# Nearly Optimal Stochastic Approximation for Online Principal Subspace Estimation

Xin Liang* Zhen-Chen Guo† Ren-Cang Li‡ Wen-Wei Lin§

August 14, 2018

**Abstract**

Processing streaming data as they arrive is often necessary for high dimensional data analysis. In this paper, we analyse the convergence of a subspace online PCA iteration, as a followup of the recent work of Li, Wang, Liu, and Zhang [*Math. Program., Ser. B*, DOI 10.1007/s10107-017-1182-z] who considered the case for the most significant principal component only, i.e., a single vector. Under the sub-Gaussian assumption, we obtain a finite-sample error bound that closely matches the minimax information lower bound of Vu and Lei [*Ann. Statist.* **41**:6 (2013), 2905–2947].

**Key words.** Principal component analysis, Principal component subspace, Stochastic approximation, High-dimensional data, Online algorithm, Finite-sample analysis

**AMS subject classifications**. 62H25, 68W27, 60H10, 60H15

# Contents

**1 Introduction** **2**

**2 Canonical Angles** **4**

**3 Online PCA for Principal Subspace** **5**

**4 Main Results** **7**

**5 Comparisons with Previous Results** **10**

**6 Proof of Theorem 4.1** **12**

6.1 Simplification . . . . . 12
6.2 Increments of One Iteration . . . . . 14
6.3 Quasi-Power Iteration Process . . . . . 20
6.4 Proof of Theorem 4.1 . . . . . 28

**7 Proofs of Theorems 4.2 and 4.3** **30**

**8 Conclusion** **36**

*Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C. E-mail: liangxinslm@pku.edu.cn. Supported in part by the MOST.

†Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C. E-mail: guozhch06@gmail.com. Supported in part by the MOST.

‡Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019, USA. E-mail: rcli@uta.edu. Supported in part by NSF grants CCF-1527104, DMS-1719620, and NSFC grant 11428104.

§Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C. E-mail: wwlin@am.nctu.edu.tw. Supported in part by the MOST, the NCTS, and the ST Yau Centre at the National Chiao Tung University.

# 1 Introduction

Principal component analysis (PCA) introduced in [8, 17] is one of the most well-known and popular methods for dimensional reduction in high-dimensional data analysis.

Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector with mean $\mathrm{E}\{\boldsymbol{X}\}$ and covariance matrix

$$\Sigma = \mathrm{E}\big\{(\boldsymbol{X} - \mathrm{E}\{\boldsymbol{X}\})(\boldsymbol{X} - \mathrm{E}\{\boldsymbol{X}\})^{\mathrm{T}}\big\}.$$

To reduce the dimension of $\boldsymbol{X}$ from $d$ to $p$, PCA looks for a $p$-dimensional linear subspace that is closest to the centered random vector $\boldsymbol{X} - \mathrm{E}\{\boldsymbol{X}\}$ in a mean squared sense, through the independent and identically distributed samples $X^{(1)}, \ldots, X^{(n)}$.

Denote by $\mathbb{G}_p(\mathbb{R}^d)$ the Grassmann manifold of $p$-planes in $\mathbb{R}^d$, or equivalently, the set of all $p$-dimensional subspaces of $\mathbb{R}^d$. Without loss of generality, we assume $\mathrm{E}\{\boldsymbol{X}\} = 0$. Then PCA corresponds to a stochastic optimization problem

$$\min_{\mathcal{U} \in \mathbb{G}_p(\mathbb{R}^d)} \mathrm{E}\big\{\|(I_d - \Pi_{\mathcal{U}})\boldsymbol{X}\|_2^2\big\}, \tag{1.1}$$

where $I_d$ is the $d \times d$ identity matrix, and $\Pi_{\mathcal{U}}$ is the orthogonal projector onto the subspace $\mathcal{U}$. Let $\Sigma = U \Lambda U^{\mathrm{T}}$ be the spectral decomposition of $\Sigma$, where

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \text{ with } \lambda_1 \geq \cdots \geq \lambda_d \geq 0, \text{ and orthogonal } U = [u_1, \ldots, u_d]. \tag{1.2}$$

If $\lambda_p > \lambda_{p+1}$, then the unique solution to the optimization problem (1.1), namely the $p$-dimensional principal subspace of $\Sigma$, is $\mathcal{U}_* = \mathcal{R}([u_1, \ldots, u_p])$, the subspace spanned by $u_1, \ldots, u_p$.

In practice, $\Sigma$ is unknown, and we must use sample data to estimate $\mathcal{U}_*$. The classical PCA does it by the spectral decomposition of the empirical covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X^{(i)}(X^{(i)})^{\mathrm{T}}$. Specifically, the classical PCA uses $\widehat{\mathcal{U}}_* = \mathcal{R}([\widehat{u}_1, \ldots, \widehat{u}_p])$ to estimate $\mathcal{U}_*$, where $\widehat{u}_i$ is the corresponding eigenvectors of $\widehat{\Sigma}$. The important quantity is the distance between $\mathcal{U}_*$ and $\widehat{\mathcal{U}}_*$. Vu and Lei [23, Theorem 3.1] proved that if $p(d-p)\frac{\sigma_*^2}{n}$ is bounded, then

$$\inf_{\widetilde{\mathcal{U}}_* \in \mathbb{G}_p(\mathbb{R}^d)} \sup_{\boldsymbol{X} \in \mathcal{P}_0(\sigma_*^2, d)} \mathrm{E}\Big\{\|\sin\Theta(\widehat{\mathcal{U}}_*, \mathcal{U}_*)\|_{\mathrm{F}}^2\Big\} \geq cp(d-p)\frac{\sigma_*^2}{n}, \tag{1.3}$$

where $c > 0$ is an absolute constant, and $\mathcal{P}_0(\sigma_*^2, d)$ is the set of all $d$-dimensional sub-Gaussian distributions for which the eigenvalues of the covariance matrix satisfy

$$\frac{\lambda_1 \lambda_{p+1}}{(\lambda_p - \lambda_{p+1})^2} \leq \sigma_*^2. \tag{1.4}$$

Note that its left-hand side is the effective noise variance. In the classical PCA, obtaining the empirical covariance matrix has time complexity $\mathrm{O}(nd^2)$ and space complexity $\mathrm{O}(d^2)$. So storing and calculating a large empirical covariance matrix are very expensive when the data are of high dimension, not to mention the cost of computing its eigenvalues and eigenvectors.

To reduce both the time and space complexities, Oja [15] proposed an online PCA iteration

$$\widetilde{u}^{(n)} = u^{(n-1)} + \beta^{(n-1)} X^{(n)} (X^{(n)})^{\mathrm{T}} u^{(n-1)}, \qquad u^{(n)} = \widetilde{u}^{(n)} \|\widetilde{u}^{(n)}\|_2^{-1}, \tag{1.5}$$

to approximate the most significant principal component, where $\beta^{(n)} > 0$ is a stepsize. Later Oja and Karhunen [16] proposed a subspace online PCA iteration

$$\widetilde{U}^{(n)} = U^{(n-1)} + X^{(n)}(X^{(n)})^{\mathrm{T}} U^{(n-1)} \,\mathrm{diag}(\beta_1^{(n-1)}, \ldots, \beta_p^{(n-1)}), \qquad U^{(n)} = \widetilde{U}^{(n)} R^{(n)}, \tag{1.6}$$

to approximate the principal subspace $\mathcal{U}_*$, where $\beta_i^{(n)} > 0$ for $1 \leq i \leq p$ are stepsizes, and $R^{(n)}$ is a normalization matrix to make $U^{(n)}$ have orthonormal columns. One such an $R^{(n)}$ is

$$R^{(n)} = [(\widetilde{U}^{(n)})^{\mathrm{T}} \widetilde{U}^{(n)}]^{-1/2}. \tag{1.7}$$

Usually, we can use a fixed stepsize $\beta$. It can be seen that these methods update the approximations incrementally by processing data one at a time as soon as it comes in, and calculating the empirical covariance matrix explicitly is completely avoided. In the online PCA, obtaining the principal subspace has time complexity $\mathrm{O}(p^2nd)$ and space complexity $\mathrm{O}(pd)$, which is much less than those required by the classical PCA.

Although the online PCA iteration (1.5) was proposed over 30 years ago, its convergence analysis is rather scarce. Some recent works [2, 9, 18] studied the convergence of the online PCA for the most significant principal component, i.e., $u_1$, from different points of view and obtained some results for the case where the samples are almost surely uniformly bounded. For such a case, De Sa, Olukotun, and Ré [4] studied a different but closely related problem, in which the angular part is equivalent to the online PCA, and obtained some convergence results. In contrast, for the distributions with sub-Gaussian tails (note that the samples of this kind of distributions may be unbounded), Li, Wang, Liu, and Zhang [11] proved a nearly optimal convergence rate for the iteration (1.5): if the initial guess $u^{(0)}$ is randomly chosen according to a uniform distribution and the stepsize $\beta$ is chosen in accordance with the sample size $n$, then there exists a high-probability event $\mathbb{A}_*$ with $\mathrm{P}\{\mathbb{A}_*\} \geq 1-\delta$ such that

$$\mathrm{E}\Big\{|\tan\Theta(u^{(n)}, u_*)|^2 \;\Big|\; \mathbb{A}_*\Big\} \leq C(d,n,\delta)\frac{\ln n}{n}\frac{1}{\lambda_1-\lambda_2}\sum_{i=2}^{d}\frac{\lambda_1\lambda_i}{\lambda_1-\lambda_i} \tag{1.8a}$$

$$\leq C(d,n,\delta)\frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_2)^2}\frac{(d-1)\ln n}{n}, \tag{1.8b}$$

where $\delta \in [0,1)$, $u_* = u_1$ in (1.2), and $C(d,n,\delta)$ can be approximately treated as a constant. It can be seen that this bound matches the minimax low bound (1.3) up to a logarithmic factor of $n$, hence, *nearly optimal*. Also, the convergence rate holds true as long as the initial approximation satisfies

$$|\tan\Theta(u^{(0)}, u_*)| \leq cd, \tag{1.9}$$

for some constant $c > 0$, which means *nearly global*. It is significant because a uniformly distributed initial value is nearly orthogonal to the principal component with high probability when $d$ is large. This result is more general than previous ones in [2, 9, 18], because it is for distributions that can possibly be unbounded, and the convergence rate is nearly optimal and nearly global. For more details of comparison, the reader is referred to [11].

However, there is still no convergence result for the subspace online PCA, namely the subspace iteration (1.6). Garber et al. [6] use shift-and-invert technique to deal with the convergence of a faster variant of subspace PCA, but only the result on the top eigenvector is analyzed. Our main purpose in this paper is to analyze the convergence of the subspace online PCA iteration (1.6), similarly to the effort in [11] which is for the special case $p = 1$. One of our results for the convergence rate states that: if the initial guess $U^{(0)}$ is randomly chosen to satisfy that $\mathcal{R}(U^{(0)})$ is uniformly sampled from $\mathbb{G}_p(\mathbb{R}^d)$, and the stepsize $\beta_i^{(n)}$ is chosen the same for $1 \leq i \leq p$ and in accordance with the sample size $n$, then there exists a high-probability event $\mathbb{H}_*$ with $\mathrm{P}\{\mathbb{H}_*\} \geq 1-2\delta^{p^2}$, such that

$$\mathrm{E}\Big\{\|\tan\Theta(U^{(n)}, U_*)\|_{\mathrm{F}}^2 \;\Big|\; \mathbb{H}_*\Big\} \leq C(d,n,\delta)\frac{\ln n}{n}\frac{1}{\lambda_p-\lambda_{p+1}}\sum_{j=1}^{p}\sum_{i=p+1}^{d}\frac{\lambda_j\lambda_i}{\lambda_j-\lambda_i} \tag{1.10a}$$

$$\leq C(d,n,\delta)\frac{\lambda_p\lambda_{p+1}}{(\lambda_p-\lambda_{p+1})^2}\frac{p(d-p)\ln n}{n}, \tag{1.10b}$$

where the constant $C(d,n,\delta) \to 24\psi^4/(1-\delta^{p^2})$ as $d \to \infty$ and $n \to \infty$, and $\psi$ is $\boldsymbol{X}$'s Orlicz norm. This is also *nearly optimal*, *nearly global*, and valid for any sub-Gaussian distribution. When $p = 1$, it degenerates to (1.8), as it should be. Although this result of ours look like a straightforward generalization, its proof, however, turns out to be nontrivially much more complicated. Also note that the factor in our result is

$$\frac{\lambda_p\lambda_{p+1}}{(\lambda_p-\lambda_{p+1})^2} \quad vs. \quad \frac{\lambda_1\lambda_{p+1}}{(\lambda_p-\lambda_{p+1})^2}.$$

The second quantity appeared in (1.8b). The first quantity is always smaller but both are of similar order if $\lambda_1$ and $\lambda_p$ are of similar order. However, their magnitude can differ greatly when $\lambda_p \ll \lambda_1$.

The rest of this paper is organized as follows. Section 2 reviews the basics about the canonical angles and the canonical angle matrix between two $k$-dimensional subspaces, the metrics on $\mathbb{G}_p(\mathbb{R}^d)$, and proves a lemma on the tangent of the canonical angle matrix, which will be used in later proofs. In section 3, we reformulate the subspace online PCA iteration (1.6) for the case $\beta_i^{(n)} = \beta$ for all $1 \le i \le p$, which will be the version to be analyzed. Our main results, one of which leads to (1.10), are stated in section 4. We compare our results for $p = 1$ with the recent results in [11] and outline the technical differences in proof between ours and those from [11] in section 5. Due to their complexities, the proofs of these results are deferred to sections 6 and 7. Finally, section 8 summarizes the results of the paper.

We point out in passing that this paper improves its earlier version available online [12] in that an assumption, namely Assumption 4.2 in [12] is no longer necessary and thus removed, which is made possible by the new quasi-bounded event (6.5).

*Notations.* $\mathbb{R}^{n\times m}$ is the set of all $n \times m$ real matrices, $\mathbb{R}^n = \mathbb{R}^{n\times 1}$, and $\mathbb{R} = \mathbb{R}^1$. $I_n$ (or simply $I$ if its dimension is clear from the context) is the $n \times n$ identity matrix and $e_j$ is its $j$th column (usually with dimension determined by the context). For a matrix $X$, $\sigma(X)$, $\|X\|_\infty$, $\|X\|_2$ and $\|X\|_{\mathrm{F}}$ are the multiset of the singular values, the $\ell_\infty$-operator norm, the spectral norm, and the Frobenius norm of $X$, respectively. $\mathcal{R}(X)$ is the column space spanned by the columns of $X$, $X_{(i,j)}$ is the $(i,j)$th entry of $X$, and $X_{(k:\ell,:)}$ and $X_{(:,i:j)}$ are two submatrices of $X$ consisting of its row $k$ to row $\ell$ and column $i$ to column $j$, respectively. $X \circ Y$ is the Hadamard, i.e., entrywise, product of matrices (vector) $X$ and $Y$ of the same size.

For any vector or matrix $X$, $Y$, $X \le Y$ ($X < Y$) means $X_{(i,j)} \le Y_{(i,j)}$ ($X_{(i,j)} < Y_{(i,j)}$) for any $i,j$. $X \ge Y$ ($X > Y$) if $-X \le -Y$ ($-X < -Y$); $X \le \alpha$ ($X < \alpha$) for a scalar $\alpha$ means $X_{(i,j)} \le \alpha$ ($X_{(i,j)} < \alpha$) for any $i,j$; similarly $X \ge \alpha$ and $X > \alpha$.

For a subset or an event $\mathbb{A}$, $\mathbb{A}^{\mathrm{c}}$ is the complement set of $\mathbb{A}$. By $\sigma\{\mathbb{A}_1, \dots, \mathbb{A}_p\}$ we denote the $\sigma$-algebra generated by the events $\mathbb{A}_1, \dots, \mathbb{A}_p$. $\mathbb{N} = \{1, 2, 3, \dots\}$. $\mathrm{E}\{\boldsymbol{X};\, \mathbb{A}\} := \mathrm{E}\{\boldsymbol{X}\mathbf{1}_{\mathbb{A}}\}$ denotes the expectation of a random variable $\boldsymbol{X}$ over event $\mathbb{A}$. Note that

$$\mathrm{E}\{\boldsymbol{X};\, \mathbb{A}\} = \mathrm{E}\{\boldsymbol{X} \mid \mathbb{A}\}\,\mathrm{P}\{\mathbb{A}\}. \tag{1.11}$$

For a random vector or matrix $\boldsymbol{X}$, $\mathrm{E}\{\boldsymbol{X}\} := \big[\mathrm{E}\{\boldsymbol{X}_{(i,j)}\}\big]$. Note that $\|\mathrm{E}\{\boldsymbol{X}\}\|_{\mathrm{ui}} \le \mathrm{E}\{\|\boldsymbol{X}\|_{\mathrm{ui}}\}$ for $\mathrm{ui} = 2, \mathrm{F}$. Write $\mathrm{cov}_\circ(\boldsymbol{X}, \boldsymbol{Y}) := \mathrm{E}\{[\boldsymbol{X} - \mathrm{E}\{\boldsymbol{X}\}] \circ [\boldsymbol{Y} - \mathrm{E}\{\boldsymbol{Y}\}]\}$ and $\mathrm{var}_\circ(\boldsymbol{X}) := \mathrm{cov}_\circ(\boldsymbol{X}, \boldsymbol{X})$.

# 2 Canonical Angles

For two subspaces $\mathcal{X}$, $\mathcal{Y} \in \mathbb{G}_p(\mathbb{R}^d)$, let $X$, $Y \in \mathbb{C}^{d\times p}$ be the basis matrices of $\mathcal{X}$ and $\mathcal{Y}$, respectively, i.e., $\mathcal{X} = \mathcal{R}(X)$ and $\mathcal{Y} = \mathcal{R}(Y)$, and denote by $\sigma_j$ for $1 \le j \le k$ in nondecreasing order, i.e., $\sigma_1 \le \cdots \le \sigma_p$, the singular values of

$$(X^{\mathrm{T}}X)^{-1/2}X^{\mathrm{T}}Y(Y^{\mathrm{T}}Y)^{-1/2}.$$

The $k$ *canonical angles* $\theta_j(\mathcal{X}, \mathcal{Y})$ *between* $\mathcal{X}$ *to* $\mathcal{Y}$ are defined by

$$0 \le \theta_j(\mathcal{X}, \mathcal{Y}) := \arccos \sigma_j \le \frac{\pi}{2} \quad \text{for } 1 \le j \le p. \tag{2.1}$$

They are in non-increasing order, i.e., $\theta_1(\mathcal{X}, \mathcal{Y}) \ge \cdots \ge \theta_p(\mathcal{X}, \mathcal{Y})$. Set

$$\Theta(\mathcal{X}, \mathcal{Y}) = \mathrm{diag}(\theta_1(\mathcal{X}, \mathcal{Y}), \dots, \theta_p(\mathcal{X}, \mathcal{Y})). \tag{2.2}$$

It can be seen that angles so defined are independent of the basis matrices $X$ and $Y$, which are not unique. With the definition of canonical angles,

$$\|\sin \Theta(\mathcal{X}, \mathcal{Y})\|_{\mathrm{ui}} \quad \text{for } \mathrm{ui} = 2, \mathrm{F}$$

are metrics on $\mathbb{G}_p(\mathbb{R}^d)$ [19, Section II.4].

In what follows, we sometimes place a vector or matrix in one or both arguments of $\theta_j(\,\cdot\,,\,\cdot\,)$ and $\Theta(\,\cdot\,,\,\cdot\,)$ with the understanding that it is about the subspace spanned by the vector or the columns of the matrix argument.

For any $X \in \mathbb{R}^{d\times p}$, if $X_{(1:p,:)}$ is nonsingular, then we can define

$$\mathscr{T}(X) := X_{(p+1:d,:)}X_{(1:p,:)}^{-1}. \tag{2.3}$$

**Lemma 2.1.** *For $X \in \mathbb{R}^{d\times p}$ with nonsingular $X_{(1:p,:)}$, we have for* $\mathrm{ui} = 2, \mathrm{F}$

$$\left\|\tan\Theta(X, \begin{bmatrix} I_p \\ 0 \end{bmatrix})\right\|_{\mathrm{ui}} = \|\mathscr{T}(X)\|_{\mathrm{ui}}. \tag{2.4}$$

*Proof.* Let $Y = \begin{bmatrix} I_p \\ 0 \end{bmatrix} \in \mathbb{R}^{d\times p}$. By definition, $\sigma_j = \cos\theta_j(X,Y)$ for $1 \le j \le p$ are the singular values of

$$\left[I + \mathscr{T}(X)^{\mathrm{T}}\mathscr{T}(X)\right]^{-1/2}\begin{bmatrix} I \\ \mathscr{T}(X) \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} I \\ 0 \end{bmatrix} = \left[I + \mathscr{T}(X)^{\mathrm{T}}\mathscr{T}(X)\right]^{-1/2}.$$

So if $\tau_1 \ge \tau_2 \ge \cdots \ge \tau_p$ are the singular values of $\mathscr{T}(X)$, then

$$\sigma_j = (1+\tau_j^2)^{-1/2} \quad \Rightarrow \quad \tau_j = \frac{\sqrt{1-\sigma_j^2}}{\sigma_j} = \tan\theta_j(X,Y),$$

and hence the identity (2.4). □

# 3 Online PCA for Principal Subspace

Let $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_d]^{\mathrm{T}}$ be a random vector in $\mathbb{R}^d$. Assume $\mathrm{E}\{\boldsymbol{X}\} = 0$. Its covariance matrix $\Sigma := \mathrm{E}\left\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}$ has the spectral decomposition

$$\Sigma = U\Lambda U^{\mathrm{T}} \quad \text{with} \quad U = [u_1, u_2, \dots, u_d],\ \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d), \tag{3.1}$$

where $U \in \mathbb{R}^{d\times d}$ is orthogonal, and $\lambda_i$ for $1 \le i \le d$ are the eigenvalues of $\Sigma$, arranged for convenience in non-increasing order. Assume

$$\lambda_1 \ge \cdots \ge \lambda_p > \lambda_{p+1} \ge \cdots \ge \lambda_d > 0. \tag{3.2}$$

In section 1, we mention the subspace online PCA iteration (1.6) of Oja and Karhunen [16] for computing the principal subspace of dimension $p$

$$\mathcal{U}_* = \mathcal{R}(U_{(:,1:p)}) = \mathcal{R}([u_1, u_2, \dots, u_p]). \tag{3.3}$$

In this paper, we will use a fixed stepsize $\beta$ for all $\beta_i^{(n)}$ there. Then $U^{(n)}$ can be stated in a more explicit manner with the help of the following lemma.

**Lemma 3.1.** *Let $V \in \mathbb{R}^{d\times p}$ with $V^{\mathrm{T}}V = I_p$, $y \in \mathbb{R}^d$ with $y^{\mathrm{T}}y = 1$, and $0 < \beta \in \mathbb{R}$, and let*

$$W := V + \beta yy^{\mathrm{T}}V = (I_d + \beta yy^{\mathrm{T}})V, \quad V_+ := W(W^{\mathrm{T}}W)^{-1/2}.$$

*If $V^{\mathrm{T}}y \ne 0$, then*

$$V_+ = V + \beta yy^{\mathrm{T}}V - [1-(1+\alpha)^{-1/2}](V + \beta yy^{\mathrm{T}}V)zz^{\mathrm{T}},$$

*where $\gamma = \|V^{\mathrm{T}}y\|_2$, $z = V^{\mathrm{T}}y/\gamma$, and $\alpha = \beta(2+\beta)\gamma^2$. In particular, $V_+^{\mathrm{T}}V_+ = I_p$.*

**Algorithm 3.1** Subspace Online PCA

1: Choose $U^{(0)} \in \mathbb{R}^{d\times p}$ with $(U^{(0)})^{\mathrm{T}} U^{(0)} = I$, and choose stepsize $\beta > 0$.
2: **for** $n = 1, 2, \ldots$ until convergence **do**
3: Take an $\boldsymbol{X}$'s sample $X^{(n)}$;
4: $Z^{(n)} = (U^{(n-1)})^{\mathrm{T}} X^{(n)}$, $\alpha^{(n)} = \beta\big[2 + \beta(X^{(n)})^{\mathrm{T}} X^{(n)}\big](Z^{(n)})^{\mathrm{T}} Z^{(n)}$, $\widetilde{\alpha}^{(n)} = (1 + \alpha^{(n)})^{-1/2}$;
5: $U^{(n)} = U^{(n-1)} + \beta\widetilde{\alpha}^{(n)} X^{(n)} (Z^{(n)})^{\mathrm{T}} - \frac{1-\widetilde{\alpha}^{(n)}}{(Z^{(n)})^{\mathrm{T}} Z^{(n)}} U^{(n-1)} Z^{(n)} (Z^{(n)})^{\mathrm{T}}$.
6: **end for**

*Proof.* Let $z = V^{\mathrm{T}} y \in \mathbb{R}^p$. We have

$$W^{\mathrm{T}} W = V^{\mathrm{T}} [I_d + \beta y y^{\mathrm{T}}]^2 V = V^{\mathrm{T}} [I_d + \beta(2+\beta) y y^{\mathrm{T}}]^2 V = I_p + \alpha z z^{\mathrm{T}}.$$

Let $Z_\perp \in \mathbb{R}^{p\times(p-1)}$ such that $[z, Z_\perp]^{\mathrm{T}} [z, Z_\perp] = I_p$. The eigen-decomposition of $W^{\mathrm{T}} W$ is

$$W^{\mathrm{T}} W = [z, Z_\perp] \begin{bmatrix} 1+\alpha & \\ & I_{p-1} \end{bmatrix} [z, Z_\perp]^{\mathrm{T}}$$

which yields

$$\begin{aligned} (W^{\mathrm{T}} W)^{-1/2} &= [z, Z_\perp] \begin{bmatrix} (1+\alpha)^{-1/2} & \\ & I_{p-1} \end{bmatrix} [z, Z_\perp]^{\mathrm{T}} \\ &= (1+\alpha)^{-1/2} z z^{\mathrm{T}} + Z_\perp Z_\perp^{\mathrm{T}} \\ &= (1+\alpha)^{-1/2} z z^{\mathrm{T}} + I_p - z z^{\mathrm{T}} \\ &= I_p - [1 - (1+\alpha)^{-1/2}] z z^{\mathrm{T}}. \end{aligned}$$

Therefore,

$$\begin{aligned} V_+ &= (V + \beta y y^{\mathrm{T}} V)\{I_p - [1 - (1+\alpha)^{-1/2}] z z^{\mathrm{T}}\} \\ &= V + \beta y y^{\mathrm{T}} V - [1 - (1+\alpha)^{-1/2}](V + \beta y y^{\mathrm{T}} V) z z^{\mathrm{T}}, \end{aligned}$$

as expected. ☐

With the help of this lemma, for a fixed stepsize $\beta_i^{(n)} = \beta$ for all $1 \le i \le p$, we outline in Algorithm 3.1 a subspace online PCA algorithm based on (1.6) and (1.7). The iteration at its line 5 combines (1.6) and (1.7) as one. This seems like a minor reformulation, but it turns out to be one of the keys that make our analysis go through. The rest of this paper is devoted to analyze its convergence.

**Remark 3.1.** A couple of comments are in order for Algorithm 3.1.

1. Vectors $X^{(n)} \in \mathbb{R}^d$ for $n = 1, 2, \ldots$ are independent and identically distributed samples of $\boldsymbol{X}$.

2. If the algorithm converges, it is expected that

$$U^{(n)} \to U_* := U \begin{bmatrix} I_p \\ 0 \end{bmatrix} = [u_1, u_2, \ldots, u_p]$$

in the sense that $\|\sin\Theta(U^{(n)}, U_*)\|_{\mathrm{ui}} \to 0$ as $n \to \infty$.

Notations introduced in this section, except those in Lemma 3.1 will be adopted throughout the rest of this paper.

# 4 Main Results

We point out that any statement we will make is meant to hold *almost surely*.

We are concerned with random variables/vectors that have a sub-Gaussian distribution which we will define next. To that end, we need to introduce the Orlicz $\psi_\alpha$-norm of a random variable/vector. More details can be found in [21].

**Definition 4.1.** The *Orlicz $\psi_\alpha$-norm* of a random variable $\boldsymbol{X} \in \mathbb{R}$ is defined as

$$\|\boldsymbol{X}\|_{\psi_\alpha} := \inf\left\{\xi > 0 : \mathrm{E}\left\{\exp\left(\left|\frac{\boldsymbol{X}}{\xi}\right|^\alpha\right)\right\} \leq 2\right\},$$

and the *Orlicz $\psi_\alpha$-norm* of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ is defined as

$$\|\boldsymbol{X}\|_{\psi_\alpha} := \sup_{\|v\|_2=1} \|v^{\mathrm{T}}\boldsymbol{X}\|_{\psi_\alpha}.$$

We say that random variable/vector $\boldsymbol{X}$ follows a *sub-Gaussian distribution* if $\|\boldsymbol{X}\|_{\psi_2} < \infty$.

By the definition, we conclude that any bounded random variable/vector follows a sub-Gaussian distribution. To prepare our convergence analysis, we make a few assumptions.

**Assumption 4.1.** $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_d]^{\mathrm{T}} \in \mathbb{R}^d$ is a random vector.

(A-1) $\mathrm{E}\{\boldsymbol{X}\} = 0$, and $\Sigma := \mathrm{E}\left\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}$ has the spectral decomposition (3.1) satisfying (3.2);

(A-2) $\psi := \|\Sigma^{-1/2}\boldsymbol{X}\|_{\psi_2} < \infty$.

The principal subspace $\mathcal{U}_*$ in (3.3) is uniquely determined under (A-1) of Assumption 4.1. On the other hand, (A-2) of Assumption 4.1 ensures that all 1-dimensional marginals of $\boldsymbol{X}$ have sub-Gaussian tails, or equivalently, $\boldsymbol{X}$ follows a sub-Gaussian distribution. This is also an assumption that is used in [11].

In what follows, we will state our main results under the assumption and leave their proofs to later sections because of their high complexity. To that end, first we introduce some quantities. The eigenvalue gap is

$$\gamma := \lambda_p - \lambda_{p+1}.$$

The sum of top $i$ eigenvalues is

$$\eta_i := \lambda_1 + \cdots + \lambda_i, \quad i = 1, \dots, d.$$

The dominance of the top $i$ eigenvalues is defined as

$$\mu_i := \frac{\eta_i}{\eta_d} \in \left[\frac{i}{d}, 1\right],$$

For $s > 0$ and the stepsize $\beta < 1$ such that $\beta\gamma < 1$, define

$$N_s(\beta) := \min\left\{n \in \mathbb{N} : (1-\beta\gamma)^n \leq \beta^s\right\} = \left\lceil \frac{s \ln \beta}{\ln(1-\beta\gamma)} \right\rceil, \tag{4.1}$$

where $\lceil\cdot\rceil$ is the ceiling function taking the smallest integer that is no smaller than its argument, and for $0 < \varepsilon < 1/7$,

$$M(\varepsilon) := \min\left\{m \in \mathbb{N} : \beta^{7\varepsilon/2-1/2} \leq \beta^{(1-2^{1-m})(3\varepsilon-1/2)}\right\} = 2 + \left\lceil \frac{\ln \frac{1/2-3\varepsilon}{\varepsilon}}{\ln 2} \right\rceil \geq 2. \tag{4.2}$$

**Theorem 4.1.** *Given*

$$\varepsilon \in (0, 1/7), \quad \omega \in (0, 1), \quad \phi > 0, \quad \kappa > 6^{[M(\varepsilon)-1]/2} \max\{2(\sqrt{2}-1)^{1/2}\phi\lambda_1^{-1/2}\omega^{1/2}, \sqrt{2}\}, \tag{4.3a}$$

$$0 < \beta < \min\left\{1, \left(\frac{1}{8\kappa\eta_p}\right)^{\frac{2}{1-4\varepsilon}}, \left(\frac{\gamma}{130\kappa^2\eta_p^2}\right)^{\frac{1}{\varepsilon}}\right\}. \tag{4.3b}$$

*Let $U^{(n)}$ for $n = 1, 2, \ldots$ be the approximations of $U_*$ generated by* Algorithm 3.1. *Under* Assumption 4.1, *if $\|\tan \Theta(U^{(0)}, U_*)\|_2^2 \le \phi^2 d - 1$, and*

$$(\sqrt{2}+1)\lambda_1 d\beta^{1-7\varepsilon} \le \omega, \quad K > N_{3/2-37\varepsilon/4}(\beta), \tag{4.4}$$

*then there exist absolute constants*[1] $C_\psi$, $C_\nu$, $C_\circ$ *and a high-probability event* $\mathbb{H}$ *with*

$$\mathrm{P}\{\mathbb{H}\} \ge 1 - K[(2+\mathrm{e})d + p + 1]\exp(-C_{\nu\psi}\beta^{-\varepsilon})$$

*such that for any $n \in [N_{3/2-37\varepsilon/4}(\beta), K]$*

$$\begin{aligned} \mathrm{E}\Big\{\|\tan \Theta(U^{(n)}, U_*)\|_{\mathrm{F}}^2;\, \mathbb{H}\Big\} &\le (1-\beta\gamma)^{2(n-1)}p\phi^2 d \\ &\quad + \frac{32\psi^4\beta}{2-\lambda_1\beta}\varphi(p, d; \Lambda) + C_\circ\kappa^4\mu_p^{-2}\eta_p^2\gamma^{-1}p\sqrt{d-p}\,\beta^{3/2-7\varepsilon}, \end{aligned} \tag{4.5}$$

*where $\mathrm{e} = \exp(1)$ is Euler's number, $C_{\nu\psi} = \max\{C_\nu\mu_p, C_\psi \min\{\psi^{-1}, \psi^{-2}\}\}$, and*[2]

$$\varphi(p, d; \Lambda) := \sum_{j=1}^{p}\sum_{i=p+1}^{d} \frac{\lambda_j\lambda_i}{\lambda_j - \lambda_i} \in \left[\frac{p(d-p)\lambda_1\lambda_d}{\lambda_1 - \lambda_d}, \frac{p(d-p)\lambda_p\lambda_{p+1}}{\lambda_p - \lambda_{p+1}}\right]. \tag{4.6}$$

The conclusion of Theorem 4.1 holds for any given $U^{(0)}$ satisfying $\|\tan \Theta(U^{(0)}, U_*)\|_2^2 \le \phi^2 d - 1$. However, it is not easy, if at all possible, to verify this condition. Next we consider a randomly selected $U^{(0)}$.

Suppose that on $\mathbb{G}_p(\mathbb{R}^d)$ we use a uniform distribution, the one with the Haar invariant probability measure (see [3, Section 1.4] and [10, Section 4.6]). We refer the reader to [3, Section 2.2] on how to generate such a uniform distribution on $\mathbb{G}_p(\mathbb{R}^d)$. Our assumption for a randomly selected $U^{(0)}$ is

$$\boxed{\text{randomly selected } U^{(0)} \text{ satisfies that } \mathcal{R}(U^{(0)}) \text{ is uniformly sampled from } \mathbb{G}_p(\mathbb{R}^d).} \tag{4.7}$$

**Theorem 4.2.** *Under* Assumption 4.1, *for sufficiently large $d$ and any $\beta$ satisfying* (4.3b) *with $\kappa = 6^{[M(\varepsilon)-1]/2}\max\{2C_p, \sqrt{2}\}$, and*

$$p < (d+1)/2, \quad \varepsilon \in (0, 1/7), \quad \delta \in (0, 2^{-1/p^2}), \quad K > N_{3/2-37\varepsilon/4}(\beta),$$

*where $C_p$ is a constant only dependent on $p$, if* (4.7) *holds, and*

$$d\beta^{1-3\varepsilon} \le \delta^2, \qquad K[(2+\mathrm{e})d + p + 1]\exp(-C_{\nu\psi}\beta^{-\varepsilon}) \le \delta^{p^2},$$

*then there exists a high-probability event $\mathbb{H}_*$ with $\mathrm{P}\{\mathbb{H}_*\} \ge 1 - 2\delta^{p^2}$ such that*

[1] We attach each with a subscript for the convenience of indicating their associations. They don't change as the values of the subscript variables vary, by which we mean *absolute constants*. Later in (6.6), we explicitly bound these absolute constants.

[2] To see the inclusion in (4.6), we note the following: if $0 \le a \le c < d \le b$, then

$$0 \le \frac{1}{b} \le \frac{1}{d} < \frac{1}{c} \le \frac{1}{a} \;\Rightarrow\; \frac{dc}{d-c} = \frac{1}{\frac{1}{c} - \frac{1}{d}} \ge \frac{1}{\frac{1}{a} - \frac{1}{b}} = \frac{ab}{b-a}.$$

$$\mathrm{E}\Big\{\|\tan\Theta(U^{(n)},U_*)\|_{\mathrm{F}}^2;\ \mathbb{H}_*\Big\} \le (1-\beta\gamma)^{2(n-1)}p\,C_p^2\delta^{-2}d + \frac{32\psi^4\beta}{2-\lambda_1\beta}\varphi(p,d;\varLambda) + C_\circ\kappa^4\mu_p^{-2}\eta_p^2\gamma^{-1}p\sqrt{d-p}\beta^{3/2-7\varepsilon} \tag{4.8}$$

*for any* $n\in[N_{3/2-37\varepsilon/4}(\beta),K]$*, where* $\varphi(p,d;\varLambda)$ *is as in* (4.6)*.*

Finally, suppose that the number of principal components $p$ and the eigenvalue gap $\gamma=\lambda_p-\lambda_{p+1}$ is known in advance, and the sample size is fixed at $N_*$. We must choose a proper $\beta$ to obtain the principal components as accurately as possible. A good choice turns out to be

$$\beta=\beta_*:=\frac{3\ln N_*}{2\gamma N_*}. \tag{4.9}$$

**Theorem 4.3.** *Under* Assumption 4.1*, for sufficiently large* $d\ge 2p$ *and sufficiently large* $N_*$*,* $\varepsilon\in(0,1/7)$*,* $\delta\in(0,2^{-1/p^2})$ *satisfying*

$$d\beta_*^{1-3\varepsilon}\le\delta^2,\qquad N_*[(2+\mathrm{e})d+p+1]\exp(-C_{\nu\psi}\beta^{-\varepsilon})\le\delta^{p^2}, \tag{4.10}$$

*where* $\beta_*$ *is given by* (4.9)*, if* (4.7) *holds, then there exists a high-probability event* $\mathbb{H}_*$ *with* $\mathrm{P}\{\mathbb{H}_*\}\ge 1-2\delta^{p^2}$*, such that*

$$\mathrm{E}\Big\{\|\tan\Theta(U^{(N_*)},U_*)\|_{\mathrm{F}}^2;\ \mathbb{H}_*\Big\}\le C_*(d,N_*,\delta)\frac{\varphi(p,d;\varLambda)}{\lambda_p-\lambda_{p+1}}\frac{\ln N_*}{N_*}, \tag{4.11}$$

*where the constant* $C_*(d,N_*,\delta)\to 24\psi^4$ *as* $d\to\infty, N_*\to\infty$*, and* $\varphi(p,d;\varLambda)$ *is as in* (4.6)*.*

In Theorems 4.1, 4.2 and 4.3, the conclusions are stated in term of the expectation of $\|\tan\Theta(U^{(n)},U_*)\|_{\mathrm{F}}^2$ over some highly probable event. These expectations can be turned into conditional expectations, thanks to the relation (1.11). In fact, (1.10) is a consequence of (4.11) and (1.11).

The proofs of the three theorems are given in sections 6 and 7. Although overall our proofs follow the same structure of those in Li, Wang, Liu, and Zhang [11], there are inherently critical subtleties in going from one-dimension ($p=1$) to multi-dimension ($p>1$). In fact, one of key steps in proof works for $p=1$ does not seem to work for $p>1$. More detail will be discussed in the next section.

Now we observe the effect of the scaling on the random vector $\boldsymbol{X}$. Let

$$\check{\boldsymbol{X}}=\xi\boldsymbol{X},\quad \check{U}^{(0)}=U^{(0)},\quad \check{\beta}=\xi^{-2}\beta. \tag{4.12}$$

Then we can examine that

$$\check{X}^{(n)}=\xi X^{(n)},\quad \check{Z}^{(n)}=\xi Z^{(n)},\quad \check{\alpha}^{(n)}=\alpha^{(n)},\quad \check{U}^{(n)}=U^{(n)},$$

which means that Algorithm 3.1 will produce the same sequence $\{U^{(n)}\}$ under the scaling (4.12). Also, we have

$$\check{\varSigma}=\xi^2\varSigma,\quad \check{\psi}=\psi,\quad \check{U}=U,\quad \check{\varLambda}=\xi^2\varLambda,\quad \check{\lambda}_i=\xi^2\lambda_i,\quad \check{\gamma}=\xi^2\gamma,\quad \check{\mu}=\mu,\quad \check{\nu}=\nu.$$

Considering a scaling $\xi=\beta^{-\zeta/2}$ with $\zeta$ an integer, we can see

$$\check{N}_s(\check{\beta})=(1+\zeta)N_s(\beta),\quad \check{\beta}_*=\beta^\zeta\beta_*.$$

In other words, using the scaling technique, we can use a much smaller stepsize (e.g. from $\beta$ to $\beta^{1+\zeta}$) while the number of steps to converge does not increase too much (from $N$ to $(1+\zeta)N$); on the other hand, with the same sample size $N_*$, using the scaling technique, we can use a much smaller stepsize (e.g. from $\beta_*$ to $\beta_*^{1+\zeta}$) in order to make the stepsize satisfy (4.10). This could be very helpful when choosing a proper stepsize in practice.

## 5 Comparisons with Previous Results

Our three theorems in the previous section, namely Theorems 4.1, 4.2 and 4.3, are the analogs for $p > 1$ of Li, Wang, Liu, and Zhang's three theorems [11, Theorems 1, 2, and 3] which are for $p = 1$ only. Naturally, we would like to know how our results when applied to the case $p = 1$ and our proofs would stand against those in [11]. In what follows, we will do a fairly detailed comparison. But before we do that, let us state their theorems (in our notation).

**Theorem 5.1** ([11, Theorem 1]). *Under* Assumption 4.1 *and* $p = 1$*, suppose there exists a constant* $\phi > 1$ *such that* $\tan\Theta(U^{(0)}, U_*) \le \phi^2 d$*. Then for any* $\varepsilon \in (0, 1/8)$*, stepsize* $\beta > 0$ *satisfying* $d[\lambda_1^2\gamma^{-1}\beta]^{1-2\varepsilon} \le b_1\phi^{-2}$*, and any* $t > 1$*, there exists an event* $\mathbb{H}$ *with*

$$\mathrm{P}\{\mathbb{H}\} \ge 1 - 2(d+2)\widehat{N}^{\mathrm{o}}(\beta,\phi)\exp\left(-C_0[\lambda_1^2\gamma^{-1}\beta]^{-2\varepsilon}\right) - 4d\widehat{N}_t(\beta)\exp\left(-C_1[\lambda_1^2\gamma^{-1}\beta]^{-2\varepsilon}\right),$$

*such that for any* $n \in [\widehat{N}_1(\beta) + \widehat{N}^{\mathrm{o}}(\beta,\phi), \widehat{N}_t(\beta)]$

$$\mathrm{E}\Big\{\tan^2\Theta(U^{(n)}, U_*);\ \mathbb{H}\Big\} \le (1-\beta\gamma)^{2[n-\widehat{N}^{\mathrm{o}}(\beta,\phi)]} + C_2\beta\varphi(1, d; \Lambda) + C_2\sum_{i=2}^{d}\frac{\lambda_1-\lambda_2}{\lambda_1-\lambda_i}[\lambda_1^2\gamma^{-1}\beta]^{3/2-4\varepsilon}, \tag{5.1}$$

*where* $b_1 \in (0, \ln^2 2/16)$*,* $C_0, C_1, C_2$ *are absolute constants, and*

$$\widehat{N}^{\mathrm{o}}(\beta,\phi) := \min\left\{n\in\mathbb{N} : (1-\beta\gamma)^n \le [4\phi^2 d]^{-1}\right\} = \left\lceil\frac{-\ln[4\phi^2 d]}{\ln(1-\beta\gamma)}\right\rceil,$$
$$\widehat{N}_s(\beta) := \min\left\{n\in\mathbb{N} : (1-\beta\gamma)^n \le [\lambda_1^2\gamma^{-1}\beta]^s\right\} = \left\lceil\frac{s\ln[\lambda_1^2\gamma^{-1}\beta]}{\ln(1-\beta\gamma)}\right\rceil.$$

What we can see that Theorem 4.1 for $p = 1$ is essentially the same as Theorem 5.1. In fact, since $(1-\beta\gamma)^{1-\widehat{N}^{\mathrm{o}}(\beta,\phi)} \le 4\phi^2 d \le (1-\beta\gamma)^{-\widehat{N}^{\mathrm{o}}(\beta,\phi)}$, the upper bounds by (4.5) for $p = 1$ and by (5.1) are comparable in the sense that they are in the same order in $d, \beta, \delta$.

Naturally one may try to generalize the proving techniques in [11] which is for the one-dimensional case ($p = 1$) to handle the multi-dimensional case ($p > 1$). Indeed, we tried but didn't succeed, due to we believe insurmountable obstacles. We now explain. The basic structure of the proof in [11] is to split the Grassmann manifold $\mathbb{G}_p(\mathbb{R}^d)$, where the initial guess comes from, into two regions: the *cold region* and *warm region*. Roughly speaking, an approximation $U^{(n)}$ in the warm region means that $\|\tan\Theta(U^{(n)}, U_*)\|_{\mathrm{F}}$ is small while it in the cold region means that $\|\tan\Theta(U^{(n)}, U_*)\|_{\mathrm{F}}$ is not that small. $U_*$ sits at the "*center*" of the warm region which is wrapped around by the cold region. The proof is divided into two cases: the first case is when the initial guess is in the warm region and the other one is when it is in the code region. For the first case, they proved that the algorithm will produce a sequence convergent to the principal subspace (which is actually the most significant principal component because it is for $p = 1$) with high probability. For the second case, they first proved that the algorithm will produce a sequence of approximations that, after a finite number of iterations, will fall into the warm region with high probability, and then use the conclusion proved for the first case to conclude the proof because of the Markov property.

For our situation $p > 1$, we still structure our proof in the same way, i.e., dividing the whole proof into two cases of $U^{(0)}$ coming from the *cold region* or *warm region*. The proof in [11] for the warm region case can be carried over with a little extra effort, as we will see later, but we didn't find that it was possible to use a similar argument in [11] to get the job done for the cold region case. Three major difficulties are as follows. In [11], essentially $\|\cot\Theta(U^{(n)}, U_*)\|_{\mathrm{F}}$ was used to track the behavior of a martingale along with the power iteration. Note $\cot\Theta(U^{(n)}, U_*)$ is $p \times p$. Thus it is a scalar when $p = 1$, perfectly well-conditioned if treated as a matrix, but for $p > 1$, it is a genuine matrix and, in fact, an inverse of a random matrix in the proof. The first difficulty is how to estimate the inverse because it may not even exist! We tried to separate the flow of $U^{(n)}$ into two subflows: the ill-conditioned flow and the well-conditioned flow, and estimate the related quantities separately. Here the ill-conditioned flow at each step represents the subspace generated by the singular vectors of $\cot\Theta(U^{(n)},U_*)$ whose corresponding singular values are tiny, while the well-conditioned flow at each step represents the subspace generated by the other singular vectors, of which the inverse (restricted to this subspace) is well conditioned. Unfortunately, tracking the two flows can be an impossible task because, due to the randomness, some elements in the ill-conditioned flow could jump to the well-conditioned flow during the iteration, and vice versa. This is the second difficulty. The third one is to build a martingale to go along with a proper power iteration, or equivalently, to find the Doob decomposition of the process, because the recursion formula of the main part of the inverse — the drift in the Doob decomposition, even if limited to the well-conditioned flow, is not a linear operator, which makes it impossible to build a proper power iteration.

In the end, to deal with the cold region, we gave up the idea of estimating $\|\cot\Theta(U^{(n)},U_*)\|_{\mathrm{F}}$. Instead, we invent another method: cutting the cold region into many layers, each wrapped around by another with the innermost one around the warm region. We prove the initial guess in any layer will produce a sequence of approximations that will fall into its inner neighbor layer (or the warm region if the layer is innermost) in a finite number of iterations with high probability. Therefore eventually, any initial guess in the cold region will lead to an approximation in the warm region within a finite number of iterations with high probability, returning to the case of initial guesses coming from the warm region because of the Markov property. This enables us to completely avoid the difficulties mentioned above. This technique can also be used for the one-dimensional case to simplify the proof in [11].

**Theorem 5.2** ([11, Theorem 2]). *Under* Assumption 4.1 *and $p=1$, suppose that $U^{(0)}$ is uniformly sampled from the unit sphere. Then for any $\varepsilon\in(0,1/8)$, stepsize $\beta>0,\delta>0$ satisfying*

$$d[\lambda_1^2\gamma^{-1}\beta]^{1-2\varepsilon}\le b_2\delta^2,\qquad 4d\widehat{N}_2(\beta)\exp\left(-C_3[\lambda_1^2\gamma^{-1}\beta]^{-2\varepsilon}\right)\le\delta,$$

*there exists an event $\mathbb{H}_*$ with $\mathrm{P}\{\mathbb{H}_*\}\ge 1-2\delta$ such that for any $n\in[\widehat{N}_2(\beta),\widehat{N}_3(\beta)]$*

$$\mathrm{E}\Big\{\tan^2\Theta(U^{(n)},U_*);\,\mathbb{H}_*\Big\}\le C_4(1-\beta\gamma)^{2n}\delta^{-4}d^2+C_4\beta\varphi(1,d;\Lambda)+C_4\sum_{i=2}^{d}\frac{\lambda_1-\lambda_2}{\lambda_1-\lambda_i}[\lambda_1^2\gamma^{-1}\beta]^{3/2-4\varepsilon},\tag{5.2}$$

*where $b_2,C_3,C_4$ are absolute constants.*

**Theorem 5.3** ([11, Theorem 3]). *Under* Assumption 4.1 *and $p=1$, suppose that $U^{(0)}$ is uniformly sampled from the unit sphere and let $\beta_*=\frac{2\ln N_*}{\gamma N}$. Then for any $\varepsilon\in(0,1/8)$, $N_*\ge1$, $\delta>0$ satisfying*

$$d[\lambda_1^2\gamma^{-1}\beta_*]^{1-2\varepsilon}\le b_3\delta^2,\qquad 4d\widehat{N}_2(\beta_*)\exp\left(-C_6[\lambda_1^2\gamma^{-1}\beta_*]^{-2\varepsilon}\right)\le\delta,$$

*there exists an event $\mathbb{H}_*$ with $\mathrm{P}\{\mathbb{H}_*\}\ge 1-2\delta$ such that*

$$\mathrm{E}\Big\{\tan^2\Theta(U^{(N_*)},U_*);\,\mathbb{H}_*\Big\}\le C_*(d,N_*,\delta)\frac{\varphi(1,d;\Lambda)}{\lambda_1-\lambda_2}\frac{\ln N_*}{N_*},\tag{5.3}$$

*where the constant $C_*(d,N_*,\delta)\to C_5$ as $d\to\infty,N_*\to\infty$, and $b_3,C_5,C_6$ are absolute constants.*

Our Theorems 4.2 and 4.3 when applied to the case $p=1$ do not exactly yield Theorems 5.2 and 5.3, respectively. But the resulting conditions and upper bounds have the same orders in variables $d,\beta,\delta$, and the coefficients of $\beta$ and $\frac{\ln N_*}{N}$ in the upper bounds are comparable. But we note that the first term in right-hand side of (4.8) is proportional to $d$, not $d^2$ as in (5.2); so ours is tighter.

The proofs here for Theorems 4.2 and 4.3 are nearly the same as those in [11] for Theorems 5.2 and 5.3 owing to the fact that the difficult estimates have already been taken care of by either Theorem 4.1 or Theorem 5.1. But still there are some extras for $p>1$, namely, the need to estimate the marginal probability for the uniform distribution on the Grassmann manifold of dimension higher than 1. We cannot find it in the literature, and thus have to build it ourselves with the help of the theory of special functions of a matrix argument, rarely used in the statistical community.

It may also be worth pointing out that all absolute constants, except $C_p$ which has an explicit expression in (7.4) and $C_\psi$, in our theorems are concretely bounded as in (6.6), whereas those in Theorems 5.1 to 5.3 are not.

# 6 Proof of Theorem 4.1

In this section, we will prove Theorem 4.1. For that purpose, we build a quite amount of preparation material in subsections 6.1, 6.2 and 6.3 before we prove the theorem in subsection 6.4. Figure 6.1 shows a pictorial description of our proof process.

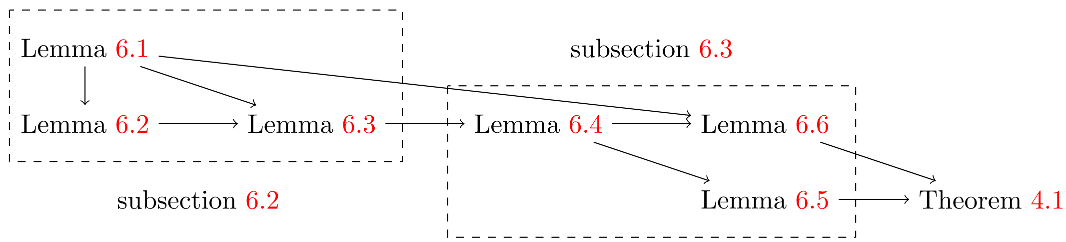


Figure 6.1: Proof process for Theorem 4.1

## 6.1 Simplification

Without loss of generality, we may assume that the covariance matrix $\Sigma$ diagonal. Otherwise, we can perform a (constant) orthogonal transformation as follows. Recall the spectral decomposition $\Sigma = U\Lambda U^{\mathrm{T}}$ in (3.1). Instead of the random vector $\boldsymbol{X}$, we equivalently consider

$$\boldsymbol{Y} \equiv [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n]^{\mathrm{T}} := U^{\mathrm{T}}\boldsymbol{X}.$$

Accordingly, perform the same orthogonal transformation on all involved quantities:

$$Y^{(n)} = U^{\mathrm{T}}X^{(n)}, \quad V^{(n)} = U^{\mathrm{T}}U^{(n)}, \quad V_* = U^{\mathrm{T}}U_* = \begin{bmatrix} I_p \\ 0 \end{bmatrix}. \tag{6.1}$$

As consequences, we will have the equivalent versions of Algorithm 3.1, Theorems 4.1, 4.2 and 4.3. Firstly, because

$$(V^{(n-1)})^{\mathrm{T}}Y^{(n)} = (U^{(n-1)})^{\mathrm{T}}X^{(n)}, \quad (Y^{(n)})^{\mathrm{T}}Y^{(n)} = (X^{(n)})^{\mathrm{T}}X^{(n)},$$

the equivalent version of Algorithm 3.1 is obtained by symbolically replacing all letters $X$, $U$ by $Y$, $V$ while keeping their respective superscripts. If the algorithm converges, it is expected that $\mathcal{R}(V^{(n)}) \to \mathcal{R}(V_*)$. Secondly, noting

$$\|\Sigma^{-1/2}\boldsymbol{X}\|_{\psi_2} = \|U\Lambda^{-1/2}U^{\mathrm{T}}\boldsymbol{X}\|_{\psi_2} = \|\Lambda^{-1/2}\boldsymbol{Y}\|_{\psi_2},$$

we can restate Assumption 4.1 equivalently as

(A-1′) $\mathrm{E}\{\boldsymbol{Y}\} = 0, \mathrm{E}\left\{\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\right\} = \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ with (3.2);

(A-2′) $\psi := \|\Lambda^{-1/2}\boldsymbol{Y}\|_{\psi_2} < \infty$.

Thirdly, all canonical angles between two subspaces are invariant under the orthogonal transformation. Therefore the equivalent versions of Theorems 4.1, 4.2 and 4.3 for $\boldsymbol{Y}$ can be simply obtained by replacing all letters $X$, $U$ by $Y$, $V$ while keeping their respective superscripts.

In the rest of this section, we will prove the mentioned equivalent version of Theorem 4.1. Likewise in the next section, we will prove the equivalent versions of Theorems 4.2 and 4.3.

In what follows, we assume that $\Sigma$ is diagonal.

To facilitate our proof, we introduce new notations for two particular submatrices of any $V \in \mathbb{R}^{d\times p}$:

$$\bar{V} = V_{(1:p,:)}, \quad \underline{V} = V_{(p+1:d,:)}. \tag{6.2}$$

In particular, $\mathscr{T}(V) = \underline{V}\bar{V}^{-1}$ for the operator $\mathscr{T}$ defined in (2.3), provided $\bar{V}$ is nonsingular. Set

$$\bar{\varLambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_p), \quad \underline{\varLambda} = \operatorname{diag}(\lambda_{p+1}, \dots, \lambda_d). \tag{6.3}$$

Although the assignments to $\bar{\varLambda}$ and $\underline{\varLambda}$ are not consistent with the extractions defined by (6.2), they don't seem to cause confusions in our later presentations.

For $\kappa > 1$, define $\mathbb{S}(\kappa) := \{V \in \mathbb{R}^{d\times p} : \sigma(\bar{V}) \subset [\frac{1}{\kappa}, 1]\}$. It can be verified that

$$V \in \mathbb{S}(\kappa) \Leftrightarrow \|\mathscr{T}(V)\|_2 \le \sqrt{\kappa^2 - 1}. \tag{6.4}$$

For the sequence $V^{(n)}$, define

$$N_{\text{out}}\{\mathbb{S}(\kappa)\} := \min\{n : V^{(n)} \notin \mathbb{S}(\kappa)\}, \quad N_{\text{in}}\{\mathbb{S}(\kappa)\} := \min\{n : V^{(n)} \in \mathbb{S}(\kappa)\}.$$

$N_{\text{out}}\{\mathbb{S}(\kappa)\}$ is the first step of the iterative process at which $V^{(n)}$ jumps from $\mathbb{S}(\kappa)$ to outside, and $N_{\text{in}}\{\mathbb{S}(\kappa)\}$ is the first step of the iterative process at which $V^{(n)}$ jumps from outside to $\mathbb{S}(\kappa)$. Write

$$\widetilde{\lambda}_i := \lambda_i \beta^{-2\varepsilon}, \quad \widetilde{\eta}_i := \widetilde{\lambda}_1 + \cdots + \widetilde{\lambda}_i = \eta_i \beta^{-2\varepsilon},$$

and define

$$N_{\text{qb}}\{\varLambda\} := \max\Big\{n \ge 1 : \|Z^{(n)}\|_2 \le \widetilde{\eta}_p^{1/2}, |Y_i^{(n)}| \le \widetilde{\lambda}_i^{1/2}, i = 1, \dots n\Big\} + 1. \tag{6.5}$$

$N_{\text{qb}}\{\varLambda\}$ is the first step of the iterative process at which either $|Y_i^{(n)}| > \widetilde{\lambda}_i^{1/2}$ for some $i$ or the norm of $Z^{(n)}$ exceeds $\widetilde{\eta}_p^{1/2}$. For $n < N_{\text{qb}}\{\varLambda\}$, we have

$$\|Y^{(n)}\|_2 \le \widetilde{\eta}_d^{1/2} = \nu^{1/2}\widetilde{\eta}_p^{1/2}, \quad \|Z^{(n)}\|_2 \le \widetilde{\eta}_p^{1/2},$$

where $\nu = 1/\mu_p$.

For convenience, we will set $T^{(n)} = \mathscr{T}(V^{(n)})$, and let $\mathbb{F}_n = \sigma\{Y^{(1)}, \dots, Y^{(n)}\}$ be the $\sigma$-algebra filtration, i.e., the information known by step $n$. Also, since in this section $\varepsilon$, $\beta$ are fixed, we suppress the dependency information of $M(\varepsilon)$ on $\varepsilon$ and $N_s(\beta)$ on $\beta$ to simply write $M$ for $M(\varepsilon)$ and $N_s$ for $N_s(\beta)$.

Lastly, we discuss some of the important implications of the conditions:

$$0 < \beta < \min\left\{1, \left(\frac{1}{8\kappa\eta_p}\right)^{\frac{2}{1-4\varepsilon}}, \left(\frac{\gamma}{130\kappa^2\eta_p^2}\right)^{\frac{1}{\varepsilon}}\right\}, \tag{4.3b}$$

$$(\sqrt{2}+1)\lambda_1 d\beta^{1-7\varepsilon} \le \omega, \quad K > N_{3/2-37\varepsilon/4}(\beta) \tag{4.4}$$

of Theorem 4.1. They guarantee that

($\beta$-1) $\beta < 1$;

($\beta$-2) $\beta\gamma \le \beta\widetilde{\eta}_p \le \nu\beta\widetilde{\eta}_p = \beta\widetilde{\eta}_d \le d\beta\widetilde{\lambda}_1 = d\lambda_1\beta^{1-2\varepsilon} \le (\sqrt{2}-1)\omega \le \sqrt{2}-1$.

Set

$$C_V = \frac{5}{2} + \frac{7}{2}(\nu\widetilde{\eta}_p\beta) + \frac{15}{8}(\nu\widetilde{\eta}_p\beta)^2 + \frac{3}{8}(\nu\widetilde{\eta}_p\beta)^3 \le \frac{16+13\sqrt{2}}{8} \approx 4.298; \tag{6.6a}$$

$$C_\varDelta = 2 + \frac{1}{2}(\nu\widetilde{\eta}_p\beta) + C_V\widetilde{\eta}_p\beta \le \frac{22+7\sqrt{2}}{8} \approx 3.987; \tag{6.6b}$$

$$C_T = C_V + 2C_\Delta + 2C_\Delta C_V \widetilde{\eta}_p \beta \le \frac{251+122\sqrt{2}}{16} \approx 26.471; \tag{6.6c}$$

$$C_\kappa = \frac{(3-\sqrt{2})C_\Delta^2}{64(C_T+2C_\Delta)^2} \le \frac{565+171\sqrt{2}}{21504} \approx 0.038; \tag{6.6d}$$

$$C_\nu = 4\sqrt{2} C_T C_\kappa \le \frac{223702+183539\sqrt{2}}{86016} \approx 5.618; \tag{6.6e}$$

$$\begin{aligned} C_\circ &= \frac{29+8\sqrt{2}}{16(3-\sqrt{2})} + \frac{4C_T}{(3-\sqrt{2})C_\Delta}\beta^{3\varepsilon} + [C_T + \frac{29+8\sqrt{2}}{32}]\beta^{1/2-3\varepsilon} \\ &\quad + \frac{3C_T^2}{2(3-\sqrt{2})C_\Delta^2}\beta^{1/2+3\varepsilon} + \frac{2C_T}{C_\Delta}\beta^{1-3\varepsilon} + \frac{C_T^2}{2C_\Delta^2}\beta^{3/2-3\varepsilon} \\ &\le \frac{2582968+1645155\sqrt{2}}{14336} \approx 342.464. \end{aligned} \tag{6.6f}$$

The condition (4.3b) also guarantees that

($\beta$-3) $2C_\Delta \widetilde{\eta}_p \beta^{1/2}\kappa = 2C_\Delta \eta_p \beta^{1/2-2\varepsilon}\kappa \le \frac{2C_\Delta}{8} < 1$, and thus $2C_\Delta \widetilde{\eta}_p \beta\kappa < 1$;

($\beta$-4) $4\sqrt{2}C_T\kappa^2\widetilde{\eta}_p^2\gamma^{-1}\beta^{5\varepsilon} \le 1$, and thus $4\sqrt{2}C_T\kappa^2\widetilde{\eta}_p^2\gamma^{-1}\beta^{1/2+\chi} \le 1$ for $\chi \in [-1/2+5\varepsilon, 0]$.

## 6.2 Increments of One Iteration

**Lemma 6.1.** *For any fixed $K \ge 1$,*

$$\mathrm{P}\{N_{\mathrm{qb}}\{\Lambda\} > K\} \ge 1 - K(\mathrm{e}d + p + 1)\exp\left(-C_\psi \min\{\psi^{-1}, \psi^{-2}\}\beta^{-2\varepsilon}\right),$$

*where $C_\psi$ is an absolute constant.*

*Proof.* Since

$$\{N_{\mathrm{qb}}\{\Lambda\} \le K\} \subset \bigcup_{n\le K}\left(\left\{\|Z^{(n)}\|_2 \ge \widetilde{\eta}_p^{1/2}\right\} \cup \bigcup_{1\le i\le d}\left\{|e_i^{\mathrm{T}} Y^{(n)}| \ge \widetilde{\lambda}_i^{1/2}\right\}\right),$$

we know

$$\mathrm{P}\{N_{\mathrm{qb}}\{\Lambda\} \le K\} \le \sum_{n\le K}\left(\mathrm{P}\left\{\|Z^{(n)}\|_2 \ge \widetilde{\eta}_p^{1/2}\right\} + \sum_{1\le i\le d}\mathrm{P}\left\{|e_i^{\mathrm{T}} Y^{(n)}| \ge \widetilde{\lambda}_i^{1/2}\right\}\right). \tag{6.7}$$

First,

$$\begin{aligned} \mathrm{P}\left\{|e_i^{\mathrm{T}} Y^{(n)}| \ge \widetilde{\lambda}_i^{1/2}\right\} &= \mathrm{P}\left\{\left|\frac{(\Lambda^{1/2}e_i)^{\mathrm{T}}}{\|\Lambda^{1/2}e_i\|_2}\Lambda^{-1/2}Y^{(n)}\right| \ge \frac{\widetilde{\lambda}_i^{1/2}}{\|\Lambda^{1/2}e_i\|_2}\right\} \\ &\le \exp\left(1 - \frac{C_{\psi,i}\frac{\widetilde{\lambda}_i}{e_i^{\mathrm{T}}\Lambda e_i}}{\|\frac{(\Lambda^{1/2}e_i)^{\mathrm{T}}}{\|\Lambda^{1/2}e_i\|_2}\Lambda^{-1/2}Y^{(n)}\|_{\psi_2}}\right) \qquad \text{by [22, (5.10)]} \\ &\le \exp\left(1 - \frac{C_{\psi,i}\widetilde{\lambda}_i}{\|\Lambda^{-1/2}Y^{(n)}\|_{\psi_2}\lambda_i}\right) = \exp\left(1 - C_{\psi,i}\psi^{-1}\beta^{-2\varepsilon}\right), \end{aligned} \tag{6.8}$$

where $C_{\psi,i}, i = 1, \dots, d$ are absolute constants [22, (5.10)]. Next, we claim

$$\mathrm{P}\left\{\|Z^{(n)}\|_2 \ge \widetilde{\eta}_p^{1/2}\right\} \le (p+1)\exp\left(-C_{\psi,d+1}\psi^{-2}\beta^{-2\varepsilon}\right). \tag{6.9}$$

Together, (6.7) – (6.9) yield

$$\mathrm{P}\{N_{\mathrm{qb}}\{\Lambda\} \le K\} = \sum_{n\le K}\sum_{1\le i\le d}\exp\left(1 - C_{\psi,i}\psi^{-1}\beta^{-2\varepsilon}\right) + \sum_{n\le K}(p+1)\exp\left(-C_{\psi,d+1}\psi^{-2}\beta^{-2\varepsilon}\right)$$

$$\leq K(\mathrm{e}d+p+1)\exp\left(-C_\psi \min\{\psi^{-1},\psi^{-2}\}\beta^{-2\varepsilon}\right),$$

where $C_\psi = \min_{1\leq i\leq d+1} C_{\psi,i}$. Finally, use $\mathrm{P}\{N_{\mathrm{qb}}\{\Lambda\}>K\} = 1-\mathrm{P}\{N_{\mathrm{qb}}\{\Lambda\}\leq K\}$ to complete the proof.

It remains to prove the claim (6.9). To avoid the cluttered superscripts, we drop the superscript "$\cdot^{(n-1)}$" on $V$, and drop the superscript "$\cdot^{(n)}$" on $Y$, $Z$. Consider

$$W := \begin{bmatrix} 0 & Z \\ Z^{\mathrm{T}} & 0\end{bmatrix} = \begin{bmatrix} & V^{\mathrm{T}}Y \\ Y^{\mathrm{T}}V & \end{bmatrix} = \sum_{k=1}^d Y_k \begin{bmatrix} & & & v_{k1} \\ & & & \vdots \\ & & & v_{kp} \\ v_{k1} & \cdots & v_{kp} & 0\end{bmatrix} =: \sum_{k=1}^d Y_k W_k,$$

where $v_{ij}$ is the $(i,j)$-entry of $V$. By the matrix version of master tail bound [20, Theorem 3.6], for any $\alpha>0$,

$$\mathrm{P}\{\|Z\|_2\geq\alpha\} = \mathrm{P}\{\lambda_{\max}(W)\geq\alpha\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta\alpha}\operatorname{trace}\exp\left(\sum_{k=1}^d \ln \mathrm{E}\{\exp(\theta Y_k W_k)\}\right).$$

$Y$ is sub-Gaussian and $\mathrm{E}\{Y\}=0$, and so is $Y_k$. Moreover,

$$\|Y_k\|_{\psi_2} = \|e_k^{\mathrm{T}}\Lambda^{1/2}\|_2 \left\|\frac{e_k^{\mathrm{T}}\Lambda^{1/2}}{\|e_k^{\mathrm{T}}\Lambda^{1/2}\|_2}\Lambda^{-1/2}Y\right\|_{\psi_2} \leq \lambda_k^{1/2}\|\Lambda^{-1/2}Y\|_{\psi_2} = \lambda_k^{1/2}\psi.$$

Also, by [22, (5.12)],

$$\mathrm{E}\{\exp(\theta W_k Y_k)\} \leq \exp(C_{\psi,d+k}\theta^2 W_k\circ W_k \|Y_k\|_{\psi_2}^2) \leq \exp(c_{\psi,k}\theta^2\lambda_k\psi^2 W_k\circ W_k),$$

where $c_{\psi,k}, k=1,\dots,d$ are absolute constants. Therefore, writing $[4C_{\psi,d+1}]^{-1} = \max_{1\leq k\leq d} c_{\psi,k}$ and $W_\psi := \sum_{k=1}^d \lambda_k W_k\circ W_k$ with the spectral decomposition $W_\psi = V_\psi\Lambda_\psi V_\psi^{\mathrm{T}}$, we have

$$\begin{aligned}
\operatorname{trace}\exp\left(\sum_{k=1}^d \ln\mathrm{E}\{\exp(\theta Y_k W_k)\}\right) &\leq \operatorname{trace}\exp\left(\sum_{k=1}^d c_{\psi,k}\theta^2\lambda_k\psi^2 W_k\circ W_k\right) \\
&\leq \operatorname{trace}\exp([4C_{\psi,d+1}]^{-1}\theta^2\psi^2 W_\psi) \\
&= \operatorname{trace}\exp([4C_{\psi,d+1}]^{-1}\theta^2\psi^2 V_\psi\Lambda_\psi V_\psi^{\mathrm{T}}) \\
&= \operatorname{trace}\left(V_\psi\exp([4C_{\psi,d+1}]^{-1}\theta^2\psi^2\Lambda_\psi)V_\psi^{\mathrm{T}}\right) \\
&= \operatorname{trace}\exp([4C_{\psi,d+1}]^{-1}\theta^2\psi^2\Lambda_\psi) \\
&\leq (p+1)\exp([4C_{\psi,d+1}]^{-1}\theta^2\psi^2\lambda_{\max}(\Lambda_\psi)) \\
&= (p+1)\exp([4C_{\psi,d+1}]^{-1}\theta^2\psi^2\lambda_{\max}(W_\psi)).
\end{aligned}$$

Note that

$$W_\psi = \begin{bmatrix} & & & \sum_{k=1}^d \lambda_k v_{k1}^2 \\ & & & \vdots \\ & & & \sum_{k=1}^d \lambda_k v_{kp}^2 \\ \sum_{k=1}^d\lambda_k v_{k1}^2 & \cdots & \sum_{k=1}^d \lambda_k v_{kp}^2 & 0\end{bmatrix} = \begin{bmatrix} & & & e_1^{\mathrm{T}}V^{\mathrm{T}}\Lambda V e_1 \\ & & & \vdots \\ & & & e_p^{\mathrm{T}}V^{\mathrm{T}}\Lambda V e_p \\ e_1^{\mathrm{T}}V^{\mathrm{T}}\Lambda V e_1 & \cdots & e_p^{\mathrm{T}}V^{\mathrm{T}}\Lambda V e_p & 0\end{bmatrix},$$

and thus

$$\lambda_{\max}(W_\psi) = \left\|\begin{bmatrix} e_1^{\mathrm{T}}V^{\mathrm{T}}\Lambda V e_1 \\ \vdots \\ e_p^{\mathrm{T}}V^{\mathrm{T}}\Lambda V e_p\end{bmatrix}\right\|_2 \leq \sum_{k=1}^p e_k^{\mathrm{T}}V^{\mathrm{T}}\Lambda V e_k = \operatorname{trace}(V^{\mathrm{T}}\Lambda V)$$

$$\leq \max_{V^{\mathrm{T}}V=I_p} \operatorname{trace}(V^{\mathrm{T}}\Lambda V) = \sum_{k=1}^{p} \lambda_k = \eta_p.$$

In summary, we have

$$\begin{aligned}
\mathrm{P}\{\|Z\|_2 \geq \alpha\} &\leq (p+1) \inf_{\theta>0} \exp([4C_{\psi,d+1}]^{-1}\theta^2\psi^2\eta_p - \theta\alpha) \\
&= (p+1)\exp\left(-\frac{C_{\psi,d+1}\alpha^2}{\psi^2\eta_p}\right).
\end{aligned}$$

Substituting $\alpha = \widetilde{\eta}_p^{1/2}$, we have the claim (6.9). □

**Lemma 6.2.** *Suppose that the conditions of* Theorem 4.1 *hold. If $n < N_{\mathrm{qb}}\{\Lambda\}$, then*

$$\begin{aligned}
V^{(n+1)} = V^{(n)} &+ \beta Y^{(n+1)}(Z^{(n+1)})^{\mathrm{T}} \\
&- \beta\Big[1 + \frac{\beta}{2}(Y^{(n+1)})^{\mathrm{T}}Y^{(n+1)}\Big]V^{(n)}Z^{(n+1)}(Z^{(n+1)})^{\mathrm{T}} + R^{(n)}(Z^{(n+1)})^{\mathrm{T}}, \quad (6.10)
\end{aligned}$$

*where $R^{(n)} \in \mathbb{R}^d$ is a random vector with $\|R^{(n)}\|_2 \leq C_V \nu^{1/2}\widetilde{\eta}_p^{3/2}\beta^2$ and $C_V$ is as in (6.6a).*

*Proof.* To avoid the cluttered superscripts, in this proof, we drop the superscript "$\cdot^{(n)}$" and use the superscript "$\cdot^{+}$" to replace "$\cdot^{(n+1)}$" on $V$, and drop the superscript "$\cdot^{(n+1)}$" on $Y$, $Z$.

On the set $\{N_{\mathrm{qb}}\{\Lambda\} > n\}$, by (4.4) and ($\beta$-2), we have

$$\alpha = \beta(2+\beta Y^{\mathrm{T}}Y)Z^{\mathrm{T}}Z \leq \beta(2+\nu\widetilde{\eta}_p\beta)\widetilde{\eta}_p \leq (2+\sqrt{2}-1)(\sqrt{2}-1)/\nu < 1.$$

By Taylor's expansion, there exists $0 < \xi < \alpha$ such that

$$\begin{aligned}
(1+\alpha)^{-1/2} &= 1 - \frac{1}{2}\alpha + \frac{3}{8}\frac{1}{(1+\xi)^{5/2}}\alpha^2 \\
&= 1 - \beta Z^{\mathrm{T}}Z - \frac{\beta^2}{2}Y^{\mathrm{T}}YZ^{\mathrm{T}}Z + \beta^2(Z^{\mathrm{T}}Z)^2\zeta,
\end{aligned}$$

where $\zeta = \frac{3}{8}\frac{1}{(1+\xi)^{5/2}}(2+\beta Y^{\mathrm{T}}Y)^2 \leq \frac{3}{8}(2+\nu\beta\widetilde{\eta}_p)^2$. Thus

$$\begin{aligned}
V^{+} &= (V + \beta YZ^{\mathrm{T}})\left(I - \left[\beta Z^{\mathrm{T}}Z + \frac{\beta^2}{2}Y^{\mathrm{T}}YZ^{\mathrm{T}}Z - \beta^2(Z^{\mathrm{T}}Z)^2\zeta\right]\frac{ZZ^{\mathrm{T}}}{Z^{\mathrm{T}}Z}\right) \\
&= V + \beta YZ^{\mathrm{T}} - \beta VZZ^{\mathrm{T}} - \frac{\beta^2}{2}(Y^{\mathrm{T}}Y)VZZ^{\mathrm{T}} + RZ^{\mathrm{T}},
\end{aligned}$$

where $R = -\frac{\beta^2}{2}(Z^{\mathrm{T}}Z)(2+\beta Y^{\mathrm{T}}Y)Y + \zeta\beta^2(Z^{\mathrm{T}}Z)VZ + \zeta\beta^3(Z^{\mathrm{T}}Z)^2Y$ for which

$$\begin{aligned}
\|R\|_2 &\leq \frac{\beta^2}{2}\widetilde{\eta}_p(2+\beta\nu\widetilde{\eta}_p)(\nu\widetilde{\eta}_p)^{1/2} + \zeta\beta^2\widetilde{\eta}_p^{3/2} + \zeta\beta^3\widetilde{\eta}_p^2(\nu\widetilde{\eta}_p)^{1/2} \\
&= \left[\frac{1}{2}(2+\beta\nu\widetilde{\eta}_p) + \frac{3}{8}(2+\beta\nu\widetilde{\eta}_p)^2 + \frac{3}{8}(2+\beta\nu\widetilde{\eta}_p)^2(\beta\widetilde{\eta}_p)\right]\nu^{1/2}\widetilde{\eta}_p^{3/2}\beta^2 \\
&= C_V\nu^{1/2}\widetilde{\eta}_p^{3/2}\beta^2,
\end{aligned}$$

as expected. □

**Lemma 6.3.** *Suppose that the conditions of* Theorem 4.1 *hold. Let $\tau = \|T^{(n)}\|_2$, and $C_T$ be as in (6.6c). If $n < \min\{N_{\mathrm{qb}}\{\Lambda\}, N_{\mathrm{out}}\{\mathbb{S}(\kappa)\}\}$, then the following statements hold.*

1. *$T^{(n)}$ and $T^{(n+1)}$ are well-defined.*

2. *Define* $E_T^{(n)}(V^{(n)})$ *by* $\mathrm{E}\{T^{(n+1)} - T^{(n)} \mid \mathbb{F}_n\} = \beta(\underline{\Lambda}T^{(n)} - T^{(n)}\bar{\Lambda}) + E_T^{(n)}(V^{(n)})$*. Then*
   (a) $\sup_{V\in\mathbb{S}(\kappa)}\|E_T^{(n)}(V)\|_2 \le C_T\nu^{1/2}(\widetilde{\eta}_p\beta)^2(1+\tau^2)^{3/2}$;
   (b) $\|T^{(n+1)} - T^{(n)}\|_2 \le \nu^{1/2}(\widetilde{\eta}_p\beta)(1+\tau^2) + C_T\nu^{1/2}(\widetilde{\eta}_p\beta)^2(1+\tau^2)^{3/2}$.
3. *Define* $R_\circ$ *by* $\mathrm{var}_\circ\big(T^{(n+1)} - T^{(n)} \mid \mathbb{F}_n\big) = \beta^2 H_\circ + R_\circ$*. Then*
   (a) $H_\circ = \mathrm{var}_\circ\big(\underline{Y}\bar{Y}^{\mathrm{T}}\big) \le 16\psi^4 H$*, where* $H = [\eta_{ij}]_{(d-p)\times p}$ *with* $\eta_{ij} = \lambda_{p+i}\lambda_j$ *for* $i = 1,\dots,d-p$, $j = 1,\dots,p$;
   (b) $\|R_\circ\|_2 \le (\nu\widetilde{\eta}_p\beta)^2\tau(1+\frac{11}{2}\tau+\tau^2+\frac{1}{4}\tau^3) + 4C_T\nu(\widetilde{\eta}_p\beta)^3(1+\tau^2)^{5/2} + 2C_T^2\nu(\widetilde{\eta}_p\beta)^4(1+\tau^2)^3$.

*Proof.* For readability, we will drop the superscript "$\cdot^{(n)}$", and use the superscript "$\cdot^{+}$" to replace "$\cdot^{(n+1)}$" for $V, R$, drop the superscript "$\cdot^{(n+1)}$" on $Y$, $Z$, and drop the conditional sign "$|\,\mathbb{F}_n$" in the computation of $\mathrm{E}\{\cdot\}, \mathrm{var}(\cdot), \mathrm{cov}(\cdot)$ with the understanding that they are conditional with respect to $\mathbb{F}_n$. Finally, for any expression or variable $F$, we define $\Delta F := F^+ - F$.

Consider item 1. Since $n < N_{\mathrm{out}}\{\mathbb{S}(\kappa)\}$, we have $V \in \mathbb{S}(\kappa)$ and $\tau = \|T\|_2 \le (\kappa^2-1)^{1/2}$. Thus, $\|\bar{V}^{-1}\|_2 \le \kappa$ and $T = \underline{V}\bar{V}^{-1}$ is well-defined. Recall (6.10) and the partitioning

$$Y = \begin{matrix} & 1 \\ \begin{matrix} p \\ d-p \end{matrix} & \begin{bmatrix} \bar{Y} \\ \underline{Y} \end{bmatrix} \end{matrix}, \quad R = \begin{matrix} & 1 \\ \begin{matrix} p \\ d-p \end{matrix} & \begin{bmatrix} \bar{R} \\ \underline{R} \end{bmatrix} \end{matrix}.$$

We have $\Delta\bar{V} = \beta(\bar{Y}Z^{\mathrm{T}} - (1+\frac{\beta}{2}Y^{\mathrm{T}}Y)\bar{V}ZZ^{\mathrm{T}}) + \bar{R}Z^{\mathrm{T}}$, and

$$\bar{R} = -\frac{\beta^2}{2}(Z^{\mathrm{T}}Z)(2+\beta Y^{\mathrm{T}}Y)\bar{Y} + \zeta\beta^2(Z^{\mathrm{T}}Z)\bar{V}Z + \zeta\beta^3(Z^{\mathrm{T}}Z)^2\bar{Y}.$$

Noticing $\|\bar{Y}\|_2 \le \widetilde{\eta}_p^{1/2}$, we find

$$\begin{aligned} \|\Delta\bar{V}\|_2 &\le \beta\widetilde{\eta}_p + \beta(1+\frac{\beta}{2}\nu\widetilde{\eta}_p)\widetilde{\eta}_p + C_V\widetilde{\eta}_p^2\beta^2 \\ &\le \left[2 + \frac{\beta}{2}\nu\widetilde{\eta}_p + C_V\widetilde{\eta}_p\beta\right]\widetilde{\eta}_p\beta = C_\Delta\widetilde{\eta}_p\beta, \end{aligned}$$

where $C_\Delta$ is as in (6.6b). Thus $\|\Delta\bar{V}\bar{V}^{-1}\|_2 \le \|\Delta\bar{V}\|_2\|\bar{V}^{-1}\|_2 \le C_\Delta\widetilde{\eta}_p\beta\kappa \le 1/2$ by ($\beta$-3). As a result, $\bar{V}^+$ is nonsingular, and

$$\|(\bar{V}^+)^{-1}\|_2 \le \frac{\|\bar{V}^{-1}\|_2}{1 - \|\bar{V}^{-1}\Delta\bar{V}\|_2} \le 2\|\bar{V}^{-1}\|_2.$$

In particular, $T^+ = \underline{V}^+(\bar{V}^+)^{-1}$ is well-defined. This proves item 1.

For item 2, using the Sherman-Morrison-Woodbury formula [5, p. 95], we get

$$\begin{aligned} \Delta T &= \underline{V}^+(\bar{V}^+)^{-1} - \underline{V}\bar{V}^{-1} \\ &= (\underline{V} + \Delta\underline{V})(\bar{V} + \Delta\bar{V})^{-1} - \underline{V}\bar{V}^{-1} \\ &= (\underline{V} + \Delta\underline{V})(\bar{V}^{-1} - \bar{V}^{-1}\Delta\bar{V}(\bar{V} + \Delta\bar{V})^{-1}) - \underline{V}\bar{V}^{-1} \\ &= \Delta\underline{V}\bar{V}^{-1} - \underline{V}\bar{V}^{-1}\Delta\bar{V}(\bar{V} + \Delta\bar{V})^{-1} - \Delta\underline{V}\bar{V}^{-1}\Delta\bar{V}(\bar{V} + \Delta\bar{V})^{-1} \\ &= \Delta\underline{V}\bar{V}^{-1} - \underline{V}\bar{V}^{-1}\Delta\bar{V}(\bar{V}^{-1} - \bar{V}^{-1}\Delta\bar{V}(\bar{V} + \Delta\bar{V})^{-1}) - \Delta\underline{V}\bar{V}^{-1}\Delta\bar{V}(\bar{V} + \Delta\bar{V})^{-1} \\ &= \Delta\underline{V}\bar{V}^{-1} - T\Delta\bar{V}\bar{V}^{-1} + T\Delta\bar{V}\bar{V}^{-1}\Delta\bar{V}(\bar{V}^+)^{-1} - \Delta\underline{V}\bar{V}^{-1}\Delta\bar{V}(\bar{V}^+)^{-1} \\ &= [\Delta\underline{V} - T\Delta\bar{V}][I - (\bar{V}^+)^{-1}\Delta\bar{V}]\bar{V}^{-1}. \end{aligned}$$

Write $T_L = \begin{bmatrix} -T & I \end{bmatrix}$ and $T_R = \begin{bmatrix} I \\ T \end{bmatrix}$, and then $T_L V = 0$ and $V = T_R\bar{V}$. Thus,

$$\Delta T = T_L\Delta V[I - (\bar{V}^+)^{-1}\Delta\bar{V}]V^{\mathrm{T}}T_R.$$

Since $\varDelta V$ is rank-1, $\varDelta T$ is also rank-1. By Lemma 6.2,

$$
\begin{aligned}
\varDelta T &= T_L\left[\beta Y Z^{\mathrm{T}} - \beta(1+\frac{\beta}{2}Y^{\mathrm{T}}Y)VZZ^{\mathrm{T}} + RZ^{\mathrm{T}}\right]\left[I-(\bar V^{+})^{-1}\varDelta \bar V\right]V^{\mathrm{T}}T_R\\
&= T_L\left[\beta YY^{\mathrm{T}}V + RZ^{\mathrm{T}}\right]\left[I-(\bar V^{+})^{-1}\varDelta \bar V\right]V^{\mathrm{T}}T_R\\
&= T_L(\beta YY^{\mathrm{T}}VV^{\mathrm{T}} + R_T)T_R\\
&= T_L(\beta YY^{\mathrm{T}} + R_T)T_R,
\end{aligned}
\tag{6.11}
$$

where $R_T = RZ^{\mathrm{T}}V^{\mathrm{T}} - (\beta Y + R)Z^{\mathrm{T}}(\bar V^{+})^{-1}\varDelta\bar V V^{\mathrm{T}}$. Note that

$$
T_L YY^{\mathrm{T}}T_R = \underline{Y}\bar Y^{\mathrm{T}} - T\bar Y\underline{Y}^{\mathrm{T}}T - T\bar Y\bar Y^{\mathrm{T}} + \underline{Y}\underline{Y}^{\mathrm{T}}T,
\tag{6.12}
$$

and

$$
\mathrm{E}\{\underline{Y}\bar Y^{\mathrm{T}}\} = 0,\quad \mathrm{E}\{T\bar Y\bar Y^{\mathrm{T}}\} = T\,\mathrm{E}\{\bar Y\bar Y^{\mathrm{T}}\} = T\bar\varLambda,
\tag{6.13a}
$$

$$
\mathrm{E}\{T\bar Y\underline{Y}^{\mathrm{T}}T\} = T\,\mathrm{E}\{\bar Y\underline{Y}^{\mathrm{T}}\}T = 0,\quad \mathrm{E}\{\underline{Y}\underline{Y}^{\mathrm{T}}T\} = \mathrm{E}\{\underline{Y}\underline{Y}^{\mathrm{T}}\}T = \underline{\varLambda}T.
\tag{6.13b}
$$

Thus, $\mathrm{E}\{\varDelta T\} = \beta(\underline{\varLambda}T - T\bar\varLambda) + E_T(V)$, where $E_T(V) = \mathrm{E}\{T_L R_T T_R\}$.

Since $V\in\mathbb{S}(\kappa)$, $\|T\|_2 \le (\kappa^2-1)^{1/2}$ by (6.4). Thus

$$
\begin{aligned}
\|R_T\|_2 &\le \|R\|_2\widetilde\eta_p^{1/2} + [(\nu\widetilde\eta_p)^{1/2}\beta + \|R\|_2]\widetilde\eta_p^{1/2}2(1+\|T\|_2^2)^{1/2}C_{\varDelta}\widetilde\eta_p\beta\\
&\le C_V\nu^{1/2}\widetilde\eta_p^2\beta^2 + (1+\|T\|_2^2)^{1/2}[1+C_V\widetilde\eta_p\beta]2C_{\varDelta}\nu^{1/2}\widetilde\eta_p^2\beta^2\\
&\le C_T\nu^{1/2}(\widetilde\eta_p\beta)^2(1+\|T\|_2^2)^{1/2},
\end{aligned}
\tag{6.14}
$$

where $C_T = C_V + 2C_{\varDelta}(1+C_V\widetilde\eta_p\beta)$. Therefore,

$$
\|E_T(V)\|_2 \le \mathrm{E}\{\|T_L R_T T_R\|_2\} \le (1+\|T\|_2^2)\,\mathrm{E}\{\|R_T\|_2\}.
$$

Item 2(a) holds. For item 2(b), we have

$$
\begin{aligned}
\|\varDelta T\|_2 &\le (1+\|T\|_2^2)(\beta\|YY^{\mathrm{T}}VV^{\mathrm{T}}\|_2 + \|R_T\|_2)\\
&\le \beta(\nu\widetilde\eta_p)^{1/2}\widetilde\eta_p^{1/2}(1+\|T\|_2^2) + C_T\nu^{1/2}(\widetilde\eta_p\beta)^2(1+\|T\|_2^2)^{3/2}\\
&\le \nu^{1/2}\widetilde\eta_p\beta(1+\|T\|_2^2) + C_T\nu^{1/2}(\widetilde\eta_p\beta)^2(1+\|T\|_2^2)^{3/2}.
\end{aligned}
$$

Now we turn to item 3. We have

$$
\mathrm{var}_\circ(\varDelta T) = \mathrm{var}_\circ\big(T_L(\beta YY^{\mathrm{T}} + R_T)T_R\big) = \beta^2\,\mathrm{var}_\circ\big(T_LYY^{\mathrm{T}}T_R\big) + 2\beta R_{\circ,1} + R_{\circ,2},
\tag{6.15}
$$

where $R_{\circ,1} = \mathrm{cov}_\circ\big(T_LYY^{\mathrm{T}}T_R, T_LR_TT_R\big)$, and $R_{\circ,2} = \mathrm{var}_\circ(T_LR_TT_R)$. By (6.12),

$$
\mathrm{var}_\circ\big(T_LYY^{\mathrm{T}}T_R\big) = \mathrm{var}_\circ\big(\underline{Y}\bar Y^{\mathrm{T}}\big) + R_{\circ,0},
\tag{6.16}
$$

where

$$
\begin{aligned}
R_{\circ,0} = {}& \mathrm{var}_\circ\big(T\bar Y\underline{Y}^{\mathrm{T}}T\big) + \mathrm{var}_\circ\big(T\bar Y\bar Y^{\mathrm{T}}\big) + \mathrm{var}_\circ\big(\underline{Y}\underline{Y}^{\mathrm{T}}T\big)\\
&- 2\,\mathrm{cov}_\circ\big(\underline{Y}\bar Y^{\mathrm{T}}, T\bar Y\underline{Y}^{\mathrm{T}}T\big) - 2\,\mathrm{cov}_\circ\big(\underline{Y}\bar Y^{\mathrm{T}}, T\bar Y\bar Y^{\mathrm{T}}\big) + 2\,\mathrm{cov}_\circ\big(\underline{Y}\bar Y^{\mathrm{T}}, \underline{Y}\underline{Y}^{\mathrm{T}}T\big)\\
&+ 2\,\mathrm{cov}_\circ\big(T\bar Y\underline{Y}^{\mathrm{T}}T, T\bar Y\bar Y^{\mathrm{T}}\big) - 2\,\mathrm{cov}_\circ\big(T\bar Y\underline{Y}^{\mathrm{T}}T, \underline{Y}\underline{Y}^{\mathrm{T}}T\big) - 2\,\mathrm{cov}_\circ\big(T\bar Y\bar Y^{\mathrm{T}}, \underline{Y}\underline{Y}^{\mathrm{T}}T\big).
\end{aligned}
$$

Examine (6.15) and (6.16) together to get $H_\circ = \mathrm{var}_\circ\big(\underline{Y}\bar Y^{\mathrm{T}}\big)$ and $R_\circ = \beta^2R_{\circ,0} + 2\beta R_{\circ,1} + R_{\circ,2}$. We note

$$
\begin{aligned}
Y_j &= e_j^{\mathrm{T}}Y = e_j^{\mathrm{T}}\varLambda^{1/2}\varLambda^{-1/2}Y = \lambda_j^{1/2}e_j^{\mathrm{T}}\varLambda^{-1/2}Y,\\
e_i^{\mathrm{T}}\,\mathrm{var}_\circ\big(\underline{Y}\bar Y^{\mathrm{T}}\big)\,e_j &= \mathrm{var}(e_i^{\mathrm{T}}\underline{Y}\bar Y^{\mathrm{T}}e_j) = \mathrm{var}(Y_{p+i}Y_j) = \mathrm{E}\{Y_{p+i}^2Y_j^2\}.
\end{aligned}
$$

By [22, (5.11)],

$$\mathrm{E}\{Y_j^4\} = \lambda_j^2\,\mathrm{E}\Big\{(e_j^{\mathrm{T}}\mathit{\Lambda}^{-1/2}Y)^4\Big\} \le 16\lambda_j^2\|e_j^{\mathrm{T}}\mathit{\Lambda}^{-1/2}Y\|_{\psi_2}^4 \le 16\lambda_j^2\|\mathit{\Lambda}^{-1/2}Y\|_{\psi_2}^4 = 16\lambda_j^2\psi^4.$$

Therefore

$$e_i^{\mathrm{T}}\,\mathrm{var}_\circ\big(\underline{Y}\bar{Y}^{\mathrm{T}}\big)\,e_j \le [\mathrm{E}\{Y_{p+i}^4\}\,\mathrm{E}\{Y_j^4\}]^{1/2} \le 16\lambda_{p+i}\lambda_j\psi^4,$$

i.e., $H_\circ = \mathrm{var}_\circ\big(\underline{Y}\bar{Y}^{\mathrm{T}}\big) \le 16\psi^4 H$. This proves item 3(a). To show item 3(b), first we bound the entrywise variance and covariance. For any matrices $A_1, A_2$, by Schur's inequality (which was generalized to all unitarily invariant norm in [7, Theorem 3.1]),

$$\|A_1\circ A_2\|_2 \le \|A_1\|_2\|A_2\|_2, \tag{6.17}$$

we have

$$\begin{aligned}
\|\mathrm{cov}_\circ(A_1,A_2)\|_2 &= \|\mathrm{E}\{A_1\circ A_2\} - \mathrm{E}\{A_1\}\circ\mathrm{E}\{A_2\}\|_2\\
&\le \mathrm{E}\{\|A_1\circ A_2\|_2\} + \|\mathrm{E}\{A_1\}\circ\mathrm{E}\{A_2\}\|_2\\
&\le \mathrm{E}\{\|A_1\|_2\|A_2\|_2\} + \|\mathrm{E}\{A_1\}\|_2\|\mathrm{E}\{A_2\}\|_2, && (6.18a)\\
\|\mathrm{var}_\circ(A_1)\|_2 &\le \mathrm{E}\{\|A_1\|_2^2\} + \|\mathrm{E}\{A_1\}\|_2^2. && (6.18b)
\end{aligned}$$

Apply (6.18) to $R_{\circ,1}$ and $R_{\circ,2}$ to get

$$\|R_{\circ,1}\|_2 \le 2C_T\nu\widetilde{\eta}_p^3\beta^2(1+\|T\|_2^2)^{5/2}, \quad \|R_{\circ,2}\|_2 \le 2C_T^2\nu(\widetilde{\eta}_p\beta)^4(1+\|T\|_2^2)^3, \tag{6.19}$$

upon using

$$\begin{aligned}
\|T_LYY^{\mathrm{T}}T_R\|_2 = \|T_LYY^{\mathrm{T}}VV^{\mathrm{T}}T_R\|_2 &\le \nu^{1/2}\widetilde{\eta}_p(1+\|T\|_2^2),\\
\|T_LR_TT_R\|_2 &\le C_T\nu^{1/2}(\widetilde{\eta}_p\beta)^2(1+\|T\|_2^2)^{3/2}.
\end{aligned}$$

For $R_{\circ,0}$, by (6.13), we have

$$\begin{aligned}
\|\mathrm{cov}_\circ\big(\underline{Y}\bar{Y}^{\mathrm{T}}, T\bar{Y}\underline{Y}^{\mathrm{T}}T\big)\|_2 &\le \mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2^2\}\|T\|_2^2,\\
\|\mathrm{cov}_\circ\big(\underline{Y}\bar{Y}^{\mathrm{T}}, T\bar{Y}\bar{Y}^{\mathrm{T}}\big)\|_2 &\le \mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2\|\bar{Y}\bar{Y}^{\mathrm{T}}\|_2\}\|T\|_2,\\
\|\mathrm{cov}_\circ\big(\underline{Y}\bar{Y}^{\mathrm{T}}, \underline{Y}\underline{Y}^{\mathrm{T}}T\big)\|_2 &\le \mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2\|\underline{Y}\underline{Y}^{\mathrm{T}}\|_2\}\|T\|_2,\\
\|\mathrm{cov}_\circ\big(T\bar{Y}\underline{Y}^{\mathrm{T}}T, T\bar{Y}\bar{Y}^{\mathrm{T}}\big)\|_2 &\le \mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2\|\bar{Y}\bar{Y}^{\mathrm{T}}\|_2\}\|T\|_2^3,\\
\|\mathrm{cov}_\circ\big(T\bar{Y}\underline{Y}^{\mathrm{T}}T, \underline{Y}\underline{Y}^{\mathrm{T}}T\big)\|_2 &\le \mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2\|\underline{Y}\underline{Y}^{\mathrm{T}}\|_2\}\|T\|_2^3,\\
\|\mathrm{var}_\circ\big(T\bar{Y}\underline{Y}^{\mathrm{T}}T\big)\|_2 &\le \mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2^2\}\|T\|_2^4,\\
\|\mathrm{var}_\circ\big(T\bar{Y}\bar{Y}^{\mathrm{T}}\big)\|_2 &\le \mathrm{E}\{\|\bar{Y}\bar{Y}^{\mathrm{T}}\|_2^2\}\|T\|_2^2 + \|T\bar{\mathit{\Lambda}}\|_2^2,\\
\|\mathrm{var}_\circ\big(\underline{Y}\underline{Y}^{\mathrm{T}}T\big)\|_2 &\le \mathrm{E}\{\|\underline{Y}\underline{Y}^{\mathrm{T}}\|_2^2\}\|T\|_2^2 + \|\underline{\mathit{\Lambda}}T\|_2^2,\\
\|\mathrm{cov}_\circ\big(T\bar{Y}\bar{Y}^{\mathrm{T}}, \underline{Y}\underline{Y}^{\mathrm{T}}T\big)\|_2 &\le \mathrm{E}\{\|\bar{Y}\bar{Y}^{\mathrm{T}}\|_2\|\underline{Y}\underline{Y}^{\mathrm{T}}\|_2\}\|T\|_2^2 + \|T\bar{\mathit{\Lambda}}\|_2\|\underline{\mathit{\Lambda}}T\|_2.
\end{aligned}$$

Since

$$\begin{aligned}
\|\bar{Y}\bar{Y}^{\mathrm{T}}\|_2 + \|\underline{Y}\underline{Y}^{\mathrm{T}}\|_2 &= \bar{Y}^{\mathrm{T}}\bar{Y} + \underline{Y}^{\mathrm{T}}\underline{Y} = Y^{\mathrm{T}}Y \le \nu\widetilde{\eta}_p,\\
\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2 = (\bar{Y}^{\mathrm{T}}\bar{Y})^{1/2}(\underline{Y}^{\mathrm{T}}\underline{Y})^{1/2} &\le \frac{\bar{Y}^{\mathrm{T}}\bar{Y} + \underline{Y}^{\mathrm{T}}\underline{Y}}{2} \le \frac{\nu\widetilde{\eta}_p}{2},
\end{aligned}$$

we have

$$\begin{aligned}
\|R_{\circ,0}\|_2 &\le \mathrm{E}\{2\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2^2 + (\|\bar{Y}\bar{Y}^{\mathrm{T}}\|_2 + \|\underline{Y}\underline{Y}^{\mathrm{T}}\|_2)^2\}\|T\|_2^2 + (\|T\bar{\mathit{\Lambda}}\|_2 + \|\underline{\mathit{\Lambda}}T\|_2)^2\\
&\quad + 2\,\mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2(\|\bar{Y}\bar{Y}^{\mathrm{T}}\|_2 + \|\underline{Y}\underline{Y}^{\mathrm{T}}\|_2)\}\,(\|T\|_2 + \|T\|_2^3) + \mathrm{E}\{\|\underline{Y}\bar{Y}^{\mathrm{T}}\|_2^2\}\|T\|_2^4\\
&\le (\nu\widetilde{\eta}_p)^2\|T\|_2 + \Big[\frac{3}{2}(\nu\widetilde{\eta}_p)^2 + (\lambda_1+\lambda_{p+1})^2\Big]\|T\|_2^2 + (\nu\widetilde{\eta}_p)^2\|T\|_2^3 + \frac{1}{4}(\nu\widetilde{\eta}_p)^2\|T\|_2^4\\
&\le (\nu\widetilde{\eta}_p)^2\|T\|_2\Big(1 + \frac{11}{2}\|T\|_2 + \|T\|_2^2 + \frac{1}{4}\|T\|_2^3\Big).
\end{aligned} \tag{6.20}$$

Finally collecting (6.19) and (6.20) yields the desired bound on $R_\circ = \beta^2R_{\circ,0} + 2\beta R_{\circ,1} + R_{\circ,2}$. □

## 6.3 Quasi-Power Iteration Process

Define $D^{(n+1)} = T^{(n+1)} - \mathrm{E}\{T^{(n+1)} \mid \mathbb{F}_n\}$. It can be seen that

$$T^{(n)} - \mathrm{E}\Big\{T^{(n)} \,\Big|\, \mathbb{F}_n\Big\} = 0, \quad \mathrm{E}\Big\{D^{(n+1)} \,\Big|\, \mathbb{F}_n\Big\} = 0,$$
$$\mathrm{E}\Big\{D^{(n+1)} \circ D^{(n+1)} \,\Big|\, \mathbb{F}_n\Big\} = \mathrm{var}_\circ\Big(T^{(n+1)} - T^{(n)} \mid \mathbb{F}_n\Big).$$

By item 2 of Lemma 6.3, we have

$$\begin{aligned} T^{(n+1)} &= D^{(n+1)} + T^{(n)} + \mathrm{E}\Big\{T^{(n+1)} - T^{(n)} \,\Big|\, \mathbb{F}_n\Big\} \\ &= D^{(n+1)} + T^{(n)} + \beta(\Lambda T^{(n)} - T^{(n)}\bar{\Lambda}) + E_T^{(n)}(V^{(n)}) \\ &= \mathcal{L}T^{(n)} + D^{(n+1)} + E_T^{(n)}(V^{(n)}), \end{aligned}$$

where $\mathcal{L}\colon T \mapsto T + \beta\Lambda T - \beta T\bar{\Lambda}$ is a bounded linear operator. It can be verified that $\mathcal{L}T = L \circ T$, the Hadamard product of $L$ and $T$, where $L = [\lambda_{ij}]_{(d-p)\times p}$ with $\lambda_{ij} = 1 + \beta\lambda_{p+i} - \beta\lambda_j$. Moreover, it can be shown that[3] $\|\mathcal{L}\|_{\mathrm{ui}} = \rho(\mathcal{L}) = 1 - \beta\gamma$, where $\|\mathcal{L}\|_{\mathrm{ui}} = \sup_{\|T\|_{\mathrm{ui}}=1}\|\mathcal{L}T\|_{\mathrm{ui}}$ is an operator norm induced by the matrix norm $\|\cdot\|_{\mathrm{ui}}$. Recursively,

$$T^{(n)} = \mathcal{L}^n T^{(0)} + \sum_{s=1}^{n} \mathcal{L}^{n-s} D^{(s)} + \sum_{s=1}^{n} \mathcal{L}^{n-s} E_T^{(s-1)}(V^{(s-1)}) =: J_1 + J_2 + J_3. \tag{6.21}$$

Define events $\mathbb{M}_n(\chi)$, $\mathbb{T}_n(\chi)$, and $\mathbb{Q}_n$ as

$$\mathbb{M}_n(\chi) = \Big\{\|T^{(n)} - \mathcal{L}^n T^{(0)}\|_2 \le \frac{1}{2}(\kappa^2\beta^{2\chi-1} - 1)^{1/2}\beta^{\chi-3\varepsilon}\Big\}, \tag{6.22}$$

$$\mathbb{T}_n(\chi) = \Big\{\|T^{(n)}\|_2 \le (\kappa^2\beta^{2\chi-1} - 1)^{1/2}\beta^{\chi-3\varepsilon}\Big\}, \quad \mathbb{Q}_n = \{n < N_{\mathrm{qb}}\{\Lambda\}\}. \tag{6.23}$$

**Lemma 6.4.** *Suppose that the conditions of* Theorem 4.1 *hold. If* $n < \min\{N_{\mathrm{qb}}\{\Lambda\}, N_{\mathrm{out}}\{\mathbb{S}(\kappa\beta^\chi)\}\}$ *and* $V^{(0)} \in \mathbb{S}(\kappa\beta^\chi)$, *then for any* $\chi \in (5\varepsilon - 1/2, 0]$ *and* $\kappa > \sqrt{2}$, *we have*

$$\mathrm{P}\{\mathbb{M}_n(\chi + 1/2)\} \ge 1 - 2d\exp(-C_\kappa\gamma\kappa^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon}), \tag{6.24}$$

*where* $C_\kappa$ *is as in* (6.6d).

*Proof.* Since $\kappa > \sqrt{2}$, we have $\kappa^2\beta^{2\chi} > 2$ and $\kappa\beta^\chi < [2(\kappa^2\beta^{2\chi} - 1)]^{1/2}$. Thus, by ($\beta$-4),

$$4C_T\kappa^3\widetilde{\eta}_p^2\gamma^{-1}\beta^{1+3\chi}(\kappa^2\beta^{2\chi} - 1)^{-1/2}\beta^{-1/2-\chi} \le 4\sqrt{2}C_T\kappa^2\widetilde{\eta}_p^2\gamma^{-1}\beta^{1/2+\chi} \le 1.$$

For any $n < \min\{N_{\mathrm{qb}}\{\Lambda\}, N_{\mathrm{out}}\{\mathbb{S}(\kappa\beta^\chi)\}\}$, $V^{(n)} \in \mathbb{S}(\kappa\beta^\chi)$ and thus $\|T^{(n)}\|_2 \le \sqrt{\kappa^2\beta^{2\chi} - 1}$ by (6.4). Therefore, by item 2(b) of Lemma 6.3, we have

$$\begin{aligned} \|D^{(n+1)}\|_2 &= \Big\|T^{(n+1)} - T^{(n)} - \mathrm{E}\Big\{T^{(n+1)} - T^{(n)} \,\Big|\, \mathbb{F}_n\Big\}\Big\|_2 \\ &\le \|T^{(n+1)} - T^{(n)}\|_2 + \mathrm{E}\Big\{\|T^{(n+1)} - T^{(n)}\|_2 \,\Big|\, \mathbb{F}_n\Big\} \\ &\le 2\nu^{1/2}\widetilde{\eta}_p\beta(1 + \|T^{(n)}\|_2^2)[1 + C_T\widetilde{\eta}_p\beta(1 + \|T^{(n)}\|_2^2)^{1/2}] \\ &\le 2\kappa^2\nu^{1/2}\widetilde{\eta}_p\beta^{1+2\chi}[1 + C_T\kappa\widetilde{\eta}_p\beta^{1+\chi}]. \end{aligned} \tag{6.25}$$

---

[3] Since $\lambda(\mathcal{L}) = \{\lambda_{ij} : i = 1, \dots, d-p,\ j = 1, \dots, p\}$, we have the spectral radius $\rho(\mathcal{L}) = 1 - \beta(\lambda_p - \lambda_{p+1})$. Thus for any $T$,

$$\begin{aligned} \|\mathcal{L}T\|_{\mathrm{ui}} &= \|T(I - \beta\bar{\Lambda}) + \beta\Lambda T\|_{\mathrm{ui}} \\ &\le \|I - \beta\bar{\Lambda}\|_2\|T\|_{\mathrm{ui}} + \|\beta\Lambda\|_2\|T\|_{\mathrm{ui}} \\ &= (1 - \beta\lambda_p + \beta\lambda_{p+1})\|T\|_{\mathrm{ui}} = \rho(\mathcal{L})\|T\|_{\mathrm{ui}}, \end{aligned}$$

which means $\|\mathcal{L}\|_{\mathrm{ui}} \le \rho(\mathcal{L})$. This ensures $\|\mathcal{L}\|_{\mathrm{ui}} = \rho(\mathcal{L})$.

For any $n < \min\{N_{\text{qb}}\{\Lambda\}, N_{\text{out}}\{\mathbb{S}(\kappa\beta^{\chi})\}\}$,

$$\begin{aligned}
\|J_3\|_2 &\leq \sum_{s=1}^{n} \|\mathcal{L}\|_2^{n-s} \|E_T^{(s-1)}(V^{(s-1)})\|_2 \\
&\leq C_T \nu^{1/2} \kappa^3 \widetilde{\eta}_p^2 \beta^{2+3\chi} \sum_{s=1}^{n} (1-\beta\gamma)^{n-s} \\
&\leq \frac{C_T \nu^{1/2} \kappa^3 \widetilde{\eta}_p^2 \beta^{2+3\chi}}{\beta\gamma} = C_T \nu^{1/2} \kappa^3 \widetilde{\eta}_p^2 \gamma^{-1} \beta^{1+3\chi} \\
&\leq \frac{1}{4} \nu^{1/2} (\kappa^2 \beta^{2\chi} - 1)^{1/2} \beta^{1/2+\chi}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|J_2\|_2 &\leq \sum_{s=1}^{n} \|\mathcal{L}\|_2^{n-s} \|D^{(s)}\|_2 \\
&\leq \frac{2\kappa^2 \nu^{1/2} \widetilde{\eta}_p \beta^{2\chi} (1 + C_T \kappa \widetilde{\eta}_p \beta^{1+\chi})}{\gamma} \\
&\leq \frac{2\kappa^2 \nu^{1/2} \widetilde{\eta}_p \beta^{2\chi}}{\gamma} + \frac{1}{2} \nu^{1/2} (\kappa^2 \beta^{2\chi} - 1)^{1/2} \beta^{1/2+\chi}.
\end{aligned}$$

Also, $\|J_1\|_2 \leq \|\mathcal{L}\|_2^n \|T^{(0)}\|_2 \leq \|T^{(0)}\|_2 \leq \nu^{1/2}(\kappa^2\beta^{2\chi} - 1)^{1/2}$. For fixed $n > 0$ and $\beta > 0$,

$$\left\{ M_0^{(n)} := \mathcal{L}^n T^{(0)}, M_t^{(n)} := \mathcal{L}^n T^{(0)} + \sum_{s=1}^{\min\{t, N_{\text{out}}\{\mathbb{S}(\kappa)\}-1\}} \mathcal{L}^{n-s} D^{(s)} : t = 1, \ldots, n \right\}$$

forms a martingale with respect to $\mathbb{F}_t$, because

$$\mathrm{E}\left\{ \|M_t^{(n)}\|_2 \right\} \leq \|J_1\|_2 + \|J_2\|_2 < +\infty,$$

and

$$\mathrm{E}\left\{ M_{t+1}^{(n)} - M_t^{(n)} \,\middle|\, \mathbb{F}_t \right\} = \mathrm{E}\left\{ \mathcal{L}^{n-t-1} D^{(t+1)} \,\middle|\, \mathbb{F}_t \right\} = \mathcal{L}^{n-t-1}\, \mathrm{E}\left\{ D^{(t+1)} \,\middle|\, \mathbb{F}_t \right\} = 0.$$

Use the matrix version of Azuma's inequality [20, Section 7.2] to get, for any $\alpha > 0$,

$$\mathrm{P}\left\{ \|M_n^{(n)} - M_0^{(n)}\|_2 \geq \alpha \right\} \leq 2d \exp(-\frac{\alpha^2}{2\sigma^2}),$$

where

$$\begin{aligned}
\sigma^2 &= \sum_{s=1}^{\min\{n, N_{\text{out}}\{\mathbb{S}(\kappa)\}-1\}} \|\mathcal{L}^{n-s} D^{(s)}\|_2^2 \\
&\leq [2\kappa^2 \nu^{1/2} \widetilde{\eta}_p \beta^{1+2\chi} (1 + C_T \kappa \widetilde{\eta}_p \beta^{1+\chi})]^2 \sum_{s=1}^{\min\{n, N_{\text{out}}\{\mathbb{S}(\kappa)\}-1\}} (1-\beta\gamma)^{2(n-s)} \\
&\leq \frac{4\kappa^4 \nu \widetilde{\eta}_p^2 \beta^{2+4\chi} (1 + C_T \kappa \widetilde{\eta}_p \beta^{1+\chi})^2}{\beta\gamma[2 - \beta\gamma]} \\
&\leq \frac{4\kappa^4 \nu \widetilde{\eta}_p^2 \gamma^{-1} \beta^{1+4\chi} (1 + \frac{C_T}{2C_\Delta})^2}{3 - \sqrt{2}} \qquad \text{by } (\beta\text{-3}),\ \widetilde{\eta}_p \beta^{1/2} \leq \frac{1}{2\kappa C_\Delta} \\
&= C_\sigma \kappa^4 \nu \gamma^{-1} \widetilde{\eta}_p^2 \beta^{1+4\chi},
\end{aligned}$$

and $C_\sigma = \frac{(C_T + 2C_\Delta)^2}{(3-\sqrt{2})C_\Delta^2}$. Thus, noticing $J_2 = M_n^{(n)} - M_0^{(n)}$ for $n \leq N_{\text{out}}\{\mathbb{S}(\kappa)\} - 1$, we have

$$\mathrm{P}\{\|J_2\|_2 \geq \alpha\} \leq 2d \exp\left( -\frac{\alpha^2}{2C_\sigma \kappa^4 \nu \gamma^{-1} \widetilde{\eta}_p^2 \beta^{1+4\chi}} \right).$$

Choosing $\alpha = \frac{1}{4}(\kappa^2\beta^{2\chi} - 1)^{1/2}\beta^{\chi+1/2-3\varepsilon}$ and noticing $\|J_3\|_2 \le \frac{1}{4}(\kappa^2\beta^{2\chi} - 1)^{1/2}\beta^{\chi+1/2-3\varepsilon}$ and $T^{(n)} - \mathcal{L}^n T^{(0)} = J_2 + J_3$, we have

$$\begin{aligned}
\mathrm{P}\{\mathbb{M}_n(\chi+1/2)^{\mathrm{c}}\} &= \mathrm{P}\bigg\{\|T^{(n)} - \mathcal{L}^n T^{(0)}\|_2 \ge \frac{1}{2}(\kappa^2\beta^{2\chi} - 1)^{1/2}\beta^{\chi+1/2-3\varepsilon}\bigg\} \\
&\le \mathrm{P}\bigg\{\|J_2\|_2 \ge \frac{1}{4}(\kappa^2\beta^{2\chi} - 1)^{1/2}\beta^{\chi+1/2-3\varepsilon}\bigg\} \\
&\le 2d\exp\left(-\frac{\kappa^2\beta^{2\chi} - 1}{32C_\sigma\kappa^4\nu\gamma^{-1}\widetilde{\eta}_p^2\beta^{2\chi}}\beta^{-6\varepsilon}\right) \\
&\le 2d\exp\left(-\frac{\kappa^2\beta^{2\chi}}{64C_\sigma\kappa^4\nu\gamma^{-1}\widetilde{\eta}_p^2\beta^{2\chi}}\beta^{-6\varepsilon}\right) \\
&= 2d\exp(-C_\kappa\gamma\kappa^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon}),
\end{aligned}$$

where $C_\kappa = \frac{1}{64C_\sigma}$ which is the same as in (6.6d). □

**Lemma 6.5.** *Suppose that the conditions of* Theorem 4.1 *hold. If*

$$N_{2^{-m}(1-6\varepsilon)} < \min\{N_{\mathrm{qb}}\{\Lambda\}, N_{\mathrm{out}}\{\mathbb{S}(\kappa\beta^\chi)\}\}$$

*and* $V^{(0)} \in \mathbb{S}(\beta^{(1-2^{1-m})(3\varepsilon-1/2)}\kappa_m/2)$ *with* $m \ge 2$*, then for* $\kappa_m > \sqrt{2}$

$$\mathrm{P}\{\mathbb{H}_m\} \ge 1 - 2dN_{2^{-m}(1-6\varepsilon)}\exp(-C_\kappa\gamma\kappa_m^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon}),$$

*where* $\mathbb{H}_m = \Big\{N_{\mathrm{in}}\Big\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\kappa_m)\Big\} \le N_{2^{-m}(1-6\varepsilon)}\Big\}$.

*Proof.* By the definition of the event $\mathbb{T}_n$,

$$\mathbb{T}_n(2^{-m}[1-6\varepsilon]+3\varepsilon) = \Big\{\|T^{(n)}\|_2 \le (\kappa_m^2 - \beta^{(1-2^{1-m})(1-6\varepsilon)})^{1/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\Big\}.$$

For $n \ge N_{2^{-m}(1-6\varepsilon)}$ and $V^{(0)} \in \mathbb{S}(\beta^{(1-2^{1-m})(3\varepsilon-1/2)}\kappa_m/2)$,

$$\mathbb{M}_n(2^{-m}(1-6\varepsilon)+3\varepsilon) \subset \mathbb{T}_n(2^{-m}(1-6\varepsilon)+3\varepsilon)$$

because

$$\begin{aligned}
\|T^{(n)}\|_2 &\le \|T^{(n)} - \mathcal{L}^n T^{(0)}\|_2 + \|\mathcal{L}\|_2^n\|T^{(0)}\|_2 \\
&\le \frac{1}{2}\Big(\kappa_m^2 - \beta^{(1-2^{1-m})(1-6\varepsilon)}\Big)^{1/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)} \\
&\quad + \beta^{2^{-m}(1-6\varepsilon)}\Big(\frac{\kappa_m^2}{4} - \beta^{(1-2^{1-m})(1-6\varepsilon)}\Big)^{1/2}\beta^{(1-2^{1-m})(3\varepsilon-1/2)} \\
&\le \Big(\kappa_m^2 - \beta^{(1-2^{1-m})(1-6\varepsilon)}\Big)^{1/2}\beta^{(1-2^{1-m})(3\varepsilon-1/2)}.
\end{aligned}$$

Therefore, noticing

$$\begin{aligned}
\Big(\kappa_m^2 - \beta^{(1-2^{1-m})(1-6\varepsilon)}\Big)^{1/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)} &= \Big(\beta^{(1-2^{2-m})(6\varepsilon-1)}\kappa_m^2 - \beta^{2^{1-m}(1-6\varepsilon)}\Big)^{1/2} \\
&\le \Big(\frac{3}{2}\beta^{(1-2^{2-m})(6\varepsilon-1)}\kappa_m^2 - 1\Big)^{1/2},
\end{aligned}$$

we get

$$\mathbb{M}_{N_{2^{-m}(1-6\varepsilon)}}(2^{-m}(1-6\varepsilon)+3\varepsilon) \subset \Big\{N_{\mathrm{in}}\Big\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\kappa_m)\Big\} \le N_{2^{-m}(1-6\varepsilon)}\Big\} =: \mathbb{H}_m.$$

Since

$$\bigcap_{n\le \min\left\{N_{2^{-m}(1-6\varepsilon)},N_{\text{in}}\left\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\kappa_m)\right\}-1\right\}} \mathbb{M}_n(2^{-m}(1-6\varepsilon)+3\varepsilon)\cap \mathbb{H}_m^{\mathrm{c}}$$
$$\subset \bigcap_{n\le N_{2^{-m}(1-6\varepsilon)}} \mathbb{M}_n(2^{-m}(1-6\varepsilon)+3\varepsilon)\subset \mathbb{M}_{N_{2^{-m}(1-6\varepsilon)}}(2^{-m}(1-6\varepsilon)+3\varepsilon),$$

we have

$$\bigcap_{n\le \min\left\{N_{2^{-m}(1-6\varepsilon)},N_{\text{in}}\left\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\kappa_m)\right\}-1\right\}} \mathbb{M}_n(2^{-m}(1-6\varepsilon)+3\varepsilon)\subset \mathbb{H}_m.$$

By Lemma 6.4 with $\chi=2^{-m}(1-6\varepsilon)+3\varepsilon-\frac{1}{2}=2^{-m}(1-2^{m-1})(1-6\varepsilon)$, we get

$$\begin{aligned}
\mathrm{P}\{\mathbb{H}_m^{\mathrm{c}}\}&\le \mathrm{P}\left\{\bigcup_{n\le \min\left\{N_{2^{-m}(1-6\varepsilon)},N_{\text{in}}\left\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\kappa_m)\right\}-1\right\}} \mathbb{M}_n(2^{-m}(1-6\varepsilon)+3\varepsilon)^{\mathrm{c}}\right\}\\
&\le \min\left\{N_{2^{-m}(1-6\varepsilon)},N_{\text{in}}\left\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\kappa_m)\right\}-1\right\}\\
&\quad\times 2d\exp(-C_\kappa\gamma\kappa_m^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon})\\
&\le 2dN_{2^{-m}(1-6\varepsilon)}\exp(-C_\kappa\gamma\kappa_m^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon}),
\end{aligned}$$

as expected. □

**Lemma 6.6.** *Suppose that the conditions of* Theorem 4.1 *hold. If $V^{(0)}\in\mathbb{S}(\kappa/2)$ with $\kappa>2\sqrt{2}$, $K>N_{1-6\varepsilon}$, then there exists a high-probability event $\mathbb{H}_1\cap\mathbb{Q}_K=\bigcap_{n\in[N_{1/2-3\varepsilon},K]}\mathbb{T}_n(1/2)\cap\mathbb{Q}_K$ satisfying*

$$\mathrm{P}\{\mathbb{H}_1\cap\mathbb{Q}_K\}\ge 1-2dK\exp(-C_\kappa\gamma\kappa^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon})-K(\mathrm{e}d+p+1)\exp\left(-C_\psi\min\{\psi^{-1},\psi^{-2}\}\beta^{-2\varepsilon}\right),$$

*such that for any $n\in[N_{1-6\varepsilon},K]$,*

$$\mathrm{E}\left\{T^{(n)}\circ T^{(n)};\,\mathbb{H}_1\cap\mathbb{Q}_K\right\}\le \mathcal{L}^{2n}T^{(0)}\circ T^{(0)}+2\beta^2[I-\mathcal{L}^2]^{-1}[I-\mathcal{L}^{2n}]H_\circ+R_E,$$

*where $\|R_E\|_2\le C_\circ\kappa^4\gamma^{-1}\nu^2\widetilde{\eta}_p^2\beta^{3/2-3\varepsilon}$, $H_\circ=\operatorname{var}_\circ\big(\underline{Y}\bar{Y}^{\mathrm{T}}\big)\le 16\psi^4H$ is as in* item 3(a) *of* Lemma 6.3, *and $C_\circ$ is as in* (6.6f).

*Proof.* First we estimate the probability of the event $\mathbb{H}_1$. We know $\mathbb{T}_n(1/2)\subset\left\{\|T^{(n)}\|_2\le(\kappa^2-1)^{1/2}\right\}$. If $K\ge N_{\text{out}}\{\mathbb{S}(\kappa)\}$, then there exists some $n\le K$, such that $V^{(n)}\notin\mathbb{S}(\kappa)$, i.e., $\|T^{(n)}\|_2>(\kappa^2-1)^{1/2}$ by (6.4). Thus,

$$\{K\ge N_{\text{out}}\{\mathbb{S}(\kappa)\}\}\subset\bigcup_{n\le K}\left\{\|T^{(n)}\|_2>(\kappa^2-1)^{1/2}\right\}\subset\bigcup_{n\le K}\mathbb{T}_n(1/2)^{\mathrm{c}}.$$

On the other hand, for $n\ge N_{1/2-3\varepsilon}$ and $V^{(0)}\in\mathbb{S}(\kappa/2)$, $\mathbb{M}_n(1/2)\subset\mathbb{T}_n(1/2)$ because

$$\begin{aligned}
\|T^{(n)}\|_2&\le\|T^{(n)}-\mathcal{L}^nT^{(0)}\|_2+\|\mathcal{L}\|_2^n\|T^{(0)}\|_2\\
&\le\frac{1}{2}(\kappa^2-1)^{1/2}\beta^{1/2-3\varepsilon}+\beta^{1/2-3\varepsilon}\left(\frac{\kappa^2}{4}-1\right)^{1/2}\\
&\le(\kappa^2-1)^{1/2}\beta^{1/2-3\varepsilon}.
\end{aligned}\tag{6.26}$$

Therefore,

$$\bigcap_{n\in[N_{1/2-3\varepsilon},K]}\mathbb{M}_n(1/2)\subset\bigcap_{n\in[N_{1/2-3\varepsilon},K]}\mathbb{T}_n(1/2)\subset\{K\le N_{\text{out}}\{\mathbb{S}(\kappa)\}-1\},$$

and so

$$
\begin{aligned}
\bigcap_{n \le \min\{K, N_{\mathrm{out}}\{\mathbb{S}(\kappa)\}-1\}} \mathbb{M}_n(1/2) &\subset \bigcap_{n \in [N_{1/2-3\varepsilon}, \min\{K, N_{\mathrm{out}}\{\mathbb{S}(\kappa)\}-1\}]} \mathbb{M}_n(1/2) \\
&= \bigcap_{n \in [N_{1/2-3\varepsilon}, K]} \mathbb{M}_n(1/2) \\
&\subset \bigcap_{n \in [N_{1/2-3\varepsilon}, K]} \mathbb{T}_n(1/2) \\
&=: \mathbb{H}_1.
\end{aligned}
$$

By Lemma 6.4 with $\chi = 0$, we have

$$
\begin{aligned}
&\mathrm{P}\left\{ \bigcup_{n \le \min\{K, N_{\mathrm{out}}\{\mathbb{S}(\kappa)\}-1\}} \mathbb{M}_n(1/2)^{\mathrm{c}} \cap \mathbb{Q}_K \right\} \\
&\qquad \le \min\{K, N_{\mathrm{out}}\{\mathbb{S}(\kappa)\}-1\} \cdot 2d \exp(-C_\kappa \gamma \kappa^{-2} \nu^{-1} \eta_p^{-2} \beta^{-2\varepsilon}) \\
&\qquad = 2dK \exp(-C_\kappa \gamma \kappa^{-2} \nu^{-1} \eta_p^{-2} \beta^{-2\varepsilon}).
\end{aligned}
$$

Thus, by Lemma 6.1,

$$
\begin{aligned}
\mathrm{P}\{(\mathbb{H}_1 \cap \mathbb{Q}_K)^{\mathrm{c}}\} &= \mathrm{P}\{\mathbb{H}_1^{\mathrm{c}} \cup \mathbb{Q}_K^{\mathrm{c}}\} \\
&= \mathrm{P}\{\mathbb{H}_1^{\mathrm{c}} \cap \mathbb{Q}_K\} + \mathrm{P}\{\mathbb{Q}_K^{\mathrm{c}}\} \\
&\le \mathrm{P}\left\{ \bigcup_{n \le \min\{K, N_{\mathrm{out}}\{\mathbb{S}(\kappa)\}-1\}} \mathbb{M}_n(1/2)^{\mathrm{c}} \cap \mathbb{Q}_K \right\} + \mathrm{P}\{\mathbb{Q}_K^{\mathrm{c}}\} \\
&\le 2dK \exp(-C_\kappa \gamma \kappa^{-2} \nu^{-1} \eta_p^{-2} \beta^{-2\varepsilon}) + K(\mathrm{e}d + p + 1) \exp\left(-C_\psi \min\{\psi^{-1}, \psi^{-2}\} \beta^{-2\varepsilon}\right).
\end{aligned}
$$

Next we estimate the expectation. Since

$$
\mathbb{H}_1 = \bigcap_{n \in [N_{1/2-3\varepsilon}, K]} \mathbb{T}_n(1/2) \subset \bigcap_{n \in [N_{1/2-3\varepsilon}, K]} \left\{ \mathbf{1}_{\mathbb{T}_{n-1}} D^{(n)} = D^{(n)} \right\},
$$

we have for $n \in [N_{1/2-3\varepsilon}, K]$

$$
\begin{aligned}
T^{(n)} \mathbf{1}_{\mathbb{H}_1 \cap \mathbb{Q}_K} &= \mathbf{1}_{\mathbb{Q}_K} \Big( \mathcal{L}^n T^{(0)} + \sum_{s=1}^{N_{1/2-3\varepsilon}-1} \mathcal{L}^{n-s} D^{(s)} + \sum_{s=N_{1/2-3\varepsilon}}^{n} \mathcal{L}^{n-s} D^{(s)} \mathbf{1}_{\mathbb{T}_{s-1}} + \sum_{s=1}^{n} \mathcal{L}^{n-s} E_T^{(s-1)}(V^{(s-1)}) \Big) \\
&=: \widetilde{J}_1 + \widetilde{J}_{21} + \widetilde{J}_{22} + \widetilde{J}_3.
\end{aligned}
$$

In what follows, we simply write $E_T^{(n)} = E_T^{(n)}(V^{(n)})$ for convenience. Then,

$$
\begin{aligned}
\mathrm{E}\Big\{ T^{(n)} \circ T^{(n)};\, \mathbb{H}_1 \cap \mathbb{Q}_K \Big\} &= \mathrm{E}\Big\{ T^{(n)} \circ T^{(n)} \mathbf{1}_{\mathbb{H}_1 \cap \mathbb{Q}_K} \Big\} \\
&= \mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_1 \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_{21} \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_{22} \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_3 \Big\} \\
&\quad + \mathrm{E}\Big\{ [\widetilde{J}_{21} + \widetilde{J}_{22}] \circ [\widetilde{J}_{21} + \widetilde{J}_{22}] \Big\} + 2\,\mathrm{E}\Big\{ [\widetilde{J}_{21} + \widetilde{J}_{22}] \circ \widetilde{J}_3 \Big\} + \mathrm{E}\Big\{ \widetilde{J}_3 \circ \widetilde{J}_3 \Big\} \\
&\le \mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_1 \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_{21} \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_{22} \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_3 \Big\} \\
&\quad + 2\,\mathrm{E}\Big\{ \widetilde{J}_{21} \circ \widetilde{J}_{21} \Big\} + 4\,\mathrm{E}\Big\{ \widetilde{J}_{21} \circ \widetilde{J}_{22} \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_{22} \circ \widetilde{J}_{22} \Big\} + 2\,\mathrm{E}\Big\{ \widetilde{J}_3 \circ \widetilde{J}_3 \Big\}.
\end{aligned}
$$

In the following, we estimate each summand above for $n \in [N_{1-6\varepsilon}, K]$. We have the following.

1. $\mathrm{E}\Big\{ \widetilde{J}_1 \circ \widetilde{J}_1 \Big\} = \mathcal{L}^{2n} T^{(0)} \circ T^{(0)}$.

2. $\mathrm{E}\Big\{\widetilde{J}_1 \circ \widetilde{J}_{21}\Big\} = \sum_{s=1}^{N_{1/2-3\varepsilon}-1} \mathcal{L}^{2n-s} T^{(0)} \circ \mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K}\Big\} = 0$, because

$$\mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K}\Big\} = \mathrm{E}\Big\{\mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{F}_{s-1}\Big\}\Big\} = 0.$$

3. $\mathrm{E}\Big\{\widetilde{J}_1 \circ \widetilde{J}_{22}\Big\} = \sum_{s=N_{1/2-3\varepsilon}}^{n} \mathcal{L}^{2n-s} T^{(0)} \circ \mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{T}_{s-1}} \mathbf{1}_{\mathbb{Q}_K}\Big\} = 0$, because $\mathbb{T}_{s-1} \subset \mathbb{F}_{s-1}$ and so

$$\begin{aligned}
\mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{T}_{s-1}} \mathbf{1}_{\mathbb{Q}_K}\Big\} &= \mathrm{P}\{\mathbb{T}_{s-1}\}\, \mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{T}_{s-1}\Big\} \\
&= \mathrm{P}\{\mathbb{T}_{s-1}\}\, \mathrm{E}\Big\{\mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{F}_{s-1}\Big\} \,\Big|\, \mathbb{T}_{s-1}\Big\} = 0.
\end{aligned}$$

4. $\mathrm{E}\Big\{\widetilde{J}_1 \circ \widetilde{J}_3\Big\} = \sum_{s=1}^{n} \mathcal{L}^{2n-s} T^{(0)} \circ \mathrm{E}\Big\{E_T^{(s-1)} \mathbf{1}_{\mathbb{Q}_K}\Big\}$. Recall (6.17). By item 2(a) of Lemma 6.3, we have

$$\begin{aligned}
\|\mathrm{E}\Big\{\widetilde{J}_1 \circ \widetilde{J}_3\Big\}\|_2 &\le \sum_{s=1}^{n} \|\mathcal{L}\|_2^{2n-s} \|T^{(0)}\|_2 \|\mathrm{E}\Big\{E_T^{(s-1)} \mathbf{1}_{\mathbb{Q}_K}\Big\}\|_2 \\
&\le \sum_{s=1}^{n} (1-\beta\gamma)^{2n-s} \Big(\frac{\kappa^2}{4} - 1\Big)^{1/2} C_T \nu^{1/2} (\widetilde{\eta}_p \beta)^2 \kappa^3 \\
&\le (1-\beta\gamma)^n \frac{(\kappa^2-1)^{1/2} C_T \nu^{1/2} \widetilde{\eta}_p^2 \beta^2 \kappa^3}{2\beta\gamma} \\
&\le \frac{1}{2} \beta^{1-6\varepsilon} C_T \nu^{1/2} \widetilde{\eta}_p^2 \gamma^{-1} \beta \kappa^4, \qquad \text{by } n \ge N_{1-6\varepsilon}.
\end{aligned}$$

5. $\mathrm{E}\Big\{\widetilde{J}_{21} \circ \widetilde{J}_{22}\Big\} = \sum_{s=1}^{N_{1/2-3\varepsilon}-1} \sum_{s'=N_{1/2-3\varepsilon}}^{n} \mathcal{L}^{2n-s-s'}\, \mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \circ D^{(s')} \mathbf{1}_{\mathbb{T}_{s'-1}} \mathbf{1}_{\mathbb{Q}_K}\Big\} = 0$, because $s < s'$ and

$$\begin{aligned}
\mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \circ D^{(s')} \mathbf{1}_{\mathbb{T}_{s'-1}} \mathbf{1}_{\mathbb{Q}_K}\Big\} &= \mathrm{E}\Big\{D^{(s)} \circ D^{(s')} \mathbf{1}_{\mathbb{T}_{s'-1}} \mathbf{1}_{\mathbb{Q}_K}\Big\} \\
&= \mathrm{P}\{\mathbb{T}_{s'-1}\}\, \mathrm{E}\Big\{D^{(s)} \circ D^{(s')} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{T}_{s'-1}\Big\} \\
&= \mathrm{P}\{\mathbb{T}_{s'-1}\}\, \mathrm{E}\Big\{\mathrm{E}\Big\{D^{(s)} \circ D^{(s')} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{F}_{s'-1}\Big\} \,\Big|\, \mathbb{T}_{s'-1}\Big\} \\
&= \mathrm{P}\{\mathbb{T}_{s'-1}\}\, \mathrm{E}\Big\{\mathrm{E}\Big\{D^{(s')} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{F}_{s'-1}\Big\} \circ D^{(s)} \,\Big|\, \mathbb{T}_{s'-1}\Big\} \\
&= 0.
\end{aligned}$$

6. For $\mathrm{E}\Big\{\widetilde{J}_{21} \circ \widetilde{J}_{21}\Big\}$, we have

$$\begin{aligned}
\mathrm{E}\Big\{\widetilde{J}_{21} \circ \widetilde{J}_{21}\Big\} &= \sum_{s=1}^{N_{1/2-3\varepsilon}-1} \sum_{s'=1}^{N_{1/2-3\varepsilon}-1} \mathcal{L}^{2n-s-s'}\, \mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \circ D^{(s')} \mathbf{1}_{\mathbb{Q}_K}\Big\} \\
&= \sum_{s=1}^{N_{1/2-3\varepsilon}-1} \mathcal{L}^{2(n-s)}\, \mathrm{E}\Big\{D^{(s)} \circ D^{(s)} \mathbf{1}_{\mathbb{Q}_K}\Big\},
\end{aligned}$$

because for $s \ne s'$,

$$\mathrm{E}\Big\{D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \circ D^{(s')} \mathbf{1}_{\mathbb{Q}_K}\Big\} = \mathrm{E}\Big\{D^{(s)} \circ D^{(s')} \mathbf{1}_{\mathbb{Q}_K}\Big\}$$

$$= \mathrm{E}\Big\{\mathrm{E}\Big\{D^{(\max\{s,s'\})}\mathbf{1}_{\mathbb{Q}_K}\;\Big|\;\mathbb{F}_{\max\{s,s'\}-1}\Big\}\circ D^{(\min\{s,s'\})}\Big\}$$
$$= 0.$$

Use items 3(a) and 3(b) of Lemma 6.3 to get

$$\begin{aligned}
\mathrm{E}\Big\{D^{(s)}\circ D^{(s)}\mathbf{1}_{\mathbb{Q}_K}\Big\} &= \mathrm{E}\Big\{\mathrm{E}\Big\{D^{(s)}\circ D^{(s)}\mathbf{1}_{\mathbb{Q}_K}\;\Big|\;\mathbb{F}_{s-1}\Big\}\Big\}\\
&= \mathrm{E}\Big\{\mathrm{var}_\circ\Big([T^{(n+1)}-T^{(n)}]\mathbf{1}_{\mathbb{Q}_K}\;\Big|\;\mathbb{F}_{s-1}\Big)\Big\}\\
&= \mathrm{E}\big\{\beta^2 H_\circ + R_\circ\big\}\\
&= \beta^2 H_\circ + \mathrm{E}\{R_\circ\}.
\end{aligned}$$

Therefore $\mathrm{E}\Big\{\widetilde{J}_{21}\circ\widetilde{J}_{21}\Big\} = \beta^2\sum_{s=1}^{N_{1/2-3\varepsilon}-1}\mathcal{L}^{2(n-s)}H_\circ + \sum_{s=1}^{N_{1/2-3\varepsilon}-1}\mathcal{L}^{2(n-s)}\,\mathrm{E}\{R_\circ\}$. We have

$$\begin{aligned}
\|R_\circ\|_2 &\le (\nu\widetilde{\eta}_p\beta)^2\tau_{s-1}(1+\frac{11}{2}\tau_{s-1}+\tau_{s-1}^2+\frac{1}{4}\tau_{s-1}^3) + 4C_T\kappa^5\nu(\widetilde{\eta}_p\beta)^3 + 2C_T^2\kappa^6\nu(\widetilde{\eta}_p\beta)^4\\
&\le (\nu\widetilde{\eta}_p\beta)^2\tau_{s-1}(\kappa^2+\frac{21}{4}\kappa+\frac{1}{4}\kappa^3) + 4C_T\kappa^5\nu(\widetilde{\eta}_p\beta)^3 + 2C_T^2\kappa^6\nu(\widetilde{\eta}_p\beta)^4\\
&\le \frac{29+8\sqrt{2}}{32}\kappa^3\nu^2(\widetilde{\eta}_p\beta)^2\tau_{s-1} + 4C_T\kappa^5\nu(\widetilde{\eta}_p\beta)^3 + 2C_T^2\kappa^6\nu(\widetilde{\eta}_p\beta)^4 \qquad \text{for } \kappa > 2\sqrt{2},
\end{aligned}$$

where $\tau_{s-1} = \|T^{(s-1)}\|_2 \le (\kappa^2-1)^{1/2}$. Write $E_{21} := \sum_{s=1}^{N_{1/2-3\varepsilon}-1}\mathcal{L}^{2(n-s)}\,\mathrm{E}\{R_\circ\}$. Since $2N_{1/2-3\varepsilon}-1 \le N_{1-6\varepsilon} \le 2N_{1/2-3\varepsilon}$ by definition, we get

$$\begin{aligned}
\|E_{21}\|_2 &\le \sum_{s=1}^{N_{1/2-3\varepsilon}-1}\|\mathcal{L}\|_2^{2(n-s)}\,\mathrm{E}\{\|R_\circ\|_2\}\\
&\le \frac{(1-\beta\gamma)^{2(n+1-N_{1/2-3\varepsilon})}}{\beta\gamma[2-\beta\gamma]}\,\mathrm{E}\{\|R_\circ\|_2\}\\
&\le \frac{1-\beta\gamma}{2-\beta\gamma}\,\frac{(1-\beta\gamma)^n}{\beta\gamma}\,\mathrm{E}\{\|R_\circ\|_2\}\\
&\le \frac{1}{2}\beta^{1-6\varepsilon}\gamma^{-1}\beta\kappa^4\nu\widetilde{\eta}_p^2\left(\frac{29+8\sqrt{2}}{32}\nu + 4C_T\kappa(\widetilde{\eta}_p\beta) + 2C_T^2\kappa^2(\widetilde{\eta}_p\beta)^2\right)\\
&\le \left(\frac{29+8\sqrt{2}}{64} + 2C_T\kappa(\widetilde{\eta}_p\beta) + C_T^2\kappa^2(\widetilde{\eta}_p\beta)^2\right)\gamma^{-1}\kappa^4\nu^2\widetilde{\eta}_p^2\beta^{2-6\varepsilon}.
\end{aligned}$$

7. For $\mathrm{E}\Big\{\widetilde{J}_{22}\circ\widetilde{J}_{22}\Big\}$, we have

$$\begin{aligned}
\mathrm{E}\Big\{\widetilde{J}_{22}\circ\widetilde{J}_{22}\Big\} &= \sum_{s=N_{1/2-3\varepsilon}}^{n}\mathcal{L}^{2(n-s)}\,\mathrm{E}\Big\{D^{(s)}\mathbf{1}_{\mathbb{Q}_K}\mathbf{1}_{\mathbb{T}_{s-1}}\circ D^{(s)}\mathbf{1}_{\mathbb{Q}_K}\mathbf{1}_{\mathbb{T}_{s-1}}\Big\}\\
&= \beta^2\sum_{s=N_{1/2-3\varepsilon}}^{n}\mathcal{L}^{2(n-s)}H_\circ + \sum_{s=N_{1/2-3\varepsilon}}^{n}\mathcal{L}^{2(n-s)}\,\mathrm{E}\big\{R_\circ\mathbf{1}_{\mathbb{T}_{s-1}}\big\},
\end{aligned}$$

because for $s\neq s'$,

$$\begin{aligned}
\mathrm{E}\Big\{D^{(s)}\mathbf{1}_{\mathbb{Q}_K}\mathbf{1}_{\mathbb{T}_{s-1}}\circ D^{(s')}\mathbf{1}_{\mathbb{Q}_K}\mathbf{1}_{\mathbb{T}_{s'-1}}\Big\} &= \mathrm{E}\Big\{D^{(s)}\circ D^{(s')}\mathbf{1}_{\mathbb{Q}_K}\mathbf{1}_{\mathbb{T}_{s-1}}\mathbf{1}_{\mathbb{T}_{s'-1}}\Big\}\\
&= \mathrm{P}\{\mathbb{T}_{s-1}\cap\mathbb{T}_{s'-1}\}\,\mathrm{E}\Big\{D^{(s)}\circ D^{(s')}\mathbf{1}_{\mathbb{Q}_K}\;\Big|\;\mathbb{T}_{s-1}\cap\mathbb{T}_{s'-1}\Big\}
\end{aligned}$$

$$= \mathrm{P}\{\mathbb{T}_{s-1} \cap \mathbb{T}_{s'-1}\}\, \mathrm{E}\Big\{ \mathrm{E}\Big\{ D^{(\max\{s,s'\})} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{F}_{\max\{s,s'\}-1} \Big\} \circ D^{(\min\{s,s'\})} \,\Big|\, \mathbb{T}_{s-1} \cap \mathbb{T}_{s'-1} \Big\}$$
$$= 0,$$

and

$$\mathrm{E}\Big\{ D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \mathbf{1}_{\mathbb{T}_{s-1}} \circ D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \mathbf{1}_{\mathbb{T}_{s-1}} \Big\} = \mathrm{E}\Big\{ D^{(s)} \circ D^{(s')} \mathbf{1}_{\mathbb{Q}_K} \mathbf{1}_{\mathbb{T}_{s-1}} \Big\}$$
$$= \mathrm{P}\{\mathbb{T}_{s-1}\}\, \mathrm{E}\Big\{ \mathrm{E}\Big\{ D^{(s)} \circ D^{(s)} \mathbf{1}_{\mathbb{Q}_K} \,\Big|\, \mathbb{F}_{s-1} \Big\} \,\Big|\, \mathbb{T}_{s-1} \Big\}$$
$$\le \beta^2 H_\circ + \mathrm{E}\big\{ R_\circ \mathbf{1}_{\mathbb{T}_{s-1}} \big\}.$$

We have

$$\|R_\circ \mathbf{1}_{\mathbb{T}_{s-1}}\|_2 \le \frac{29 + 8\sqrt{2}}{32} \kappa^3 \nu^2 (\widetilde{\eta}_p \beta)^2 \tau_{s-1} + 4C_T \kappa^5 \nu (\widetilde{\eta}_p \beta)^3 + 2C_T^2 \kappa^6 \nu (\widetilde{\eta}_p \beta)^4$$
$$\le \frac{29 + 8\sqrt{2}}{32} \kappa^3 \nu^2 (\widetilde{\eta}_p \beta)^2 (\kappa^2 - 1)^{1/2} \beta^{1/2 - 3\varepsilon} + 4C_T \kappa^5 \nu (\widetilde{\eta}_p \beta)^3 + 2C_T^2 \kappa^6 \nu (\widetilde{\eta}_p \beta)^4$$
$$\le \frac{29 + 8\sqrt{2}}{32} \kappa^4 \nu^2 (\widetilde{\eta}_p \beta)^2 \beta^{1/2 - 3\varepsilon} + 4C_T \kappa^5 \nu (\widetilde{\eta}_p \beta)^3 + 2C_T^2 \kappa^6 \nu (\widetilde{\eta}_p \beta)^4.$$

Write $E_{22} := \sum_{s=N_{1/2-3\varepsilon}}^{n} \mathcal{L}^{2(n-s)}\, \mathrm{E}\big\{ R_\circ \mathbf{1}_{\mathbb{T}_{s-1}} \big\}$ for which we have

$$\|E_{22}\|_2 \le \sum_{s=N_{1/2-3\varepsilon}}^{n} \|\mathcal{L}\|_2^{2(n-s)}\, \mathrm{E}\big\{ \|R_\circ \mathbf{1}_{\mathbb{T}_{s-1}}\|_2 \big\}$$
$$\le \frac{1}{\beta\gamma[2 - \beta\gamma]}\, \mathrm{E}\big\{ \|R_\circ \mathbf{1}_{\mathbb{T}_{s-1}}\|_2 \big\}$$
$$\le \frac{1}{3 - \sqrt{2}} \gamma^{-1} \kappa^4 \nu \widetilde{\eta}_p^2 \beta \left( \frac{29 + 8\sqrt{2}}{32} \nu \beta^{1/2 - 3\varepsilon} + 4C_T \kappa (\widetilde{\eta}_p \beta) + 2C_T^2 \kappa^2 (\widetilde{\eta}_p \beta)^2 \right)$$
$$\le \frac{1}{3 - \sqrt{2}} \left( \frac{29 + 8\sqrt{2}}{32} + 4C_T \kappa \widetilde{\eta}_p \beta^{1/2 + 3\varepsilon} + 2C_T^2 \kappa^2 \widetilde{\eta}_p^2 \beta^{3/2 + 3\varepsilon} \right) \gamma^{-1} \kappa^4 \nu^2 \widetilde{\eta}_p^2 \beta^{3/2 - 3\varepsilon}.$$

8. $\mathrm{E}\Big\{ \widetilde{J}_3 \circ \widetilde{J}_3 \Big\} = \sum_{s=1}^{n} \mathcal{L}^{2(n-s)}\, \mathrm{E}\Big\{ E_T^{(s-1)} \mathbf{1}_{\mathbb{Q}_K} \circ E_T^{(s-1)} \mathbf{1}_{\mathbb{Q}_K} \Big\}$. Also, by (6.17),

$$\|\mathrm{E}\Big\{ \widetilde{J}_3 \circ \widetilde{J}_3 \Big\}\|_2 \le \sum_{s=1}^{n} \|\mathcal{L}\|_2^{2(n-s)}\, \mathrm{E}\Big\{ \|E_T^{(s-1)} \mathbf{1}_{\mathbb{Q}_K}\|_2^2 \Big\}$$
$$\le \sum_{s=1}^{n} (1 - \beta\gamma)^{2(n-s)} [C_T \nu^{1/2} (\widetilde{\eta}_p \beta)^2 \kappa^3]^2$$
$$\le \frac{C_T^2 \nu (\widetilde{\eta}_p \beta)^4 \kappa^6}{\beta\gamma[2 - \beta\gamma]}$$
$$\le \frac{1}{3 - \sqrt{2}} C_T^2 \nu \widetilde{\eta}_p^4 \gamma^{-1} \kappa^6 \beta^3.$$

Collecting all estimates together, we obtain

$$\mathrm{E}\Big\{ T^{(n)} \circ T^{(n)};\, \mathbb{H}_1 \cap \mathbb{Q}_K \Big\} \le \mathcal{L}^{2n} T^{(0)} \circ T^{(0)} + 2\beta^2 \sum_{s=1}^{n} \mathcal{L}^{2(n-s)} H_\circ + R_E$$
$$\le \mathcal{L}^{2n} T^{(0)} \circ T^{(0)} + 2\beta^2 [I - \mathcal{L}^2]^{-1} [I - \mathcal{L}^{2n}] H_\circ + R_E,$$

where, by ($\beta$-3), $2C_\Delta\kappa\widetilde{\eta}_p\beta^{1/2}\le 1$, and

$$
\begin{aligned}
\|R_E\|_2 &\le 2\Bigg[\frac{C_T}{2}\kappa^4\nu^{1/2}\widetilde{\eta}_p^2\gamma^{-1}\beta^{2-6\varepsilon}+\left(\frac{29+8\sqrt{2}}{64}+2C_T\kappa\widetilde{\eta}_p\beta+C_T^2\kappa^2(\widetilde{\eta}_p\beta)^2\right)\kappa^4\nu^2\widetilde{\eta}_p^2\gamma^{-1}\beta^{2-6\varepsilon}\\
&\qquad+\frac{2}{3-\sqrt{2}}\left(\frac{29+8\sqrt{2}}{64}+2C_T\kappa\widetilde{\eta}_p\beta^{1/2+3\varepsilon}+C_T^2\kappa^2\widetilde{\eta}_p^2\beta^{3/2+3\varepsilon}\right)\kappa^4\nu^2\widetilde{\eta}_p^2\gamma^{-1}\beta^{3/2-3\varepsilon}\\
&\qquad+\frac{C_T^2}{3-\sqrt{2}}\nu\widetilde{\eta}_p^4\gamma^{-1}\kappa^6\beta^3\Bigg]\\
&\le 2\Bigg[\frac{C_T}{2}\beta^{1/2-3\varepsilon}+\left(\frac{29+8\sqrt{2}}{64}+2C_T\kappa\widetilde{\eta}_p\beta+C_T^2\kappa^2(\widetilde{\eta}_p\beta)^2\right)\beta^{1/2-3\varepsilon}+\frac{C_T^2}{3-\sqrt{2}}\widetilde{\eta}_p^2\kappa^2\beta^{3/2+3\varepsilon}\\
&\qquad+\frac{2}{3-\sqrt{2}}\left(\frac{29+8\sqrt{2}}{64}+2C_T\kappa\widetilde{\eta}_p\beta^{1/2+3\varepsilon}+C_T^2\kappa^2\widetilde{\eta}_p^2\beta^{3/2+3\varepsilon}\right)\Bigg]\kappa^4\gamma^{-1}\nu^2\widetilde{\eta}_p^2\beta^{3/2-3\varepsilon}\\
&\le 2\Bigg[\frac{C_T}{2}\beta^{1/2-3\varepsilon}+\left(\frac{29+8\sqrt{2}}{64}+\frac{C_T}{C_\Delta}\beta^{1/2}+\frac{C_T^2}{4C_\Delta^2}\beta\right)\beta^{1/2-3\varepsilon}+\frac{C_T^2}{4(3-\sqrt{2})C_\Delta^2}\beta^{1/2+3\varepsilon}\\
&\qquad+\frac{2}{3-\sqrt{2}}\left(\frac{29+8\sqrt{2}}{64}+\frac{C_T}{C_\Delta}\beta^{3\varepsilon}+\frac{C_T^2}{4C_\Delta^2}\beta^{1/2+3\varepsilon}\right)\Bigg]\kappa^4\gamma^{-1}\nu^2\widetilde{\eta}_p^2\beta^{3/2-3\varepsilon}\\
&= C_\circ\kappa^4\gamma^{-1}\nu^2\widetilde{\eta}_p^2\beta^{3/2-3\varepsilon},
\end{aligned}
$$

where $C_\circ$ is as given in (6.6f). □

## 6.4 Proof of Theorem 4.1

Write $\widetilde{N}_s=\frac{s\ln\beta}{\ln(1-\beta\gamma)}$. Then $(1-\beta\gamma)^{\widetilde{N}_s}=\beta^s$ and $N_s=\lceil\widetilde{N}_s\rceil$, where $N_s$ is defined in (4.1). It can be verified that $\widetilde{N}_{s_1}+\widetilde{N}_{s_2}=\widetilde{N}_{s_1+s_2}$ for any $s_1,s_2$.

Write $\kappa_m=6^{(1-m)/2}\kappa$ for $m=1,\dots,M\equiv M(\epsilon)$. Since $d\beta^{1-7\varepsilon}\le(\sqrt{2}-1)\lambda_1^{-1}\omega$, we know

$$\phi d^{1/2}\le\phi\omega^{1/2}\beta^{7\varepsilon/2-1/2}\le\beta^{(1-2^{1-M})(3\varepsilon-1/2)}\kappa_M/2.$$

The key to our proof is to divide the whole process into $M$ segments of iterations. Thanks to the strong Markov property of the process, we can use the final value of current segment as the initial guess of the very next one. By Lemma 6.5, after the first segment of

$$n_1:=\min\Big\{N_{\rm in}\Big\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-M})(3\varepsilon-1/2)}\kappa_1)\Big\},N_{2^{-M}(1-6\varepsilon)}\Big\}$$

iterations, $V^{(n_1)}$ lies in $\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-M})(3\varepsilon-1/2)}\kappa_1)=\mathbb{S}(\beta^{(1-2^{2-M})(3\varepsilon-1/2)}\kappa_2/2)$ with high probability, which will be a good initial guess for the second segment. In general, the $i$th segment of iterations starts with $V^{(n_{i-1})}$ and ends with $V^{(n_i)}$, where

$$n_i=\min\left\{N_{\rm in}\Big\{\mathbb{S}(\beta^{(1-2^{i+1-M})(3\varepsilon-1/2)}\kappa_{i+1}/2)\Big\},\left\lceil\sum_{m=M+1-i}^{M}\widetilde{N}_{2^{-m}(1-6\varepsilon)}\right\rceil\right\}.$$

At the end of the $(M-1)$st segment of iterations, $V^{(n_{M-1})}$ is produced and it is going to be used as an initial guess for the last step, at which we can apply Lemma 6.6. Now $n_{M-1}=\min\Big\{N_{\rm in}\{\mathbb{S}(\kappa_M/2)\},\widehat{K}\Big\}$, where $\widehat{K}=\left\lceil\sum_{m=2}^{M}\widetilde{N}_{2^{-m}(1-6\varepsilon)}\right\rceil=\left\lceil\widetilde{N}_{(1-2^{1-M})(1/2-3\varepsilon)}\right\rceil$. By $2^{2-M}\ge\frac{\varepsilon/2}{1/2-3\varepsilon}\ge 2^{1-M}$, we have

$$N_{1/2-7\varepsilon/2}=\left\lceil\widetilde{N}_{1/2-7\varepsilon/2}\right\rceil\le\widehat{K}\le\left\lceil\widetilde{N}_{1/2-13\varepsilon/4}\right\rceil\le N_{1/2-13\varepsilon/4}.$$

Let

$$\widetilde{\mathbb{H}}_m = \Big\{ N_{\rm in}\Big\{\mathbb{S}(\sqrt{3/2}\beta^{(1-2^{2-m})(3\varepsilon-1/2)}\kappa_{M+1-m})\Big\} \le \widetilde{N}_{2^{-m}(1-6\varepsilon)} + n_{M-m}\Big\} \quad \text{for } m=2,\dots,M,$$

$$\widetilde{\mathbb{H}}_1 = \bigcap_{n\in[N_{1/2-3\varepsilon}, K-N_{\rm in}\{\mathbb{S}(\kappa_M/2)\}]} \mathbb{T}_{n+N_{\rm in}\{\mathbb{S}(\kappa_M/2)\}}(1/2),$$

$$\mathbb{H} = \bigcap_{m=1}^{M} \widetilde{\mathbb{H}}_m \cap \mathbb{Q}_K,$$

where $n_0 = 0$. We have

$$\begin{aligned}
\mathrm{P}\{\mathbb{H}^{\rm c}\} &= \mathrm{P}\Big\{\bigcup_{m=1}^{M} \widetilde{\mathbb{H}}_m^{\rm c} \cup \mathbb{Q}_K^{\rm c}\Big\} \le \sum_{m=1}^{M} \mathrm{P}\Big\{\widetilde{\mathbb{H}}_m^{\rm c} \cup \mathbb{Q}_K^{\rm c}\Big\}\\
&\le \sum_{m=2}^{M} 2dN_{2^{-m}(1-6\varepsilon)} \exp(-C_\kappa\gamma\kappa_{M+1-m}^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon})\\
&\quad + 2d\Big(K - \sum_{m=2}^{M} N_{2^{-m}(1-6\varepsilon)}\Big) \exp(-C_\kappa\gamma\kappa_M^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon})\\
&\quad + K(\mathrm{e}d+p+1)\exp\big(-C_\psi \min\{\psi^{-1},\psi^{-2}\}\beta^{-2\varepsilon}\big)\\
&\le 2dK\exp(-C_\kappa\gamma\kappa^{-2}\nu^{-1}\eta_p^{-2}\beta^{-2\varepsilon}) + K(\mathrm{e}d+p+1)\exp\big(-C_\psi \min\{\psi^{-1},\psi^{-2}\}\beta^{-2\varepsilon}\big)\\
&\le 2dK\exp(-C_\kappa 4\sqrt{2}C_T\nu^{-1}\beta^{\varepsilon}\beta^{-2\varepsilon}) + K(\mathrm{e}d+p+1)\exp\big(-C_\psi \min\{\psi^{-1},\psi^{-2}\}\beta^{-2\varepsilon}\big) \qquad \text{by } (\beta\text{-}4)\\
&\le 2dK\exp(-4\sqrt{2}C_TC_\kappa\nu^{-1}\beta^{-\varepsilon}) + K(\mathrm{e}d+p+1)\exp\big(-C_\psi \min\{\psi^{-1},\psi^{-2}\}\beta^{-\varepsilon}\big)\\
&\le K[(2+\mathrm{e})d+p+1]\exp(-\max\{C_\nu\nu^{-1}, C_\psi\min\{\psi^{-1},\psi^{-2}\}\}\beta^{-\varepsilon}),
\end{aligned}$$

where $C_\nu = 4\sqrt{2}C_TC_\kappa$ is as given in (6.6e).

Set $\mathbb{H}'_{n'} := \{N_{\rm in}\{\mathbb{S}(\kappa/2)\} = n'\}$. If $n' > \widehat{K}$, then $\mathbb{H}\cap\mathbb{H}'_{n'} = \emptyset$. Otherwise, by Lemma 6.5, $V^{(n')} \in \mathbb{S}(\kappa_M/2)$ and then $\|T^{(n')}\|_{\rm F}^2 \le p((\frac{\kappa_M}{2})^2 - 1)$. Thus,

$$\phi^2 d(1-\beta\gamma)^{2(n'-1)} \ge \phi^2 d(1-\beta\gamma)^{2(\widehat{K}-1)} > \Big(\frac{\kappa_M}{2}\Big)^2 \ge \frac{1}{p}\|T^{(n')}\|_{\rm F}^2.$$

Hence, for any $n \in [N_{1-6\varepsilon} + N_{\rm in}\{\mathbb{S}(\kappa/2)\}, K] \subset [N_{1-6\varepsilon}+n', K+n']$, by Lemma 6.6, we have

$$\mathrm{E}\Big\{T^{(n)}\circ T^{(n)}\mathbf{1}_{\mathbb{H}} \,\Big|\, \mathbb{H}'_{n'}\cap\mathbb{F}_{n'}\Big\} \le \mathcal{L}^{2(n-n')}T^{(n')}\circ T^{(n')} + 2\beta^2[I-\mathcal{L}^2]^{-1}[I-\mathcal{L}^{2(n-n')}]H_\circ + R_E.$$

Introduce $\mathrm{sum}(A)$ for the sum of all the entries of $A$. In particular, $\mathrm{sum}(A\circ A) = \|A\|_{\rm F}^2$. We have

$$\begin{aligned}
\mathrm{E}\Big\{\|T^{(n)}\|_{\rm F}^2\mathbf{1}_{\mathbb{H}} \,\Big|\, \mathbb{H}'_{n'}\Big\} &= \mathrm{E}\Big\{\mathrm{E}\Big\{\|T^{(n)}\|_{\rm F}^2\mathbf{1}_{\mathbb{H}} \,\Big|\, \mathbb{H}'_{n'}\cap\mathbb{F}_{n'}\Big\}\Big\}\\
&\le \mathrm{E}\Big\{(1-\beta\gamma)^{2(n-n')}\|T^{(n')}\|_{\rm F}^2 + 2\beta^2\,\mathrm{sum}([I-\mathcal{L}^2]^{-1}H_\circ) + \mathrm{sum}(R_E)\Big\}\\
&\le (1-\beta\gamma)^{2(n-1)}p\phi^2 d + 2\beta^2\,\mathrm{sum}([I-\mathcal{L}^2]^{-1}H_\circ) + \sqrt{p(d-p)}\|R_E\|_{\rm F}\\
&\le (1-\beta\gamma)^{2(n-1)}p\phi^2 d + 2\beta^2\frac{1}{\beta(2-\lambda_1\beta)}\,\mathrm{sum}(G\circ H_\circ)\\
&\quad + \sqrt{p(d-p)}C_\circ\sqrt{p}\kappa^4(\nu\widetilde{\eta}_p)^2\gamma^{-1}\beta^{3/2-3\varepsilon},
\end{aligned}$$

where $G = [\gamma_{ij}]_{(d-p)\times p}$ with $\gamma_{ij} = \frac{1}{\lambda_j - \lambda_{p+i}}$. Putting all together, we get

$$\begin{aligned}
\mathrm{E}\Big\{\|T^{(n)}\|_{\rm F}^2;\, \mathbb{H}\Big\} &= \mathrm{E}\Big\{\mathrm{E}\Big\{\|T^{(n)}\|_{\rm F}^2\mathbf{1}_{\mathbb{H}} \,\Big|\, \mathbb{H}'_{n'}\Big\}\Big\}\\
&\le (1-\beta\gamma)^{2(n-1)}p\phi^2 d + \frac{2\beta}{2-\lambda_1\beta}\,\mathrm{sum}(G\circ H_\circ) + C_\circ\kappa^4\nu^2\widetilde{\eta}_p^2 p\sqrt{d-p}\gamma^{-1}\beta^{3/2-3\varepsilon}.
\end{aligned}$$

Note that on $\mathbb{H}$, $N_{\text{in}}\{\mathbb{S}(\kappa/2)\} \le \widehat{K}$. So the expectation is valid for any $n \in [N_{1-2\varepsilon}+\widehat{K}, K]$. Finally we estimate $\operatorname{sum}(G \circ H_\circ)$. By Lemma 6.3, $H_\circ \le 16\psi^4 H$, and hence

$$\operatorname{sum}(G \circ H_\circ) \le \sum_{j=1}^{p}\sum_{i=1}^{d-p} \frac{16\psi^4 \lambda_{p+i}\lambda_j}{\lambda_j - \lambda_{p+i}} = 16\psi^4 \varphi(p,d;\Lambda).$$

This completes the proof.

# 7 Proofs of Theorems 4.2 and 4.3

To prove Theorem 4.2, we will first prove that it is a high-probability event that $V^{(0)}$ satisfies the initial condition there, which is the result of Lemma 7.2 below. Then, together with Theorem 4.1, we will have its conclusion. During estimating the probability, we need a property on the Gaussian hypergeometric function of a matrix argument, as in Lemma 7.1.

The gamma function and the multivariate gamma function are

$$\Gamma(x) := \int_0^\infty t^{x-1}\exp(-t)\,\mathrm{d}t, \quad \Gamma_m(x) := \pi^{m(m-1)/4}\prod_{i=1}^{m}\Gamma\Big(x - \frac{i-1}{2}\Big),$$

respectively. Denote by ${}_2F_1$ the Gaussian hypergeometric function of matrix argument (see [14, Definition 7.3.1]), and also by ${}_1F_0$ and ${}_1F_1$ the generalized hypergeometric functions that will be used later.

**Lemma 7.1.** *For any scalar $a, b, c$ and a symmetric matrix $T \in \mathbb{R}^{m\times m}$,*

$$\begin{aligned}
{}_2F_1(a,b;c;T) &= \frac{\Gamma_m(c-a-b)\Gamma_m(c)}{\Gamma_m(c-a)\Gamma_m(c-b)}\,{}_2F_1(a,b;a+b-c+\frac{m+1}{2};I-T)\\
&\quad + \frac{\Gamma_m(a+b-c)\Gamma_m(c)}{\Gamma_m(a)\Gamma_m(b)}\det(I-T)^{c-a-b}\,{}_2F_1(c-a,c-b;c-a-b+\frac{m+1}{2};I-T). \quad (7.1)
\end{aligned}$$

*Proof.* The proof is the same as that for the case $p=1$ by Kummer's solutions of the hypergeometric differential equation (see, e.g., [13, Section 3.8]). Let the eigenvalues of $T$ be $\mu_1, \dots, \mu_m$. Since ${}_2F_1(a,b;c;T)$ is defined on the spectrum of $T$, it is a function of $\mu_1,\dots,\mu_m$. When treated as such, by [14, Theorem 7.5.5], ${}_2F_1(a,b;c;T)$ is the unique solution of partial differential equations,

$$\begin{aligned}
\mu_i(1-\mu_i)\frac{\partial^2 F}{\partial \mu_i^2} + \left(c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2}\sum_{1\le j\le m}^{j\ne i}\frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j}\right)\frac{\partial F}{\partial \mu_i}\\
- \frac{1}{2}\sum_{1\le j\le m}^{j\ne i}\frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j}\frac{\partial F}{\partial \mu_j} - abF = 0, \quad (7.2)
\end{aligned}$$

subject to the conditions that $F$ is a symmetric function of $\mu_1,\dots,\mu_m$, analytic at $(\mu_1,\dots,\mu_m) = (0,\dots,0)$, and $F(0,\dots,0)=1$.

We claim that $\widetilde{F}(\mu_1,\dots,\mu_m) := {}_2F_1(a,b;a+b-c+\frac{m+1}{2};I-T)$ satisfies (7.2). In fact, letting $\widetilde{\mu}_i = 1-\mu_i$ for $1 \le i \le m$ which are the eigenvalues of $I-T$, we have

$$\begin{aligned}
\mu_i(1-\mu_i)\frac{\partial^2 \widetilde{F}}{\partial \mu_i^2} + \left(c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2}\sum_{1\le j\le m}^{j\ne i}\frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j}\right)\frac{\partial \widetilde{F}}{\partial \mu_i}\\
- \frac{1}{2}\sum_{1\le j\le m}^{j\ne i}\frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j}\frac{\partial \widetilde{F}}{\partial \mu_j} - ab\widetilde{F}
\end{aligned}$$

$$
\begin{aligned}
&= (1-\widetilde{\mu}_i)\widetilde{\mu}_i \frac{\partial^2 \widetilde{F}}{\partial \widetilde{\mu}_i^2} + \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{(1-\widetilde{\mu}_j)\widetilde{\mu}_j}{(1-\widetilde{\mu}_i)-(1-\widetilde{\mu}_j)} \frac{\partial \widetilde{F}}{\partial \widetilde{\mu}_j} - ab\widetilde{F} \\
&\qquad - \left( c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})(1-\widetilde{\mu}_i) + \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{(1-\widetilde{\mu}_i)\widetilde{\mu}_i}{(1-\widetilde{\mu}_i)-(1-\widetilde{\mu}_j)} \right) \frac{\partial \widetilde{F}}{\partial \widetilde{\mu}_i} \\
&= (1-\widetilde{\mu}_i)\widetilde{\mu}_i \frac{\partial^2 \widetilde{F}}{\partial \widetilde{\mu}_i^2} - \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{(1-\widetilde{\mu}_j)\widetilde{\mu}_j}{\widetilde{\mu}_i-\widetilde{\mu}_j} \frac{\partial \widetilde{F}}{\partial \widetilde{\mu}_j} - ab\widetilde{F} \\
&\qquad + \left( -c + \frac{m+1}{2} + a + b - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\widetilde{\mu}_i + \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{(1-\widetilde{\mu}_i)\widetilde{\mu}_i}{\widetilde{\mu}_i-\widetilde{\mu}_j} \right) \frac{\partial \widetilde{F}}{\partial \widetilde{\mu}_i} \\
&= 0,
\end{aligned}
$$

where the last equality holds because $\widetilde{F}(\mu_1,\dots,\mu_m) = {}_2F_1(a,b;a+b-c+\frac{m+1}{2};I-T)$ satisfies a version of (7.2) after substitutions: $\mu_i \to \widetilde{\mu}_i$ for all $i$ and $c \to a+b-c+\frac{m+1}{2}$.

$\widehat{F}(\mu_1,\dots,\mu_m) := \det(T)^{\frac{m+1}{2}-c} {}_2F_1(a-c+\frac{m+1}{2}, b-c+\frac{m+1}{2}; m+1-c; T)$ satisfies (7.2), too. Set $t = \frac{m+1}{2} - c$ and write $G(\mu_1,\dots,\mu_m) = {}_2F_1(a+t,b+t;c+2t;T)$. We have

$$
\begin{aligned}
\frac{\partial \widehat{F}}{\partial \mu_i} &= \frac{t}{\mu_i}\det(T)^t G + \det(T)^t \frac{\partial G}{\partial \mu_i}, \\
\frac{\partial^2 \widehat{F}}{\partial \mu_i^2} &= \frac{t(t-1)}{\mu_i^2}\det(T)^t G + 2\frac{t}{\mu_i}\det(T)^t \frac{\partial G}{\partial \mu_i} + \det(T)^t \frac{\partial^2 G}{\partial \mu_i^2},
\end{aligned}
$$

and thus

$$
\begin{aligned}
&\mu_i(1-\mu_i)\frac{\partial^2 \widehat{F}}{\partial \mu_i^2} + \left( c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j} \right) \frac{\partial \widehat{F}}{\partial \mu_i} \\
&\qquad\qquad - \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j} \frac{\partial \widehat{F}}{\partial \mu_j} - ab\widehat{F} \\
&= \mu_i(1-\mu_i) \left( \frac{t(t-1)}{\mu_i^2}\det(T)^t G + 2\frac{t}{\mu_i}\det(T)^t \frac{\partial G}{\partial \mu_i} + \det(T)^t \frac{\partial^2 G}{\partial \mu_i^2} \right) \\
&\quad + \left( c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j} \right) \left( \frac{t}{\mu_i}\det(T)^t G + \det(T)^t \frac{\partial G}{\partial \mu_i} \right) \\
&\qquad\qquad - \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j} \left( \frac{t}{\mu_i}\det(T)^t G + \det(T)^t \frac{\partial G}{\partial \mu_i} \right) - ab\det(T)^t G \\
&= \det(T)^t \Bigg\{ \mu_i(1-\mu_i)\frac{\partial^2 G}{\partial \mu_i^2} - \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j} \frac{\partial G}{\partial \mu_j} \\
&\quad + \left( 2\mu_i(1-\mu_i)\frac{t}{\mu_i} + c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j} \right) \frac{\partial G}{\partial \mu_i} \\
&\quad + \Bigg[ \mu_i(1-\mu_i)\frac{t(t-1)}{\mu_i^2} + \left( c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j} \right) \frac{t}{\mu_i} \\
&\qquad\qquad - \frac{1}{2} \sum_{1\le j\le m}^{j\neq i} \frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j} \frac{t}{\mu_j} - ab \Bigg] G \Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
&= \det(T)^t \Bigg\{ \mu_i(1-\mu_i)\frac{\partial^2 G}{\partial \mu_i^2} - \frac{1}{2}\sum_{1\le j\le m}^{j\neq i} \frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j}\frac{\partial G}{\partial \mu_j} \\
&\qquad + \left( 2(1-\mu_i)t + c - \frac{m-1}{2} - (a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2}\sum_{1\le j\le m}^{j\neq i} \frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j}\right)\frac{\partial G}{\partial \mu_i} \\
&\qquad + \left[ \frac{t(t-1)}{\mu_i} - t(t-1) + (c-\frac{m-1}{2})\frac{t}{\mu_i} - (a+b+1-\frac{m-1}{2})t + \frac{1}{2}\sum_{1\le j\le m}^{j\neq i}(-1)t - ab \right] G \Bigg\} \\
&= \det(T)^t \Bigg\{ \mu_i(1-\mu_i)\frac{\partial^2 G}{\partial \mu_i^2} - \frac{1}{2}\sum_{1\le j\le m}^{j\neq i} \frac{\mu_j(1-\mu_j)}{\mu_i-\mu_j}\frac{\partial G}{\partial \mu_j} \\
&\qquad + \left( 2t + c - \frac{m-1}{2} - (2t+a+b+1-\frac{m-1}{2})\mu_i + \frac{1}{2}\sum_{1\le j\le m}^{j\neq i} \frac{\mu_i(1-\mu_i)}{\mu_i-\mu_j}\right)\frac{\partial G}{\partial \mu_i} \\
&\qquad\qquad - \left[t^2 + (a+b)t + ab\right] G \Bigg\} \\
&= 0,
\end{aligned}
$$

where the last equality holds because $G(\mu_1,\dots,\mu_m) = {}_2F_1(a+t, b+t; c+2t; T)$ satisfies a version of (7.2) after substitutions: $a \to a+t$, $b \to b+t$, and $c \to c+2t$.

Similarly $\widetilde{\widetilde{F}}(\mu_1,\dots,\mu_m) := \det(I-T)^{c-a-b}{}_2F_1(c-b, c-a; c-a-b+\frac{m+1}{2}; I-T)$ satisfies (7.2). Thus, any linear combination of $\widetilde{F}$ and $\widetilde{\widetilde{F}}$, such as the right-hand side of (7.1), also satisfies (7.2). It can be verified that the combination is symmetric with respect to $\mu_1,\dots,\mu_m$, and analytic at $T=0$. Therefore, by the uniqueness and $F(0)=1$, similarly to the discussion in [13, Section 3.9], we have (7.1). □

**Lemma 7.2.** *Suppose $p < (d+1)/2$. If $V^{(0)}$ satisfies the condition that $\mathcal{R}(V^{(0)})$ is uniformly sampled from $\mathbb{G}_p(\mathbb{R}^d)$, then for sufficiently large $d$ and $\delta \in [0,1]$, there exists a constant $C_p$, independent of $\delta$ and $d$, such that*

$$
\mathrm{P}\left\{V^{(0)} \in \mathbb{S}(C_p\delta^{-1}d^{1/2})\right\} \ge 1 - \delta^{p^2}. \tag{7.3}
$$

*Proof.* Let $1 \ge \sigma_1 \ge \cdots \ge \sigma_p \ge 0$ be the singular value of $\bar{V}^{(0)}$, and then $\sigma_i = \cos\theta_i$, where $\theta_i$ are the canonical angles between $\mathcal{R}(V^{(0)})$ and $\mathcal{R}(V_*)$ (recall (6.1)). By [1, Theorem 1], since $p < (d+1)/2$, the probability distribution function of $\sigma_p$ is

$$
\begin{aligned}
\mathrm{P}\left\{V^{(0)} \in \mathbb{S}(1/x)\right\} &= \mathrm{P}\{\sigma_p \ge x\} = \mathrm{P}\{\theta_p \le \arccos x\} \\
&= \frac{\mathit{\Gamma}(\frac{p+1}{2})\mathit{\Gamma}(\frac{d-p+1}{2})}{\mathit{\Gamma}(\frac{1}{2})\mathit{\Gamma}(\frac{d+1}{2})}(1-x^2)^{p(d-p)/2}{}_2F_1\left(\frac{d-p}{2}, \frac{1}{2}; \frac{d+1}{2}; (1-x^2)I_p\right).
\end{aligned}
$$

Set

$$
f_d := \frac{\mathit{\Gamma}_p(\frac{d+1}{2})\mathit{\Gamma}_p(\frac{p}{2})}{\mathit{\Gamma}_p(\frac{p+1}{2})\mathit{\Gamma}_p(\frac{d}{2})}, \quad g_d := \frac{\mathit{\Gamma}_p(\frac{d+1}{2})\mathit{\Gamma}_p(-\frac{p}{2})}{\mathit{\Gamma}_p(\frac{d-p}{2})\mathit{\Gamma}_p(\frac{1}{2})}.
$$

Here in defining $g_d$, although $\mathit{\Gamma}_p(-\frac{p}{2})$ and $\mathit{\Gamma}_p(\frac{1}{2})$ may be $\infty$, by analytic continuation, $\mathit{\Gamma}_p(-\frac{p}{2})/\mathit{\Gamma}_p(\frac{1}{2})$ is well-defined because

$$
\frac{\mathit{\Gamma}_p(-\frac{p}{2}+\epsilon)}{\mathit{\Gamma}_p(\frac{1}{2}+\epsilon)} = \prod_{i=1}^{p} \frac{\mathit{\Gamma}(-\frac{p}{2}-\frac{i-1}{2}+\epsilon)}{\mathit{\Gamma}(\frac{1}{2}-\frac{i-1}{2}+\epsilon)}
$$

$$
= \begin{cases} \prod_{i=1}^{p} \prod_{j=1}^{(p-1)/2} \frac{1}{\frac{-i}{2} - j + 1 + \epsilon} & \text{for odd } p, \\ \frac{\Gamma(\frac{1-2p}{2} + \epsilon)}{\Gamma(\frac{1}{2} + \epsilon)} \prod_{i=1}^{p-1} \prod_{j=1}^{p/2} \frac{1}{\frac{-i-1}{2} - j + 1 + \epsilon} & \text{for even } p, \end{cases}
$$

$$
\xrightarrow{\epsilon \to 0} \begin{cases} \prod_{i=1}^{p} \prod_{j=1}^{(p-1)/2} \frac{-2}{i + 2j - 2}, \\ \prod_{k=1}^{p} \frac{1}{\frac{-1}{2} - k + 1} \prod_{i=1}^{p-1} \prod_{j=1}^{p/2} \frac{-2}{i + 2j - 1} \end{cases}
$$

$$
= \begin{Bmatrix} \prod_{i=1}^{p} \prod_{j=1}^{(p-1)/2} \frac{-2}{i + 2j - 2} \\ \prod_{i=1}^{p+1} \prod_{j=1}^{p/2} \frac{-2}{i + 2j - 2} \end{Bmatrix} = \prod_{i=1}^{2\lfloor p/2 \rfloor + 1} \prod_{j=1}^{\lfloor p/2 \rfloor} \frac{-2}{i + 2j - 2}.
$$

Also,

$$
\frac{\Gamma_p(\frac{p}{2})}{\Gamma_p(\frac{p+1}{2})} = \prod_{i=1}^{p} \frac{\Gamma(\frac{p}{2} - \frac{i-1}{2})}{\Gamma(\frac{p+1}{2} - \frac{i-1}{2})} = \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{p+1}{2})}, \quad \frac{\Gamma_p(\frac{d}{2})}{\Gamma_p(\frac{d+1}{2})} = \prod_{i=1}^{p} \frac{\Gamma(\frac{d}{2} - \frac{i-1}{2})}{\Gamma(\frac{d+1}{2} - \frac{i-1}{2})} = \frac{\Gamma(\frac{d-p+1}{2})}{\Gamma(\frac{d+1}{2})},
$$

which implies $f_d = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{d+1}{2})}{\Gamma(\frac{p+1}{2})\Gamma(\frac{d-p+1}{2})}$. We have

$$
f_d^{-1} g_d = \frac{\Gamma_p(\frac{p+1}{2})\Gamma_p(\frac{d}{2})\Gamma_p(-\frac{p}{2})}{\Gamma_p(\frac{p}{2})\Gamma_p(\frac{d-p}{2})\Gamma_p(\frac{1}{2})} = \frac{\Gamma(\frac{p+1}{2})\Gamma_p(\frac{d}{2})\Gamma_p(-\frac{p}{2})}{\Gamma(\frac{1}{2})\Gamma_p(\frac{d-p}{2})\Gamma_p(\frac{1}{2})}.
$$

Note that

$$
\frac{\Gamma_p(\frac{d}{2})}{\Gamma_p(\frac{d-p}{2})} = \prod_{i=1}^{p} \frac{\Gamma(\frac{d}{2} - \frac{i-1}{2})}{\Gamma(\frac{d-p}{2} - \frac{i-1}{2})} = \begin{cases} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-p}{2})} \prod_{i=1}^{p} \prod_{j=1}^{(p-1)/2} \left( \frac{d-i}{2} - j \right) & \text{for odd } p, \\ \prod_{i=1}^{p} \prod_{j=1}^{p/2} \left( \frac{d-i}{2} - j \right) & \text{for even } p, \end{cases}
$$

and by $\lim_{n \to \infty} \frac{\Gamma(n+\alpha)}{\Gamma(n) n^{\alpha}} = 1$ for any $\alpha$ (see, e.g., [13, (16) of section 2.1]),

$$
\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-p}{2})} = \begin{cases} \frac{\Gamma(\frac{d-1}{2})(\frac{d-1}{2})^{1/2}[1 + \mathrm{o}(1)]}{\Gamma(\frac{d-1}{2})(\frac{d-1}{2})^{(1-p)/2}[1 + \mathrm{o}(1)]}, & \text{for odd } d, \\ \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2})(\frac{d}{2})^{-p/2}[1 + \mathrm{o}(1)]}, & \text{for even } d \end{cases} = \begin{cases} \left(\frac{d-1}{2}\right)^{p/2}[1 + \mathrm{o}(1)], \\ \left(\frac{d}{2}\right)^{p/2}[1 + \mathrm{o}(1)] \end{cases}
$$

which implies

$$
\frac{\Gamma_p(\frac{d}{2})}{\Gamma_p(\frac{d-p}{2})} = \left(\frac{d}{2}\right)^{p^2/2} [1 + \mathrm{o}(1)] \qquad \text{as } d \to \infty.
$$

Now we return to calculate the probability. By (7.1), we have

$$
{}_2F_1\left(\frac{d-p}{2}, \frac{1}{2}; \frac{d+1}{2}; (1 - x^2) I_p\right)
$$

$$= f_d\, {}_2F_1\left(\frac{d-p}{2}, \frac{1}{2}; \frac{1}{2}; x^2 I_p\right) + g_d \det(x^2 I_p)^{p/2}{}_2F_1(\frac{p+1}{2}, \frac{d}{2}; \frac{2p+1}{2}; x^2 I_p).$$

Also, [14, Definition 7.3.1 and Corollary 7.3.5] give us

$${}_2F_1\left(\frac{d-p}{2}, \frac{1}{2}; \frac{1}{2}; x^2 I_p\right) = {}_1F_0\left(\frac{d-p}{2}; x^2 I_p\right) = \det(I_p - x^2 I_p)^{-(d-p)/2} = (1-x^2)^{-p(d-p)/2}.$$

Therefore,

$$\mathrm{P}\Big\{V^{(0)} \in \mathbb{S}(1/x)\Big\} = 1 + f_d^{-1} g_d\, (1-x^2)^{p(d-p)/2} x^{p^2}{}_2F_1\left(\frac{p+1}{2}, \frac{d}{2}; \frac{2p+1}{2}; x^2 I_p\right).$$

Substituting $x = (\delta^{-1} d^{1/2})^{-1}$ and by [14, (8) of Section 7.4], we get as $d \to \infty$

$$\begin{aligned}
&\mathrm{P}\Big\{V^{(0)} \notin \mathbb{S}(\delta^{-1} d^{1/2})\Big\}\\
&= -f_d^{-1} g_d (1-\delta^2 d^{-1})^{p(d-p)/2} (\delta^2 d^{-1})^{p^2/2}{}_2F_1\left(\frac{p+1}{2}, \frac{d}{2}; \frac{2p+1}{2}; \frac{\delta^2}{d} I_p\right)\\
&= \frac{\Gamma(\frac{p+1}{2})\Gamma_p(-\frac{p}{2})}{-\Gamma(\frac{1}{2})\Gamma_p(\frac{1}{2})} \frac{\Gamma_p(\frac{d}{2})}{\Gamma_p(\frac{d-p}{2})} \left(1 - \frac{\delta^2}{d}\right)^{pd/2} \left(\frac{d}{\delta^2} - 1\right)^{-p^2/2} \left[{}_1F_1\left(\frac{p+1}{2}; \frac{2p+1}{2}; \frac{\delta^2}{2} I_p\right) + \mathrm{o}(1)\right]\\
&= \frac{\Gamma(\frac{p+1}{2})\Gamma_p(-\frac{p}{2})}{-\Gamma(\frac{1}{2})\Gamma_p(\frac{1}{2})} \left(\frac{d}{2}\right)^{p^2/2} [1 + \mathrm{o}(1)] \left[\exp\left(-\frac{p\delta^2}{2}\right) + \mathrm{o}(1)\right] \left[\frac{\delta^{p^2}}{d^{p^2/2}} + \mathrm{o}(1)\right]\\
&\qquad \times \left[{}_1F_1\left(\frac{p+1}{2}; \frac{2p+1}{2}; \frac{\delta^2}{2} I_p\right) + \mathrm{o}(1)\right]\\
&= \frac{\Gamma(\frac{p+1}{2})\Gamma_p(-\frac{p}{2})}{-\Gamma(\frac{1}{2})\Gamma_p(\frac{1}{2})} \exp\left(-\frac{p\delta^2}{2}\right) {}_1F_1\left(\frac{p+1}{2}; \frac{2p+1}{2}; \frac{\delta^2}{2} I_p\right) \delta^{p^2} [1 + \mathrm{o}(1)]\\
&\le \frac{\Gamma(\frac{p+1}{2})\Gamma_p(-\frac{p}{2})}{-\Gamma(\frac{1}{2})\Gamma_p(\frac{1}{2})} {}_1F_1\left(\frac{p+1}{2}; \frac{2p+1}{2}; \frac{1}{2} I_p\right) \delta^{p^2} 2\\
&=: C_p^{p^2} \delta^{p^2},
\end{aligned} \tag{7.4}$$

where the only inequality is guaranteed by ${}_1F_1\left(\frac{p+1}{2}; \frac{2p+1}{2}; \frac{\delta^2}{2} I_p\right) \le {}_1F_1\left(\frac{p+1}{2}; \frac{2p+1}{2}; \frac{1}{2} I_p\right)$, according to [14, Theorem 7.5.6]. Substituting $\delta/C_p$ for $\delta$, we infer from (7.4) that $\mathrm{P}\big\{V^{(0)} \notin \mathbb{S}(C_p \delta^{-1} d^{1/2})\big\} \le \delta^{p^2}$. The claim (7.3) is now a simple consequence. □

Now we are ready to prove Theorem 4.2.

*Proof of* Theorem 4.2. Define the event $\mathbb{H}'_* = \big\{V^{(0)} \in \mathbb{S}(C_p \delta^{-1} d^{1/2})\big\}$. Since $\mathcal{R}(V^{(0)})$ is uniformly sampled from $\mathbb{G}_p(\mathbb{R}^d)$, Lemma 7.2 says $\mathrm{P}\{\mathbb{H}'_*\} \ge 1 - \delta^{p^2}$. In the following, we will apply Theorem 4.1 with $\phi = C_p \delta^{-1}$, $\omega = (\sqrt{2}+1)\lambda_1 \delta^2$. Since Theorem 4.1 is valid on $\mathbb{H}'_*$, and

$$K[(2+\mathrm{e})d + p + 1] \exp(-C_{\nu\psi} \beta^{-\varepsilon}) \le \delta^{p^2},$$

there exists an event $\mathbb{H}$ with

$$\mathrm{P}\{\mathbb{H} \mid \mathbb{H}'_*\} \ge 1 - K[(2+\mathrm{e})d + p + 1] \exp(-C_{\nu\psi} \beta^{-\varepsilon}) \ge 1 - \delta^{p^2},$$

such that for any $n \in [N_{3/2 - 37\varepsilon/4}(\beta), K]$,

$$\begin{aligned}
\mathrm{E}\Big\{\|T^{(n)}\|_{\mathrm{F}}^2;\ \mathbb{H} \cap \mathbb{H}'_*\Big\} &= \mathrm{P}\{\mathbb{H}'_*\}\, \mathrm{E}\Big\{\|T^{(n)}\|_{\mathrm{F}}^2 \mathbf{1}_{\mathbb{H}} \,\Big|\, \mathbb{H}'_*\Big\}\\
&\le \mathrm{E}\Big\{\|T^{(n)}\|_{\mathrm{F}}^2 \mathbf{1}_{\mathbb{H}} \,\Big|\, \mathbb{H}'_*\Big\}
\end{aligned}$$

$$\le (1-\beta\gamma)^{2(n-1)} p C_p^2 \delta^{-2} d + \frac{32\psi^4\beta}{2-\lambda_1\beta}\varphi(p,d;\mathit{\Lambda}) + C_\circ \kappa^4\nu^2\eta_p^2\gamma^{-1} p\sqrt{d-p}\,\beta^{3/2-5\varepsilon}.$$

Let $\mathbb{H}_* = \mathbb{H} \cap \mathbb{H}'_*$ for which $\mathrm{P}\{\mathbb{H}_*\} = \mathrm{P}\{\mathbb{H} \mid \mathbb{H}'_*\}\,\mathrm{P}\{\mathbb{H}'_*\} \ge (1-\delta^{p^2})^2 \ge 1-2\delta^{p^2}$, as expected. □

Finally, we prove Theorem 4.3.

*Proof of* Theorem 4.3. First we examine the conditions of Theorem 4.2 to make sue that they are satisfied. It can be seen $\beta_* \to 0$ as $N_* \to \infty$. Thus, $\beta_*$ satisfies (4.3b) for sufficiently large $N_*$. We have

$$\begin{aligned}(1-\beta_*\gamma)^{N_*} &= \left(1-\frac{3\ln N_*}{2N_*}\right)^{N_*} = \exp(-\frac{3}{2}\ln N_*)[1+\mathrm{o}(1)] = N_*^{-3/2}[1+\mathrm{o}(1)]\\ &= \left(\frac{3\ln N_*}{2\gamma\beta_*}\right)^{-3/2}[1+\mathrm{o}(1)] = \frac{\beta_*^{3/2}\gamma^{3/2}}{(3/2)^{3/2}(\ln N_*)^{3/2}}[1+\mathrm{o}(1)] \le \beta_*^{3/2},\end{aligned}$$

which implies $N_* \ge N_{3/2}(\beta) \ge N_{3/2-9\varepsilon}(\beta)$.

The conclusion of the theorem will be a straightforward consequence if

$$\widetilde{C}(d,N_*,\delta) := \frac{(1-\beta_*\gamma)^{2(N_*-1)} p C_p^2\delta^{-2} d + \frac{32\psi^4\beta_*}{2-\lambda_1\beta_*}\varphi(p,d;\mathit{\Lambda}) + C_\circ\kappa^4\nu^2\eta_p^2\gamma^{-1}p\sqrt{d-p}\,\beta_*^{3/2-7\varepsilon}}{\frac{\varphi(p,d;\mathit{\Lambda})}{\lambda_p-\lambda_{p+1}}\frac{\ln N_*}{N_*}}$$

is bounded, say by $C_*(d,N_*,\delta)$ to be defined. In fact,

$$\begin{aligned}&\widetilde{C}(d,N_*,\delta)\\ &= \gamma\frac{N_*}{\ln N_*}\left[(1-\beta_*\gamma)^{2(N_*-1)} C_p^2\delta^{-2}\frac{pd}{\varphi(p,d;\mathit{\Lambda})} + \frac{32\psi^4\beta_*}{2-\lambda_1\beta_*} + \frac{C_\circ\kappa^4\nu^2\eta_p^2\gamma^{-1}p\sqrt{d-p}}{\varphi(p,d;\mathit{\Lambda})}\beta_*^{3/2-7\varepsilon}\right]\\ &\le \gamma\frac{N_*}{\ln N_*}\left[\frac{\beta_*^3}{(1-\beta_*\gamma)^2}C_p^2\delta^{-2}\frac{pd}{\varphi(p,d;\mathit{\Lambda})} + \frac{32\psi^4\beta_*}{2-\lambda_1\beta_*} + \frac{C_\circ\kappa^4\nu^2\eta_p^2\gamma^{-1}p\sqrt{d-p}}{\varphi(p,d;\mathit{\Lambda})}\beta_*^{3/2-7\varepsilon}\right]\\ &\qquad\qquad\left(\text{by } N_* \ge N_{3/2}\text{, or equivalently, } (1-\beta_*\gamma)^{N_*} \le \beta_*^{3/2}\right)\\ &\le \gamma\frac{N_*}{\ln N_*}\beta_*\left[\frac{\beta_*^2}{(1-\beta_*\gamma)^2}C_p^2\delta^{-2}\frac{d}{p}\frac{1}{\frac{\lambda_1\lambda_d}{\lambda_1-\lambda_d}} + \frac{32\psi^4}{2-\lambda_1\beta_*} + \frac{C_\circ\kappa^4\nu^2\eta_p^2\gamma^{-1}}{\sqrt{p}\frac{\lambda_1\lambda_d}{\lambda_1-\lambda_d}}\beta_*^{1/2-7\varepsilon}\right]\\ &\qquad\qquad\left(\text{by } \varphi(p,d;\mathit{\Lambda}) \ge \frac{p(d-p)\lambda_1\lambda_d}{\lambda_1-\lambda_d} \text{ and } d \ge 2p\right)\\ &\le \frac{3}{2}\left[\frac{\beta_*^{1+3\varepsilon}}{(1-\beta_*\gamma)^2}\frac{C_p^2}{p}\frac{\lambda_1-\lambda_d}{\lambda_1\lambda_d} + \frac{32\psi^4}{2-\lambda_1\beta_*} + C_\circ\kappa^4\nu^2\eta_p^2\gamma^{-1}p^{-1/2}\frac{\lambda_1-\lambda_d}{\lambda_1\lambda_d}\beta_*^{1/2-7\varepsilon}\right]\\ &\qquad\qquad\left(\text{by } d\beta_*^{1-3\varepsilon} \le \delta^2\right)\\ &=: C_*(d,N_*,\delta).\end{aligned}$$

Since $\beta_* \le 1$ and $\beta_*\gamma \le \lambda_1\beta_* \le \sqrt{2}-1$, we have

$$C_*(d,N_*,\delta) \le \frac{3}{2}\left[\frac{C_p^2}{2(3-2\sqrt{2})p}\frac{\lambda_1-\lambda_d}{\lambda_1\lambda_d} + \frac{32\psi^4}{3-\sqrt{2}} + \frac{C_\circ\kappa^4\nu^2\eta_p^2(\lambda_1-\lambda_d)}{p^{1/2}\gamma\lambda_1\lambda_d}\right],$$

and also $C_*(d,N_*,\delta) \to 24\psi^4$ as $d\to\infty, N_*\to\infty$, as was to be shown. □

# 8 Conclusion

We have presented a detailed convergence analysis for the multidimensional subspace online PCA iteration with sub-Gaussian samples, following the recent work [11] by Li, Wang, Liu, and Zhang who considered only the one-dimensional case, i.e., the most significant principal component. Our results bear similar forms to theirs and when applied to the one-dimensional case yield estimates of essentially the same quality, as expected. As we embarked on the analysis presented in this paper, we found that a straightforward extension of the analysis in [11] was not possible because of the involvement of a cot-matrix of dimension higher than 1 in the multidimensional case but just a scalar in the one-dimensional case.

Our results yields an explicit convergence rate, and it is *nearly optimal* because it nearly attains the minimax information lower bound for sub-Gaussian PCA under a constraint, as well as *nearly global* because the finite sample error bound holds with high probability if the initial value is uniformly sampled from the Grassmann manifold.

# References


[1] P.-A. Absil, A. Edelman, and P. Koev, *On the largest principal angle between random subspaces*, Linear Algebra Appl., 414 (2006), pp. 288–294.

[2] A. Balsubramani, S. Dasgupta, and Y. Freund, *The Fast Convergence of Incremental PCA*, ArXiv e-prints, (2015). arXiv: 1501.03796.

[3] Y. Chikuse, *Statistics on special manifolds*, no. 174 in Lecture Notes in Statistics, Springer, New York, 2003.

[4] C. De Sa, K. Olukotun, and C. Ré, *Global Convergence of Stochastic Gradient Descent for Some Non-convex Matrix Problems*, ArXiv e-prints, (2014). arXiv: 1411.1134.

[5] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[6] D. Garber, E. Hazan, C. Jin, S. M. Kakade, C. Musco, P. Netrapalli, and A. Sidford, *Faster eigenvector computation via shift-and-invert preconditioning*, ArXiv e-prints, (2016). arXiv: 1605.08754.

[7] R. A. Horn and R. Mathias, *An analog of the Cauchy-Schwarz inequality for Hadamard products and unitarily invariant norms*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 481–498.

[8] H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, 24 (1933), pp. 417–441.

[9] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford, *Streaming PCA: Matching Matrix Bernstein and Near-Optimal Finite Sample Guarantees for Oja's Algorithm*, ArXiv e-prints, (2016). arXiv: 1602.06929.

[10] A. James, *Normal multivariate analysis and the orthogonal group*, Ann. Math. Statist., 25 (1954), pp. 40–75.

[11] C. J. Li, M. Wang, H. Liu, and T. Zhang, *Near-optimal stochastic approximation for online principal component estimation*, Math. Program., Ser. B, (2017). DOI: 10.1007/s10107-017-1182-z.

[12] X. Liang, Z.-C. Guo, R.-C. Li, and W.-W. Lin, *Nearly optimal stochastic approximation for online principal subspace estimation*, ArXiv e-prints, (2017). 35 pages, arXiv: 1711.06644v1.

[13] Y. L. Luke, *The special functions and their approximations*, no. 53, vol. I in Mathematics in Science and Engineering, Academic Press, New York, 1969.

[14] R. J. Muirhead, *Aspects of multivariate statistical theory*, Wiley series in probability and mathematical statistics, John Wiley & Sons, New York, 1982.

[15] E. Oja, *Simplified neuron model as a principal component analyzer*, J. Mathematical Biology, 15 (1982), pp. 267–273.

[16] E. Oja and J. Karhunen, *On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix*, J. Math. Anal. and Appl., 106 (1985), pp. 69–84.

[17] K. Pearson, F.R.S., *On lines and planes of closest fit to systems of points in space*, Philosophical Magazine, 2 (1901), pp. 559–572.

[18] O. Shamir, *Convergence of Stochastic Gradient Descent for PCA*, ArXiv e-prints, (2015). arXiv: 1509.09002.

[19] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

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

[21] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer, New York, 1996.

[22] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed sensing: theory and applications, Y. C. Eldar and G. Kutyniok, eds., Cambridge University Press, New York, 2012, ch. 5, pp. 210–268.

[23] V. Q. Vu and J. Lei, *Minimax sparse principal subspace estimation in high dimensions*, Ann. Statist., 41 (2013), pp. 2905–2947.