A Discussion on the Cumulative Metric

As discussed on the Section 3, we use $d(\mathrm{ran}(\mathbf{A}_T), \phi_{\mathcal{X}})$ as the performance metric. We can make results with respect to the cumulative error $\sum_{t=1}^{T} \mathrm{coeff}^{T-t}\mathrm{metric}(\mathbf{A}_t, \phi_{\mathcal{X}})$ ($\mathrm{coeff} \in (0,1]$) or $\sum_{t=1}^{T} \mathrm{metric}(\mathbf{A}_t, \phi_{\mathcal{X}})/T$ from our results of $d(\mathbf{A}_T, \phi_{\mathcal{X}})$. Note that the distance between actual and estimated space (check the Figure 3b for visualization) does not tend to 0 as $T \to \infty$ on both algorithms when $\Gamma > 0$, which is inevitable as the non-zero lower bound of the Theorem 1 suggests.

To extend our results to the cumulative metric, for the Theorem 1, we can construct a new $\mathcal{A}_1, \dots \mathcal{A}_{M'}$ for estimating the fundamental lower bound; and for the Theorem 2 and the Theorem 3, our discussion can easily be applied to each $t \in [T]$ (not only to the termination time T) using the union bound. However, we have to multiply $\log T$ to the result of $d(\mathbf{A}_T, \phi_{\mathcal{X}})$ since we use the union bound to have bounded noise matrices from column blocks and the number of blocks linearly increases to T. Therefore, when $T \to \infty$, our results become the trivial bound $\sum_{t=1}^{T} \mathrm{metric}(\mathbf{A}_t, \phi_{\mathcal{X}})/T \le 1$.

B Rate Optimality

Our Assumption 1 at the Section 5 allows us to write the clear statement on the performance of two algorithms. Furthermore, by simply truncating the multivariate Gaussian distribution of $\mathbf{x}_t$ ($\forall t$) with high probability $1-1/T$ (we denote this event as $\mathfrak{E}$), we may adopt Assumption 1 on the model with equation (1), for suitable $\mathcal{M}$ and $\mathcal{V}$.

Here we will discuss about the construction of $\mathfrak{E}$, $\mathcal{M}$, and $\mathcal{V}$. To apply the Assumption 1 on the Gaussian model, we define the high-probability($\ge 1-1/T$) event $\mathfrak{E}$ under $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p})$, where $(\mathbf{A}_t)_{t=1}^T \in \mathrm{Tu}(\delta, \Gamma)$. Precisely, we define the event $\mathfrak{E}$ at the main paper as below:

Definition 4. Formal Version. *Let* $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p})$ *and* $\mathrm{SVD}(\mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p}) = \mathbf{U}_t\mathbf{D}_t\mathbf{U}_t^\top$. *We define the event* $\mathfrak{E}$ *as follows:*

$$\mathfrak{E} := \forall t \in [T] \ : \ \mathbf{z}_t = \mathbf{D}_t^{-1/2}\mathbf{U}_t^\top\mathbf{x}_t \in [-\rho, \rho]^p \ , \ \text{for} \quad \rho = \sqrt{2\log(2pT^2)}\,.$$

Under the event $\mathfrak{E}$, $\|\mathbf{z}_t\|^2$ is bounded by $p\rho^2 = \tilde{\Theta}(p)$. Furthermore, by well-known formula on the variance of truncated normal distribution [13], we have $\mathbb{E}[\mathbf{z}_t\mathbf{z}_t^\top|\mathfrak{E}] = (1-\nu(pT^2))\mathbf{I}_{p\times p}$ and $\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}] = (1-\nu(pT^2))\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top]$ where:

$$\nu(x) := \frac{1}{2\sqrt{\pi}}\frac{\sqrt{\log(x/2)}}{x-1} + \frac{1}{2\pi}\frac{x^2}{(x-1)^4}\,, \tag{13}$$

because $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p\times p})$. Since $\nu(x) = \mathcal{O}(x^{-1}\log(x/2))$ for $x \gg 1$, we have:

$$\nu(pT^2) = \mathcal{O}\left(\frac{\log(pT^2)}{pT^2}\right).$$

B.1 Properties under $\mathfrak{E}$

Under the truncation event $\mathfrak{E}$, the expectation of the covariance estimator becomes different. Therefore, the properties which affect to the convergence are also differed, but those are not significant for the sufficiently large T. We first define:

$$\mathbf{A}_t^{\mathfrak{E}} := \sqrt{1-\nu(pT^2)}\mathbf{A}_t = \sqrt{1-\mathcal{O}\left(\frac{\log(pT^2)}{pT^2}\right)}\mathbf{A}_t\,.$$

Let us first define new parameters as:

$$\{(\sigma^{\mathfrak{E}})^2, \delta^{\mathfrak{E}}, \tilde{\delta}^{\mathfrak{E}}, \Gamma^{\mathfrak{E}}\} = (1-\nu(pT^2))\{\sigma^2, \delta, \tilde{\delta}, \Gamma\}$$

Then, since $\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}] = (1-\nu(pT^2))(\mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p}) = \mathbf{A}_t^{\mathfrak{E}}(\mathbf{A}_t^{\mathfrak{E}})^\top + (\sigma^{\mathfrak{E}})^2\mathbf{I}_{p\times p}$, noise magnitude ($\sigma^2$), spectral gap of $\mathbf{A}_t\mathbf{A}_t^\top$ (δ), largest spectrum of $\mathbf{A}_t\mathbf{A}_t^\top$ ($\tilde{\delta}$), and distance between covariance matrix (Γ) should be replaced by $(1-\nu(pT^2))$-scaled new parameters. However, if T is sufficiently large, we restore the original parameters with logarithmic multiplicative factor.

B.2 Applying Assumption 1 under $\mathfrak{E}$

In this section, we will show that **under the event $\mathfrak{E}$, the equation (1) satisfies the Assumption 1 (for probability greater than $1-1/T$)**, with parameters:

1. $\mathcal{M} = (p\tilde{\delta} + k\sigma^2)\rho^2 + \tilde{\delta}^{\mathfrak{E}} + (\sigma^{\mathfrak{E}})^2$,
2. $\mathcal{V} = (\tilde{\delta}^{\mathfrak{E}} + (\sigma^{\mathfrak{E}})^2)\mathcal{M}$.

For the first statement, under $\mathfrak{E}$ we have:

$$\begin{aligned}\|\mathbf{x}_t\mathbf{x}_t^\top - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}]\| &= \|\mathbf{D}_t^{1/2}(\mathbf{z}_t\mathbf{z}_t^\top - \mathbb{E}[\mathbf{z}_t\mathbf{z}_t^\top|\mathfrak{E}])\mathbf{D}_t^{1/2}\| \\ &\le \|\mathbf{D}_t^{1/2}\mathbf{z}_t\|^2 + \|\mathbf{D}_t^{1/2}(1-\gamma(\mathfrak{E}/pT))\mathbf{I}_{p\times p}\mathbf{D}_t^{1/2}\| \\ &\le (p\tilde{\delta} + k\sigma^2)(\rho)^2 + \tilde{\delta}^{\mathfrak{E}} + (\sigma^{\mathfrak{E}})^2\,.\end{aligned}$$

Finally, for second argument:

$$\begin{aligned}&\|\mathbb{E}[(\mathbf{x}_t\mathbf{x}_t^\top - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}])(\mathbf{x}_t\mathbf{x}_t^\top - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}])|\mathfrak{E}]\| \\ &\quad= \|\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{x}_t\mathbf{x}_t^\top - \mathbf{x}_t\mathbf{x}_t^\top\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}]|\mathfrak{E}]\| \\ &\quad\le \|\mathbf{x}_t\|^2\|\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}]\| + \|\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top|\mathfrak{E}]\|^2 \\ &\quad\le (p\tilde{\delta} + k\sigma^2)(\rho)^2(\tilde{\delta}^{\mathfrak{E}} + (\sigma^{\mathfrak{E}})^2) + (\tilde{\delta}^{\mathfrak{E}} + (\sigma^{\mathfrak{E}})^2)^2 = (\tilde{\delta}^{\mathfrak{E}} + (\sigma^{\mathfrak{E}})^2)\mathcal{M}\,.\end{aligned}$$

B.3 Discussion

The temporal uncertainty set $\mathrm{Tu}(\delta, \Gamma)$ does not have any information or clue about $\tilde{\delta}$, which is the upper bound for the first singular value of $\mathbf{A}_t\mathbf{A}_t^\top$. Therefore, let us assume $\tilde{\delta} = \Theta(\delta)$. Then we have:

$$\begin{aligned}\mathcal{M} &\le 2(p\delta + k\sigma^2)\log(2pT^2) \\ \mathcal{V} &\le 2(\delta+\sigma^2)(p\delta + k\sigma^2)\log(2pT^2)\end{aligned}$$

On this case, the first term of the upper bound for the noisy power method becomes:

$$\mathcal{O}\left(\frac{((\delta+\sigma^2)(p\delta+k\sigma^2)\Gamma\log(2pT^2)\log(2pT^2))^{1/3}}{\delta}\right), \text{ or } \tilde{\mathcal{O}}\left(\frac{((\delta+\sigma^2)(p\delta+k\sigma^2)\Gamma)^{1/3}}{\delta}\right).$$

By similar procedure, we can also find that our guarantee for Oja's algorithm is sub-optimal.

C Preparation for Detailed Proofs

C.1 Notation Table

Table 1: Table of Notations throughout the appendix. We omit notations what we already defined at the Section 1.

Parameters related to the environment:	
T	time horizon length
k	number of principal components
p	dimension of observation vectors
σ	magnitude of observation noise
$\mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p}$	covariance matrix at the time t
δ	lower bound of spectral gap between kth and $k+1$th singular value of $\mathbf{A}_t\mathbf{A}_t^\top$
$\tilde{\delta}$	upper bound of the largest spectrum of $\mathbf{A}_t\mathbf{A}_t^\top$
Γ	upper bound of the $\|\mathbf{A}_t\mathbf{A}_t^\top - \mathbf{A}_{t+1}\mathbf{A}_{t+1}^\top\|$
Parameter for the algorithms:	
B	block size for the noisy power method
ζ	learning rate for the Oja's algorithm
ζ_{opt} , B_{opt}	optimal learning parameter when there exists covariance shifts (Tu(δ,Γ))
$L(\simeq T/B)$	number of iteration in the noisy power method / for the Oja's algorithm, B is always the virtual block size corresponding to the optimal parameter ζ_{opt}
Related to the rate optimality (Section B):	
$\mathfrak{E}$	high-probability event under the spiked covariance model setting ($\mathbb{P}[\mathfrak{E}] \geq 1 - 1/T$), for bounding the norm of observation vectors
$\delta^{\mathfrak{E}}$, $\sigma^{\mathfrak{E}}$, $\Gamma^{\mathfrak{E}}$, , $\tilde{\delta}^{\mathfrak{E}}$	corresponding parameters when we assume the event $\mathfrak{E}$ (since we have different expectation for $\mathbf{x}_t\mathbf{x}_t^\top$ under the $\mathfrak{E}$)
$\mathcal{M}$, $\mathcal{V}$	probabilistic upper bound, which plays the role of $\mathcal{M}$ and $\mathcal{V}$ on the spiked covariance model setting
Related to the proof of Theorem 1 (Section D):	
$\mathrm{St}_k(\mathbb{R}^p)$	Stiefel manifold, which consists with the matrix $\mathbf{M} \in \mathbb{R}^{p\times k}$ satisfying $\mathbf{M}^\top\mathbf{M} = \mathbf{I}_{k\times k}$
$\mathcal{G}_k(\mathbb{R}^p)$	Grassmann manifold, which is the Riemannian manifold with k-dimensional subspace in the $\mathbb{R}^p$
$[\mathbf{M}] \in \mathcal{G}_k(\mathbb{R}^p)$	k-dimensional subspace generated with the columns of $\mathbf{M} \in \mathrm{St}_k(\mathbb{R}^p)$
$\mathcal{G}_{[\mathbf{M}]\to[\mathbf{N}]}(\Psi')$	principal rotation from $[\mathbf{M}]$ to $[\mathbf{N}]$
s	latent value for the lower bound

C.2 Technical Lemmas

Lemma 3 (Theorem 2.6.1, [28]). *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathbb{R}^p$, such that* $\dim(\mathcal{S}_1)$=$\dim(\mathcal{S}_2)$. *We define the distance between these two subspaces ($\mathcal{S}_1$,$\mathcal{S}_2$) by $\|\mathbf{P}_1 - \mathbf{P}_2\|$, where $\mathbf{P}_i, i = 1, 2$ is the orthogonal projection onto $\mathcal{S}_i$ $(i = 1, 2)$. Moreover, suppose $\mathbf{M} = [\underset{k}{\mathbf{M}_1}\ \underset{p-k}{\mathbf{M}_2}]$, $\mathbf{N} = [\underset{k}{\mathbf{N}_1}\ \underset{p-k}{\mathbf{N}_2}]$ are $p \times p$ orthogonal matrices. If $\mathcal{S}_1 = \mathrm{ran}(\mathbf{M}_1)$ and $\mathcal{S}_2 = \mathrm{ran}(\mathbf{N}_1)$, then:*

$$\mathrm{dist}(\mathcal{S}_1, \mathcal{S}_2) = \|\mathbf{M}_1^\top \mathbf{N}_2\| = \|\mathbf{M}_2^\top \mathbf{N}_1\| .$$

Lemma 4 (Davis-Kahan $\sin(\theta)$ theorem; Theorem VII.3.1, [8]). *For given symmetric matrices $\mathbf{M}, \mathbf{N}$ with singular value decomposition* $\mathrm{SVD}(\mathbf{M}) = \mathbf{UDU}$ *and* $\mathrm{SVD}(\mathbf{M} + \mathbf{N}) = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{U}}$, *we have:*

$$\|\mathbf{U}_{1:k}\mathbf{U}_{1:k}^\top - \hat{\mathbf{U}}_{1:k}\hat{\mathbf{U}}_{1:k}^\top\| \leq \frac{\|\mathbf{N}\|}{s_k(\mathbf{M}) - s_{k+1}(\mathbf{M}) + \|\mathbf{N}\|} .$$

Lemma 5 (Weyl's theorem). *For any $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{p\times k}$ and $1 \leq i \leq \min(p, k)$,*

$$s_i(\mathbf{M} + \mathbf{N}) \leq s_i(\mathbf{M}) + s_1(\mathbf{N}) .$$

Lemma 6 (Sub-additivity of rank). *For any $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{p\times k}$,*

$$\mathrm{rk}(\mathbf{M} + \mathbf{N}) \leq \mathrm{rk}(\mathbf{M}) + \mathrm{rk}(\mathbf{N}) .$$

C.3 Grassmann Manifold

To effectively handle the k-dimensional subspace of Euclidean space, we would like to consider the Grassmann manifold [7, 19, 20, 41, 47].

Definition 4 (Grassmann manifold). Grassmann manifold $\mathcal{G}_k(\mathbb{R}^p)$ is the $k(p-k)$ dimensional Riemannian manifold with k-dimensional subspace in $\mathbb{R}^p$ as elements. For example, $\mathbb{RP}^{p-1}$ is topologically isomorphic with Grassmann manifold $\mathcal{G}_1(\mathbb{R}^p)$. The elements in $\mathcal{G}_k(\mathbb{R}^p)$ are often expressed as the equivalence class $[\mathbf{M}]$ of $p \times k$ orthogonal matrix ($\mathbf{M} \in \mathrm{St}_k(\mathbb{R}^p)$). Here, each class is a collection of orthogonal matrices sharing the same column space. A necessary and sufficient condition for both elements of $\mathrm{St}_k(\mathbb{R}^p)$ to have the same column space is that the associated projection matrices are the same. That is,

$$\mathbf{M}_1 \sim \mathbf{M}_2\ (\mathbf{M}_1,\ \mathbf{M}_2 \in \mathrm{St}_k(\mathbb{R}^p)) \Longleftrightarrow \mathbf{M}_1\mathbf{M}_1^\top = \mathbf{M}_2\mathbf{M}_2^\top .$$

Definition 5 (Principal angle). We can define **k principal angles** between two elements in $\mathcal{G}_k(\mathbb{R}^p)$. This is clear generalization of an angle $\in [0, \pi/2]$ between two 1d-lines in $\mathbb{R}^p$.

(a) Let us assume $[\mathbf{M}], [\mathbf{N}] \in \mathcal{G}_k(\mathbb{R}^p)$ ($\mathbf{M}, \mathbf{N} \in \mathrm{St}_k(\mathbb{R}^p)$). Then we define the principal angle by the inverse cosine of the diagonal matrix $\mathbf{\Sigma}$ in $\mathrm{SVD}(\mathbf{M}^\top\mathbf{N}) = \mathbf{U\Sigma V}^\top$. Therefore, the principal angle can be treated as the k-dimensional vector in $[0, \pi/2]^k$.

(b) The principle angle is well defined in terms of the fact that:

$$\text{for } \mathbf{M}_1, \mathbf{M}_2, \mathbf{N}_1, \mathbf{N}_2 \in \mathrm{St}_k(\mathbb{R}^p) \quad \text{s.t} \quad \mathbf{M}_1 \sim \mathbf{M}_2 \ \text{ and } \ \mathbf{N}_1 \sim \mathbf{N}_2 ,$$

$\mathbf{M}_1^\top\mathbf{N}_1$ and $\mathbf{M}_2^\top\mathbf{N}_2$ have same set of singular values.

(c) Generally, we mean the set of angles or k dimensional vector or $k \times k$ diagonal matrix when we denote principal angle. Specifically, We denote $\Psi = \cos^{-1}(\mathrm{diag}(\mathbf{\Sigma}))$ as the vector in $\mathbb{R}^k$ with the principal angles as elements. On the other hand, when we apply trigonometric function on Ψ, we treat the result as a diagonal matrix.

To measure the distance between two elements in $\mathcal{G}_k(\mathbb{R}^p)$, we define the projection 2-distance $d_2(\cdot, \cdot)$, which is the operator 2-norm between projectors of element in $\mathcal{G}_k(\mathbb{R}^p)$. We summarize about the projection distance as follows.

Definition 6 (Projection 2-distance). Let $[\mathbf{U}], [\mathbf{V}] \in \mathcal{G}_k(\mathbb{R}^p)$ ($\mathbf{U}, \mathbf{V} \in \mathrm{St}_k(\mathbb{R}^p)$). Assumte that the principal angles between $[\mathbf{U}]$ and $[\mathbf{V}]$ are $\Psi = (\psi_i)_i \in [0, \pi/2]^k$. We define the projection 2-distance as:

$$d_2([\mathbf{U}], [\mathbf{V}]) := \|\mathbf{UU}^\top - \mathbf{VV}^\top\| = \|\sin\Psi\|_\infty = \max_{1\leq i\leq k} \sin\psi_i .$$

Note also that the following equality holds for $\mathbf{U}$ and $\mathbf{V}$ in $\mathrm{St}_k(\mathbb{R}^p)$:

$$d_2([\mathbf{U}]\,,[\mathbf{V}]) = d(\mathbf{U},\mathbf{V})(= d(\mathrm{ran}(\mathbf{U}),\mathrm{ran}(\mathbf{V}))\,.$$

We consider the r-ball covering with respect to the projection 2-distance. This result can be derived from the slight variation of the proof on the chordal metric r-ball [20] (Precisely, we can get the proof by redefining the integral domain D_r).

Proposition 1 (Projection 2-distance ball in $\mathcal{G}_k(\mathbb{R}^p)$)**.** *Let us define $\mathcal{B}([\mathbf{M}],r) \in \mathcal{G}_k(\mathbb{R}^p)(r \in (0,1))$ as the d_2-ball with radius r and center $[\mathbf{M}]$. Then there exists a measure μ on $\mathcal{G}_k(\mathbb{R}^p)$ such that $\mu(\mathcal{B}([\mathbf{M}],r))=\mu(r)$ for all $\mathbf{M}] \in \mathcal{G}_k(\mathbb{R}^p)$ where $\mu(r)$ satisfies:*

$$c_{p,k}r^{k(p-k)} \leq \mu(r) \leq \frac{c_{p,k}r^{k(p-k)}}{(1-r^2)^{k/2}}\,.$$

Here, $c_{p,k}$ is a constant determined with p and k.

While the rotation between two vectors is self-explanatory, the rotation between two orthogonal planes is not intuitive. Because we have **k** principal angles, we may consider **k-dimensional rotation**. Let $\Psi = \mathrm{diag}(\psi_i)_{i=1}^k$ be the principal angles between $[\mathbf{M}]$ and $[\mathbf{N}]$ ($\mathbf{M}\,,\mathbf{N} \in \mathrm{St}_k(\mathbb{R}^p)$). The following definition defines $\mathcal{G}$-mapping, which faithfully generalize traditional rotation.

Definition 7 (Principal rotation)**.** Let us assume $[\mathbf{M}]\,,[\mathbf{N}] \in \mathcal{G}_k(\mathbb{R}^p)$. Then, we have $\mathcal{G}$-mapping

$$\mathcal{G}_{[\mathbf{M}]\to[\mathbf{N}]} \;:\; [0\,,\psi_1] \times [0\,,\psi_2] \times \cdots \times [0\,,\psi_k] \to \mathcal{G}_k(\mathbb{R}^p)\,,$$

which satisfies:

- $\Psi = (\psi_i)_{i=1}^k$ is principal angle between $[\mathbf{M}]$ and $[\mathbf{N}]$.
- $\mathcal{G}_{[\mathbf{M}]\to[\mathbf{N}]}(0) = [\mathbf{M}]\,$, and $\mathcal{G}_{[\mathbf{M}]\to[\mathbf{N}]}(\Psi) = [\mathbf{N}]$.
- Principal angle between $\mathcal{G}(\Psi^1)$ and $\mathcal{G}(\Psi^2)$ is $|\,\Psi^1 - \Psi^2\,|$.

From the last property, for $0 \preceq \Psi^1 = (\psi_i^1)_{i=1}^k$, $\Psi^2 = (\psi_i^2)_{i=1}^k \preceq \Psi$, we have:

- $d_2(\mathcal{G}(\Psi^1)\,,\mathcal{G}(\Psi^2)) = \max_{1\leq i\leq k} |\sin(\psi_i^1 - \psi_i^2)|\,.$

Proof. Consider the following singular value decomposition:

$$\mathbf{M}^\top\mathbf{N} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \Rightarrow (\mathbf{M}\mathbf{U})^\top(\mathbf{N}\mathbf{V}) = \mathbf{\Sigma} = \cos\Psi\,. \quad (0 \preceq \Psi \preceq \pi/2)$$

Note that $\mathbf{M}\mathbf{U} \sim \mathbf{M}\,,\mathbf{N}\mathbf{V} \sim \mathbf{N}$. We provide the $\mathcal{G}$-mapping from $[\mathbf{M}]$ to $[\mathbf{N}]$:

$$\mathcal{G}_{[\mathbf{M}]\to[\mathbf{N}]}(\Psi) = [\,\mathbf{M}\mathbf{U}\cos\Psi' + (-\mathbf{M}\mathbf{U}\cot\Psi + \mathbf{N}\mathbf{V}\csc\Psi)\sin\Psi'\,]\,. \quad (0 \preceq \Psi' \preceq \Psi)$$

If $\psi_i = \psi_i' = 0$, we treat $\sin\psi_i'\,/\sin\psi_i$ as 1. Since the first two requirements are obvious, we will show that the above formulation satisfies the third condition. From

$$(\mathbf{M}\mathbf{U}\cos\Psi')^\top(-\mathbf{M}\mathbf{U}\cot\Psi + \mathbf{N}\mathbf{V}\csc\Psi)\sin\Psi' = \cos\Psi'(-\cot\Psi + \cos\Psi\csc\Psi)\sin\Psi' = 0\,,$$

and

$$\begin{aligned}
&\sin\Psi^1(-\mathbf{M}\mathbf{U}\cot\Psi + \mathbf{N}\mathbf{V}\csc\Psi)^\top(-\mathbf{M}\mathbf{U}\cot\Psi + \mathbf{N}\mathbf{V}\csc\Psi)\sin\Psi^2\\
=&\sin\Psi^1\csc\Psi(-\mathbf{M}\mathbf{U}\cos\Psi + \mathbf{N}\mathbf{V})^\top(-\mathbf{M}\mathbf{U}\cos\Psi + \mathbf{N}\mathbf{V})\csc\Psi\sin\Psi^2\\
=&\sin\Psi^1\csc\Psi(\mathbf{I}_{p\times p} - \cos^2\Psi)\csc\Psi\sin\Psi^2 = \sin\Psi\sin\Psi^2\,,
\end{aligned}$$

we have:

$$\mathcal{G}(\Psi^1)^\top\mathcal{G}(\Psi^2) = \cos\Psi^1\cos\Psi^2 + \sin\Psi^1\sin\Psi^2 = \cos(\Psi^1 - \Psi^2)\,.$$

Therefore, the principal angles between $\mathcal{G}(\Psi^1)$ and $\mathcal{G}(\Psi^2)$ are $|\Psi^1 - \Psi^2|$. The last property is immediate from the third. □

D Proof of Theorem 1

D.1 Reduction to finite number of hypotheses

Let us define

$$s := \sqrt[3]{\frac{\log(3/2)}{2160}} \left(\left(\frac{\Gamma}{\delta}\right)^{1/3} \left(\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2}\right)^{1/3} + \frac{1}{\sqrt{T}} \left(\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2}\right)^{1/2} \right) > 0,$$

and assume $s < 1/3$, and $p > 2k+1$ (Note that we assume $p \gg k$). Then we have:

$$\mathbb{E}_{\mathcal{X}\sim\mathbb{P}_{\mathcal{A}}} \Big(d(\text{ran}(\mathbf{A}_T)\,,\,\phi_{\mathcal{X}}) \Big) \geq s\,\mathbb{P}_{\mathcal{A}} \big(d(\text{ran}(\mathbf{A}_T)\,,\,\phi_{\mathcal{X}}) \geq s \big)\,,$$

and

$$\mathcal{R}^* = \inf_{\phi} \sup_{\mathcal{A}\in \text{Tu}(\delta,\Gamma)} \mathbb{E}_{\mathcal{X}\sim\mathbb{P}_{\mathcal{A}}} \big(d\big(\text{ran}(\mathbf{A}_T)\,,\,\phi_{\mathcal{X}}\big) \big) \geq s \cdot \inf_{\phi} \sup_{\mathcal{A}\in \text{Tu}(\delta,\Gamma)} \mathbb{P}_{\mathcal{A}} \Big(d\big(\text{ran}(\mathbf{A}_T)\,,\,\phi_{\mathcal{X}}\big) \geq s \Big)\,.$$

Since we are considering supremum over sequences in $\text{Tu}(\delta,\Gamma)$, we immediately have that

$$\inf_{\phi} \sup_{\mathcal{A}\in \text{Tu}(\delta,\Gamma)} \mathbb{P}_{\mathcal{A}} \big(d(\text{ran}(\mathbf{A}_T)\,,\,\phi_{\mathcal{X}}) \geq s \big) \geq \inf_{\phi} \sup_{\mathcal{A}\in\{\mathcal{A}_0,\dots,\mathcal{A}_M\}} \mathbb{P}_{\mathcal{A}} \big(d(\text{ran}(\mathbf{A}_T)\,,\,\phi_{\mathcal{X}}) \geq s \big)\,,$$

where $\Big\{ \mathcal{A}_i = \{\mathbf{A}_1^{(i)}, \mathbf{A}_2^{(i)}, \dots, \mathbf{A}_T^{(i)}\} \Big\}_{i=1}^{T}$ is an appropriately chosen subset of $\text{Tu}(\delta,\Gamma)$ of size $(M+1)$ which will be defined with the construction below (Section D.2).

D.2 Constructing $\mathbf{A}_t^{(i)}$

On the construction, we construct a set with size $M+1$, $\{\mathcal{A}_i\}_{i=0}^{M} \subset \text{Tu}(\delta,\Gamma)$, where $\mathbf{A}_t^{(i)}$ ($\forall t \in [T]$ and $\forall i \in [M]\cup\{0\}$) to satisfy:

$$s_1(\mathbf{A}_t^{(i)}\mathbf{A}_t^{(i)\top}) = s_2(\mathbf{A}_t^{(i)}\mathbf{A}_t^{(i)\top}) = \cdots = s_k(\mathbf{A}_t^{(i)}\mathbf{A}_t^{(i)\top}) = \delta\,.$$

Therefore, for the rest of Section D, **we treat $\mathbf{A}_t^{(i)}$ as orthogonal matrix in $\mathbb{R}^{p\times k}$ for simplicity**. We consider the notation $\mathbf{A}_t^{(i)}$ as the element of $\mathcal{G}_k(\mathbb{R}^p)$(; equivalence class of orthogonal matrices) or the particular orthogonal matrix in $\text{St}_k(\mathbb{R}^p)$. As the matrices in hypothesis are the $\sqrt{\delta}$-scaled orthogonal matrix, notation overloading does not harm the rigorousness of the proof. In summary, if there is no conflict, we denote $\mathbf{A}_t^{(i)}$ as orthogonal matrix, or its equivalent class.

We initialize with:

$$\mathcal{A}_0 = \{\mathbf{A}_1^{(0)}, \mathbf{A}_2^{(0)}, \dots, \mathbf{A}_T^{(0)}\}\,,$$

where $\mathbf{A}_1^{(0)} = \mathbf{A}_2^{(0)} = \cdots = \mathbf{A}_T^{(0)}$ with $\mathbf{A}_T^{(0)} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k) \in \mathbb{R}^{p\times k}$.

Using the above terminology, we first identify the orthogonal matrix $\mathbf{A}_T^{(i)}$ and then construct other elements of $\mathcal{A}_i$ from $\mathbf{A}_T^{(i)}$. First, we show that we can define $M \sim (3/2)^{k(p-k)}$ sequences, to bound projection 2-distance between $\mathbf{A}_T^{(i)}$ and $\mathbf{A}_T^{(j)}$ in the range $[2s\,,6s]$.

Goal 1. Assume that we have sufficiently small $s < 1/3$. We want to construct $\mathbf{A}_T^{(i)} \in \mathcal{G}_k(\mathbb{R}^p)$ ($i \in [M]$), where $M \sim (3/2)^{k(p-k)}$ and :

$$\forall (i,j) \text{ s.t. } i,j \geq 1 \text{ and } i \neq j \;:\; 2s \leq d_2([\mathbf{A}_T^{(i)}]\,,[\mathbf{A}_T^{(j)}]) \leq 6s\,. \tag{14}$$

$\Longrightarrow$ Let $\mathbf{A}_T^{(i)}$ be the elements in $\mathcal{G}_k(\mathbb{R}^p)$. We define $\mathcal{S}_c$ and $\mathcal{S}_i$ as follows:

$$\mathcal{S}_c = \mathcal{B}(\mathbf{A}_T^{(0)}\,,3s)\,,\; \mathcal{S}_i = \mathcal{B}(\mathbf{A}_T^{(i)}\,,2s)\,,$$

where $\mathcal{B}(x\,,r)$ is a projection 2-norm ball in $\mathcal{G}_k(\mathbb{R}^p)$ with radius r. Now, we may choose maximal M which keeps the inequality $M\cdot\mu(\mathcal{B}(2s)) \leq \mu(\mathcal{B}(3s))$. By using the Proposition 1

on the Appendix C.3, we have:

$$\frac{\mu(\mathcal{B}(\mathbf{A}_T^{(0)}, 3s))}{\mu(\mathcal{B}(\mathbf{A}_T^{(i)}, 2s))} \geq \left(1-(3s)^2\right)^{k/2} (3/2)^{k(p-k)} \geq M = (3/2)^{k(p-k-1/2)} .$$

The last inequality comes from the assumption $s < 1/3$. If we fix $\mathbf{A}_T^{(0)} \in \mathcal{G}_k(\mathbb{R}^p)$, we can select $\mathbf{A}_T^{(1)}$ in $\mathcal{S}_c$ and exclude $\mathcal{S}_1$ from $\mathcal{S}_c$. By repeating this process, we can select at least M k-dimensional plane $\mathbf{A}_T^{(i)}$ while ensuring that $\mathcal{S}_c - \bigcup_{1\leq i\leq M-1} \mathcal{S}_i$ is non-empty. From the construction, the condition by the equation (14) is satisfied by the triangle inequality.

For the next step, we construct $\mathbf{A}_t^{(i)} (t < T\,, 0 < i)$ satisfying the following second goal:

Goal 2. Let us assume that we constructed $\mathbf{A}_T^{(i)} \in \mathcal{G}_k(\mathbb{R}^p)$ to satisfy equation (14). We want to construct $\mathbf{A}_t^{(i)}$ for every $t = 1 \cdots T$ and $i = 1 \cdots M$, satisfying:

$$d_2([\mathbf{A}_{t-1}^{(i)}]\,, [\mathbf{A}_t^{(i)}]) \leq \frac{\Gamma}{\delta}\,. \tag{15}$$

$\Longrightarrow$ Let us define the mapping $\mathcal{G}^{(i)}(\Psi(t))$ $\left(0 \preceq \Psi(t) \preceq \Psi^{(i)} = (\psi_j^{(i)})_{j=1}^k\right)$ as $\mathcal{G}_{[\mathbf{A}_T^{(0)}]\to[\mathbf{A}_T^{(i)}]}$ in the Definition 7 on the Appendix C.3. For $t \in [0, T]$, we define $\mathbf{A}_t^{(i)}$ as $\mathcal{G}^{(i)}(\Psi(t))$, where $\Psi(t) = (\psi_1(t), \ldots, \psi_k(t))$ is:

$$\psi_j(t) = \begin{cases} 0 & t \leq T - \left\lfloor d_2([\mathbf{A}_T^{(0)}]\,, [\mathbf{A}_T^{(i)}])\,\delta/\Gamma \right\rfloor \\ \psi_j^{(i)} \max\left(1 - (T-t)\frac{\Gamma}{\delta\, d_2([\mathbf{A}_T^{(0)}]\,, [\mathbf{A}_T^{(i)}])}\,, 0\right) & t > T - \left\lfloor d_2([\mathbf{A}_T^{(0)}]\,, [\mathbf{A}_T^{(i)}])\,\delta/\Gamma \right\rfloor . \end{cases}$$

From the property of principal rotation, for every $t \in [T]$, we have

$$d_2([\mathbf{A}_{t-1}^{(i)}]\,, [\mathbf{A}_t^{(i)}]) \leq \max_{1\leq j\leq k} \left(\frac{\Gamma}{\delta} \frac{\psi_j^{(i)}}{d_2([\mathbf{A}_T^{(0)}]\,, [\mathbf{A}_T^{(i)}])}\right) = \frac{\Gamma}{\delta}\,.$$

On the last step, we will bound the distance between different hypotheses at arbitrary t with $6s$ as in the case of $t = T$:

Goal 3.

$$d_2([\mathbf{A}_t^{(i)}]\,, [\mathbf{A}_t^{(j)}]) \leq 6s\,. \tag{16}$$

$\Longrightarrow$ Here, we bound the above distance as:

$$\begin{aligned} d_2([\mathbf{A}_t^{(i)}]\,, [\mathbf{A}_t^{(j)}]) &\leq d_2([\mathbf{A}_t^{(i)}]\,, [\mathbf{A}_T^{(i)}]) + d_2([\mathbf{A}_t^{(j)}]\,, [\mathbf{A}_T^{(j)}]) \qquad (\mathbf{A}_0^{(i)} = \mathbf{A}_0^{(j)} = \mathbf{A}_T^{(0)}) \\ &\overset{(\bigstar)}{\leq} d_2([\mathbf{A}_T^{(i)}]\,, [\mathbf{A}_0^{(i)}]) + d_2([\mathbf{A}_T^{(j)}]\,, [\mathbf{A}_0^{(j)}]) \\ &\leq 6s\,, \end{aligned}$$

where $(\bigstar)$ follows from the construction on the proof of **Goal 2**.

D.3 Reduction to error probability

Recall that by the construction of the sequence $\{\mathcal{A}_i\}$ (**Goal 1**), we have:

$$\forall (i, j) \;\text{ s.t. }\; i, j \geq 1 \text{ and } i \neq j \;:\; 2s \leq d_2([\mathbf{A}_T^{(i)}]\,, [\mathbf{A}_T^{(j)}])\,.$$

Therefore, for any estimation of the top eigenvectors $\phi_{\mathcal{X}}$, by triangle inequality, we have that

$$\mathbb{P}_{\mathcal{A}_j}\left[d\big(\text{ran}(\mathbf{A}_T^{(j)}), \phi_{\mathcal{X}}\big) \geq s\right] \geq \mathbb{P}_{\mathcal{A}_j}\big(\xi^* \neq j\big)\,,$$

where $\xi^* : \mathcal{X} \mapsto \mathcal{A}$ denotes the minimum distance test defined by

$$\xi^* = \text{argmin}_{1\le i\le M} d(\text{ran}(\mathbf{A}_T^{(i)}), \phi_{\mathcal{X}})\,.$$

The above equations imply that:

$$\mathbb{P}_{\mathcal{A}_j}\left[d(\text{ran}(\mathbf{A}_T^{(j)}), \phi_{\mathcal{X}}) \ge s\right] \ge \mathbb{P}_{\mathcal{A}_j}\left(\xi^* \ne j\right) \ge p_{e,M}\,,$$

where $p_{e,M} = \inf_\xi \max_{0\le j\le M} \mathbb{P}_{\mathcal{A}_j}(\xi \ne j)$ and the infimum is over all possible tests ξ. To analyze and bound $p_{e,M}$, we use the following Lemma 7.

Lemma 7 (Theorem 2.5 in [57]). *Assume that* $\text{Tu}(\delta, \Gamma)$ *contains elements* $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_M$ $(M \ge 2)$ *such that*

$$\frac{1}{M}\sum_{j=1}^{M} \text{KL}\big(\mathbb{P}_{\mathcal{A}_j} \| \mathbb{P}_{\mathcal{A}_0}\big) \le \alpha \log M.$$

Then, we have

$$p_{e,M} \ge \frac{\sqrt{M}}{1+\sqrt{M}}\Big(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\Big).$$

To apply Lemma 7, we first bound $\text{KL}\big(\mathbb{P}_{\mathcal{A}_i} \| \mathbb{P}_{\mathcal{A}_j}\big)$ for all $i \ne j$. Since $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are independent,

$$\text{KL}\big(\mathbb{P}_{\mathcal{A}_i} \| \mathbb{P}_{\mathcal{A}_j}\big) = \sum_{t=1}^{T} \text{KL}\big(\mathbf{A}_t^{(i)} \| \mathbf{A}_t^{(j)}\big) \le \min\Big\{T,\, d_2(\mathbf{A}_T^{(0)}\,, \mathbf{A}_T^{(i)})\,\delta/\Gamma\Big\}\,\text{KL}\big(\mathbf{A}_T^{(i)} \| \mathbf{A}_T^{(j)}\big),$$

where $\text{KL}\big(\mathbf{A}_t^{(i)} \| \mathbf{A}_t^{(j)}\big)$ is the KL-divergence between two spiked covariance models defined with $\mathbf{A}_t^{(i)}$ and $\mathbf{A}_t^{(j)}$. We first study the KL-divergence between models with $\mathbf{A}_T^{(i)}$ and $\mathbf{A}_T^{(j)}$ as below:

$$\begin{aligned}
&\text{KL}\big(\mathbf{A}_T^{(i)} \| \mathbf{A}_T^{(j)}\big)\\
=&\log\left(\frac{|\delta\mathbf{A}_T^{(i)}\mathbf{A}_T^{(i)\top} + \sigma^2\mathbf{I}_{p\times p}|}{|\delta\mathbf{A}_T^{(j)}\mathbf{A}_T^{(j)\top} + \sigma^2\mathbf{I}_{p\times p}|}\right) - p + \text{tr}\Big((\delta\mathbf{A}_T^{(i)}\mathbf{A}_T^{(i)\top} + \sigma^2\mathbf{I}_{p\times p})^{-1}(\delta\mathbf{A}_T^{(j)}\mathbf{A}_T^{(j)\top} + \sigma^2\mathbf{I}_{p\times p})\Big)\\
=&-p + \text{tr}\Big(\frac{1}{\sigma^2}\big(\mathbf{I}_{p\times p} - \frac{\delta}{(\sigma^2+\delta)}\mathbf{A}_T^{(i)}\mathbf{A}_T^{(i)\top}\big)(\sigma^2\mathbf{I}_{p\times p} + \delta\mathbf{A}_T^{(j)}\mathbf{A}_T^{(j)\top})\Big)\\
=&-p + \text{tr}\Big(\frac{1}{\sigma^2}\big(\sigma^2\mathbf{I}_{p\times p} + \delta\mathbf{A}_T^{(j)}\mathbf{A}_T^{(j)\top} - \frac{\delta\sigma^2}{(\sigma^2+\delta)}\mathbf{A}_T^{(i)}\mathbf{A}_T^{(i)\top} - \frac{\delta^2}{(\sigma^2+\delta)}\mathbf{A}_T^{(i)}\mathbf{A}_T^{(i)\top}\mathbf{A}_T^{(j)}\mathbf{A}_T^{(j)\top}\big)\Big)\\
=&\frac{\delta^2}{\sigma^2(\sigma^2+\delta)}\text{tr}\Big(\mathbf{A}_T^{(i)}\mathbf{A}_T^{(i)\top}(\mathbf{I}_{p\times p} - \mathbf{A}_T^{(j)}\mathbf{A}_T^{(j)\top})\Big)\\
=&\frac{1}{2}\frac{\delta^2}{\sigma^2(\sigma^2+\delta)}\|\mathbf{U}_T^{(i)}\mathbf{U}_T^{(i)\top} - \mathbf{U}_T^{(j)}\mathbf{U}_T^{(j)\top}\|_{\text{F}}^2 \overset{(\bigstar)}{\le} \frac{\delta^2}{\sigma^2(\sigma^2+\delta)}k\|\mathbf{U}_T^{(i)}\mathbf{U}_T^{(i)\top} - \mathbf{U}_T^{(j)}\mathbf{U}_T^{(j)\top}\|^2\\
\le&\frac{k\delta^2}{\sigma^2(\sigma^2+\delta)}(6s)^2.
\end{aligned}$$

The $(\bigstar)$ follows from the subadditivity of rank(Lemma 6) and the relation between two norms($\|\mathbf{M}\|_{\text{F}} \le \sqrt{\text{rk}(\mathbf{M})}\|\mathbf{M}\|$). Therefore, we have the following upper bound:

$$\begin{aligned}
\text{KL}\big(\mathbb{P}_{\mathcal{A}_i} \| \mathbb{P}_{\mathcal{A}_j}\big) &\le \frac{36k\delta^2}{\sigma^2(\sigma^2+\delta)}\min\Big\{T,\, d_2(\mathbf{A}_T^{(0)}\,, \mathbf{A}_T^{(i)})\,\delta/\Gamma\Big\}s^2\\
&\le \frac{36k\delta^2}{\sigma^2(\sigma^2+\delta)}\min\Big\{T,\, \frac{3s\delta}{\Gamma}\Big\}s^2\,,
\end{aligned} \tag{17}$$

where the last inequality comes from the construction of **Goal 1**.

D.4 Proving the Theorem

Now, we establish Theorem 1 based on the progress so far. As mentioned at the last section, we are using the notation $\mathbf{A}_t^{(i)}$ to represent an orthogonal matrix(or its class), not scaled with $\sqrt{\delta}$. We first start from the result in the Appendix D.1:

$$1/3 > \forall s > 0\ :\ \mathcal{R}^* \ge s \inf_{\phi} \sup_{\mathcal{A}\in\{\mathcal{A}_i\}_{i=0}^M} \mathbb{P}_{\mathcal{A}}\left[d(\text{ran}(\mathbf{A}_T)\,, \phi_{\mathcal{X}}) \ge s\right]\,,$$

where $d_2(\cdot\,,\cdot)$ is distance defined by operator norm between projectors. From the construction of **Goal 1**$\sim$**3** on the Appendix D.2, we bound the KL divergence between hypothesis at the equation (17):

$$\mathrm{KL}\big(\mathbb{P}_{\mathcal{A}_i}\|\mathbb{P}_{\mathcal{A}_j}\big) \le \frac{36k\delta^2}{\sigma^2(\sigma^2+\delta)} \min\left\{T,\, \frac{3s\delta}{\Gamma}\right\} s^2\,.$$

Now, we find the particular range of s satisfying:

$$\mathrm{KL}(\mathbb{P}_{\mathcal{A}_i}\|\mathbb{P}_{\mathcal{A}_j}) \le \frac{36\delta^2}{\sigma^2(\sigma^2+\delta)} k \min\left(T, \frac{3\delta}{\Gamma}\, s\right) s^2 \overset{(\bigstar 1)}{\le} \frac{1}{10}\log M = \frac{1}{10}\log\left((3/2)^{k(p-k-1/2)}\right)\,.$$

Note that ($\bigstar$1) is satisfied when the ($\bigstar$2) of the following inequality holds:

$$\frac{360\delta^2}{\sigma^2(\sigma^2+\delta)} \min\left(T, \frac{3\delta}{\Gamma}\, s\right) s^2 \overset{(\bigstar 2)}{\le} \frac{1}{2} p \log\frac{3}{2} \le (p-k-\frac{1}{2})\log\frac{3}{2}\,.$$

Therefore, we have a constant $(\log(3/2)/2160)^{1/3} \simeq 0.05726 > 0$ such that with

$$s = \sqrt[3]{\frac{\log(3/2)}{2160}} \left((\frac{\Gamma}{\delta})^{1/3} (\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2})^{1/3} + \frac{1}{\sqrt{T}} (\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2})^{1/2} \right)\,, \tag{18}$$

we have

$$\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{\mathcal{A}_j}\|\mathbb{P}_{\mathcal{A}_0}) \le \frac{1}{10}\log M\,.$$

Finally, we use the reduction to error probability argument at the Appendix D.1. For $p_{e,M} = \inf_{\xi}\max_{0\le j\le M}\mathbb{P}_{\mathcal{A}_j}(\xi \neq j)$,

$$\mathbb{P}_{\mathcal{A}_j}\left[d(\mathrm{ran}(\mathbf{A}_T)\,, \phi_{\mathcal{X}}) \ge s\right] \ge \mathbb{P}_{\mathcal{A}_j}(\xi \neq j) \ge p_{e,M}\,.$$

Now, with s in the equation (18), we have the following:

$$\begin{aligned}
\mathcal{R}^* &\ge s \cdot \inf_{\phi} \sup_{\mathcal{A}\in\{\mathcal{A}_0\,,\mathcal{A}_1\,,\ldots,\mathcal{A}_M\}} \mathbb{P}_{\mathcal{A}}\left[d(\mathrm{ran}(\mathbf{A}_T)\,, \phi_{\mathcal{X}}) \ge s\right] \\
&\ge s\cdot p_{e,M} = \sqrt[3]{\frac{\log(3/2)}{2160}} \left((\frac{\Gamma}{\delta})^{1/3} (\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2})^{1/3} + \frac{1}{\sqrt{T}} (\frac{p\sigma^2(\sigma^2+\delta)}{\delta^2})^{1/2} \right) p_{e,M}\,,
\end{aligned}$$

where the $p_{e,M}$ is lower bounded by ~ 0.8 since $M \ge 1.5^{pk/2}$ is sufficiently large . ■

E Proof of Lemma 1

We prove the Lemma 1 under the condition $\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] = \mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p}$ and Assumption 1 defined at the Section 5 holds. Later at the Appendix B, we apply this result on the our original model. We first start from the decomposition of $\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\mathbf{x}_t\mathbf{x}_t^\top$,

$$\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\mathbf{x}_t\mathbf{x}_t^\top = \mathbb{E}[\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top] + \mathcal{E}(\ell) = \mathbf{A}_{\ell B}\mathbf{A}_{\ell B}^\top + \sigma^2\mathbf{I}_{p\times p} + \mathcal{E}(\ell) = \mathbf{M}(\ell) + \mathcal{E}(\ell)\,.$$

Note that formulation for $\mathcal{E}(\ell)$ is following:

$$\mathcal{E}(\ell) = \frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\mathbf{x}_t\mathbf{x}_t^\top - \mathbb{E}[\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top]\,.$$

We would decompose $\mathcal{E}(\ell)$ in two terms as following.

$$\begin{aligned}
\mathcal{E}(\ell) &= \frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\Big(\mathbf{x}_t\mathbf{x}_t^\top - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] + \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] - \mathbb{E}[\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top]\Big) \\
&= \underbrace{\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\Big(\mathbf{x}_t\mathbf{x}_t^\top - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t]\Big)}_{\mathcal{E}_1(\ell)} + \underbrace{\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\Big(\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] - \mathbb{E}[\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top]\Big)}_{\mathcal{E}_2(\ell)}\,.
\end{aligned}$$

E.1 Bounding $\max_\ell \|\mathcal{E}_1(\ell)\|$ with probability $1 - 1/T$

First, we bound the $\mathcal{E}_1(\ell)$ using the following matrix Bernstein inequality.

Theorem 4 (Matrix Bernstein Inequality [56]). *Let $\mathbf{X}_1, \dots, \mathbf{X}_B \in \mathbb{R}^{p\times p}$ be independent, centered, symmetric real random variables, and assume that each one is uniformly bounded:*

$$\mathbb{E}[\mathbf{X}_i] = \mathbf{0} \quad \textit{and} \quad \|\mathbf{X}_i\| \le \mathcal{M} \quad \textit{for each } i = 1, \dots, B\,.$$

and let $\mathcal{V}$ denote upper bound for the matrix variance statistics of the sum:

$$\mathcal{V} \ge \|\mathbb{E}[\mathbf{X}_i^2]\|\,.$$

Then

$$\mathbb{P}\left\{\left\|\frac{1}{B}\sum_{i=1}^{B}\mathbf{X}_i\right\| \ge x\right\} \le 2p\exp\left\{\frac{-Bx^2}{2(\mathcal{V}+\mathcal{M}x/3)}\right\}\,.$$

We set $\mathbf{X}_t$ as $\mathbf{x}_t\mathbf{x}_t^\top - \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top]$ for all $(\ell-1)B+1 \le t \le \ell B$ and overload the notation $\mathcal{M}$ and $\mathcal{V}$. Consider the following inequality:

$$\log(2pT^2) \le \frac{Bx^2}{2(\mathcal{V}+\mathcal{M}x/3)}\,.$$

Then we have following sufficient condition for the above inequality:

$$x \ge \frac{\mathcal{M}\log(2pT^2)}{3B}\left[1+\sqrt{1+\frac{18\mathcal{V}}{\mathcal{M}^2}\frac{B}{\log 2pT^2}}\right]\,.$$

From the inequality $\sqrt{1+x} \le 1+\sqrt{x}(x\ge 0)$, we get the following argument.

$$\text{If } x = \frac{\mathcal{M}}{3}\frac{\log 2pT^2}{B} + \sqrt{2\mathcal{V}}\sqrt{\frac{\log 2pT^2}{B}}\,, \text{ then } \mathbb{P}\left\{\left\|\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\mathbf{X}_t\right\| \ge x\right\} \le \frac{1}{T^2}\,.$$

Let us assume the condition $\mathcal{M}^2\log(2pT^2)/\mathcal{V} \le B$. Then, with probability greater than $1-1/T^2$, we have that:

$$\|\mathcal{E}_1(\ell)\| = \left\|\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\mathbf{X}_t\right\| \le \frac{1+3\sqrt{2}}{3}\sqrt{\frac{\mathcal{V}\log(2pT^2)}{B}}\,.$$

Now, we use the union bound argument. That is, for probability greater than $1-1/T$ ($\le 1-(T/B)/T^2$),

$$\max_{1\le\ell\le L(=T/B)}\|\mathcal{E}_1(\ell)\| \le \frac{1+3\sqrt{2}}{3}\sqrt{\frac{\mathcal{V}\log(2pT^2)}{B}}\,.$$

E.2 Bounding $\|\mathcal{E}_2(\ell)\|$ for all ℓ

Since the our model limits the perturbation amount of covariance matrix, we may bound the $\|\mathcal{E}_2(\ell)\|$ as follows:

$$\begin{aligned}
\|\mathcal{E}_2(\ell)\| &\le \left\|\frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\left(\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] - \mathbb{E}[\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top]\right)\right\| \le \frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\|\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] - \mathbb{E}[\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top]\| \\
&= \frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}\|\mathbf{A}_t\mathbf{A}_t^\top - \mathbf{A}_{\ell B}\mathbf{A}_{\ell B}^\top\| \le \frac{1}{B}\sum_{t=(\ell-1)B+1}^{\ell B}(\ell B - t)\Gamma \\
&\le \frac{1}{B}\frac{B(B-1)}{2}\Gamma \le \frac{B\Gamma}{2}\,.
\end{aligned}$$

E.3 Bounding $\|\mathcal{E}(\ell)\|$ for all ℓ, with high probability

On the Appendix E.1, we bounded $\max_\ell\|\mathcal{E}_1(\ell)\|$ for the probability greater than $1-1/T$. Therefore, for probability greater than $1-1/T$,

$$\max_\ell\|\mathcal{E}(\ell)\| \le \max_\ell\|\mathcal{E}_1(\ell)\| + \max_\ell\|\mathcal{E}_2(\ell)\| \le \frac{1+3\sqrt{2}}{3}\sqrt{\frac{\mathcal{V}\log(2pT^2)}{B}} + \frac{B\Gamma}{2}\,. \tag{19}$$

■

F Proof of Theorem 2

We prove the Theorem 2 under the condition defined by the equation (19), which holds for probability greater than $1 - 1/T$.

F.1 Deriving optimal learning block size B

Consider the upper bound (for probability greater than $1-1/T$) on $\max_\ell \|\mathcal{E}(\ell)\|$ from the Appendix E:

$$\max_\ell \|\mathcal{E}(\ell)\| \leq C_{\text{NPM}} \sqrt{\frac{\log 2pT^2}{B}} + \frac{B\Gamma}{2}\,, \text{ where } C_{\text{NPM}} = \frac{1+3\sqrt{2}}{3}\sqrt{\mathcal{V}}\,.$$

By differentiating and find the critical point, we have the following optimal block size:

$$B_{\text{opt}} = \frac{{C_{\text{NPM}}}^{2/3} \log(2pT^2)^{1/3}}{\Gamma^{2/3}} = \Omega\left(\frac{\mathcal{V}^{2/3}\log(2pT^2)^{1/3}}{\Gamma^{2/3}}\right)\,.$$

In this case, the uniform upper bound for error matrix becomes:

$$\max_\ell \|\mathcal{E}(\ell)\| \leq \frac{3}{2} C_{\text{NPM}}^{2/3} \log(2pT^2)^{1/3}\Gamma^{1/3}\,.$$

F.2 Defining Regime and Parameters

Let us set $B = B_{\text{opt}}$ and consider the regime:

(A) $36 B_{\text{opt}}\Gamma = 24\, C_{\text{NPM}}^{2/3} \log(2pT^2)^{1/3}\Gamma^{1/3} \leq \delta$, from $\Gamma = \mathcal{O}\left(\frac{\delta^3}{C_{\text{NPM}}^2 \log(2pT^2)}\right)$.

(B) $\delta \geq \frac{12}{17}\sigma^2$.

For this regime, we define $\Lambda\,, \epsilon\,, \eta > 0$ as:

(a) $\Lambda := \frac{3}{2} B_{\text{opt}}\Gamma = \frac{3}{2} C_{\text{NPM}}^{2/3} \log(2pT^2)^{1/3}\Gamma^{1/3}$
$(\geq \max_\ell \|\mathcal{E}(\ell)\|$ on probability greater than $1 - 1/T,$ shown at the Appendix F.1$)$.

(b) $\epsilon := \frac{4\Lambda}{\delta} \leq \frac{1}{4}$.

(c) $\eta := \frac{B_{\text{opt}}\Gamma}{\delta - B_{\text{opt}}\Gamma}$.

Note that from the item (A) and (c) above, we have:

$$\frac{\eta}{\epsilon} = \frac{B_{\text{opt}}\Gamma/(\delta - B_{\text{opt}}\Gamma)}{4\Lambda/\delta} = \frac{B_{\text{opt}}\Gamma/(\delta - B_{\text{opt}}\Gamma)}{6B_{\text{opt}}\Gamma/\delta} = \frac{1}{6}\frac{\delta}{\delta - B_{\text{opt}}\Gamma} \leq \frac{6}{35} \leq \frac{1}{5}\,.$$

With this parameters, we show the following lemma:

Lemma 8. *Assume the regime in the above. Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a positive definite matrix and SVD$(\mathbf{M}) = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ with $s_k(\mathbf{M}) \geq \delta + \sigma^2$ and $s_{k+1}(\mathbf{M}) = \sigma^2$. Then we have:*

- $\beta := (1-(\eta+\epsilon)^2)\frac{\delta+\sigma^2-\Lambda/\sqrt{1-(\eta+\epsilon)^2}}{\sigma^2+\Lambda} \geq 1.4465 > \frac{1}{0.7} > 1\,.$
- $\frac{\epsilon}{\sqrt{1-\epsilon^2}}\frac{s_k(\mathbf{M})-\Lambda/\epsilon}{s_{k+1}(\mathbf{M})+\Lambda/\sqrt{1-\epsilon^2}} \geq \frac{\epsilon}{\sqrt{1-\epsilon^2}}\frac{0.75s_k(\mathbf{M})+0.25s_{k+1}(\mathbf{M})}{0.25s_k(\mathbf{M})+0.75s_{k+1}(\mathbf{M})} \overset{(\spadesuit)}{>} \frac{\epsilon+\eta}{\sqrt{1-(\epsilon+\eta)^2}}\,.$
- $\frac{s_k(\mathbf{M})-\Lambda/\sqrt{1-(\epsilon+\eta)^2}}{s_{k+1}(\mathbf{M})+\Lambda/(\epsilon+\eta)}\frac{\sqrt{1-(\epsilon+\eta)^2}}{\epsilon+\eta} > \frac{\sqrt{1-\epsilon^2}}{\epsilon}\,.$

Proof. Here we provide the bound for $(\epsilon+\eta)$: $\epsilon+\eta \leq 6\epsilon/5 \leq 3/10$. For the first item, we have:

$$\beta \geq \frac{91}{100}\frac{\delta+\sigma^2-\sqrt{100/91}\,\Lambda}{\sigma^2+\Lambda} \geq \frac{91}{100}\frac{\delta+\sigma^2-\sqrt{100/91}\,\delta/16}{\sigma^2+\delta/16} = \frac{91}{100}\frac{\left(1-\sqrt{\frac{100}{91}}\frac{1}{16}\right)+\frac{\sigma^2}{\delta}}{\frac{1}{16}+\frac{\sigma^2}{\delta}} \geq 1.4465\,.$$

where the inequalities follows from $\epsilon + \eta \le 3/10$, $\Lambda \le \delta/16$, and $\delta \ge 12\sigma^2/17$ respectively.

For the second one, the first inequality is immediate from $\epsilon \le 1/4$. We now show the inequality (♠). We provide the sufficient condition as:

$$\begin{aligned}
&\left(\frac{\epsilon}{\sqrt{1-\epsilon^2}}\frac{\sigma^2+\frac{3}{4}\delta}{\sigma^2+\frac{1}{4}\delta} = \right) \frac{\epsilon}{\sqrt{1-\epsilon^2}}\frac{\frac{\sigma^2}{\delta}+\frac{3}{4}}{\frac{\sigma^2}{\delta}+\frac{1}{4}} \ge \frac{\epsilon+\eta}{\sqrt{1-(\epsilon+\eta)^2}} \\
\Longleftarrow\quad &\left(\frac{\epsilon}{\sqrt{1-\epsilon^2}}\frac{17/12+3/4}{17/12+1/4} = \right) \frac{13}{10}\frac{\epsilon}{\sqrt{1-\epsilon^2}} \ge \frac{\epsilon+\eta}{\sqrt{1-(\epsilon+\eta)^2}} \qquad (\delta \ge \frac{12\sigma^2}{17}) \\
\Longleftrightarrow\quad &1.69\left(\frac{\epsilon}{\epsilon+\eta}\right)^2 - 1 \ge 0.69\epsilon^2 \\
\Longleftarrow\quad &1.69\left(\frac{\epsilon}{\epsilon+\eta}\right)^2 - 1 \ge \frac{0.69}{16} \qquad (\epsilon \le 1/4) \\
\Longleftarrow\quad &\eta \le 0.288\,\epsilon\,.
\end{aligned}$$

The final condition is automatically satisfied the inequality $\eta \le \epsilon/5$ (from the definition of η). Therefore, we proved the second inequality.

For the last inequality, enough to show:

$$\frac{\epsilon+\eta}{\sqrt{1-(\epsilon+\eta)^2}}\frac{\sigma^2 \;+\; \delta\epsilon/4(\epsilon+\eta)}{\sigma^2+\delta \;-\; \delta\epsilon/4\sqrt{1-(\epsilon+\eta)^2}} \le \frac{\epsilon}{\sqrt{1-\epsilon^2}}\,.$$

We upper bound second term of LHS:

$$\begin{aligned}
\frac{\sigma^2 \;+\; \delta\epsilon/4(\epsilon+\eta)}{\sigma^2+\delta \;-\; \delta\epsilon/4\sqrt{1-(\epsilon+\eta)^2}} &= \frac{\sigma^2/\delta \;+\; \epsilon/4(\epsilon+\eta)}{\sigma^2/\delta + (1 \;-\; \epsilon/4\sqrt{1-(\epsilon+\eta)^2})} \\
&\overset{(\blacksquare)}{\le} \frac{17/12 \;+\; \epsilon/4(\epsilon+\eta)}{17/12 + (1 \;-\; \epsilon/4\sqrt{1-(\epsilon+\eta)^2})} \\
&\overset{(\blacklozenge)}{\le} \frac{17/12 \;+\; \epsilon/4(\epsilon+\eta)}{29/12 \;-\; \epsilon/4(\epsilon+\eta)} \\
&\le \frac{17/3 \;+\; 1/(1+\eta/\epsilon)}{29/3 \;-\; 1/(1+\eta/\epsilon)} \le \frac{10}{13}\,,
\end{aligned}$$

where the (■) comes from:

$$\begin{aligned}
\epsilon+\eta \le \frac{1}{\sqrt{2}} \le \sqrt{1-(\epsilon+\eta)^2} \Longrightarrow\; &\frac{\epsilon}{4}\left(\frac{1}{\epsilon+\eta} + \frac{1}{\sqrt{1-(\epsilon+\eta)^2}}\right) \le \frac{\epsilon}{4}\frac{2}{\epsilon+\eta} \le \frac{1}{2} \\
\Longrightarrow\; &1 - \frac{\epsilon}{4\sqrt{1-(\epsilon+\eta)^2}} \ge \frac{\epsilon}{4(\epsilon+\eta)}\,,
\end{aligned}$$

and the (♦) is immediate from $\epsilon+\eta \le \frac{1}{\sqrt{2}} \le \sqrt{1-(\epsilon+\eta)^2}$. Finally, the only left part is :

$$\frac{10}{13}\frac{\epsilon+\eta}{\sqrt{1-(\epsilon+\eta)^2}} \le \frac{\epsilon}{\sqrt{1-\epsilon^2}}\,,$$

which was shown (from $\eta \le \epsilon/5$), when we proved the inequality (♠) above. □

F.3 Lemmas for $\mathbf{N}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$

Lemma 9 (Orthogonal amplification)**.** *Assume the regime in the Appendix F.2. Let $\mathbf{M} \in \mathbb{R}^{p\times p}$ be a positive definite matrix and $\mathrm{SVD}(\mathbf{M}) = \mathbf{U}\mathbf{D}\mathbf{U}^\top$. Moreover, let $\mathcal{E} \in \mathbb{R}^{p\times p}$ with $\|\mathcal{E}\| \le \Lambda$. For $0 < k < p$, let $\mathcal{Y}$ be the set of $\mathbf{Y} \in \mathrm{St}_{p-k}(\mathbb{R}^p)$ such that $s_1\left(\mathbf{U}_{1:k}^\top \mathbf{Y}\right) \le \epsilon + \eta$. For every given $\mathbf{Y} \in \mathcal{Y}$, there exists a $\overline{\mathbf{N}} \in \mathrm{St}_{p-k}(\mathbb{R}^p)$ such that*

$$\mathrm{ran}\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) \subseteq \mathrm{ran}\left(\mathbf{Y}\right),\; s_1\left(\mathbf{U}_{1:k}^\top \overline{\mathbf{N}}\right) \le \epsilon, \tag{20}$$

$$s_1\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) \le \frac{\sigma^2+\Lambda}{\sqrt{1-(\epsilon+\eta)^2}}. \tag{21}$$

Proof. We first show that for any positive definite matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, $\mathcal{E} \in \mathbb{R}^{p\times p}$ and $\mathbf{Y} \in \mathcal{Y}$, there exists $\overline{\mathbf{N}} \in \mathrm{St}_{p-k}(\mathbb{R}^p)$ such that $\mathrm{ran}\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) \subseteq \mathrm{ran}\,(\mathbf{Y})$ as follows:

1. When $\mathbf{M}+\mathcal{E}$ is a full-rank matrix, $\mathrm{ran}\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) = \mathrm{ran}\,(\mathbf{Y})$ with $\overline{\mathbf{N}} = b\left((\mathbf{M}+\mathcal{E})^{-1}\mathbf{Y}\right)$.

2. When the rank of $\mathbf{M}+\mathcal{E}$ is $r \le k$, every $\overline{\mathbf{N}}$ such that $(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}} = \mathbf{0}$ satisfies that $\mathrm{ran}\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) = \emptyset \subseteq \mathrm{ran}\,(\mathbf{Y})$.

3. Assume that the rank of $\mathbf{M}+\mathcal{E}$ is r, $k < r < p$. We identify $\overline{\mathbf{N}}$ in parts by identifying the first $(r-k)$ columns and then the remaining columns. Let $(\mathbf{M}+\mathcal{E}) = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\top$ and $\mathbf{Y}^\top\tilde{\mathbf{U}}_{1:r} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}^\top$ be the singular value decomposition of $(\mathbf{M}+\mathcal{E})$ and $\mathbf{Y}^\top\tilde{\mathbf{U}}_{1:r}$ respectively.

   Observe that $\mathbf{Y}$ has $(p-k)$ columns and $\tilde{\mathbf{U}}_{\mathbf{1:r}}$ has r columns and these vectors form a basis for $(p-k)$ dimensional subspace and r dimensional subspace of $\mathbb{R}^p$ respectively. Since $(p-k)+r > p$, the column spaces of $\mathbf{Y}$ and $\tilde{\mathbf{U}}_{1:r}$ overlap on a subspace of dimension at least $r-k$. Therefore, we can find $(r-k)$ orthonormal vectors in this shared subspace, say, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{r-k} \in \mathbb{R}^p$. For $1 \le j \le r$, let $\mathbf{f}_j \in \mathbb{R}^r$ be such that

   $$\tilde{\mathbf{U}}_{1:r}\mathbf{f}_j = \mathbf{v}_j$$

   i.e. $\mathbf{f}_j = \tilde{\mathbf{U}}_{1:r}^\top\mathbf{v}_j$. Thus the $\mathbf{f}_j$ are orthonormal, and since $\{\mathbf{v}_j\}_{j=1}^{r-k}$ are orthonormal and contained in the column space of $\mathbf{Y}$, for $1 \le j \le r$ we have $1 = \|\mathbf{Y}^\top\mathbf{v}_j\| = \|\mathbf{Y}^\top\tilde{\mathbf{U}}_{1:r}\mathbf{f}_j\| = \|\tilde{\mathbf{U}}_{1:r}\mathbf{f}_j\|$. Thus, $\{\mathbf{f}_j\}_{j=1}^{r-k}$ are right-singular vectors of $\mathbf{Y}^\top\tilde{\mathbf{U}}_{1:r}$ with singular value 1 (which is the maximum singular value of $\mathbf{Y}^\top\tilde{\mathbf{U}}_{1:r}$) and therefore without loss of generality, they are the first $r-k$ columns of $\hat{\mathbf{V}}$. To identify these $\mathbf{v}_j$ we use the above paragraph, that is to say,

   $$\{\mathbf{v}_j\}_{j=1}^{r-k} = \tilde{\mathbf{U}}_{1:r-k}(\hat{\mathbf{V}})_{1:r} = (\mathbf{M}+\mathcal{E})(\tilde{\mathbf{V}}\tilde{\mathbf{D}}^{-1})_{1:r}(\hat{\mathbf{V}})_{1:r-k}.$$

   Hence, since the $\mathbf{v}_j$ are spanned by the columns of $\mathbf{Y}$,

   $$(\mathbf{M}+\mathcal{E})(\tilde{\mathbf{V}}\tilde{\mathbf{D}}^{-1})_{1:r}(\hat{\mathbf{V}})_{1:r-k} \subseteq \mathrm{ran}\,(\mathbf{Y})\,.$$

   Define $\{\mathbf{z}_i\}_{i=1}^{r-k}$ to be an orthonormal basis of the column space of $(\tilde{\mathbf{V}}\tilde{\mathbf{D}}^{-1})_{1:r}(\hat{\mathbf{V}})_{1:r-k}$ i.e. $\{\mathbf{z}_i\}_{i=1}^{r-k} = b((\tilde{\mathbf{V}}\tilde{\mathbf{D}}^{-1})_{1:r}(\hat{\mathbf{V}})_{1:r-k})$. The first $r-k$ columns of $\overline{\mathbf{N}}$ are defined to be $\{\mathbf{z}_i\}_{i=1}^{r-k}$. At this point we have identified only $r-k$ columns for $\overline{\mathbf{N}}$. The remaning $(p-r)$ columns are picked from the null space of $(\mathbf{M}+\mathcal{E})$. A vector $\mathbf{f}$ in the null space $(\mathbf{M}+\mathcal{E})\mathbf{f} = 0$ is also a right singular vector of $(\mathbf{M}+\mathcal{E})$ whose singular value is 0. Since $\mathbf{M}+\mathcal{E}$ has rank r, there are $p-r$ right singular vectors of $\mathbf{M}+\mathcal{E}$ with zero singular value and we use them to define the remaining $r-k$ columns of $\overline{\mathbf{N}}$. Thus, when

   $$\overline{\mathbf{N}} = \left[b\left((\tilde{\mathbf{V}}\tilde{\mathbf{D}}^{-1})_{1:r}(\hat{\mathbf{V}})_{1:r-k}\right), \tilde{\mathbf{V}}_{r+1:p}\right],$$

   we have $\mathrm{ran}\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) \subseteq \mathrm{ran}\,(\mathbf{Y})$.

We establish the second part of (20) by contradiction. To show that $\mathrm{ran}\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) \subseteq \mathrm{ran}\,(\mathbf{Y}) \Rightarrow s_1\left(\mathbf{U}_{1:k}^\top\overline{\mathbf{N}}\right) \le \epsilon$, we will show that:

$$\text{If} \quad \mathbf{f} \in \mathrm{ran}(\overline{\mathbf{N}}),\ \ \|\mathbf{f}\| = 1,\ \text{ and }\ \left\|\mathbf{U}_{1:k}^\top\mathbf{f}\right\| > \epsilon\,, \quad \text{then} \quad (\mathbf{M}+\mathcal{E})\mathbf{f} \notin \mathrm{ran}\,(\mathbf{Y})\ .$$

To show this, when $\left\|\mathbf{U}_{1:k}^\top\mathbf{f}\right\| > \epsilon$,

- $\|\mathbf{U}_{1:k}^\top(\mathbf{M}+\mathcal{E})\mathbf{f}\| \overset{(i)}{\ge} \|\mathbf{U}_{1:k}^\top\mathbf{M}\mathbf{f}\| - \|\mathbf{U}_{1:k}^\top\mathcal{E}\mathbf{f}\| > s_k(\mathbf{M})\epsilon - \Lambda$
- $\|\mathbf{U}_{k+1:p}^\top(\mathbf{M}+\mathcal{E})\mathbf{f}\| \overset{(ii)}{\le} \|\mathbf{U}_{k+1:p}^\top\mathbf{M}\mathbf{f}\| + \|\mathbf{U}_{k+1:p}^\top\mathcal{E}\mathbf{f}\| \le s_{k+1}(\mathbf{M})\sqrt{1-\epsilon^2} + \Lambda$

where (i) and (ii) follows from triangle inequality for matrix norms. $\|\mathbf{U}_{1:k}^\top \mathbf{Y}\| \leq \epsilon + \eta$ is equivalent to $\frac{\|\mathbf{U}_{1:k}^\top \mathbf{v}\|}{\|\mathbf{U}_{k+1:p}^\top \mathbf{v}\|} \leq \frac{\epsilon+\eta}{\sqrt{1-(\epsilon+\eta)^2}}$ for any unit-norm $\mathbf{v} \in \mathrm{ran}(\mathbf{Y})$. Thus, using (i) and (ii), we obtain $(\mathbf{M}+\mathcal{E})\mathbf{f} \notin \mathrm{ran}(\mathbf{Y})$ since the inequality (22) follows from the Lemma 8.

$$\frac{\epsilon}{\sqrt{1-\epsilon^2}} \frac{s_k(\mathbf{M}) - \Lambda/\epsilon}{s_{k+1}(\mathbf{M}) + \Lambda/\sqrt{1-\epsilon^2}} \geq \frac{\epsilon}{\sqrt{1-\epsilon^2}} \frac{0.75 s_k(\mathbf{M}) + 0.25 s_{k+1}(\mathbf{M})}{0.25 s_k(\mathbf{M}) + 0.75 s_{k+1}(\mathbf{M})} > \frac{\epsilon+\eta}{\sqrt{1-(\epsilon+\eta)^2}}. \tag{22}$$

We can derive (21) from (20) as follows:

$$\begin{aligned} s_1\left((\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\right) &= \sup_{\mathbf{y}\in\mathbb{R}^{p-k}:\|\mathbf{y}\|=1} \left\|(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\right\| \\ &\overset{(iii)}{\leq} \sup_{\mathbf{y}\in\mathbb{R}^{p-k}:\|\mathbf{y}\|=1} \frac{\left\|\mathbf{U}_{k+1:p}^\top(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\right\|}{\sqrt{1-(\epsilon+\eta)^2}} \leq \frac{\left\|\mathbf{U}_{k+1:p}^\top(\mathbf{M}+\mathcal{E})\right\|}{\sqrt{1-(\epsilon+\eta)^2}} \\ &\overset{(iv)}{\leq} \frac{s_{k+1}(\mathbf{M})+\Lambda}{\sqrt{1-(\epsilon+\eta)^2}}. \end{aligned}$$

For validating (iii), observe:

- $(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y} \in \mathrm{ran}(\mathbf{Y})$
- $\|\mathbf{U}_{1:k}^\top(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\| \leq (\epsilon+\eta)\|(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\|$
- $\|\mathbf{U}_{k+1:p}^\top(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\|^2 = \|(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\|^2 - \|\mathbf{U}_{1:k}^\top(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\|^2$

Then we have,

$$\|(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\|^2 \leq \frac{\|\mathbf{U}_{k+1:p}^\top(\mathbf{M}+\mathcal{E})\overline{\mathbf{N}}\mathbf{y}\|^2}{(1-(\epsilon+\eta)^2)}.$$

Finally (iv) follows from (20) where we have $\left\|\mathbf{U}_{k+1:p}^\top(\mathbf{M}+\mathcal{E})\right\| \leq s_{k+1}(\mathbf{M}) + \Lambda$. □

Next, we provide the second lemma for $\mathbf{W}^{(\ell)}$.

Lemma 10 (Amplification)**.** *Assume the regime in the Appendix F.2. Let $\mathbf{M} \in \mathbb{R}^{p\times p}$ be a positive definite matrix and $\mathrm{SVD}(\mathbf{M}) = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, and let $\mathbf{W} \in \mathrm{St}_k(\mathbb{R}^p)$. When $d\left(\mathbf{U}_{1:k}, \mathbf{W}\right) \leq \epsilon + \eta$,*

$$s_k\left((\mathbf{M}+\mathcal{E})\mathbf{W}\right) \geq \sqrt{1-(\epsilon+\eta)^2}(\delta+\sigma^2) - \Lambda\,, \;\; \textit{and} \;\; d\left(\mathbf{U}_{1:k}, (\mathbf{M}+\mathcal{E})\mathbf{W}\right) \leq \epsilon.$$

Proof. First, we show that $s_k\left((\mathbf{M}+\mathcal{E})\mathbf{W}\right) \geq \sqrt{1-(\epsilon+\eta)^2}s_k(\mathbf{M}) - \Lambda$:

$$\begin{aligned} s_k\left((\mathbf{M}+\mathcal{E})\mathbf{W}\right) &\overset{(i)}{\geq} s_k\left(\mathbf{U}_{1:k}^\top(\mathbf{M}+\mathcal{E})\mathbf{W}\right) \\ &\overset{(ii)}{\geq} s_k\left(\mathbf{U}_{1:k}^\top\mathbf{M}\mathbf{W}\right) - \|\mathcal{E}\mathbf{W}\| \\ &\overset{(iii)}{\geq} s_k\left(\mathbf{M}\right) s_k\left(\mathbf{U}_{1:k}^\top\mathbf{W}\right) - \|\mathcal{E}\mathbf{W}\| \\ &\overset{(iv)}{=} s_k\left(\mathbf{M}\right)\sqrt{1-(\epsilon+\eta)^2} - \|\mathcal{E}\mathbf{W}\|\,. \end{aligned}$$

Where (i) follows from the equation $\mathbf{I}_{p\times p} = \mathbf{U}_{1:k}\mathbf{U}_{1:k}^\top + \mathbf{U}_{k+1:p}\mathbf{U}_{k+1:p}^\top$, (ii) follows from the Lemma 5. (iii) follows from:

$$\begin{aligned} s_k\left(\mathbf{U}_{1:k}^\top\mathbf{M}\mathbf{W}\right) &= s_k\left(\mathrm{diag}(s_1(\mathbf{M}),\ldots s_k(\mathbf{M}))\mathbf{U}_{1:k}^\top\mathbf{W}\right) \\ &= \min_{\mathbf{f}\in\mathbb{S}^{p-1}} \|\mathrm{diag}(s_1(\mathbf{M}),\ldots s_k(\mathbf{M}))\mathbf{U}_{1:k}^\top\mathbf{W}\mathbf{f}\| \\ &\geq \min_{\tilde{\mathbf{f}}\in s_k(\mathbf{U}_{1:k}^\top\mathbf{W})\cdot\mathbb{S}^{k-1}} \|\mathrm{diag}(s_1(\mathbf{M}),\ldots s_k(\mathbf{M}))\tilde{\mathbf{f}}\| \\ &= s_k(\mathbf{M})s_k(\mathbf{U}_{1:k}^\top\mathbf{W})\,. \end{aligned}$$

To obtain (iv), let columns of $\tilde{\mathbf{W}}$ represent the space orthogonal to column space of $\mathbf{W}$ and note that both $\mathbf{W}$ and $\tilde{\mathbf{W}}$ have orthonormal columns. Then,

$$\|\mathbf{f}^\top \mathbf{U}_{1:k}^\top (\mathbf{W} + \tilde{\mathbf{W}})\|^2 = \|\mathbf{f}^\top \mathbf{U}_{1:k}^\top \mathbf{W}\|^2 + \|\mathbf{f}^\top \mathbf{U}_{1:k}^\top \tilde{\mathbf{W}}\|^2 = 1$$

and therefore,

$$\min_{\mathbf{f}\in\mathbb{S}^{k-1}} \|\mathbf{f}^\top \mathbf{U}_{1:k}^\top \mathbf{W}\|^2 = 1 - \max_{\mathbf{f}\in\mathbb{S}^{k-1}} \|\mathbf{f}^\top \mathbf{U}_{1:k}^\top \tilde{\mathbf{W}}\|^2 = 1 - (d(\mathbf{U}_{1:k}, \mathbf{W}))^2$$

(iv) now follows from the definition of largest singular value and the assumptions on this Lemma.

We now prove that $d(\mathbf{U}_{1:k}, (\mathbf{M}+\mathcal{E})\mathbf{W}) \leq \epsilon$ or equivalently $s_k(\mathbf{U}_{1:k}, b((\mathbf{M}+\mathcal{E})\mathbf{W}))^2 \geq 1-\epsilon^2$, since

$$d(\mathbf{U}_{1:k}, (\mathbf{M}+\mathcal{E})\mathbf{W}) = s_1\left(\mathbf{U}_{1:k}^\top b((\mathbf{M}+\mathcal{E})\mathbf{W})_\perp\right) = \sqrt{1 - s_k\left(\mathbf{U}_{1:k}^\top b((\mathbf{M}+\mathcal{E})\mathbf{W})\right)^2}.$$

Further,

$$\begin{aligned}
s_k\left(\mathbf{U}_{1:k}^\top b((\mathbf{M}+\mathcal{E})\mathbf{W})\right)^2 &\overset{(vi)}{=} \min_{\mathbf{f}\in\mathbb{S}^{k-1}} \frac{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2}{\|(\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2} \\
&\overset{(vii)}{=} \min_{\mathbf{f}\in\mathbb{S}^{k-1}} \frac{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2}{\|(\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2} \\
&\overset{(viii)}{=} \min_{\mathbf{f}\in\mathbb{S}^{k-1}} \frac{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2}{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2 + \|\mathbf{U}_{k+1:n}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2}
\end{aligned}$$

To obtain (vi) note that by definition of $b((\mathbf{M}+\mathcal{E})\mathbf{W})$ and $(\mathbf{M}+\mathcal{E})\mathbf{W}$ share the same column space and therefore, $\forall\ \mathbf{f} \in \mathbb{R}^k,\ \exists\ \mathbf{y} \in \mathbb{R}^k$ such that $b((\mathbf{M}+\mathcal{E})\mathbf{W})\mathbf{f} = \frac{(\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{y}}{\|(\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{y}\|}$, (vi) then follows from the definition of the largest singular value. (vii) and $(viii)$ follow by projecting $(\mathbf{M}+\mathcal{E})\mathbf{W}$ onto the column spaces of $\mathbf{U}_{1:k}$ and $\mathbf{U}_{k+1:n}$ and noting that $\|\mathbf{U}\mathbf{y}\| = \|\mathbf{y}\|$ when $\mathbf{U}$ has orthonormal columns. Using this decomposition it now suffices to show:

$$\min_{\mathbf{f}\in\mathbb{S}^{k-1}} \frac{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2}{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2 + \|\mathbf{U}_{k+1:p}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|^2} \geq 1-\epsilon^2 \tag{23}$$

To obtain the equation (23) observe that its left hand side is of the form $\min_{\mathbf{f}\in\mathbb{S}^{k-1}} \frac{\mu(f)}{1+\mu(f)}$ with $\mu(f) = \frac{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|}{\|\mathbf{U}_{k+1:p}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|}$ and it monotonically increases in μ and therefore the minimum is attained at the smallest possible value of μ. Therefore, we bound μ from below as:

$$\min_{\mathbf{f}\in\mathbb{S}^{k-1}} \frac{\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|}{\|\mathbf{U}_{k+1:p}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\|} \overset{(viii)}{\geq} \frac{s_k(\mathbf{M}) - \Lambda/\sqrt{1-(\epsilon+\eta)^2}}{s_{k+1}(\mathbf{M}) + \Lambda/(\epsilon+\eta)} \frac{\sqrt{1-(\epsilon+\eta)^2}}{\epsilon+\eta} \overset{(ix)}{>} \frac{\sqrt{1-\epsilon^2}}{\epsilon},$$

where $(viii)$ stems from the fact that for all given unit vector $\mathbf{f} \in \mathbb{R}^k$,

$$\|\mathbf{U}_{1:k}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\| \geq \|\mathbf{U}_{1:k}^\top \mathbf{M}\mathbf{W}\mathbf{f}\| - \|\mathbf{U}_{1:k}^\top \mathcal{E}\mathbf{W}\mathbf{f}\| \geq s_k(\mathbf{M})\sqrt{1-(\epsilon+\eta)^2} - \Lambda$$

and,

$$\|\mathbf{U}_{k+1:p}^\top (\mathbf{M}+\mathcal{E})\mathbf{W}\mathbf{f}\| \leq \|\mathbf{U}_{k+1:p}^\top \mathbf{M}\mathbf{W}\mathbf{f}\| + \|\mathbf{U}_{k+1:p}^\top \mathcal{E}\mathbf{W}\mathbf{f}\| \leq s_{k+1}(\mathbf{M})(\epsilon+\eta) + \Lambda$$

and (ix) can be obtained from the Lemma 8. Substituting $\mu > \frac{\sqrt{1-\epsilon^2}}{\epsilon}$ gives the desired lower bound in the equation (23). □

F.4 Proving the Theorem 2

We split the proof of theorem 2 into three steps. In the first two steps, using Lemma 9 and 10 we identify appropriate matrices $\mathbf{N}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$. Then, in the last step we bound the distance between the output of the robust power method and the real low-dimensional space by bridging them with $\mathcal{M}^{(L)}\mathbf{W}^{(1)}$.

Step 1: Constructing $\mathbf{N}^{(\ell)}$

We construct the sequence $\{\mathbf{N}^{(\ell)}\}_{1\leq\ell\leq L+1}, \mathbf{N}^{(\ell)} \in \mathrm{St}_{p-k}(\mathbb{R}^p)$ so that the following is satisfied:

N.1 $\mathbf{N}^{(L+1)} = \mathbf{U}_{k+1:p}(L)$.

N.2 $\text{ran}((\mathbf{M}(\ell) + \mathcal{E}(\ell))\mathbf{N}^{(\ell)}) \subseteq \text{ran}(\mathbf{N}^{(\ell+1)})$, $\quad \forall \ell \in [L]$.

N.3 $s_1(\mathbf{U}_{1:k}(\ell)^\top \mathbf{N}^{(\ell)}) \leq \epsilon$ and $s_1((\mathbf{M}(\ell) + \mathcal{E}(\ell))\mathbf{N}^{(\ell)}) \leq \dfrac{s_{k+1}(\mathbf{M}(\ell)) + \Lambda}{\sqrt{1-(\epsilon+\eta)^2}}$, $\quad \forall \ell \in [L]$.

N.4 $s_1\left(\mathbf{U}_{1:k}(\ell-1)^\top \mathbf{N}^{(\ell)}\right) \leq \epsilon + \eta$, $\quad 2 \leq \forall \ell \leq L+1$.

To show the existence of $\{\mathbf{N}^{(\ell)}\}_{\ell=1}^{L+1}$ satisfying N.1-N.4, we use the Lemma 9 and backward mathematical induction.

Base case: At $\ell = L+1$, $\mathbf{N}^{(L+1)} = \mathbf{U}_{k+1:p}(L)$, therefore N.4 holds from the model assumption $\|(\mathbf{U}_{k+1:p}(\ell))^\top \mathbf{U}_{1:k}(\ell-1)\| \leq \eta$. Other conditions are required for $\ell \leq L$ and hence $\mathbf{N}^{(L+1)}$ exists.

Inductive Hypothesis: Assume that there exists $\mathbf{N}^{(\ell+1)}$ satisfying N.1-N.4. We show that there exists an $\mathbf{N}^{(\ell)}$. Define $\mathbf{N}^{(\ell)}$ to be the matrix identified as $\overline{\mathbf{N}}$ in the Lemma 9 with $\mathbf{M} = \mathbf{M}(\ell)$, $\mathcal{E} = \mathcal{E}(\ell)$ and $\mathbf{Y} = \mathbf{N}^{(\ell+1)}$. Then, the Lemma 9 shows that $\mathbf{N}^{(\ell)}$ satisfies N.2 and N.3. Further, since

$$
\begin{aligned}
s_1(\mathbf{U}_{1:k}(\ell-1)\mathbf{N}^{(\ell)}) &\overset{(i)}{=} \|\mathbf{U}_{1:k}(\ell-1)\mathbf{U}_{1:k}^\top(\ell-1) - (\mathbf{I}_{p\times p} - \mathbf{N}^{(\ell)}(\mathbf{N}^{(\ell)})^\top)\| \\
&\overset{(ii)}{\leq} \|\mathbf{U}_{1:k}(\ell-1)\mathbf{U}_{1:k}^\top(l-1) - \mathbf{U}_{1:k}(\ell)\mathbf{U}_{1:k}^\top(\ell)\| + \|\mathbf{U}_{1:k}(\ell)\mathbf{U}_{1:k}^\top(\ell) - (\mathbf{I}_{p\times p} - \mathbf{N}^{(\ell)}(\mathbf{N}^{(\ell)})^\top)\| \\
&\overset{(iii)}{\leq} \eta + \epsilon
\end{aligned}
$$

where, (i) follows from Lemma 3, (ii) follows from triangle inequality and (iii) follows since $\|(\mathbf{U}_{k+1:p}(\ell))^\top \mathbf{U}_{1:k}(l-1)\| \leq \eta$. Therefore, we can conclude that there exists desired sequence $\{\mathbf{N}^{(\ell)}\}_{1\leq\ell\leq L+1}$ with properties N.1-N.4.

From the properties N.1-N.3 of $\{\mathbf{N}^{(\ell)}\}_{1\leq\ell\leq L+1}$, we have:

$$
\text{ran}(\mathcal{M}^{(L)}\mathbf{N}^{(1)}) \subseteq \text{ran}(\mathbf{U}_{k+1:p}(L)) \text{ and } \|\mathcal{M}^{(L)}\mathbf{N}^{(1)}\| \leq \prod_{\ell=1}^{L}\left(\frac{s_{k+1}(\mathbf{M}(\ell)) + \Lambda}{\sqrt{1-(\epsilon+\eta)^2}}\right). \tag{24}
$$

Step 2: Constructing $\mathbf{W}^{(\ell)}$

Next, we define the sequence $\{\mathbf{W}^{(\ell)}\}_{\ell=1}^{L+1}$, $\mathbf{W}^{(\ell)} \in \text{St}_k(\mathbb{R}^p)$ as follows:

W.1 $\mathbf{W}^{(1)} \in \text{St}_k(\mathbb{R}^p)$ be a matrix such that $(\mathbf{W}^{(1)})^\top \mathbf{N}^{(1)} = \mathbf{0}$.

W.2 $\mathbf{W}^{(\ell+1)} = \text{Gram-Schmidt}((\mathbf{M}(\ell) + \mathcal{E}(\ell))\mathbf{W}^{(\ell)})$, $\quad \forall \ell \in [L]$.

From N.4 and the triangle inequality, we have $d(\mathbf{U}_{1:k}(1), \mathbf{W}^{(1)}) \leq \epsilon + \eta$. Then, the Lemma 10 implies that:

CW.1 $s_k\left((\mathbf{M}(\ell) + \mathcal{E}(\ell))\mathbf{W}^{(\ell)}\right) \geq \sqrt{1-(\epsilon+\eta)^2}\, s_k(\mathbf{M}(\ell)) - \Lambda$, $\quad \forall \ell \in [L]$.

CW.2 $d(\mathbf{U}_{1:k}(\ell), \mathbf{W}^{(\ell+1)}) \leq \epsilon$ and since, $\|\mathbf{U}_{1:k}(\ell)^\top \mathbf{U}_{1:k}(\ell+1)\| \leq \eta$, we have $d(\mathbf{U}_{1:k}(\ell), \mathbf{W}^{(\ell)}) \leq \epsilon + \eta$. $(\forall \ell \in [L])$

From CW.1 and CW.2, we have

$$
d(\mathbf{W}^{(L+1)}, \mathbf{U}_{1:k}(L)) \leq \epsilon \text{ and } s_k(\mathcal{M}^{(L)}\mathbf{W}^{(1)}) \geq \prod_{\ell=1}^{L}\left(\sqrt{1-(\epsilon+\eta)^2}\, s_k(\mathbf{M}(\ell)) - \Lambda\right). \tag{25}
$$

This establishes the existence and properties of the sequence $\{\mathbf{N}^{(\ell)}, \mathbf{W}^{(\ell)}\}_{\ell=1}^{L+1}$. We now use this characterization to bound the distance between the k-dimensional subspace of $\mathbb{R}^p$ and $\mathcal{M}^{(L)}\hat{\mathbf{U}}_{(0)}$.

Since, N.3 bounds the distance between the $(p-k)$ dimensional subspace of $\mathbb{R}^p$ and $\mathcal{M}^{(L)}\hat{\mathbf{U}}_{(0)}$, we consider bound the distance between $\mathcal{M}^{(L)}\hat{\mathbf{U}}_{(0)}$ and $\mathbf{W}^{(L+1)}$.

Step 3: Distance between actual and recovered spaces

Now, we upper bound the distance between the output of the power method $\hat{\mathbf{U}}_{1:k}(L)$ and the first k singular vectors of the true underlying subspace $\mathbf{U}_{1:k}(L)$, $d(\mathbf{U}_{1:k}(L), \widehat{\mathbf{U}}_{1:k}(L))$. From the triangle inequality we have,

$$d(\mathbf{U}_{1:k}(L), \widehat{\mathbf{U}}_{1:k}(L)) \leq d(\mathbf{U}_{1:k}(L), \mathbf{W}^{(L+1)}) + d(\mathcal{M}^{(L)}\mathbf{W}^{(1)}, \widehat{\mathbf{U}}_{1:k}(L)). \tag{26}$$

(Note that $\mathbf{W}^{(L+1)}$ and $\mathcal{M}^{(L)}\mathbf{W}^{(1)}$ represents the same column space) From the equation (25), we have

$$d(\mathbf{U}_{1:k}(L), \mathcal{M}^{(L)}\mathbf{W}^{(1)}) \leq \epsilon\,. \tag{27}$$

To bound the second term in the RHS of the equation (26), consider the following:

$$\begin{aligned}
d(\mathcal{M}^{(L)}\mathbf{W}^{(1)}, \widehat{\mathbf{U}}_{1:k}(L)) &\overset{(i)}{=} \left\|(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp^\top \widehat{\mathbf{U}}_{1:k}(L)\right\| \\
&\overset{(ii)}{=} \left\|(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp^\top \mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0)((\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))^\top \mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))^{-1/2}\right\| \\
&\overset{(iii)}{\leq} \|(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp^\top \mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0)\|\|((\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))^\top \mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))^{-1/2}\| \\
&\overset{(iv)}{\leq} \frac{\left\|(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp^\top \mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0)\right\|}{s_k(\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))} \\
&\overset{(v)}{=} \frac{\left\|(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp^\top \mathcal{M}^{(L)}\mathbf{N}^{(1)}(\mathbf{N}^{(1)})^\top \widehat{\mathbf{U}}_{1:k}(0)\right\|}{s_k(\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))} \\
&\overset{(vi)}{\leq} \frac{\|(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp^\top\|\|\mathcal{M}^{(L)}\mathbf{N}^{(1)}\|\|(\mathbf{N}^{(1)})^\top \widehat{\mathbf{U}}_{1:k}(0)\|}{s_k(\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))}\,.
\end{aligned}$$

where, (i) follows by substituting the definition of distance function and (ii) follows by observing that due to the power iterations $\widehat{\mathbf{U}}_{1:k}(L)$ is an orthonormal basis of $\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0)$ and therefore can be written as $b(\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))$. (iii) follows by using the Cauchy-Schwarz inequality for matrix norms. (iv) follows by noting that $\|((\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))^\top \mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0))^{-1/2}\| = 1/\|\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0)\| \leq 1/s_k(M_L\widehat{\mathbf{U}}_{1:k}(0))$. To obtain (v), decompose the numerator of (iv) as $(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp \mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0) = (\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp \mathcal{M}^{(L)}(\mathbf{W}_1\mathbf{W}_1^\top + \mathbf{N}_1\mathbf{N}_1^\top)\mathcal{M}^{(L)}\widehat{\mathbf{U}}_{1:k}(0)$ and note that by orthogonality of $(\mathcal{M}^L\mathbf{W}^{(1)})_\perp$ and $(\mathcal{M}^{(L)}\mathbf{W}_1)$, $(\mathcal{M}^L\mathbf{W}^{(1)})_\perp(\mathcal{M}^{(L)}\mathbf{W}_1)\mathbf{W}_1^\top\widehat{\mathbf{U}}_{1:k}(0) = \mathbf{0}$. Finally, (vi) follows from the repeated application of Cauchy-Schwarz inequality for matrix norms.

Further, when $s_k(\mathcal{M}^{(L)}\mathbf{W}^{(1)})s_k((\mathbf{W}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)) - \|\mathcal{M}^{(L)}\mathbf{N}^{(1)}\|\left\|(\mathbf{N}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)\right\| > 0$, (Consider the amplifying/contracting singular value argument for $s_k(\mathcal{M}^{(L)}\mathbf{W}^{(1)})$ and $s_1(\mathcal{M}^{(L)}\mathbf{N}^{(1)})$ respectively.)

$$\begin{aligned}
&d(\mathcal{M}^{(L)}\mathbf{W}^{(1)}, \widehat{\mathbf{U}}_{1:k}(L)) \\
&\overset{(vii)}{\leq} \frac{\|\mathcal{M}^{(L)}\mathbf{N}^{(1)}\|\left\|(\mathbf{N}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)\right\|}{s_k(\mathcal{M}^{(L)}\mathbf{W}^{(1)})s_k((\mathbf{W}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)) - \|\mathcal{M}^{(L)}\mathbf{N}^{(1)}\|\left\|(\mathbf{N}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)\right\|} \\
&\overset{(viii)}{\leq} \frac{\|\mathcal{M}^{(L)}\mathbf{N}^{(1)}\|\left\|(\mathbf{N}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)\right\| \Big/ s_k(\mathcal{M}^{(L)}\mathbf{W}^{(1)})s_k((\mathbf{W}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0))}{1 - \|\mathcal{M}^{(L)}\mathbf{N}^{(1)}\|\left\|(\mathbf{N}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)\right\| \Big/ s_k(\mathcal{M}^{(L)}\mathbf{W}^{(1)})s_k((\mathbf{W}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0))} \\
&\overset{(ix)}{\leq} \frac{\beta^{-L}\left\|(\mathbf{N}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)\right\| \Big/ s_k((\mathbf{W}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0))}{1 - \beta^{-L}\left\|(\mathbf{N}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0)\right\| \Big/ s_k((\mathbf{W}^{(1)})^\top\widehat{\mathbf{U}}_{1:k}(0))},
\end{aligned} \tag{28}$$

where (vii) follows from $\|(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp\| = 1$ since $(\mathcal{M}^{(L)}\mathbf{W}^{(1)})_\perp$ is a projection matrix and (ix) stems from the equation (24) and (25). Putting (27) and (28) onto (26), we have

$$d(\mathbf{U}_{1:k}, \widehat{\mathbf{U}}_{1:k}) \leq \epsilon + \min\left\{1, 2\beta^{-L}\frac{\left\|\mathbf{N}^{(1)\top}\widehat{\mathbf{U}}_{1:k}(0)\right\|}{s_k(\mathbf{W}^{(1)\top}\widehat{\mathbf{U}}_{1:k}(0))}\right\}.$$

Further, by Lemma 2.5 in [29], we have

$$\frac{\|(\mathbf{N}^{(1)})\widehat{\mathbf{U}}_{1:k}(0)\|}{s_k(\mathbf{W}^{(1)}\widehat{\mathbf{U}}_{1:k})(0)} \leq \frac{c\sqrt{p}}{\sqrt{p}-\sqrt{k-1}},$$

with probability $1 - c^{\Omega(p-k+1)} - e^{-\Omega(p)}$ $(c > 0)$.

Therefore, when

$$L > \frac{\log\left(c\sqrt{p}/(\sqrt{p}-\sqrt{k-1})\right)}{\log(\beta)},$$

we have:

$$d(\mathbf{U}_{1:k}, \widehat{\mathbf{U}}_{1:k}) \leq \epsilon + \mathcal{O}(\frac{\beta^{-L}\sqrt{p}}{\sqrt{p}-\sqrt{k-1}}).$$

for probability greater than $1 - 1/T - c^{\Omega(p-k+1)} - e^{-\Omega(p)}$. The term $1/T$ stems from the probabilistic upper bound of $\|\mathcal{E}(\ell)\|$. Finally, by looking back at the order of ϵ:

$$\epsilon = \frac{4\Lambda}{\delta} \sim \frac{C_{\mathrm{NPM}}^{2/3}\log(2pT^2)^{1/3}\Gamma^{1/3}}{\delta},$$

we get our desired result.

G Proof of Lemma 2

In this section, we analyze the error matrix for Oja's algorithm. First, we consider the regime that adversary factor Γ is strictly, indeed far less than the learning rate ζ, to apply Davis-Kahan theorem (4) properly. Although the more refined calculation may eliminate this condition, we would provide a more intuitive and straightforward analysis to show our qualitative result. To simply bound the error matrix, we consider the virtual learning block with size $B = \lceil 1/\zeta \rceil$ while the total time is T is strictly larger than B. Furthermore, we assume the exact relation; $B\,\zeta = 1$ for simplicity. The following lemma controls the error caused by covariance matrix estimators on the product case.

Similar to the Lemma 1, we consider the environment with $\mathrm{Cov}[\mathbf{x}_t\mathbf{x}_t^\top] = \mathbf{A}_t\mathbf{A}_t^\top + \sigma^2\mathbf{I}_{p\times p}$ and Assumption 1 holds. Moreover, we ought to restrict our analysis to particular regime $\Gamma \ll \zeta$ for convenience. We first decompose the covariance estimator of the Oja's algorithm with $\mathbf{M}^{\mathrm{Oja}}(\ell)$ and $\mathcal{E}(\ell)$. Note that $\mathbf{M}^{\mathrm{Oja}}(\ell)$ should be the positive semi-definite matrix to apply proof arguments at the Theorem 2.

$$\prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbf{x}_t\mathbf{x}_t^\top) = \underbrace{\prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top)}_{\mathbf{M}^{\mathrm{Oja}}(\ell)} + \mathcal{E}(\ell)\left(= \mathbf{M}^{\mathrm{Oja}}(\ell) + e^{\tilde{\delta}+\sigma^2}\,\mathcal{E}'(\ell)\right).$$

By decomposing the error matrix, we get the following two terms.

$$\begin{aligned}\mathcal{E}(\ell) &= \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\mathbf{x}_t\mathbf{x}_t^\top) - \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top)\\ &= \underbrace{\prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \frac{1}{B}\mathbf{x}_t\mathbf{x}_t^\top) - \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \frac{1}{B}\,\mathbb{E}\,\mathbf{x}_t\mathbf{x}_t^\top)}_{\mathcal{E}_1(\ell)} + \underbrace{\prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_t\mathbf{x}_t^\top) - \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top)}_{\mathcal{E}_2(\ell)}.\end{aligned}$$

G.1 Bounding $\max_\ell \|\mathcal{E}_1(\ell)\|$ with probability $1 - 1/T$

For bound $\|\mathcal{E}_1(\ell)\|$, we consider the matrix multiplicative concentration inequality.

Lemma 11 ('Perturbations of the identity' in [32])**.** *Consider the independent family of matrices $\mathbf{Z}_1, \dots, \mathbf{Z}_B \in \mathbb{R}^{p\times p}$, each drawn from the distributions satisfying:*

$$\|\mathbb{E}\mathbf{Z}_t\| \le \tilde{\delta} + \sigma^2 \quad \textit{and} \quad \|\mathbf{Z}_t - \mathbb{E}\,\mathbf{Z}_t\| \le \mathcal{M} \quad (\forall\, t \in [B])\,.$$

Then, for $\Pi \ge p\, e^{-B/2\mathcal{M}^2}$, the product $\mathbf{Z} = (\mathbf{I}_{p\times p} + \mathbf{Z}_B/B)\cdots(\mathbf{I}_{p\times p} + \mathbf{Z}_1/B)$ is bounded as the below argument, with the probability greater than $1 - \Pi$:

$$\|\mathbf{Z} - \mathbb{E}\mathbf{Z}\| \le e^{\tilde{\delta}+\sigma^2}\sqrt{\frac{2e^2\,\mathcal{M}^2}{B}\log\frac{p}{\Pi}}\,.$$

In our situation $\mathbf{Z}_t$ and $\mathbb{E}\mathbf{Z}_t$ are $\mathbf{x}_{\ell B+t}\mathbf{x}_{\ell B+t}^\top$ and $\mathbf{A}_{\ell B+t}\mathbf{A}_{\ell B+t}^\top + \sigma^2\mathbf{I}_{p\times p}$. By setting $\Pi = 1/T^2$ and applying union bound on $\ell \in [L]$, we have:

$$\max_{1\le\ell\le L(=T/B)} \|\mathcal{E}_1(\ell)\| \le e^{\tilde{\delta}+\sigma^2}\sqrt{\frac{2e^2\,\mathcal{M}^2\log pT^2}{B}} = e^{\tilde{\delta}+\sigma^2}\cdot C_{\mathrm{Oja}}\sqrt{\frac{\log 2pT^2}{B}}\,,$$

for probability greater than $1 - 1/T$. We define C_{Oja} as $\sqrt{2}e\mathcal{M}$.

G.2 Bounding $\|\mathcal{E}_2(\ell)\|$ for all ℓ

Next, we present the upper bound for $\mathcal{E}_2(\ell)$, using the condition (2). We define $\mathbf{Y}_t^{(\ell)}$ as $\mathbb{E}\,\mathbf{x}_{\ell B+t}\mathbf{x}_{\ell B+t}^\top - \mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top$ and use the fact $\|\mathbf{Y}_t^{(\ell)}\| \le (\ell B - t)\,\Gamma$. We rewrite the $\mathcal{E}_2(\ell)$ as follows:

$$\begin{aligned}
\mathcal{E}_2(\ell) &= \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_t\mathbf{x}_t^\top) - \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top)\\
&= \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top + \zeta\,\mathbf{Y}_t^{(\ell)}) - \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \zeta\,\mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top)\,.
\end{aligned}$$

Therefore, we have the following expanded terms:

$$\begin{aligned}
\|\mathcal{E}_2(\ell)\| \le{}& \zeta\,(1+\zeta(\tilde{\delta}+\sigma^2))^{B-1}\sum_{t_1}\|\mathbf{Y}_{t_1}^{(\ell)}\| + \zeta^2\,(1+\zeta(\tilde{\delta}+\sigma^2))^{B-2}\sum_{t_1<t_2}\|\mathbf{Y}_{t_1}^{(\ell)}\|\|\mathbf{Y}_{t_2}^{(\ell)}\|\\
&+\cdots+\zeta^{B-1}\,(1+\zeta(\tilde{\delta}+\sigma^2))\sum_{t_1<\cdots<t_{B-1}}\|\mathbf{Y}_{t_1}^{(\ell)}\|\cdots\|\mathbf{Y}_{t_{B-1}}^{(\ell)}\|\,.
\end{aligned}$$

Then by spaciously bound the sum of products:

$$\sum_{t_1<\cdots<t_n}\|\mathbf{Y}_{t_1}^{(\ell)}\|\cdots\|\mathbf{Y}_{t_n}^{(\ell)}\| \le \sum_{t_1}\|\mathbf{Y}_{t_1}^{(\ell)}\|\sum_{t_2}\|\mathbf{Y}_{t_2}^{(\ell)}\|\cdots\sum_{t_n}\|\mathbf{Y}_{t_n}^{(\ell)}\| \le \left(\frac{B^2\,\Gamma}{2}\right)^n\,.$$

Thus, when we substitute above bound,

$$\begin{aligned}
\|\mathcal{E}_2(l)\| &\le (1+\zeta(\tilde{\delta}+\sigma^2))^B\sum_{n=1}^{B-1}\left(\frac{\zeta}{1+\zeta(\tilde{\delta}+\sigma^2)}\right)^n\left(\frac{B^2\,\Gamma}{2}\right)^n\\
&= (1+\zeta(\tilde{\delta}+\sigma^2))^B\sum_{n=1}^{B-1}\left(\frac{B\Gamma}{2(1+\zeta(\tilde{\delta}+\sigma^2))}\right)^n\\
&\le \left(1+\frac{\tilde{\delta}+\sigma^2}{B}\right)^B\frac{B\Gamma}{2(1-B\Gamma/2)} \le e^{\tilde{\delta}+\sigma^2}\cdot\frac{B\Gamma}{2(1-B\Gamma/2)}\,.
\end{aligned}$$

Finally, we have the following result:

$$\|\mathcal{E}_2(\ell)\| \le e^{\tilde{\delta}+\sigma^2}\cdot\left(\frac{B\Gamma}{2}+\epsilon_{B,\Gamma}\right)\,,$$

where $\epsilon_{B,\Gamma} = (B\Gamma)^2/(1-B\Gamma/2)$ is negligible if Γ is sufficiently small relative to ζ.

G.3 Bounding $\|\mathcal{E}(\ell)\|$ for all ℓ, with probability $1 - 1/T$

Consequently, with probability $1 - 1/T$, we have:

$$\max_{1\le\ell\le L}\|\mathcal{E}(\ell)\| \le \max_{1\le\ell\le L}\|\mathcal{E}_1(\ell)\| + \max_{1\le\ell\le L}\|\mathcal{E}_2(\ell)\|$$
$$\le e^{\tilde{\delta}+\sigma^2}\cdot\left[C_{\text{Oja}}\sqrt{\frac{\log pT^2}{B}} + \frac{B\Gamma}{2} + \epsilon_{B,\Gamma}\right].$$

Note also that we have the following equivalent formulation for scaled error matrix $\mathcal{E}'(\ell)$:

$$\max_{1\le\ell\le L}\|\mathcal{E}'(\ell)\| \le C_{\text{Oja}}\sqrt{\frac{\log pT^2}{B_\zeta}} + \frac{B_\zeta\Gamma}{2} + \epsilon_{B_\zeta,\Gamma} \quad (B_\zeta = B = 1/\zeta)\,. \tag{29}$$

H Proof of Theorem 3

For the Theorem 3, we use the same technique with the Theorem 2 for the robust power method. To apply the iterative method we used in the previous proofs, **it is enough to maintain the Lemma 8 at the Appendix F.2**. However, the problem is that the spectrum are exponentiated ($s_i(\mathbf{M}(\ell)) \to (1 + s_i(\mathbf{M}(\ell))/B)^B$), therefore δ that we defined is no longer the spectral gap of $\mathbf{M}(\ell)$ in the Oja's algorithm case. Observe:

$$\mathbf{M}^{\text{Oja}}(\ell) = \prod_{t=(\ell-1)B+1}^{\ell B}(\mathbf{I}_{p\times p} + \frac{1}{B}\mathbb{E}\,\mathbf{x}_{\ell B}\mathbf{x}_{\ell B}^\top) = \left(\mathbf{I}_{p\times p} + \frac{\mathbf{A}_{\ell B}\mathbf{A}_{\ell B}^\top + \sigma^2\mathbf{I}_{p\times p}}{B}\right)^B.$$

So we should reconsider the condition on $\mathbf{M}(\ell)$ for Oja's algorithm. Since we can use the same argument in the Appendix F.3-F.4, it is enough to reset parameters and regime.

H.1 Deriving optimal learning rate ζ

Correspondingly, we provide the guide for proving the Theorem 3 when the equation (29) holds for probability greater than $1 - 1/T$ as done in the Section F.1. Consider the upper bound (for probability greater than $1 - 1/T$) on $\max_\ell \|\mathcal{E}'(\ell)\|$ from the Section G. We would neglect the term $\epsilon_{B,\Gamma}$ since we are considering the regime $\Gamma \ll \zeta$.

$$\max_{1\le\ell\le L}\|\mathcal{E}'(\ell)\| \le C_{\text{Oja}}\sqrt{\frac{\log pT^2}{B_\zeta}} + \frac{B_\zeta\Gamma}{2}\,.$$

By differentiating and find the critical point, we have the following optimal learning rate:

$$(\zeta_{\text{opt}})^{-1} = B_{\zeta_{\text{opt}}} = \frac{{C_{\text{Oja}}}^{2/3}\log(pT^2)^{1/3}}{\Gamma^{2/3}} = \Omega\left(\frac{{C_{\text{Oja}}}^{2/3}\log(pT^2)^{1/3}}{\Gamma^{2/3}}\right).$$

In this case, the uniform upper bound for error matrix becomes:

$$\max_\ell\|\mathcal{E}'(\ell)\| \le \frac{3}{2}C_{\text{Oja}}^{2/3}\log(pT^2)^{1/3}\Gamma^{1/3}\,.$$

H.2 Defining Regime and Parameters

We define the following parameters:

- $\Lambda_{\text{Oja}} := e^{\tilde{\delta}+\sigma^2}\frac{3}{2}B_{\zeta_{\text{opt}}}\Gamma \quad (\ge \max_\ell\|\mathcal{E}(\ell)\|)$
- $\delta_{\text{Oja}} := \left(1 + \frac{\delta+\sigma^2}{B_{\zeta_{\text{opt}}}}\right)^{B_{\zeta_{\text{opt}}}} - \left(1 + \frac{\sigma^2}{B_{\zeta_{\text{opt}}}}\right)^{B_{\zeta_{\text{opt}}}} \quad (\le s_k(\mathbf{M}^{\text{Oja}}) - s_{k+1}(\mathbf{M}^{\text{Oja}}))$
- $\sigma^2_{\text{Oja}} := \left(1 + \frac{\sigma^2}{B_{\zeta_{\text{opt}}}}\right)^{B_{\zeta_{\text{opt}}}} \quad (= s_{k+1}(\mathbf{M}^{\text{Oja}}))$

- $\eta_{\text{Oja}} := \frac{B_{\zeta_{\text{opt}}}\Gamma}{\delta - B_{\zeta_{\text{opt}}}\Gamma}$

Note that we may use the same η value with the case of power method, since it measures the distance between column space, which is invariant with the transformation $\mathbf{M}^{\text{Oja}} \mapsto (\mathbf{I} + \mathbf{M}^{\text{Oja}}/B)^B$. Let us consider the following approximations to simply our regime:

$$\delta_{\text{Oja}} = \left(1 + \frac{\delta + \sigma^2}{B_{\zeta_{\text{opt}}}}\right)^{B_{\zeta_{\text{opt}}}} - \left(1 + \frac{\sigma^2}{B_{\zeta_{\text{opt}}}}\right)^{B_{\zeta_{\text{opt}}}} \simeq \delta e^{\sigma^2 + \delta} \text{ and } \sigma^2_{\text{Oja}} \simeq e^{\sigma^2}.$$

Similar to the proof of the Theorem 2, we assumed the following regime:

- $\Lambda_{\text{Oja}} := e^{\tilde{\delta}+\sigma^2}\frac{3}{2}B_{\zeta_{\text{opt}}}\Gamma = e^{\tilde{\delta}+\sigma^2}\frac{3}{2}C_{\text{Oja}}^{2/3}\log(pT^2)^{1/3}\Gamma^{1/3} \leq \frac{1}{16}\delta_{\text{Oja}}$, from the regime : $\Gamma = \mathcal{O}\left(\frac{\delta^3_{\text{Oja}}}{e^{3(\tilde{\delta}+\sigma^2)}C^2_{\text{Oja}}\log(pT^2)}\right) = \mathcal{O}\left(\delta^3/e^{3(\tilde{\delta}-\delta)}\mathcal{M}^2\log(pT^2)\right)$
- $\delta_{\text{Oja}} \geq 12\sigma^2_{\text{Oja}}/17$, and
- $\epsilon_{\text{Oja}} := \frac{4\Lambda_{\text{Oja}}}{\delta_{\text{Oja}}} \leq \frac{1}{4}$, satisfying $\eta_{\text{Oja}} \leq \frac{1}{5}\epsilon_{\text{Oja}}$

For the second item, we consider the following sufficient condition:

$$\delta_{\text{Oja}} \geq \frac{12}{17}\sigma^2_{\text{Oja}} \sim \delta e^{\sigma^2+\delta} \geq \frac{12}{17}e^{\sigma^2} \Longleftarrow \delta \geq \frac{12}{17}.$$

Assuming the below regime on the Oja's algorithm

$$36e^{\tilde{\delta}+\sigma^2}B_{\zeta_{\text{opt}}}\Gamma = 24e^{\tilde{\delta}+\sigma^2}C_{\text{Oja}}^{2/3}\log(pT^2)^{1/3}\Gamma^{1/3} \leq \delta_{\text{Oja}},$$

the Lemma 8 and the section F.3-F.4 follows. Finally, by considering ϵ_{Oja} has the order:

$$\epsilon_{\text{Oja}} \sim \frac{\Lambda_{\text{Oja}}}{\delta_{\text{Oja}}} \sim \frac{e^{\tilde{\delta}+\sigma^2}C_{\text{Oja}}^{2/3}\log(pT^2)^{1/3}\Gamma^{1/3}}{\delta e^{\sigma^2}} \sim e^{\tilde{\delta}}\frac{\mathcal{M}^{2/3}\log(pT^2)^{1/3}\Gamma^{1/3}}{\delta},$$

we obtain the desired result.

I Experiment Settings

I.1 Random matrix generation

We generate $(\mathbf{A}_t)_{t=1}^T \in \mathbb{R}^{p\times k}$ as the product of three matrices, $\mathbf{U}_t \in \text{St}_p(\mathbb{R}^p), \mathbf{D}_t \in \mathbb{R}^{p\times k}$(; diagonal), and $\mathbf{V}_t \in \text{St}_k(\mathbb{R}^k)$. We update each matrix at each iteration and multiply them to calculate $\mathbf{A}_t = \mathbf{U}_t\,\mathbf{D}_t\,\mathbf{V}_t{}^\top$. First, we generate a Gaussian random matrix and then perform QR decomposition and use the resulting right matrix as $\mathbf{V}_t$. $\mathbf{D}_t$ is a diagonal matrix, with diagonal elements uniformly sampled from $\{-\sqrt{\delta}, \sqrt{\delta}\}$. This results in $\mathbf{A}_t$ satisfying adversarial budget(Γ) and the spectral gap(δ) condition:

$$s_k(\mathbf{A}_t\mathbf{A}_t^\top) = \delta \cdot s_k\left[\mathbf{U}_t\begin{pmatrix}\mathbf{I}_{k\times k} & 0\\ 0 & 0\end{pmatrix}\mathbf{U}_t^\top\right] = \delta, \tag{30}$$

and

$$\|\mathbf{A}_t\mathbf{A}_t^\top - \mathbf{A}_{t-1}\mathbf{A}_{t-1}^\top\| = \delta\left\|\mathbf{U}_t\begin{pmatrix}\mathbf{I}_{k\times k} & 0\\ 0 & 0\end{pmatrix}\mathbf{U}_t^\top - \mathbf{U}_{t-1}\begin{pmatrix}\mathbf{I}_{k\times k} & 0\\ 0 & 0\end{pmatrix}\mathbf{U}_{t-1}^\top\right\| \leq \Gamma. \tag{31}$$

We initialize $\mathbf{U}_0$ to random orthogonal matrix and rotate $\mathbf{U}_{t-1}$ to generate $\mathbf{U}_t$. Then the first condition is automatically satisfied. For the second condition, we restrict the structure of the random rotation matrix $\mathbf{R}_t$. Assume that $\mathbf{U}_t = \mathbf{U}_{t-1}\mathbf{R}_t$ ($\mathbf{R}_t\,\mathbf{R}_t^\top = \mathbf{I}_p$). Then the second condition becomes:

$$\left\|\mathbf{R}_t\begin{pmatrix}\mathbf{I}_{k\times k} & 0\\ 0 & 0\end{pmatrix}\mathbf{R}_t^\top - \begin{pmatrix}\mathbf{I}_{k\times k} & 0\\ 0 & 0\end{pmatrix}\right\| \leq \Gamma/\delta < 1. \tag{32}$$

(We only consider the case of $\Gamma < \delta$ on the experiments.) To satisfy the above condition, consider $2N(\leq k)$ indices i_1, ..., i_{2N} among $\{1, 2, ..., k\}$ with no replacement, where only one of index for each (i_{2n-1}, i_{2n}) pair lies in $\{1, \ldots, k\}$. Next, we select $\theta_1, \ldots \theta_N$ from the range $[-\sin^{-1}(\Gamma/\delta), \sin^{-1}(\Gamma/\delta)]$.

After initializing $\mathbf{R}_t$ $(= \mathbf{I}_{p\times p})$, we write θ_n-rotation matrix on the 2×2 (i_{2n-1}, i_{2n})-submatrix. Then it can be easily shown that the singular value of the matrix in RHS of the equation (32) becomes $|\sin\theta_1|, \ldots |\sin\theta_N|, 0, \ldots, 0$. Since θ_n lies in the limited range $[-\sin^{-1}(\Gamma/\delta), \sin^{-1}(\Gamma/\delta)]$, we have the desired inequality (32).

I.2 Environments

We used the value $N = 1$, $(i_1, i_2) = (1, p)$, and $\theta_1 = \sin^{-1}(\Gamma/\delta)$ for all synthetic experiments. We implement algorithms with NumPy and NumBa library. We used the dimensions $(p, k) = (100, 5)$. We run each algorithm during $T = 144000$ for each stage. This value is compatible with the slowest convergence time among the case we tested. Generally, we consider the case (δ, σ)=(1.00, 0.15). When running each algorithm, we maintained B-list and η-list respectively as follows:

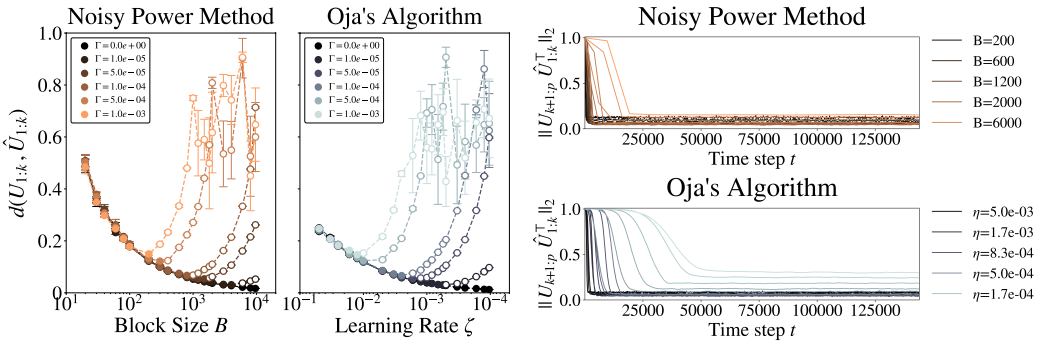
(Noisy Power Method) `B_list` $= [\,2, 3, 8, 10, 20, 30, 40, 60, 300, 400, 600, 800, 1000, 1200, 1500,$
$1800, 2000, 3000, 4000, 6000, 8000, 9600\,]$

(Oja's Algorithm) `zeta_list` $= [\,$`float`$(1/B)$ for B in `B_list`$\,]$

I.3 Additional synthetic experiment

We provide the additional result for larger perturbations in Figure 3a. We repeated five experiments for each algorithm and learning parameter. Except for the magnitude of Γ and number of repetitions, every other set was the same as Figure 1a. Qualitatively, the result shows the same tendency as the result provided in the main paper.

In the second set of experiments in Figure 3b, we show that the noisy power method converges for different values of adversarial budget Γ for different block sizes, B. The first observation from these sets of experiments is that there is an optimal block size B that attains the minimum error. Such behavior is in line with Theorem 2. Another key observation from these figures is that a smaller block size implies faster convergence; this is also in line with the dependence on the number of blocks $(L = \frac{T}{B})$ in Theorem 2.



(a) Convergence of noisy power method and Oja's algorithm. Variation with block size B and learning rate ζ for $(\sigma, \delta, p, k, \Gamma)$=$(0.15, 1.0, 100, 5, 3.0e-5)$.

(b) Convergence of noisy power method and Oja's algorithm, for bigger perturbations. (σ, δ, p, k)=$(0.15, 1.0, 100, 5)$.

Figure 3: Numerical results. We used the setting $(\sigma, \delta, p, k) = (0.15, 1.0, 100, 5)$.

I.4 Details for each experiment

In this section, we expand on the computational setup used to generate Figure 1 and Figure 3.

For the Figure 1a, we plot the distance between the estimated subspace and true space at the termination time T for two algorithms with learning parameters at the Section I.2. We used four different $\Gamma : [0.0, 1.0e{-}5, 3.0e{-}5, 5.0e{-}5]$. We run the experiment 10 times, and plot error bar on each marker.

For Figure 3a, we used same methodology with the Figure 1a, but the larger perturbations $\Gamma : [0.0, 1.0e{-}5, 5.0e{-}5, 1.0e{-}4, 5.0e{-}4, 1.0e{-}3]$ were used. We ran 5 experiments each.

For the Figure 1b, we find the empirically optimal value of the block size B of the noisy power method and the learning rate ζ of Oja's algorithm for various Γ. We first identify a lower and upper bound on the optimal value (B and ζ) from the simulations done for Figure 1a (For simulations for the Figure 1a we used learning parameters denoted in section I.2.) Once we identify upper and lower bounds for optimal value, we split the interval into 50 points and run each experiment with a fixed value of B belonging to this interval for 30 runs. Finally, we calculate the average and standard deviation on those 30 runs. We denote 50 candidates for each Γ with small dots. The optimal value incurring the least average convergence error is plotted as a big marker. Smaller markers for each Γ denotes the parameter, which has lower $\texttt{avg} + \texttt{std}$ than the optimal value's case.

In Figure 3b, we visualize the convergence of two algorithms for various learning parameters, B, and ζ. We reused the experiment result from Figure 1a. The first observation from these experiments is that there is an optimal learning parameter that attains the minimum error (such behavior is in line with Theorem 2). Another key observation from these figures is that a smaller block size implies faster convergence; this is also in line with the dependence on the number of blocks ($L = T/B$) in Theorem 2.

J Experiment on the S&P500 Stock Dataset

J.1 Non-stationary in the Setting

To observe the distribution shift in this environment, we visualized the distance between covariance matrix with various window sizes and histogram for absolute values of daily return at the Figure 4. For the covariance distance, we first split the data stream into chunks with w (window size) data each and calculated the covariance estimators. Then we plotted the operator 2-norm between the covariance matrix divided by $\sqrt{w}$. As the left figure displays, the distribution on this dataset shifts over time (the average distance is about 0.17). Furthermore, on the right, we visualized the counts for the absolute value of daily return with a logarithmic scale. We can observe a lot of zero elements and outliers. Note that both axes have a logarithmic scale.

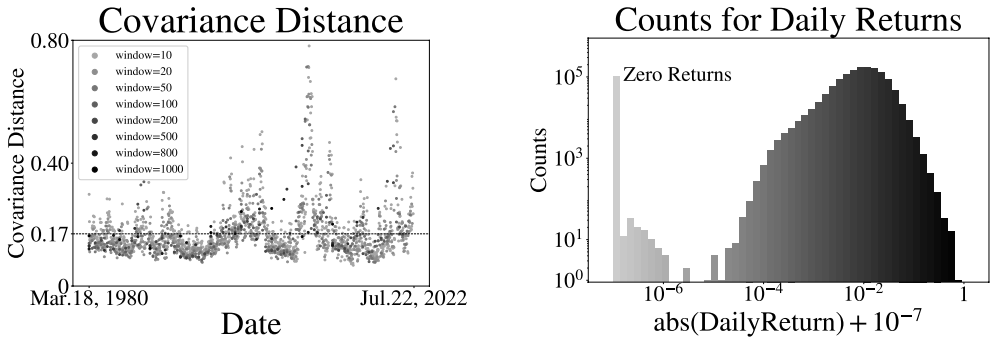


Figure 4: Non-stationary distribution characteristics of S&P500 stock market daily return. (Left): Distance of covariance matrix with window size variation and (Right): Histogram of the absolute value of daily returns.

J.2 Experimental Detail

We ran five experiments for each algorithm and $k = 1, 2, \ldots, 5$, on the various regime of learning parameters ($B : 1 - 1600$, $\eta : 10^{-3.5} - 10^{2.5}$). For the noisy power method, we just ignored the first $T \bmod B$ data to approximate the final space properly.