 1$, with probability at least $1 - C\delta$, for some $C > 0$ independent of A and $0 < \delta < 1$,*

$$\frac{u^T A^T A u}{(v^T u)^2} \leq \frac{\log(1/\delta)\ \mathrm{trace}(AA^T)}{\delta^2}$$

Proof. First, we recall the well-known fact that $u = g/\|g\|$, where $g \sim N(0, I_d)$. Therefore, $\|g\|$ cancels as follows:

$$\frac{u^T A^T A u}{(v^T u)^2} = \frac{g^T A^T A g}{(v^T g)^2}$$

Furthermore, observe that $g^T A^T A g$ may be viewed as a weighted sum of independent $\chi^2(1)$ random variables. In particular, by an eigendecomposition argument, for $\eta_1, \ldots \eta_r \sim \chi^2(1)$ and $A = VDV^T$,

$$\begin{aligned}
g^T(VDV^T)(VDV^T)g &= g^T V D^2 V^T g \\
&\stackrel{d}{=} g^T D^2 g \\
&= \sum_{r=1}^{p} \lambda_r^2 \eta_r = \psi, \text{ say}
\end{aligned}$$

where above we used the fact that $V^T g \sim N(0, I_d)$. Now observe that $\mathbb{E}[\psi] = \sum_{r=1}^{p} \lambda_r^2 = \|A\|_F^2$ and that η_r is sub-Exponential. Therefore, by by Bernstein's inequality (see for example Theorem 2.8.2 of [7]), for some $K > 0$, $C_1 > 0$, $0 < \delta < 1$,

$$\begin{aligned}
P\left(\psi - \mathbb{E}[\psi] > (\log(1/\delta) - 1)\|A\|_F^2\right) &\leq \exp\left\{-\min\left(\frac{\log^2(1/\delta)\|A\|_{\mathcal{S}_2}^4}{4K^2\|A\|_{\mathcal{S}_4}^4}, \frac{\log(1/\delta)\|A\|_{\mathcal{S}_2}^2}{2K\|A\|_{\mathcal{S}_\infty}^2}\right)\right\} \\
&\leq \exp\left\{-\min\left(\frac{\log^2(1/\delta)}{4K^2}, \frac{\log(1/\delta)}{2K}\right)\right\} \leq C_1\delta
\end{aligned}$$

where above $\|\cdot\|_{\mathcal{S}_p}$ is the pth Schatten-Norm, defined as $(\sum_{r=1}^d s_r^p)^{1/p}$, where s_r is the rth singular value and satisfies $\|\cdot\|_{\mathcal{S}_q} \leq \|\cdot\|_{\mathcal{S}_p}$ for $p \leq q$. Now for the denominator, since $v^T g \sim N(0,1)$ and $(v^T g)^2 \sim \chi^2(1)$, Proposition B.7 yields:

$$P((v^T g)^2 \leq \delta^2) \leq \frac{2\delta}{\sqrt{\pi}}$$

The result follows. □

The following anti-concentration result for weighted χ^2 distributions is also used in several places.

Proposition B.7 (Weighted χ^2 anti-concentration, [8]). *Let $a_1 \geq \cdots \geq a_p \geq 0$ such that $\sum_{r=1}^p a_i^2 = 1$ and suppose that $\xi_1, \ldots, \xi_p \sim \chi^2(1)$. Then,*

$$\sup_{t \in \mathbb{R}} P\left(t \leq \sum_{r=1}^p a_r \xi_r \leq t+h\right) \leq \sqrt{\frac{4h}{\pi}}$$

We now present a concentration result for matrix products that follow immediately from Corollary 5.4 of [3].

Lemma B.1 (Expectation bounds for operator norms of of matrix products). *Let $\mathcal{B}_k = \prod_{j=1}^k (I + \eta_n X_j X_j^T/n)$. We have,*

$$\mathbb{E}\|\mathcal{B}_k - \mathbb{E}\mathcal{B}_k\|^2 \leq \frac{M_d e \eta_n^2 (1 + 2\log d) k}{n^2} (1 + \eta_n \lambda_1/n)^{2k}. \tag{S.20}$$

For the expectation, we have, if $\frac{(1+2\log d) M_d \eta_n^2}{n} \leq 1$:

$$\mathbb{E}\|\mathcal{B}_k\|^2 \leq \exp\left(2\sqrt{2M_d \frac{k\eta_n^2}{n^2}\left(2M_d \frac{k\eta_n^2}{n^2} \vee \log d\right)}\right)(1 + \eta_n \lambda_1/n)^{2k}. \tag{S.21}$$

Proof. We invoke Corollary 5.4 in [3] with $\|\mathbb{E}(I + \eta_n/n X_i X_i^T)\| \leq 1 + \eta_n \lambda_1/n$, $\sigma_i^2 = M_d \frac{\eta_n^2}{n^2}$, and $\nu = M_d \frac{k\eta_n^2}{n^2}$. Note that for a random matrix M with Schatten norm $\|M\|_{\mathcal{S}_p}$, $\mathbb{E}\|M\| \leq \sqrt{\mathbb{E}\|M\|_{\mathcal{S}_p}^2}$ and hence the same argument as in their proof invoking Eq 5.5 and 5.6 works. □

Lemma B.2 (Concentration of the norm for the CLT). *For some $C > 0$, and any $\epsilon > 0$, $0 < \delta < 1$,*

$$\begin{aligned}
&P\left(\left|\frac{\|B_n u_0\|}{|a_1|(1+\eta_n\lambda_1/n)^n} - 1\right| \geq \epsilon\right) \\
&\leq \frac{d \exp\left(-\eta_n(\lambda_1 - \lambda_2) + \frac{\eta_n^2}{n}(\lambda_1^2 + M_d)\right) + \frac{\eta_n^2}{n} M_d \exp\left(\frac{\eta_n^2}{n}\right)}{4\log^{-1}(1/\delta)\delta^2\epsilon^2\left(1 + \frac{\eta_n^2 \lambda_1^2}{n}\right)} + \frac{e^2 \eta_n^2 M_d (1 + \log d)}{n\epsilon^2} + C\delta
\end{aligned}$$

Proof. Consider the bound:

$$\left|\frac{\|B_n u_0\|}{|a_1|(1+\eta_n\lambda_1/n)^n} - 1\right| \leq \left|\frac{\|B_n v_1 a_1\| - \|a_1 T_0 v_1\|}{|a_1|(1+\eta_n\lambda_1/n)^n}\right| + \frac{\|B_n V_\perp (V_\perp^T u_0)\|}{|a_1|(1+\eta_n\lambda_1/n)^n}$$

We will start by bounding the second term.

Using Proposition B.6, observe that, with probability at least $1 - C\delta$,

$$\frac{\|(B_n V_\perp V_\perp^T g\|^2}{|v_1^T g|^2 (1+\eta_n\lambda_1/n)^{2n}} \leq \frac{\log(1/\delta)\mathrm{trace}(V_\perp B_n B_n V_\perp^T))}{\delta^2 (1+\eta_n\lambda_1/n)^{2n}}$$

Let $\mathcal{G}$ denote the good set for which the upper bound above holds. Markov's inequality on the good set, together with Lemma 5.2 of [4] with $\mathcal{V}_n \le M_d$ yields that:

$$P\left(\frac{\|B_n V_\perp V_\perp^T g\|}{(1+\eta_n\lambda_1/n)^n} \ge \epsilon/2 \,\cap\, \mathcal{G}\right)$$
$$\le \frac{d\exp\left(-\eta_n(\lambda_1-\lambda_2)+\frac{\eta_n^2}{n}(\lambda_1^2+M_d)\right)+\frac{\eta_n^2}{n}M_d\exp\left(\frac{\eta_n^2}{n}\right)}{4\delta^2\log^{-1}(1/\delta)\,\epsilon^2\left(1+\frac{\eta_n^2\lambda_1^2}{n}\right)}$$

Now we will bound the first summand. By Lemma B.1 Eq S.20, we have by Markov's inequality,

$$P\left(\frac{\|(B_n-T_0)\|_{op}}{(1+\eta_n\lambda_1/n)^n} > \epsilon/2\right) \le \frac{e^2 M_d(1+\log d)}{n\epsilon^2}$$

Combining the two bounds and the probability of $\mathcal{G}^c$, the result follows.

□

Lemma B.3 (Negligibility of $V_\perp$ for the CLT)**.** *Let $V_\perp$ denote the matrix of eigenvectors orthogonal to v_1. Also let λ_i denote the i^{th} largest eigenvalue of Σ. For some $C>0$, and any $\epsilon>0$, $0<\delta<1$,*

$$P\left(\sqrt{\frac{n}{\eta_n}}\,\frac{\left\|V_\perp V_\perp^T B_n V_\perp V_\perp^T u_0\right\|}{|a_1|(1+\frac{\eta_n\lambda_1}{n})^n} \ge \epsilon\right)$$
$$\le \frac{nd\log(1/\delta)\exp\left\{-2\eta_n(\lambda_1-\lambda_2)+\eta_n^2(\lambda_1^2+M_d)/n\right\}}{\eta_n\epsilon^2\delta^2}+\frac{eM_d^2(1+2\log d)\eta_n^2\epsilon^{-2}\log(1/\delta)\delta^{-2}}{n2(\lambda_1-\lambda_2)+\eta_n^2(\lambda_1^2-\lambda_2^2-M_d)}+C\delta$$

Proof. We consider bounding the squared quantity. We have, with probability at least $1-C\delta$, using Proposition B.6, this quantity is upper bounded by:

$$\frac{\left\|(V_\perp V_\perp^T B_n V_\perp V_\perp)g\right\|^2}{(v_1^T g)^2(1+\eta_n\lambda_1/n)^{2n}}$$
$$\le \frac{\operatorname{trace}\left((V_\perp V_\perp^T B_n V_\perp V_\perp^T)(V_\perp V_\perp^T B_n V_\perp V_\perp^T)^T\right)}{\delta_n(v_1^T g)^2(1+\eta_n\lambda_1/n)^{2n}}$$
$$= \frac{\operatorname{trace}\left(V_\perp^T B_n V_\perp V_\perp^T B_n V_\perp\right)}{\delta_n^3(1+\eta_n\lambda_1/n)^{2n}}$$

Now we will bound the expectation of the numerator.

We will denote $\eta=\frac{\eta_n}{n}$ for simplicity. Let $U_i = I+\eta X_i X_i^T$ and $Y_i = X_i X_i^T - \Sigma$. We have that:

$$\begin{aligned}\alpha_n &:= \mathbb{E}\left\langle B_n V_\perp V_\perp^T B_n^T, V_\perp V_\perp^T\right\rangle\\ &= \mathbb{E}\left\langle B_{n-1}V_\perp V_\perp^T B_{n-1}^T, U_n V_\perp V_\perp^T U_n^T\right\rangle\\ &= \left\langle \mathbb{E}B_{n-1}V_\perp V_\perp^T B_{n-1}^T, \mathbb{E}U_n V_\perp V_\perp^T U_n^T\right\rangle \end{aligned} \tag{S.22}$$

Now we have:

$$\begin{aligned}\mathbb{E}U_n V_\perp V_\perp^T U_n^T &= \mathbb{E}\left(I+\eta\Sigma\right)V_\perp V_\perp^T\left(I+\eta\Sigma\right)^T+\eta^2\mathbb{E}Y_n V_\perp V_\perp^T Y_n^T\\ &\preceq (1+2\eta\lambda_2+\lambda_2^2\eta^2)V_\perp V_\perp^T+\eta^2 M_d(V_\perp V_\perp^T+v_1v_1^T)\\ &\preceq (1+2\eta\lambda_2+\lambda_2^2\eta^2+\eta^2 M_d^2)V_\perp V_\perp^T+\eta^2 M_d v_1 v_1^T\end{aligned} \tag{S.23}$$

Finally, using Eqs S.22 and S.23, we have:

$$\alpha_n \le \left(1+2\eta\lambda_2+\eta^2(\lambda_2^2+M_d)\right)\alpha_{n-1}+\eta^2 M_d\left\langle \mathbb{E}B_{n-1}V_\perp V_\perp^T B_{n-1}^T, v_1v_1^T\right\rangle \tag{S.24}$$

We will use the fact that,

$$\langle (I+\eta\Sigma)^{n-1}V_\perp V_\perp^T(I+\eta\Sigma)^{n-1}, v_1v_1^T\rangle = 0.$$

Thus, for some N such that the condition $\eta_n^2 M_d(1+2\log d)/n \le 1$ holds for all rows of the triangular array with index $n > N$, we have by Lemma B.1,

$$\begin{aligned}
&\left\langle \mathbb{E}B_{n-1}V_\perp V_\perp^T B_{n-1}^T, v_1 v_1^T \right\rangle \\
&= \left\langle \mathbb{E}(B_{n-1} - (I+\eta\Sigma)^{n-1})V_\perp V_\perp^T (B_{n-1} - (I+\eta\Sigma)^{n-1})^T, v_1 v_1^T \right\rangle \\
&\le \|\mathbb{E}(B_{n-1} - (I+\eta\Sigma)^{n-1})V_\perp V_\perp^T (B_{n-1} - (I+\eta\Sigma)^{n-1})^T\| \\
&\le \mathbb{E}\|B_{n-1} - (I+\eta\Sigma)^{n-1}\|^2 \\
&\le M_d e\eta^2 n(1+2\log d)(1+\eta_n\lambda_1/n)^{2(n-1)}.
\end{aligned}$$

Thus, Eq S.24 gives:

$$\begin{aligned}
\alpha_n &\le \underbrace{\left(1+2\eta\lambda_2+\eta^2(\lambda_2^2+M_d)\right)}_{c_1}\alpha_{n-1} + \eta^4 M_d^2 e(1+2\log d)\underbrace{(n-1)(1+\eta\lambda_1)^{2(n-1)}}_{(n-1)c_2^{n-1}} \\
&= c_1\alpha_{n-1} + \eta^4 M_d^2 e(1+2\log d)(n-1)c_2^{n-1} \\
&= c_1^n\alpha_0 + \eta^4 M_d^2 e(1+2\log d)\sum_i c_1^{i-1}(n-i)c_2^{n-i} \\
&\le c_2^n\left(d(c_1/c_2)^n + \frac{eM_d^2(1+2\log d)\eta^4 n}{c_2-c_1}\right) \\
&\le (1+\eta_n\lambda_1/n)^{2n}\Bigg(d(1-\lambda_1^2\eta_n^2/n)\exp\{-2\eta_n(\lambda_1-\lambda_2)+\eta_n^2(\lambda_1^2+M_d)/n\} \\
&\qquad + \frac{eM_d^2(1+2\log d)\eta_n^3/n^2}{2(\lambda_1-\lambda_2)+\eta_n^2/n(\lambda_1^2-\lambda_2^2-M_d)}\Bigg)
\end{aligned}$$

where above we used the fact $e^x(1-\frac{x^2}{n}) \le (1+\frac{x}{n})^n \le e^x$ for $|x| \le n$ to bound $(c_1/c_n)^n$.

$\square$

Lemma B.4 (Negligibility of higher-order Hoeffding projections for the CLT)**.** *Let* $\beta_n = \frac{\eta_n^2 M_d}{n}$ *and suppose that* $0 \le \beta_n \le 1$*. Then, for some* $C > 0$ *and any* $\epsilon > 0$*,*

$$P\left(\frac{\sqrt{\frac{n}{\eta_n}}\left\|V_\perp V_\perp^T \sum_{k>1} T_k v_1\right\|}{(1+\frac{\eta_n\lambda_1}{n})^n} > \epsilon\right) \le \frac{C\beta_n\eta_n}{(1-\beta_n)\epsilon^2}$$

Proof. By Markov's inequality, it follows that:

$$P\left(\frac{\frac{\sqrt{n}}{\eta_n}\left\|V_\perp V_\perp^T \sum_{k>1} T_k v_1\right\|}{(1+\frac{\eta_n\lambda_1}{n})^n} > \epsilon\right) \le \frac{\frac{n}{\eta_n^2}\mathbb{E}\left[\left\|V_\perp V_\perp^T \sum_{k>1} T_k v_1\right\|^2\right]}{\epsilon^2(1+\frac{\eta_n\lambda_1}{n})^{2n}}$$

Now, by submultiplicativity of the operator norm and the fact that $\mathbb{E}[(P_{S_1}T)^T(P_{S_2})T] = 0$ for any two Hayek projections, the numerator is upper bounded by:

$$\begin{aligned}
\left(\frac{n}{\eta_n}\right)\sum_{k=2}^n\left(\frac{\eta_n}{n}\right)^{2k}\sum_{|S|=k}\mathbb{E}\left[(v'A_S u_0)^2\right] &\le \left(\frac{n}{\eta_n}\right)\sum_{k=2}^n\sum_{|S|=k}\left(\frac{\eta_n}{n}\right)^{2k}\mathbb{E}\left[\|A_S\|_{op}^2\right] \\
&\le \left(\frac{n}{\eta_n}\right)\sum_{k=2}^n\left(\frac{\eta_n}{n}\right)^{2k}\sum_{|S|=k}\left(1+\frac{\eta_n\lambda_1}{n}\right)^{2(n-k)} M_d^k \\
&\le \eta_n M_d\left(1+\frac{\eta_n\lambda_1}{n}\right)^{2n}\sum_{k=2}^n\left(\frac{M_d\eta_n^2}{n}\right)^{k-1} \\
&\le \left(1+\frac{\eta_n\lambda_1}{n}\right)^{2n}\frac{\beta_n\eta_n M_d}{1-\beta_n}
\end{aligned}$$

The result follows.

□

C Consistency of the online bootstrap

In this section, we provide the detailed proof of Bootstrap consistency, i.e Theorem 2.

C.1 Proof of bootstrap consistency

Proof of Theorem 2. Similar to the CLT, we will establish the negligibility of remainder terms and then use anti-concentration terms to argue that the contribution to the Kolmogorov distance is small. We then show that the bootstrap covariance of the main term approaches the weighted χ^2 approximation in Theorem 1 with high probability. Let $\widehat{v}_1$ denote the leading eigenvector estimated from Oja's algorithm and let $\widehat{V}_\perp$ denote its orthogonal complement. Again, we have that:

$$\frac{n}{\eta_n}\sin^2(v_1^*, \hat{v}_1) = \frac{n}{\eta_n}\frac{(B_n^* u_0)^T \widehat{V}_\perp \widehat{V}_\perp^T (B_n^* u_0)}{\|B_n^* u_0\|^2}$$
$$= \frac{(\sqrt{n/\eta_n}\widehat{V}_\perp \widehat{V}_\perp^T B_n^* u_0)^T (\sqrt{n/\eta_n}\widehat{V}_\perp \widehat{V}_\perp^T B_n^* u_0)}{\|B_n^* u_0\|^2}$$

We aim to show that the bootstrap distribution conditional on the data is close to the weighted χ^2 approximation with high probability; therefore we may work the good set $\mathcal{A}_n$. With the a slight abuse of notation, in the remainder terms below, O_P will be on the measure restricted to $\mathcal{A}_n$.

We first approximate the norm using Lemma C.7. Analogous to the CLT, the corresponding remainder is given by:

$$R_1^* = \frac{\|B_n^* u_0\|^2}{a_1^2(1+\frac{\eta_n}{n}\lambda_1)^{2n}} = 1 - O_P\left(\sqrt{d}\exp\left(-\frac{\eta_n}{2}(\lambda_1-\lambda_2)\right) + \sqrt{\frac{\eta_n^2 M_d \log d}{n}} + \frac{\eta_n \alpha_n}{\sqrt{n}}\right)$$

Next, we bound the contribution of the higher-order Hoeffding projections. This step is different from the CLT in the sense that we handle both v_1 and $V_\perp$, using the fact that on the good set, even the Frobenius norm of certain terms are well-behaved. By Lemma C.8 we have that:

$$R_3^* := \sqrt{\frac{n}{\eta_n}} \cdot \frac{\widehat{V}_\perp \widehat{V}_\perp^T (B_n^* - T_1^*) u_0}{|a_1|(1+\eta_n/n\lambda_1)^n} = O_P\left(\exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}}\right)\sqrt{\frac{\alpha_n^4 \eta_n^3}{n}}\right)$$

Next, we bound the contribution of $V_\perp$ to the Hájek projection using Lemma C.10, as long as $\lambda_1 M_d (\log d)^2 \frac{\eta_n^2}{n} \to 0$,

$$R_2^* = \sqrt{\frac{n}{\eta_n}} \cdot \frac{\widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp V_\perp^T u_0}{|a_1|(1+\eta_n/n\lambda_1)^n} = O_P\left(\sqrt{\frac{\alpha_n M_d \eta_n^2}{n(\lambda_1-\lambda_2)}}\right)$$

The final remainder term arises from the disparity between the orthogonal complements and the residuals of matrix products from their expectation. By Lemma C.6, with $\Delta_i = X_i X_i^T - X_{i-1} X_{i-1}^T$,

$$R_4^* = \sqrt{\frac{n}{\eta_n}}\left\|\frac{\widehat{V}_\perp \widehat{V}_\perp^T T_1^* v_1 (v_1^T u_0)}{|v_1^T u_0|(1+\eta_n \lambda_1/n)^n} - \frac{\eta_n}{n}\sum_i W_i D_{i-1}\Delta_i v_1\right\| = O_P\left(\sqrt{\frac{M_d \alpha_n \eta_n^3 \log d}{n}}\right)$$

Now, define:

$$S_n^* = \sqrt{\frac{n}{\eta_n}}\frac{V_\perp V_\perp^T T_1^* v_1}{(1+\frac{\eta_n \lambda_1}{n})^n}$$

Consider the following bound:

$$
\begin{aligned}
&P\left\{\sup_{t\in\mathbb{R}}\left|P^*(n/\eta_n \sin^2(v_1^*,\widehat{v}_1)\le t)-P(Z^TZ\le t)\right|>\epsilon\right\}\\
&=P_{\mathcal{A}_n}\left\{\sup_{t\in\mathbb{R}}\left|P^*\left(R_1^*\cdot\frac{(S_n^*+R_2^*+R_3^*+R_4^*)^T(S_n^*+R_2^*+R_3^*+R_4^*)}{f}\le t\right)-P\left(\frac{Z^TZ}{f}\le t\right)\right|>\epsilon\right\}\\
&+P_{\mathcal{A}_n^c}\left\{\sup_{t\in\mathbb{R}}\left|P^*\left(R_1^*\cdot\frac{(S_n^*+R_2^*+R_3^*+R_4^*)^T(S_n^*+R_2^*+R_3^*+R_4^*)}{f}\le t\right)-P\left(\frac{Z^TZ}{f}\le t\right)\right|>\epsilon\right\}
\end{aligned}
\tag{S.25}
$$

The second term is easily upper-bounded by $P(\mathcal{A}_n^c)\to 0$, so we will bound the first term. To lower bound the Kolmogorov metric, we may follow the same reasoning used in Eqs S.12, S.14, S.15, to deduce, on the good set $\mathcal{A}_n$, we have the lower bound:

$$
\begin{aligned}
&P^*\left(\frac{S_n^{*T}S_n^*}{f}\le\frac{t}{1+\epsilon_n}-\epsilon_n\right)-P\left(\frac{Z^TZ}{f}\le\frac{t}{1+\epsilon_n}-\epsilon_n\right)\\
&+P\left(\frac{Z^TZ}{f}\le\frac{t}{1+\epsilon_n}-\epsilon_n\right)-P\left(\frac{Z^TZ}{f}\le t\right)-P^*\left(G_{boot}\cap\mathcal{A}_n\right)=I^*+II^*+III^*
\end{aligned}
$$

where G_{boot} satisfies $P(G_{boot}^c)=0$ and for some $\epsilon_n\to 0$, is defined as:

$$G_{boot}=\{|R_1^*-1|\le\epsilon_n,|R_2^*|,|R_3^*|,|R_4^*|\le\epsilon_n\ \}$$

For I, we may use Lemma 1, which establishes that bootstrap version of the covariance matrix, which consists of empirical covariances, is close to the Gaussian approximation, implying, by our Gaussian comparison result Lemma C.5:

$$I^*=O_P\left(\left(\frac{\mathbb{E}[\left\|X_iX_i^T-\Sigma\right\|^4]}{n(\lambda_1-\lambda_2)\left\|M\right\|_F^2}\right)^{1/4}\right)$$

For II^*, we may use the anti-concentration result and $P^*(G_{boot}\cap\mathcal{A}_n)\xrightarrow{P}0$ by Markov's inequality since the Lemmas hold for the unconditional measure, which is the expectation of the bootstrap measure. We may use analogous reasoning to the CLT for the upper bound and the result follows.

□

C.2 Proof of Lemma 1

Proof. Let $Y_i:=X_iX_i^T-\Sigma$. Also let $M_i=\mathbb{E}[D_{i-1}Y_iv_1v_1^TY_iD_{i-1}]$. First note that

$$
\begin{aligned}
\mathbb{E}^*ZZ^T-\bar{\mathbb{V}}_n&=\frac{\eta_n}{2n}\sum_i D_{i-1}(Y_i-Y_{i-1})v_1v_1^T(Y_i-Y_{i-1})D_{i-1}\\
&=\frac{\eta_n}{n}\sum_i\frac{\left(D_{i-1}Y_iv_1v_1^TY_iD_{i-1}-M_i\right)+\left(D_{i-1}Y_{i-1}v_1v_1^TY_{i-1}D_{i-1}-M_i\right)}{2}\\
&+\frac{\eta_n}{n}\sum_i\left(D_{i-1}Y_iv_1v_1^TY_{i-1}D_{i-1}+D_{i-1}Y_{i-1}v_1v_1^TY_iD_{i-1}\right)
\end{aligned}
\tag{S.26}
$$

We first compute trace.

$$
\begin{aligned}
\mathrm{trace}(\mathbb{E}^*ZZ^T-\bar{\mathbb{V}}_n)&=\frac{\eta_n}{2n}\sum_i\underbrace{\left(\|D_{i-1}Y_iv_1\|^2-\mathbb{E}\|D_{i-1}Y_iv_1\|^2\right)}_{U_{1,i}}\\
&+\frac{\eta_n}{2n}\sum_i\underbrace{\left(\|D_{i-1}Y_{i-1}v_1\|^2-\mathbb{E}\|D_{i-1}Y_i\|^2\right)}_{U_{2,i}}\\
&+\frac{\eta_n}{n}\sum_i\underbrace{v_1Y_iD_{2(i-1)}Y_{i-1}v_1}_{U_{3,i}}
\end{aligned}
$$

The last step is true because $D_{i-1}^2 = D_{2(i-1)}$. We start with the first term.

$$\mathbb{E}U_{i,1}^2 \leq \mathbb{E}\|D_{i-1}Y_i v_1\|^4 \leq \mathbb{E}\|Y_i\|^4 \left(\frac{1+\eta_n\lambda_2/n}{1+\eta_n\lambda_1/n}\right)^{4(i-1)}$$

$$\begin{aligned}
\mathrm{Var}(\sum_i U_{1,i}) &\leq \mathbb{E}\|Y_1\|^4 \sum_i \left(\frac{1+\eta_n\lambda_2/n}{1+\eta_n\lambda_1/n}\right)^{4(i-1)} \\
&\leq \frac{n}{\eta_n(\lambda_1-\lambda_2)} \\
&\leq \frac{n}{\eta_n}\mathbb{E}\|Y_1\|^4 \min\left(\frac{1}{\lambda_1-\lambda_2}, \eta_n\right)
\end{aligned}$$

Finally,

$$\mathbb{E}[U_{3,i}^2] \leq \mathbb{E}\left(v_1 Y_i D_{2(i-1)} Y_{i-1} v_1\right)^2 \leq M_d^2 \left(\frac{1+\eta_n\lambda_2/n}{1+\eta_n\lambda_1/n}\right)^{2(i-1)}$$

Thus, we have

$$\frac{\eta_n}{2n}\sum_i U_{1,i} = O_P\left(\sqrt{\frac{\mathbb{E}\|Y_1\|^4}{n(\lambda_1-\lambda_2)}}\right)$$

We also have,

$$\frac{\eta_n}{2n}\sum_i U_{2,i} = O_P\left(\sqrt{\frac{\mathbb{E}\|Y_1\|^4}{n(\lambda_1-\lambda_2)}}\right)$$

Also note that while $U_{3,i}$ terms are 1-dependent, they are in fact uncorrelated. Thus, we have:

$$\mathrm{Var}(\sum_i U_{3,i}) \leq \frac{M_d^2 n}{(\lambda_1-\lambda_2)},$$

and,

$$\mathrm{trace}(\mathbb{E}^* ZZ^T - \bar{\mathbb{V}}_n) = O_P\left(\sqrt{\frac{\mathbb{E}\|X_i X_i^T - \Sigma\|^4}{n(\lambda_1-\lambda_2)}}\right)$$

Now we bound the Frobenius norm. We will start with the expected Frobenius norm of the first term of Eq S.26.

$$\begin{aligned}
A_1 &= \mathbb{E}\left\|\frac{1}{2n}\sum_{i=1}^n D_{i-1}Y_i v_1 v_1^T Y_i D_{i-1} - M_i\right\|_F^2 \\
&\leq \frac{1}{4n^2}\sum_i \mathbb{E}\|D_{i-1}Y_i v_1 v_1^T Y_i D_{i-1}\|_F^2 \leq \frac{E\|Y_1\|^4}{4n\eta_n(\lambda_1-\lambda_2)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
A_2 &= \mathbb{E}\left\|\frac{1}{n}\sum_i D_{i-1}Y_i v_1 v_1^T Y_{i-1} D_{i-1}\right\|_F^2 \\
&\leq \frac{1}{n\eta_n(\lambda_1-\lambda_2)} M_d^2
\end{aligned}$$

Thus ,

$$\left\|\mathbb{E}^* ZZ^T - \bar{\mathbb{V}}_n\right\|_F = O_P\left(\sqrt{\frac{\mathbb{E}\|X_1 X_1^T - \Sigma\|^4}{n(\lambda_1-\lambda_2)}}\right)$$

$\square$

C.3 The Gaussian comparison lemma

We use the following lemma to compare to Gaussian random variables with mean 0 and different covariance matrices. Our result is related to [2], but our lemma below is easier to implement and does not require that $3\|\Sigma\|^2 \leq \|\Sigma\|_F^2$.

Lemma C.5. *[Comparison lemma for inner products of Gaussian random variables]*

Suppose that $Z \sim N(0, \mathbb{V})$, $\check{Z} \sim N(0, \check{\mathbb{V}})$, $f = \|\mathbb{V}\|_F$, and $\Delta_1 = \mathrm{tr}(\mathbb{V} - \check{\mathbb{V}})$. Then, there exists some constant $K > 0$ such that for any $\epsilon > 0$,

$$\sup_{t \in \mathbb{R}} \left| P(Z^T Z \leq t) - P(\check{Z}^T \check{Z} \leq t) \right| \lesssim \sqrt{\frac{|\Delta_1| + \epsilon}{f}} + \exp\left\{ -\left(\frac{\epsilon^2}{K^2 \|\mathbb{V} - \check{\mathbb{V}}\|_F^2} \bigwedge \frac{\epsilon}{K\|\mathbb{V} - \check{\mathbb{V}}\|} \right) \right\}$$

Proof. Let $\lambda_1 \geq \ldots \geq \lambda_p$ denote the eigenvalues $\mathbb{V}$, $\gamma \geq \ldots \geq \gamma_p$ denote the eigenvalues of $\check{\mathbb{V}}$. Recall that $Z^T Z \sim \sum_{r=1}^p \lambda_r \eta_r$, $\check{Z}^T \check{Z} \sim \sum_{r=1}^p \gamma_r \eta_r$, where $\eta_r \sim \chi^2(1)$. We upper bound the difference between the CDFs uniformly in t; the argument for the lower bound is analogous. For $\epsilon > 0$, let $t' = t - |\Delta_1| - \epsilon$. It follows that:

$$\begin{aligned}
&P(Z^T Z \leq t) - P(\check{Z}^T \check{Z} \leq t) \\
&= P\left(\frac{\sum_{r=1}^p \lambda_r \eta_r}{f} \leq \frac{t}{f} \right) - P\left(\frac{\sum_{r=1}^p \lambda_r \eta_r + \sum_{r=1}^p (\gamma_r - \lambda_r)\eta_r - \Delta_1}{f} \leq \frac{t - \Delta_1}{f} \right) \\
&\leq P\left(\frac{t'}{f} \leq \frac{\sum_{r=1}^p \lambda_r \eta_r}{f} \leq \frac{t' + |\Delta_1| + \epsilon}{f} \right) + P\left(\left| \sum_{r=1}^p (\gamma_r - \lambda_r)\eta_r - \Delta_1 \right| > \epsilon \right)
\end{aligned}$$

Observe that $\sum_{r=1}^p (\lambda_r - \gamma_r)^2 \leq \|\mathbb{V} - \check{\mathbb{V}}\|_F^2$ by Hoffman-Wielandt inequality and $\max_r |\lambda_r - \gamma_r| \leq \|\mathbb{V} - \check{\mathbb{V}}\|_{op}$ by Weyl's inequality. Since $\chi^2(1)$ is sub-Exponential, by Bernstein's inequality (see for example Theorem 2.8.2 of [7]:

$$P\left(\left| \sum_{r=1}^p (\gamma_r - \lambda_r)\eta_r - \Delta_1 \right| > \epsilon \right) \leq \exp\left\{ -\left(\frac{\epsilon^2}{K^2 \|\mathbb{V} - \check{\mathbb{V}}\|_F^2} \bigwedge \frac{\epsilon}{K\|\mathbb{V} - \check{\mathbb{V}}\|} \right) \right\}$$

The upper bound follows from an application of Proposition B.7. The lower bound is analogous.

□

C.4 Other supporting lemmas for bootstrap consistency

Before presenting our supporting lemmas, we present some events we will use frequently. Let $\mathcal{A}_{\sin}$ denote the set

$$\mathcal{A}_{\sin} := \left\{ 1 - (v_1^T \hat{v}_1)^2 \leq \frac{\epsilon_{\sin}}{\delta_n} \right\}. \tag{S.27}$$

Using Corollary 1, and the remark thereafter, we have:

$$P\left(1 - (v_1^T \hat{v}_1)^2 \geq \frac{\epsilon_{\sin}}{\delta_n} \right) \leq \delta_n, \tag{S.28}$$

where, under the assumptions of Theorem 1,

$$\epsilon_{\sin} = C_3 \frac{M_d \eta_n}{n(\lambda_1 - \lambda_2)} \tag{S.29}$$

Also let,

$$\mathcal{A}_n = \left\{ \max_{1 \leq i \leq n} \|X_i\|_2^2 \leq \alpha_n \right\} \tag{S.30}$$

Lemma C.6. *[Bounding the norm of bootstrap residual from T_1^*] Let $\Delta_i = X_iX_i^T - X_{i-1}X_{i-1}^T$ and assume the conditions in Theorem 1. Let $D_i = V_\perp \Lambda_\perp^i V_\perp^T$, where $\Lambda_\perp(k,\ell) = \frac{1+\eta_n\lambda_{k+1}/n}{1+\eta_n\lambda_1/n} 1(k=\ell)$. For any $\epsilon, \delta > 0$, we have:*

$$
\begin{aligned}
&P\left(\left\{\sqrt{\frac{n}{\eta_n}}\left\|\frac{\widehat{V}_\perp \widehat{V}_\perp^T T_1^* v_1(v_1^T u_0)}{|v_1^T u_0|(1+\eta_n\lambda_1/n)^n} - \frac{\eta_n}{n}\sum_i W_i D_{i-1}\Delta_i v_1\right\| \geq \epsilon\right\} \cap \mathcal{A}_n\right) \\
&\leq C'' \frac{\alpha_n M_d \eta_n^3 \log d}{n\epsilon^2\delta} + \delta
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\frac{\widehat{V}_\perp \widehat{V}_\perp^T T_1^* v_1(v_1^T u_0)}{|v_1^T u_0|(1+\eta_n\lambda_1/n)^{n-1}} \\
&= \operatorname{sign}(v_1^T u_0)\frac{\eta_n}{n}\sum_i W_i D_{i-1}\Delta_i v_1 \\
&+ \operatorname{sign}(v_1^T u_0)\frac{\eta_n}{n}\underbrace{(\widehat{V}_\perp \widehat{V}_\perp^T - V_\perp V_\perp^T)\sum_i W_i D_{i-1}\Delta_i v_1}_{r_1} \\
&+ \operatorname{sign}(v_1^T u_0)\frac{\eta_n}{n}\left(\underbrace{\sum_i W_i\left(\frac{R_{1,i-1}\Delta_i v_1}{(1+\lambda_1\eta_n/n)^i}\right)}_{r_2}\right. \\
&\qquad \left. + \underbrace{\frac{W_i(I+\eta_n\lambda_1/n)^{i-1}\Delta_i R_{i,n} v_1}{(1+\lambda_1\eta_n/n)^{n-1}}}_{r_3} + \underbrace{W_i\frac{R_{1,i-1}\Delta_i R_{i,n} v_1}{(1+\lambda_1\eta_n/n)^{n-1}}}_{r_4}\right)
\end{aligned}
$$

Define

$$
\mathcal{B}_{1,j} = \prod_{i=1}^{j}\left(I + \frac{\eta_n}{n}X_iX_i^T\right) \qquad \mathcal{B}_{j,n} = \prod_{i=j}^{n}\left(I + \frac{\eta_n}{n}X_iX_i^T\right) \tag{S.31}
$$

When $j=0$, $\mathcal{B}_{1,j} = I$.

Using Lemma B.1 we have:

$$
R_{1,i} = \mathcal{B}_{1,i} - (I+\eta_n\Sigma/n)^i \qquad R_{i,n} = \mathcal{B}_{i,n} - (I+\eta_n\Sigma/n)^{n-i} \tag{S.32}
$$

$$
\mathbb{E}\|R_{1,i-1}\|^2 \leq eM_d(1+2\log d)\frac{\eta_n^2}{n^2} i\,(1+\eta_n\lambda_1/n)^{2i} \tag{S.33}
$$

$$
\mathbb{E}\|R_{i,n}\|^2 \leq eM_d(1+2\log d)\frac{\eta_n^2}{n^2}(n-i)\,(1+\eta_n\lambda_1/n)^{2(n-i)} \tag{S.34}
$$

We have, on the good set $\mathcal{A}_{\sin}$,

$$
\mathbb{E}^*\|r_1\|^2 \leq n\alpha_n\frac{\epsilon_{\sin}}{\delta_n}
$$

We also have:

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{E}^*\|r_2\|^2 1(\mathcal{A}_n)\right] &\leq \frac{\eta_n^2}{n^2}\alpha_n\sum_i \mathbb{E}[\|R_{1,i}^2 1(\mathcal{A}_n)\|^2] \\
&\leq eM_d(1+2\log d)\alpha_n\eta_n^2
\end{aligned}
$$

The last step is true because $\mathbb{E}[\|R_{1,i}^2 1(\mathcal{A}_n)\|^2] \leq \mathbb{E}[\|R_{1,i}^2\|^2]$. Similarly

$$\mathbb{E}\left[\mathbb{E}^*\|r_3\|^2 1(\mathcal{A}_n)\right] \leq eM_d(1+2\log d)\alpha_n\eta_n^2$$

and

$$\mathbb{E}\left[\mathbb{E}^*\|r_4\|^2 1(\mathcal{A}_n)\right] \leq e^2 M_d^2(1+2\log d)^2\alpha_n\eta_n^4/n$$

Finally, we have:

$$\begin{aligned}
&P\left(\left\{\frac{\eta_n}{n}\|\sum_j r_j\|^2 \geq \epsilon\right\} \cap \mathcal{A}_n\right) \leq P\left(\left\{4\frac{\eta_n}{n}\sum_j \|r_j\|^2 \geq \epsilon\right\} \cap \mathcal{A}_n\right)\\
&\leq \sum_i P\left(\left\{\|r_i\|^2 \geq \frac{n\epsilon}{16\eta_n}\right\} \cap \mathcal{A}_n \cap \mathcal{A}_{\sin}\right) + \delta_n\\
&\leq C\sum_i \mathbb{E}\left[\mathbb{E}^*\|r_i\|^2 1(\mathcal{A}_n \cap \mathcal{A}_{\sin})\right] \times \frac{\eta_n}{n\epsilon} + \delta_n\\
&\overset{(i)}{\leq} C'\left(n\alpha_n\frac{\epsilon_{\sin}}{\delta_n} + M_d\log d\alpha_n\eta_n^2\right) \times \frac{\eta_n}{n\epsilon} + \delta_n\\
&\overset{(ii)}{\leq} C''\frac{\alpha_n M_d\eta_n^3\log d}{n\epsilon\delta_n} + \delta_n
\end{aligned}$$

Step (i) is true because $M_d\log d\eta_n^2/n \to 0$. Step (ii) is true because of Eq S.29. Now setting δ_n to any $\delta > 0$ gives the result.

□

Lemma C.7 (Concentration of the norm for the bootstrap)**.** *Let u_0 be uniformly distributed on $\mathbb{S}^{d-1}$ and $a_1 = u_0'v_1$ and $V_\perp V_\perp^T$ is orthogonal complement. Suppose that $(\alpha_n)_{n\geq1}$ satisfies $0 \leq \frac{(\eta_n\alpha_n)^2}{n} \leq 1$. Then, for any $\epsilon > 0, 0 < \delta < 1$ and some $C > 0$,*

$$\begin{aligned}
&P\left(\left\{\left|\frac{\|B_n^*u_0\|}{|a_1|(1+\eta_n\lambda_1/n)^n} - 1\right| \geq \epsilon_n\right\} \bigcap \mathcal{A}_n\right)\\
&\leq \frac{d\exp\left(-\eta_n(\lambda_1-\lambda_2) + \frac{\eta_n^2}{n}(\lambda_1^2+M_d)\right) + \frac{\eta_n^2}{n}M_d\exp\left(\frac{\eta_n^2}{n}\right)}{8\log^{-1}(1/\delta)\delta^2\,\epsilon^2\left(1+\frac{\eta_n^2\lambda_1^2}{n}\right)}\\
&\quad + \frac{e^2\eta_n^2 M_d(1+\log d)}{2n\epsilon^2} + \frac{C\beta_n^*\log(1/\delta)}{(1-\beta_n^*)\delta^2\epsilon^2} + C\delta,
\end{aligned}$$

where β_n^* is defined in (S.36) and $\mathcal{A}_n$ is defined in Eq S.30.

Proof. First note that we may reduce the problem as follows:

$$\begin{aligned}
&P\left(\left\{\left|\frac{\|B_n^*u_0\|}{|a_1|(1+\eta_n\lambda_1/n)^n} - 1\right| \geq \epsilon\right\} \cap \mathcal{A}_n\right)\\
&\leq P\left(\left\{\frac{\|B_n^*u_0 - B_nu_0\|_2}{|a_1|(1+\eta_n\lambda_1/n)^n} + \left|\frac{\|B_nu_0\|_2}{|a_1|(1+\eta_n\lambda_1/n)^n} - 1\right| > \epsilon\right\} \cap \mathcal{A}_n\right)\\
&\leq \mathbb{E}\left[P^*\left(\frac{\|B_n^*u_0 - B_nu_0\|_2}{|a_1|(1+\eta_n\lambda_1/n)^n} > \frac{\epsilon}{2}\right)1(\mathcal{A}_n)\right] + P\left(\left|\frac{\|B_nu_0\|_2}{|a_1|(1+\eta_n\lambda_1/n)^n} - 1\right| > \frac{\epsilon}{2}\right)
\end{aligned}$$

The bound for the second term follows from Lemma B.2. For the first term, we invoke Proposition B.6 so that, with probability at least $1 - C\delta$,

$$\frac{\|(B_n^* - B_n)g\|_2^2}{(v_1^Tg)^2(1+\eta_n\lambda_1/n)^{2n}} \leq \frac{\log(1/\delta)\,\|B_n^* - B_n\|_F^2}{\delta^2(1+\eta_n\lambda_1/n)^{2n}}$$

Now, using the fact that for any two Hayek projections P_S^* and P_T^*, $\mathbb{E}[(P_S^*)^T P_T^*] = 0$ and for any two matrices $\|AB\|_F \leq \|A\|_F \|B\|_{op}$, we have on the high probability set:

$$\begin{aligned}&\mathbb{E}^* \|B_n^* - B_n\|_F^2 \\ &\leq \sum_{k=1}^{n} \sum_{|S|=k} \left(\frac{\eta_n}{n}\right)^{2k} \prod_{i=1}^{k} \left\| X_{S[i]} X'_{S[i]} - X_{S[i]-1} X'_{S[i]-1} \right\|_F^2 \prod_{j=1}^{k+1} \left\| \mathcal{B}_{j,n}^{(S)} \right\|_{op}^2,\end{aligned}$$

where $\mathcal{B}_{j,n}^{(S)}$ denotes a contiguous block of $I + \frac{\eta_n}{n} X_i X_i^T$ only. More precisely, suppose $|S| = k$. Let $S[i]$ denote the ith element of S, with $S[0] = 0$ and $S[k+1] = n-1$. For each $1 \leq j \leq k+1$ if $S[j] > S[j-1] + 1$ define $\mathcal{B}_{j,n}$ as:

$$\mathcal{B}_{j,n}^{(S)} = \prod_{i=S[j-1]+1}^{S[j]-1} \left(I + \frac{\eta_n}{n} X_i X_i^T \right) \tag{S.35}$$

otherwise, set $\mathcal{B}_{j,n}^{(S)} = I$. Now, we may repeat arguments in Lemma C.8 equations (S.37), (S.38), and (S.39) to conclude that, for some $C > 0$,

$$P\left(\frac{\log(1/\delta)\, \|B_n^* - B_n\|_F^2}{\delta^2 (1 + \eta_n \lambda_1/n)^{2n}} > \epsilon \bigcap \mathcal{A}_n \right) \leq \frac{C \log(1/\delta) \beta_n^*}{(1-\beta_n^*)\delta^2 \epsilon^2}$$

The result follows. $\square$

Lemma C.8 (Negligibility of higher-order Hoeffding projections for the bootstrap)**.** *Suppose α_n is defined so that $0 \leq \beta_n^* \leq 1$, where*

$$\beta_n^* = \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right) \frac{4\eta_n^2 \alpha_n^2}{n} \tag{S.36}$$

Then for any $\epsilon > 0, 0 < \delta < 1$ and for some $C > 0$,

$$\begin{aligned}&P\left(\left\{ \frac{\sqrt{\frac{n}{\eta_n}} \left\| \hat{V}_\perp \hat{V}_\perp^T \sum_{k>1} T_k^* u_0 \right\|}{|a_1| (1 + \frac{\eta_n \lambda_1}{n})^n} > \epsilon_n \right\} \bigcap \mathcal{A}_n \right) \\ &\leq \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right) \frac{\log(1/\delta)}{\delta^2} \frac{\alpha_n^2 \beta_n^* \eta_n}{(1-\beta_n^*)\epsilon^2} + C\delta,\end{aligned}$$

where $\mathcal{A}_n$ is defined in Eq S.30.

Proof. Using the trace trick in Proposition B.6 again, we have that, with probability at least $1 - C\delta$ for some $C > 0$,

$$\frac{\frac{n}{\eta_n} \left\| \hat{V}_\perp \hat{V}_\perp^T \sum_{k>1} T_k^* g \right\|^2}{(v_1^T g)^2 (1 + \frac{\eta_n \lambda_1}{n})^{2n}} \leq \frac{\frac{n}{\eta_n} \log(1/\delta) \left\| \sum_{k>1} T_k \right\|_F^2}{\delta^2 (1 + \frac{\eta_n \lambda_1}{n})^{2n}}$$

The Hoeffding decomposition (Proposition A.4), together with the fact that $\|AB\|_F \leq \|A\|_F \|B\|_{op}$ implies:

$$\begin{aligned}&\mathbb{E}^* \left[\left\| \sum_{k>1} T_k^* \right\|_F^2 \right] = \mathbb{E}^* \left[\sum_{k>1} \|T_k^*\|_F^2 \right] \\ &\leq \sum_{k=2}^{n} \sum_{|S|=k} \left(\frac{\eta_n}{n}\right)^{2k} \prod_{i=1}^{k} \left\| X_{S[i]} X_{S[i]}^T - X_{S[i]-1} X_{S[i]-1}^T \right\|_F^2 \prod_{j=1}^{k+1} \left\| \mathcal{B}_{j,n}^{(S)} \right\|_{op}^2\end{aligned} \tag{S.37}$$

Now, that expectation corresponding to a given summand is given by:

$$\begin{aligned}
&\int_{\mathcal{A}_n} \left\| X_{S[i]} X_{S[i]}^T - X_{S[i]-1} X_{S[i]-1}^T \right\|_F^2 \prod_{j=1}^{k+1} \left\| \mathcal{B}_{j,n}^{(S)} \right\|^2 dP \\
&\le \int_{\mathcal{A}_n} \prod_{i=1}^{k} 4\alpha_n^2 \prod_{j=1}^{k+1} \left\| \mathcal{B}_{j,n}^{(S)} \right\|^2 dP \\
&\le \left(4\alpha_n^2\right)^k \prod_{j=1}^{k+1} \mathbb{E}\left[\left\| \mathcal{B}_{j,n}^{(S)} \right\|^2 \right]
\end{aligned} \tag{S.38}$$

where $\mathcal{B}_{j,n}^{(S)}$ is defined in Eq S.35.

To bound $\mathbb{E}\left[\left\| \mathcal{B}_{j,n}^{(S)} \right\|^2 \right]$, we invoke Lemma B.1 Eq S.21. For some $C > 0$ uniformly in S:

$$\prod_{j=1}^{k+1} \mathbb{E}\left[\left\| \mathcal{B}_{j,n}^{(S)} \right\|^2 \right] \le \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right)^{k+1} \left(1 + \frac{\eta_n \lambda_1}{n}\right)^{2(n-k)}$$

Therefore, by Markov's inequality,

$$\begin{aligned}
&P\left(\left\{ \frac{\sqrt{\frac{n}{\eta_n}} \left\| \hat{V}_\perp \hat{V}_\perp^T \sum_{k>1} T_k^* u_0 \right\|}{(1 + \frac{\eta_n \lambda_1}{n})^n} > \epsilon_n \right\} \bigcap \mathcal{A}_n \right) \\
&\le \frac{n}{\delta^3 \epsilon_n^2 \eta_n} \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right) \sum_{k=2}^{n} \left(\frac{4\eta_n^2 \alpha_n^2}{n} \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right) \right)^k \\
&\le \alpha_n^2 \eta_n \delta_n^{-3} \epsilon_n^{-2} \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right) \sum_{k=1}^{n} \left(\frac{4\eta_n^2 \alpha_n^2}{n} \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right) \right)^k \\
&\le \exp\left(\sqrt{\frac{C M_d^2 \eta_n^2 \log d}{n}} \right) \frac{\alpha_n^2 \beta_n^* \eta_n}{(1 - \beta_n^*) \epsilon_n^2 \delta_n^3}
\end{aligned} \tag{S.39}$$

where the last line follows from a geometric series argument. ☐

Lemma C.9.

$$\sum_{i=0}^{n} \left(1 - \frac{\eta_n / n(\lambda_1 - \lambda_2)}{1 + \eta_n \lambda_1 / n} \right)^{2i} \le \frac{n}{\eta_n} \min\left(\eta_n, \frac{1}{\lambda_1 - \lambda_2} \right)$$

Proof. This follows from the definition of a geometric series. ☐

Lemma C.10 (Bounding the leading Hoeffding projection for the bootstrap on $V_\perp$)**.** *Let $\lambda_1 M_d (\log d)^2 \frac{\eta_n^2}{n} \to 0$, and $nd \exp(-\eta_n(\lambda_1 - \lambda_2)) \to 0$. For any $\epsilon, \delta > 0$, and $C_1, C_2 \ge 0$, we have:*

$$P\left(\left\{ \sqrt{\frac{n}{\eta_n}} \frac{\| \widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp V_\perp^T u_0 \|}{(1 + \eta_n \lambda_1 / n)^n |v_1^T u_0|} \ge \epsilon \right\} \cap \mathcal{A}_n \right) \le \frac{C_1 \alpha_n M_d \eta_n^2 \log(1/\delta)}{n(\lambda_1 - \lambda_2)\delta^3} \frac{1}{\epsilon^2} + C_2 \delta$$

Proof. Using Proposition B.6, with probability at least $1-\delta$,

$$\frac{\|\widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp V_\perp^T u_0\|^2}{(1+\eta_n\lambda_1/n)^{2n}\|v_1^T u_0\|^2} \leq \frac{\log(1/\delta)\left\|\widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp V_\perp^T\right\|_F^2}{\delta^2(1+\eta_n\lambda_1/n)^{2n}}$$
$$= \frac{\log(1/\delta)\mathrm{trace}(\widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp V_\perp^T T_1^* \widehat{V}_\perp \widehat{V}_\perp^T)}{\delta^2(1+\eta_n\lambda_1/n)^{2n}}$$
$$= \frac{\log(1/\delta)\left\|\widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp\right\|_F^2}{\delta^2(1+\eta_n\lambda_1/n)^{2n}} \tag{S.40}$$

First note that,

$$\|V_\perp V_\perp^T - \widehat{V}_\perp \widehat{V}_\perp^T\|_F^2 = \|v_1 v_1^T - \hat{v}_1\hat{v}_1^T\|_F^2 = 2(1-(v_1^T\hat{v}_1)^2)$$

Thus, we have

$$\mathbb{E}^*\|\widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp\|_F^2$$
$$= \frac{\eta_n^2}{n^2}\sum_i \|\widehat{V}_\perp \widehat{V}_\perp^T \mathcal{B}_{1,i-1}(X_iX_i^T - X_{i-1}X_{i-1}^T)\mathcal{B}_{i+1,n}V_\perp\|_F^2$$
$$\leq 4\frac{\eta_n^2}{n^2}\sum_i\sum_{j=1}^6 \|r_{j,i}\|_F^2, \tag{S.41}$$

where $B_{1,i}$ are defined in Eq S.32, and the residual vectors $r_{k,i}$ are defined as follows. Recall the definition of $R_{1,i}$ and $R_{i,n}$ from Eq S.32. Now define the following vectors which contribute to the remainder.

$$\begin{aligned}
r_{1,i} &= \widehat{V}_\perp \widehat{V}_\perp^T R_{1,i-1}(Y_i - Y_{i-1})R_{i+1,n}V_\perp \\
r_{2,i} &= \widehat{V}_\perp \widehat{V}_\perp^T R_{1,i-1}(Y_i - Y_{i-1})(I+\eta_n/n\Sigma)^{n-i}V_\perp \\
r_{3,i} &= V_\perp V_\perp^T (I+\eta_n/n\Sigma)^{n-i}(Y_i - Y_{i-1})R_{i+1,n}V_\perp \\
r_{4,i} &= V_\perp V_\perp^T (I+\eta_n/n\Sigma)^{n-i}(Y_i - Y_{i-1})(I+\eta_n/n\Sigma)^{n-i}V_\perp \\
r_{5,i} &= (\widehat{V}_\perp \widehat{V}_\perp^T - V_\perp V_\perp^T)(I+\eta_n/n\Sigma)^{n-i}(Y_i - Y_{i-1})R_{i+1,n}V_\perp \\
r_{6,i} &= (\widehat{V}_\perp \widehat{V}_\perp^T - V_\perp V_\perp^T)(I+\eta_n/n\Sigma)^{n-i}(Y_i - Y_{i-1})(I+\eta_n/n\Sigma)^{n-i}V_\perp
\end{aligned}$$

First we will bound $\|r_{1,i}\|_F^2$. Recall the set $\mathcal{A}_n$ where the maximum norm is bounded from S.30.

$$E_{1,i} := \int_{\mathcal{A}_n} \|r_{1,i}\|_F^2 dP \leq 2\alpha_n \int_{\mathcal{A}_n} \|R_{1,i-1}\|^2 \|R_{i+1,n}\|^2 dP$$
$$\leq 2\alpha_n \int \|R_{1,i}\|^2\|R_{i+1,n}\|^2 dP \leq 2\alpha_n \mathbb{E}\|R_{1,i}\|^2\mathbb{E}\|R_{i+1,n}\|^2 \tag{S.42}$$

Similarly,

$$E_{2,i} = \int_{\mathcal{A}_n} \|r_{2,i}\|_F^2 dP \leq 2\alpha_n \left(1+\eta_n\lambda_2/n\right)^{2(n-i)}\mathbb{E}\|R_{1,i-1}\|^2 \tag{S.43}$$

$$E_{3,i} = \int_{\mathcal{A}_n} \|r_{3,i}\|_F^2 dP \leq 2\alpha_n \left(1+\eta_n\lambda_2/n\right)^{2(i-1)}\mathbb{E}\|R_{i+1,n}\|^2 \tag{S.44}$$

Similarly,

$$E_{4,i} = \int_{\mathcal{A}_n} \|r_{4,i}\|_F^2 dP \leq 2\alpha_n \left(1+\eta_n\lambda_2/n\right)^{2(n-1)} \tag{S.45}$$

Recall the set $\mathcal{A}_{\sin}$ from Eq S.27. With probability at least $1-\delta_n$,

$$
\begin{aligned}
E_{5,i} &= \int_{\mathcal{A}_n \cap \mathcal{A}_{\sin}} \|r_{5,i}\|_F^2 dP \leq 4\alpha_n \frac{\epsilon_{\sin}}{\delta_n} \left(1+\eta_n\lambda_1/n\right)^{2(i-1)} \mathbb{E}\|R_{i+1,n}\|^2 \\
E_{6,i} &= \int_{\mathcal{A}_n \cap \mathcal{A}_{\sin}} \|r_{6,i}\|_F^2 dP \leq 2\alpha_n \frac{\epsilon_{\sin}}{\delta_n} \left(1+\eta_n\lambda_1/n\right)^{2(i-1)} \left(1+\eta_n\lambda_2/n\right)^{2(n-i)}
\end{aligned}
$$

Observe that, using Eq S.32, we have,

$$
\begin{aligned}
\mathcal{E}_1 &:= \sum_i E_{1,i} \leq \frac{2\alpha_n e^2 M_d^2 (1+2\log d)^2 \eta_n^4}{n} (1+\eta_n\lambda_1/n)^{2(n-1)} \\
\mathcal{E}_2 &:= \sum_i (E_{2,i}+E_{3,i}) \leq \frac{4\alpha_n e M_d (1+2\log d)\eta_n^3}{n} \min\left(\eta_n, \frac{1}{\lambda_1-\lambda_2}\right) (1+\eta_n\lambda_1/n)^{2n-1} \\
\mathcal{E}_3 &:= \sum_i E_{4,i} \leq 2\alpha_n n \left(1+\eta_n\lambda_2/n\right)^{2n}
\end{aligned}
$$

With probability at least $1-\delta_n$, we have

$$
\begin{aligned}
\mathcal{E}_4 &:= \sum_i E_{5,i} \leq 4\alpha_n \frac{\epsilon_{\sin}}{\delta_n} e M_d (1+2\log d)\eta_n^2 (1+\eta_n\lambda_1)^{2(n-1)} \\
\mathcal{E}_5 &:= \sum_i E_{6,i} \leq 2\alpha_n \frac{\epsilon_{\sin}}{\delta_n} \frac{n}{\eta_n} \min\left(\eta_n, \frac{1}{\lambda_1-\lambda_2}\right) (1+\eta_n\lambda_1)^{2(n-1)}
\end{aligned}
$$

If $\lambda_1 M_d (\log d)^2 \frac{\eta_n^2}{n} \to 0$, then $\mathcal{E}_4 \leq C_1\mathcal{E}_5$ for some positive constant C_1. If $nd\exp(-2\eta_n(\lambda_1-\lambda_2)) \to 0$, then $\mathcal{E}_3 \leq C_2\mathcal{E}_5$.
Thus, under these conditions,

$$
\mathcal{E}_1, \mathcal{E}_2 \leq C_4\mathcal{E}_5
$$

With probability at least $1-\delta_n$, for some positive constant C',

$$
\frac{\sum_{i=1}^5 \mathcal{E}_i}{(1+\eta_n\lambda_1/n)^{2n}} \leq C'\alpha_n \frac{\epsilon_{\sin}}{\delta_n}
$$

Finally, using Eq S.41 we get:

$$
\frac{\int_{\mathcal{A}_{\sin}\cap\mathcal{A}_n} \mathbb{E}^*\|\widehat{V}_\perp \widehat{V}_\perp^T T_1^* V_\perp\|_F^2 dP}{(1+\eta_n\lambda_1/n)^{2n}} \leq C''\alpha_n \frac{\eta_n^2}{n}\frac{\epsilon_{\sin}}{\delta_n} \tag{S.46}
$$

Let $\mathcal{A}_1$ denote the set where Eq S.40 holds.

$$
\begin{aligned}
&P\left(\left\{\frac{n}{\eta_n}\frac{\|\widehat{V}_\perp\widehat{V}_\perp^T T_1^* V_\perp V_\perp^T u_0\|^2}{(1+\eta_n\lambda_1/n)^{2n}(v_1^T u_0)^2}\geq\epsilon\right\}\cap\mathcal{A}_n\right)\\
&\leq P\left(\left\{\frac{\left\|\widehat{V}_\perp\widehat{V}_\perp^T T_1^* V_\perp\right\|_F^2}{(1+\eta_n\lambda_1/n)^{2n}}\geq\frac{\epsilon\delta^2}{\log(1/\delta)}\frac{\eta_n}{n}\right\}\cap\mathcal{A}_n\cap\mathcal{A}_1\right)+2\delta\\
&\leq P\left(\left\{\frac{\left\|\widehat{V}_\perp\widehat{V}_\perp^T T_1^* V_\perp\right\|_F^2}{(1+\eta_n\lambda_1/n)^{2n}}\geq\frac{\epsilon\delta^2}{\log(1/\delta)}\frac{\eta_n}{n}\right\}\cap\mathcal{A}_n\cap\mathcal{A}_1\cap\mathcal{A}_{\sin}\right)+2\delta+\delta_n\\
&\leq\mathbb{E}\left[\frac{\mathbb{E}^*\left\|\widehat{V}_\perp\widehat{V}_\perp^T T_1^* V_\perp\right\|_F^2}{(1+\eta_n\lambda_1/n)^{2n}}\times\frac{\log(1/\delta)n}{\epsilon\delta^2\eta_n}1(\mathcal{A}_n\cap\mathcal{A}_1\cap\mathcal{A}_{\sin})\right]+2\delta+\delta_n\\
&\overset{(i)}{\leq}\frac{C''\alpha_n\eta_n\log(1/\delta)}{\delta_n\delta^2}\frac{\epsilon_{\sin}}{\epsilon}+2\delta+\delta_n\\
&\overset{(ii)}{\leq}\frac{C'''\alpha_n M_d\eta_n^2\log(1/\delta)}{n(\lambda_1-\lambda_2)\delta_n\delta^2}\frac{1}{\epsilon}+2\delta+\delta_n
\end{aligned}
$$

Step (i) follows from Eq S.46. Step (ii) follows from the definition of $\epsilon_{\sin}$ in Eq S.29. Now setting $\epsilon_{\sin}=\delta$, we get the result. □

D Proof of Proposition 1

Proof of Proposition 1. Since $\|X_{1j}\|_{\psi_2}\leq\nu_j$ it follows that $\left\|X_{1j}^2\right\|_{\psi_1}\leq\nu_j^2$. Observe that $(X_{1j}^2-\mathbb{E}X_{1j}^2)/\nu_j^2$ is sub-Exponential with parameter at most 1 since $\left\|(X_{1j}^2-\mathbb{E}[X_{1j}^2])/\nu_j^2\right\|_{\psi_1}\leq\left\|X_{1j}^2\right\|_{\psi_1}/\nu_j^2=1$. By multivariate Holder inequality with $p_j=\sum_{j=1}^d\nu_j^2/\nu_j^2$ and property (e) of Proposition 2.7.1 of [7], for $|\lambda|<1/(\sum_{i=1}^d\nu_i^2)$:

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\sum_{j=1}^d(X_{1j}^2-\mathbb{E}[X_{1j}^2])\right)\right]&\leq\prod_{j=1}^d\mathbb{E}\left[\exp\left(\lambda(X_{1j}^2-\mathbb{E}[X_{1j}^2])\right)^{\frac{\sum_{i=1}^d\nu_i^2}{\nu_j^2}}\right]^{\frac{\nu_j^2}{\sum_{i=1}^d\nu_i^2}}\\
&=\prod_{j=1}^d\mathbb{E}\left[\exp\left(\frac{\lambda(\sum_{i=1}^d\nu_i^2)(X_{1j}^2-\mathbb{E}[X_{1j}^2])}{\nu_j^2}\right)\right]^{\frac{\nu_j^2}{\sum_{i=1}^d\nu_i^2}}\\
&\leq\prod_{j=1}^d\exp\left(\frac{K\lambda^2(\sum_{i=1}^d\nu_i^2)^2\nu_j^2}{\sum_{i=1}^d\nu_i^2}\right)\\
&=\exp\left\{K\lambda^2\left(\sum_{i=1}^d\nu_i^2\right)^2\right\}
\end{aligned}
$$

Therefore, $\left\|\sum_{i=1}^d X_{1i}^2\right\|_{\psi_1}\leq\sum_{i=1}^d\nu_i^2$. Since a subexponential random variable T satisfy the tail condition:

$$P(T-\mathbb{E}[T]>t)\leq\exp(-t/K\nu)$$

for another universal constant $K>0$, the second claim follows by a union bound and noting that $\mathbb{E}[\|X_1\|_2^2]\leq\sum_{i=1}^d\nu_i^2<C_2$ since absolute summability implies square summability. □

References

[1] V. Bentkus, F. Götze, and W. R. van Zwet. An Edgeworth expansion for symmetric statistics. *The Annals of Statistics*, 25(2):851 – 896, 1997.

[2] F. Götze, A. Naumov, V. Spokoiny, and V. Ulyanov. Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli*, 25(4A):2538 – 2563, 2019.

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

[4] P. Jain, C. Jin, S. Kakade, P. Netrapalli, and A. 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.

[5] E. M. Stein and R. Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press, Princeton,NJ, 2009.

[6] A. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[7] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, Cambridge, UK, 2018.

[8] M. Xu, D. Zhang, and W. B. Wu. Pearson's chi-squared statistics: approximation theory and beyond. *Biometrika*, 106(3):716–723, 04 2019.