On the optimality of the Oja’s algorithm for online PCA

Xin Liang

Received: 9 July 2022 / Accepted: 24 February 2023 / Published online: 30 March 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
In this paper we analyze the behavior of the Oja’s algorithm for online/streaming principal component subspace estimation. It is proved that with high probability it performs an efficient, gap-free, global convergence rate to approximate an principal component subspace for any sub-Gaussian distribution. Moreover, it is the first time to show that the convergence rate, namely the upper bound of the approximation, exactly matches the lower bound of an approximation obtained by the offline/classical PCA up to a constant factor.

Keywords Principal component analysis · Stochastic approximation · High-dimensional data · Oja’s algorithm

1 Introduction

Principal component analysis (PCA) introduced by Pearson (1901) and Hotelling (1933) is one of the most well-known and popular methods for dimensional reduction in statistics, machine learning, and data science. The goal of PCA is to find out a low-dimensional linear subspace that is closest to a centered random vector in a high-dimensional subspace in the mean squared sense through finite independent and identically distributed (i.i.d.) samples of the random vector. Theoretically, given a random vector \( X \in \mathbb{R}^d \) satisfying \( E \{ X \} = 0, E \{ XX^T \} = \Sigma \), PCA looks for a subspace \( U \) with \( \dim U = p < d \), such that

\[
U = \text{arg} \min_{\dim U = p} E \left\{ \| (I_d - \Pi_u) X \|_2^2 \right\},
\]

where \( I_d \) is the identical mapping, or equivalently the \( d \times d \) identity matrix, and \( \Pi_u \) is the orthogonal projector onto \( U \). Let \( \Sigma = UAU^T \) be the spectral decomposition of \( \Sigma \), where \( A = \text{diag}(\lambda_1, \ldots, \lambda_d) \).

If \( \lambda_p > \lambda_{p+1} \), then the unique solution to the optimization problem (1), namely the \( p \)-dimensional principal subspace of \( \Sigma \), is \( U_p = \text{span}(u_1, \ldots, u_p) \), the subspace spanned by \( u_1, \ldots, u_p \).

In practice, the covariance matrix \( \Sigma \) is difficult, if not impossible, to obtain, and samples are used instead to approximate \( U_p \). The classical/offline PCA use the spectral decomposition of the empirical covariance matrix \( \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X(i)X(i)^T \). There \( \hat{U}_p = \text{span}(\hat{u}_1, \ldots, \hat{u}_p) \) is used to approximate \( U_p \), where \( \hat{u}_i \) are corresponding eigenvectors of \( \hat{\Sigma} \). Vu and Lei (2013) proved that

\[
\inf_{\dim U = p} \sup_{X \in \mathcal{P}_d(\sigma_n^2, d)} E \left\{ \| \sin(\hat{\Theta}(\hat{U}_p, U_p)) \|_F^2 \right\} \geq c p(d - p) \sigma_n^2 \frac{\sigma_n^2}{n} \geq c \frac{\lambda_1 \lambda_{p+1}}{\lambda_p - \lambda_{p+1}} \frac{p(d - p)}{n},
\]

where \( c > 0 \) is an absolute constant, and \( \mathcal{P}_d(\sigma_n^2, d) \) is the set of all \( d \)-dimensional sub-Gaussian distributions for which the eigenvalues of the covariance matrix satisfy \( \frac{\lambda_1 \lambda_{p+1}}{(\lambda_p - \lambda_{p+1})^2} \leq \sigma_n^2 \). Note that \( \frac{\lambda_1 \lambda_{p+1}}{(\lambda_p - \lambda_{p+1})^2} \) is the effective noise variance.

Due to the practical requirement that only limited memory and a single pass over the data can be implemented, amount of attention have been paid to a class of methods under these condition, called streaming/online PCA. The most natural

\[\Sigma = U\Lambda U^T \quad \text{with} \quad U = [u_1, u_2, \ldots, u_d], \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d). \]
and simple method was designed by Oja (1982), Oja and Karhunen (1985): first choose an initial guess \( U(0) \in \mathbb{R}^{d \times p} \) with \( (U(0))^T U(0) = I \), and then iteratively update
\[
U^{(n)} = \Pi \left( [I_d + \eta_n X^{(n)}(X^{(n)})^T] U^{(n-1)} \right) = [I_d + \eta_n X^{(n)}(X^{(n)})^T] U^{(n-1)} S^{(n)},
\]
where \( \Pi(A) \) is an orthonormal projector such that \( \Pi(A)^T \Pi(A) = I_p \) and \( \text{span}(A) = \text{span}(\Pi(A)) \), and \( S^{(n)} \) is used to denote the normalization matrix. There are three classes of hyperparameters:

1. the initial guess \( U^{(0)} \): usually first generate \( \tilde{U}^{(0)} \) of which each entry follows the standard Gaussian distribution \( N(0, 1) \), and then obtain \( U^{(0)} \) by QR decomposition. Note that in this setup, \( U^{(0)} \) is uniformly sampled from all the \( p \)-dimensional subspaces under the Haar invariant probability measure (see e.g. Muirhead 1982).
2. the learning rates \( \eta_n \): there are different strategies to choose them. Two common setups are the constant learning rates \( \eta_n = \eta_0 \), and harmonic learning rates \( \eta_n \propto \frac{1}{n} \).
3. the normalization matrices \( S^{(n)} \): Two common ways to obtain the orthonormal basis are QR decomposition, and polar decomposition (Abed-Meraim et al. 2000; Liang et al. 2022).

Although the Oja’s algorithm was developed nearly 40 years ago and it works well in practice, its convergence behavior is not very clear until recently. Most theoretical results come out since 2014. As was argued by Allen-Zhu and Li (2017), the convergence rate of the Oja’s method has several features:

1. efficient: the rate only depends on the dimension \( d \) logarithmically. In fact, the dependence on \( d \) can be removed.
2. gap-free: the rate is independent of the eigenvalue gap \( \gamma = \lambda_p - \lambda_{p+1} \).
3. global: the algorithm is allowed to start from a random initial guess.

Some recent works Shamir (2016a) studied the convergence of the online PCA for the most significant principal component, i.e., \( u_1 \), from different points of view and obtained some results for the case where the samples are almost surely uniformly bounded. De Sa et al. (2015) studied a different but closely related problem, in which the angular part is equivalent to the online PCA, and obtained some convergence results. Li et al. (2018) analyzed for the distributions with sub-Gaussian tails, and for this case the samples of this kind of distributions may be unbounded. For more details of comparison, the reader is referred to Li et al. (2018).

For the subspace online PCA, some recent works studied the convergence for the case where the samples are almost surely uniformly bounded. In a series of papers (Arora et al. 2012, 2013; Marinov et al. 2018; Mianjy and Arora 2018), Arora et al. studied (1) and its variations via direct optimization approaches, namely using convex relaxation and adding regularizations. The Oja’s algorithm falls into one variant of their methods. Hardt and Price (2014) and Balcan et al. (2016) treated the method as a noisy power method and analyzed its convergence. Shamir (2016b) first proved the convergence is efficient with a good initial guess. Garber et al. (2016) used the shift-and-invert technique to speed up the convergence but their analysis was only done for the top eigenvector. Allen-Zhu and Li (2017) analyzed the method and proposed a faster variant of subspace online PCA iteration, and firstly showed the gap-free feature of the convergence and also gave a lower bound for the gap-free feature. Very recently (Huang et al. 2021) analyzed the problem using the new matrix concentration inequalities and proved stronger upper bounds. Liang et al. (2022) went further along the way of Li et al. (2018) and gave an convergence analysis for sub-Gaussian distributions.

The convergence rates obtained in some previous works and this paper are presented in Table 1.

The sum of some consecutive eigenvalues is written as
\[
\lambda_{i_1} \sim \lambda_{i_2} := \lambda_{i_1} + \cdots + \lambda_{i_2}, \quad 1 \leq i_1 \leq i_2 \leq d.
\]

The listed convergence rates are read as: with probability \( 1 - \delta \), using \( n \) samples, or equivalently after \( n \) iterations, the Oja’s algorithm produces an approximation \( \text{span}(U^{(n)}) \) of the principal subspace satisfying \( \| \sin \Theta(\text{span}(U^{(n)}), U_\delta) \| \leq \gamma \) (the rate). The global convergence rate is given for the case that the initial guess is random generated, while the local convergence rate is given for the case that the initial guess satisfies \( \| \tan \Theta(\text{span}(U^{(0)}), U_\delta) \| \) is bounded by an absolute constant like 1.

The convergence rates listed in Table 1, except ours, include a poly-logarithmic factor, which leads people to say the Oja’s method is nearly optimal. However, in this paper we will show the poly-logarithmic factor can be removed. In other words, the convergence rate, namely the upper bound of the approximation, exactly matches the lower bound (3) of an approximation obtained by the offline/classical PCA up to a constant factor. Hence in some sense, we may say the Oja’s method is optimal. To the best of our knowledge, it is the first time to point out this feature of the online method.
The term \( M_d \) represents any quantity related to \( \mathbb{E} \left[ \| \bar{X} \bar{X}^T - \Sigma \| \right] \), or the fourth central moment (not necessarily the same in different results).

In some results the term \( n_0 \) appears in the global convergence, and it represents the number of samples needed in the so-called “Phase I” or “Cold Start” process.

Note that there are two types of the dependency on \( \delta \) in Table 1: one is \( \ln \frac{1}{\delta} \), which goes to infinity as \( \delta \to 0 \); the other is \( \frac{1}{\sqrt{\ln \frac{1}{\delta}}} \), which goes to 0 as \( \delta \to 0 \). Clearly the latter term can be replaced by an absolute constant, or equivalently, the rate does not explicitly rely on \( \delta \) (but implicitly, for \( n \geq O \left( (\ln \delta)^{-4} \right) \) as is shown in Lemma 6 and Theorem 8).

In some results, such as Jain et al. (2021, Theorem 1.3), there is no \( \ln d \) or \( \ln \frac{1}{\delta} \) factor, which seems to contradict with what we list in Table 1 (based on their Theorem 1.2 or 4.1 actually). However, the assumption there is much stronger: first their success probability is \( 3/4 \), rather than \( 1 - \delta \), which removes the dependence on \( \delta \); then they need \( n \geq O \left( (\ln d)^{10} \right) \) which is much larger than \( O \left( (\ln d)^{4} \right) \) here.

Other results we will show in this paper include:

- the strategy of choosing the normalization matrices does not matter much on the convergence rate. Hence we may choose a strategy that has advantages on computation or practical consideration.
- the iteration process is somehow decoupled, and thus the gap-dependent and gap-free considerations can be treated in the same framework. This would shed light on the convergence analysis of other online algorithms.
- a lower bound for sub-Gaussian distributions on the gap-free feature is given, which ensures that the Oja’s method is optimal.

The rest of this paper is organized as follows. In Sect. 2 we make preparations for discussing the convergence analysis of the Oja’s method. The main results, namely the convergence analysis, are stated in Sect. 3, while their proofs are provided in the Online Resource due to their high complexity and high similarity to the existing ones. Sect. 4 gives some concluding remarks.

### 1.1 Notation

- \( I_n \) (or simply \( I \) if its dimension is clear from the context) is the \( n \times n \) identity matrix and \( e_j \) is its \( j \)th column (usually with dimension determined by the context).
- For a matrix \( X \), \( \sigma(X) \), \( \|X\|_2 \) and \( \|X\|_F \) are the multiset of the singular values, the spectral norm, and the Frobenius norm of \( X \), respectively. For two matrices or vectors \( X, Y \), \( X \odot Y \) is the Hadamard/entrywise product of \( X \) and \( Y \) of the same size.
- For any matrix \( X \), \( X_{(i,j)} \) is the \( \langle i, j \rangle \)th entry of \( X \), and \( X_{(i,:)} \) is the submatrix of \( X \) consisting of its row \( i \) to row \( j \).
- For any vector or matrix \( X, Y \), \( X \leq Y \) means \( X_{(i,j)} \leq Y_{(i,j)} \) for any \( i, j \). ‘\( X \leq Y \)’ \( \Leftrightarrow \) ‘\( Y \geq X \)’ can be similarly understood.
- For a subset or an event \( \mathcal{A} \), \( \mathcal{A}^c \) is the complement set of \( \mathcal{A} \). By \( \sigma(\mathcal{A}_1, \ldots, \mathcal{A}_p) \) we denote the \( \sigma \)-algebra generated by the events \( \mathcal{A}_1, \ldots, \mathcal{A}_p \). \( E \{ X ; \mathcal{A} \} := E \{ X 1_{\mathcal{A}} \} \) denotes the expectation of a random variable \( X \) over event \( \mathcal{A} \). Note that \( E \{ X ; \mathcal{A} \} = E \{ X | \mathcal{A} \} P \{ \mathcal{A} \} \). For a random vector or matrix \( X \), \( E \{ X \} := E \{ X 1_{\mathcal{A}} \} \). Note that \( \| E \{ X \} \|_{ui} \leq E \{ \| X \|_{ui} \} \) for \( ui = 2, F \). Write \( \text{cov}_{\mathcal{A}}(X, Y) := E \{ (X - E \{ X \}) \circ (Y - E \{ Y \}) \} \) and \( \text{var}_{\mathcal{A}}(X) := \text{cov}_{\mathcal{A}}(X, X) \).
- For any scalar \( x, y \), \( x \vee y = \max\{x, y\}, x \wedge y = \min\{x, y\} \).
2 Preliminaries

2.1 Canonical angles between two subspaces

We are interested in the distance of two linear subspaces. So we introduce the canonical angles between them in order to give quantities to represent their distance.

Definition 1 (Björk and Golub 1973) Given two subspaces \( \mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d \) with \( \dim \mathcal{X} = p \leq \dim \mathcal{Y} = q \). The principal/canonical angles \( \theta_j \in [0, \pi/2] \) between \( \mathcal{X} \) and \( \mathcal{Y} \) are recursively defined for \( j = 1, \ldots, p \) by

\[
\cos \theta_j = \sigma_j = \max_{u \in \mathcal{X}, v \in \mathcal{Y}} u^T v = u_j^T v_j
\]

s.t. \( \|u\|_2 = \|v\|_2 = 1 \), \( u_i^T u = 0, v_i^T v = 0, i = 1, \ldots, j - 1 \).

It can be verified that \( \sigma_1 \geq \cdots \geq \sigma_p \) are the singular values of \( X^T Y \), where \( X, Y \) are orthonormal basis matrices of \( \mathcal{X}, \mathcal{Y} \) respectively. The angles are in non-decreasing order: \( \theta_1 \leq \cdots \leq \theta_p \). Moreover, it can be seen that \( \sigma_j \) or \( \theta_j \) are independent of the basis matrices, which are not unique.

Write

\[
\Theta(\mathcal{X}, \mathcal{Y}) = \text{diag}(\theta_1, \ldots, \theta_p).
\]

Here we add ``(\( \mathcal{X}, \mathcal{Y} \))'' to emphasize the quantity is defined for two subspaces \( \mathcal{X}, \mathcal{Y} \). In particular, if \( p = q \), \( \|\sin \Theta(\mathcal{X}, \mathcal{Y})\|_u \) for \( u = 2, F \) are metrics on the set consisting for all \( p \)-dimensional subspaces of \( \mathbb{R}^d \) (Stewart and Sun 1990, Section II.4).

For matrices \( X, Y \), \( \Theta(X, Y) := \Theta(\text{span}(X), \text{span}(Y)) \).

In what follows, we give a quantity easy to compute to estimate the distance between one subspace and a particular subspace.

Given \( p \leq q \), for any matrix \( X \in \mathbb{R}^{d \times p} \) with nonsingular \( X_{(1:p,:)} \), write

\[
\mathcal{T}_{p,q}(X) := X_{(p+1:q,:)} X_{(1:p,:)}^{-1},
\]

\[
\mathcal{T}_q(X) := X_{(q+1:d,:)} X_{(1:p,:)}^{-1},
\]

which are submatrices of \( \mathcal{T}(X) := \mathcal{T}_p(X) \).

Lemma 1 We have for \( u = 2, F \)

\[
\left\| \tan \Theta(X, \begin{bmatrix} I_q \\ 0 \end{bmatrix} \right\|_u \leq \| \mathcal{T}_q(X) \|_u .
\] (4)

In particular, if \( p = q \), then the inequality ``\( \leq \)'' can be replaced by ``\( = \)''.

Proof For the readability, we use \( \mathcal{T}_{p,q}, \mathcal{T}_p, \mathcal{T}_q \) only and drop ``(\( X \))''. Then \( \cos \theta_j \) for \( j = 1, \ldots, p \) are the singular values of

\[
\begin{bmatrix} I_p + \mathcal{T}_p^T \mathcal{T}_p \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} I_p \\ \mathcal{T}_p \mathcal{T}_q \end{bmatrix}^T \begin{bmatrix} I_q \\ 0 \end{bmatrix}
\]

\[
= \begin{bmatrix} I_p + \mathcal{T}_p^T \mathcal{T}_p \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} I_p \\ \mathcal{T}_p \mathcal{T}_q \end{bmatrix}.
\]

Thus, \( \tan^2 \theta_j = \sec^2 \theta_j - 1 \) for \( j = 1, \ldots, p \) are the eigenvalues of

\[
\left( I_p^2 \left( I_q^2 \right)^{-\frac{1}{2}} \left( I_q^2 \right)^{-\frac{1}{2}} \right) - 1
\]

\[
\left( I_p^2 \left( I_q^2 \right)^{-\frac{1}{2}} \left( I_q^2 \right)^{-\frac{1}{2}} \right) - 1
\]

\[
\left( I_p^2 \left( I_q^2 \right)^{-\frac{1}{2}} \left( I_q^2 \right)^{-\frac{1}{2}} \right) - 1
\]

and also the eigenvalues of

\[
\begin{bmatrix} I_p + \mathcal{T}_p^T \mathcal{T}_p \end{bmatrix}^{-\frac{1}{2}} \mathcal{T}_q \begin{bmatrix} I_p + \mathcal{T}_p^T \mathcal{T}_p \end{bmatrix}^{-\frac{1}{2}}.
\]

Let \( \tau_1 \geq \cdots \geq \tau_p \) be the eigenvalues of \( \mathcal{T}_q^T \mathcal{T}_q \). By the Ostrowski theorem (Horn and Johnson 1985, Theorem 4.5.9),

\[
\tan^2 \theta_j \leq \tau_j,
\]

which implies (4). \( \square \)

Note that (Liang et al. 2022, Lemma 3.1) is a special case of Lemma 1.

2.2 Orlicz norms

We are concerned with random variables/vectors that have a sub-Gaussian distribution. To that end, we first introduce the Orlicz \( \psi_\alpha \)-norm of a random variable/vector. More details can be found in van der Vaart and Wellner (1996).

Definition 2 The Orlicz \( \psi_\alpha \)-norm of a random variable \( X \in \mathbb{R} \) is defined as

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

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

\[
\|X\|_{\psi_\alpha} := \sup_{\|x\|_2 = 1} \|x^T X\|_{\psi_\alpha}.
\]
We say that random variable/vector $X$ follows a sub-Gaussian distribution if $\|X\|_{\psi_2} < \infty$.

By the definition, we conclude that any bounded random variable/vector follows a sub-Gaussian distribution.

The basic properties of sub-Gaussian distributions are listed in Lemma 2.

**Lemma 2** (Vershynin 2012, (5.10)–(5.12)) Every sub-Gaussian random variable $X \in \mathbb{R}$ with $\|X\|_{\psi_2} = \psi$ satisfies:

1. $P\{|X| > t\} \leq \exp(1 - c\psi^2 t^2)$ for $t \geq 0$;
2. $E\{|X|^p\} \leq \psi^p p! / 2$ for $p \geq 1$;
3. if $E\{X\} = 0$, then $E\{\exp(tX)\} \leq \exp(C\psi^2 t^2)$ for $t \in \mathbb{R}$,

where $C > 0$, $c > 0$ are absolute constants.

Moreover, if $\|X\|_{\psi_1} < \infty$, then $X$ follows a sub-exponential distribution. Our analysis below can be easily generalized to sub-exponential random vectors, and will not be discussed.

### 2.3 Detailed algorithm and assumptions

Here we write down the detailed algorithm in Algorithm 1.

**Algorithm 1** Oja’s Algorithm for Online PCA

1. Choose $U^{(0)} \in \mathbb{R}^{d \times p}$ with $(U^{(0)})^T U^{(0)} = I$, and use a regime to choose the learning rate $\eta_n = \rho_n \eta_0 > 0$.
2. for $n = 1, 2, \ldots$ until convergence do
3. Take an $X$’s sample $X^{(n)}$;
4. $Z^{(n)} = (U^{(n-1)})^T X^{(n)}$;
5. $U^{(n)} = U^{(n-1)} + \eta_n X^{(n)} (Z^{(n)})^T$;
6. Find an orthonormal basis of the subspace spanned by $\tilde{U}^{(n)}$, namely compute a column orthonormal matrix $U^{(n)} = \tilde{U}^{(n)} S^{(n)}$.
7. end for

The learning rate of the $n$-th iteration is $\eta_n = \rho_n \eta_0$. Without loss of generality, we may assume $0 < \rho_n < 1$.

The decomposition can be chosen as QR decomposition (Oja and Karhunen 1985) or polar decomposition (Abed-Meraim et al. 2000), or any other decomposition easy to compute. However, $S^{(n)}$ is always nonsingular. In fact, noticing that $\tilde{U}^{(n)} = [I + \eta_n X^{(n)} (X^{(n)})^T] U^{(n-1)}$, since $I + \eta_n X^{(n)} (X^{(n)})^T$ is positive definite and thus nonsingular, and $U^{(n)}$ is column orthonormal, we know $\tilde{U}^{(n)}$ has full column rank, which implies the fact.

Any statement we will make holds almost surely.

To prepare our convergence analysis, we make a few assumptions.

**Assumption 1** $X = [X_1, X_2, \ldots, X_d]^T \in \mathbb{R}^d$ is a random vector.

(A-1) $E\{X\} = 0$, and $\Sigma := E\{XX^T\}$ has the spectral decomposition (2) satisfying $\lambda_p > \lambda_{p+1}$;

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

The principal subspace $U_\psi$ is uniquely determined under Item A-1 of Assumption 1. On the other hand, Item A-2 of Assumption 1 ensures that all 1-dimensional marginals of $X$ have sub-Gaussian tails, or equivalently, $X$ follows a sub-Gaussian distribution.

Using the substitutions

$$\tilde{X} \leftarrow \frac{1}{\lambda_1^{-\frac{1}{2}}} X, \quad X^{(n)} \leftarrow \frac{1}{\lambda_1^{-\frac{1}{2}}} X^{(n)},$$

$$Z^{(n)} \leftarrow \frac{1}{\lambda_1^{-\frac{1}{2}}} Z^{(n)}, \quad \eta_n \leftarrow \lambda_1^{-\frac{1}{2}} \eta_n,$$

the iterations produced by Algorithm 1 and the rest terms $\tilde{U}^{(n)}, U^{(n)}$ keep the same. Hence any convergence result has to keep this homogeneous property.

Next we make a simplification on the problem.

Recall the spectral decomposition $\Sigma = U \Lambda U^T$. Instead of the random vector $X$, we equivalently consider

$$Y \equiv [Y_1, Y_2, \ldots, Y_n]^T := U^T X.$$

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

$$Y^{(n)} = U^T X^{(n)}, \quad V^{(n)} = U^T U^{(n)}, \quad V_n = U^T U_n = \begin{bmatrix} I_p \end{bmatrix}.$$

Firstly, because

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

$$(Y^{(n)})^T Y^{(n)} = (X^{(n)})^T X^{(n)},$$

the equivalent version of Algorithm 1 is obtained by symbolically replacing all letters $X$, $U$ by $Y$, $V$ while keeping their respective superscripts. If the algorithm converges, it is expected that $\text{span}(V^{(n)}) \rightarrow \text{span}(V_n)$. Secondly, noting

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

we can restate Assumption 1 equivalently as

(A-1’) $E\{Y\} = 0$, $E\{YY^T\} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geq \cdots \geq \lambda_p > \lambda_{p+1} \geq \cdots \geq \lambda_d$;

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

Thirdly, all canonical angles between two subspaces are invariant under the orthogonal transformation. Therefore the results given below holds for not only $Y$ but also $X$.
If the algorithm converges, it is expected that
\[ U^{(n)} \rightarrow U^* := U \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} \Leftrightarrow \ V^{(n)} \rightarrow V^* = \begin{bmatrix} I_p & 0 \end{bmatrix} \]

in the sense that
\[ \| \sin \Theta(U^{(n)}, U^*) \|_{ui} \rightarrow 0 \iff \| \sin \Theta(V^{(n)}, V^*) \|_{ui} \rightarrow 0 \]
as \( n \to \infty \).

By Lemma 1, it is sufficient enough to prove \( \| \mathcal{J}(V^{(n)}) \|_{ui} \rightarrow 0 \). Our results are based on this point.

To simplify the notations in our proofs, we introduce new notations for two particular submatrices of any vector \( Y \in \mathbb{R}^d \), tall matrix \( V \in \mathbb{R}^{d \times p} \) and diagonal matrix \( A \in \mathbb{R}^{d \times d} \):

\[
Y = \begin{bmatrix} \tilde{Y} \\ \bar{Y} \end{bmatrix}, \quad V = \begin{bmatrix} \tilde{V} \\ \bar{V} \end{bmatrix}, \quad A = \begin{bmatrix} \tilde{A} & \bar{A} \\ \bar{A} & \bar{A} \end{bmatrix}
\]
or equivalently

\[
\tilde{Y} = Y_{(1:p,:)}, \quad \bar{Y} = Y_{(p+1:d,:)}, \quad \tilde{V} = V_{(1:p,:)}, \quad \bar{V} = V_{(p+1:d,:)},
\]

and

\[
\tilde{A} = \text{diag}(\lambda_1, \ldots, \lambda_p), \quad \bar{A} = \text{diag}(\lambda_{p+1}, \ldots, \lambda_d).
\]

### 3 Main results

In what follows, we will state our main results. The main technique to prove the results is the same as Li et al. (2018) and Liang et al. (2022). The differences between the results are:

- the estimations are much sharper here;
- the learning rates are changing here, rather than a fixed learning rate in Li et al. (2018), Liang et al. (2022).

Some proofs are left to the Online Resource because of their high complexity and high similarity to the existing ones.

First we introduce some quantities.

For \( \kappa \geq 0 \), define \( \mathbb{S}(\kappa) := \{ V \in \mathbb{R}^{d \times p} : \sigma(\tilde{V}) \subset [\frac{1}{\sqrt{1+\kappa^2}}, 1] \} \).

It can be verified that

\[ V \in \mathbb{S}(\kappa) \Leftrightarrow \| \mathcal{J}(V) \|_2 \leq \kappa. \tag{5} \]

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

\[
N_{\text{out}}(\kappa) := \min\{ n : V^{(n)} \notin \mathbb{S}(\kappa) \},
\]

\[
N_{\text{in}}(\kappa) := \min\{ n : V^{(n)} \in \mathbb{S}(\kappa) \}.
\]

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

For \( \mu \geq 1 \), define

\[
N_{\text{qb}}(\mu) := \max \left\{ n \geq 1 : \| Z^{(n)} \|_2 \leq \frac{1}{\sqrt{1+\mu^2}} \right\} + 1.
\]

\( N_{\text{qb}}(\mu) \) is the first step of the iterative process at which either \( |Y_i^{(n)}| > \frac{1}{\sqrt{1+\mu^2}} \) for some \( i \) or the norm of \( Z^{(n)} \) exceeds \( \frac{1}{\sqrt{1+\mu^2}} \).

For convenience, we will set \( T^{(n)} := \mathcal{J}(V^{(n)}) \), and let \( \mathcal{F}_n = \sigma(Y^{(1)}, \ldots, Y^{(n)}) \) be the \( \sigma \)-algebra filtration, i.e., the information known by step \( n \).

#### 3.1 Increments of one iteration

In each iteration,

\[
\bar{V}^{(n+1)} = (\tilde{V}^{(n)} + n_{\eta+1} \tilde{Y}^{(n+1)} (Z^{(n+1)})^T) S^{(n)},
\]

\[
\bar{V}^{(n+1)} = (V^{(n)} + n_{\eta+1} \tilde{Y}^{(n+1)} (Z^{(n+1)})^T) S^{(n)},
\]

where \( S^{(n)} \) is nonsingular as is stated above. According to the Sherman-Morrison-Woodbury formula, we get \( \tilde{V}^{(n)} + n_{\eta+1} \tilde{Y}^{(n+1)} (Z^{(n+1)})^T \) or \( \bar{V}^{(n+1)} \) is nonsingular, if and only if \( 1 + n_{\eta+1} \bar{Y}^{(n+1)} \neq 0 \) where

\[
\bar{Y}^{(n+1)} := (Z^{(n+1)})^T (\tilde{V}^{(n)})^{-1} \tilde{V}^{(n+1)}.
\]

and

\[
(\tilde{V}^{(n+1)})^{-1} = (S^{(n)})^{-1} \left( I - \frac{n_{\eta+1} \tilde{V}^{(n+1)} (Z^{(n+1)})^T}{1 + n_{\eta+1} \bar{Y}^{(n+1)}} \right) \tilde{V}^{-1}.
\]

Hence

\[
T^{(n+1)} = T^{(n)} (\tilde{V}^{(n+1)})^{-1}
\]

\[
= (V^{(n)} + n_{\eta+1} \tilde{Y}^{(n+1)} (Z^{(n+1)})^T)
\]

\[
\left( I - \frac{n_{\eta+1} \tilde{V}^{(n+1)} (Z^{(n+1)})^T}{1 + n_{\eta+1} \bar{Y}^{(n+1)}} \right) \tilde{V}^{-1}.
\]

Clearly the choice of \( S^{(n)} \) does not matter on the convergence of \( T^{(n)} \). In other words, the strategy of choosing the normalization matrices does not matter much on the convergence rate.

In the following, we need to estimate \( T^{(n+1)} - T^{(n)} \), and the results are listed in Lemma 3.
Lemma 3 Suppose

\[ 2\lambda_1 - \rho \sqrt{\kappa^2 + 1} \mu \eta_{n+1} \leq 1. \tag{6} \]

Let \( \tau = \| T^{(n)} \|_2 \). If \( n < N_{\text{qb}} \{ \mu \} \wedge N_{\text{out}} \{ \kappa \} \), then the following statements hold.

1. \( T^{(n)} \) and \( T^{(n+1)} \) are well-defined.
2. \( \| T^{(n+1)} - T^{(n)} \|_2 \leq 2 \mu \eta_{n+1} [1 + 2\lambda_1 - (1 + \tau^2) + v_1 \lambda_1 - (1 + \tau^2) \tau] \),
   where \( v_1 = 1 \vee v, v = \frac{\lambda_{p+1} - \lambda}{\lambda_1 - \rho} \).
3. Define \( R^{(n)}_E \) by \( E \left\{ T^{(n+1)} - T^{(n)} \mid \mathbb{F}_n \right\} = H_{n+1} + R^{(n)}_E \). Then
   \[ \| R^{(n)}_E \|_2 \leq 2 \lambda_1 \lambda_1 - (1 + \tau^2) \frac{\mu \eta_{n+1}}{2} (1 + \tau^2)^{\frac{3}{2}}. \]
4. Define \( R^{(n)}_o \) by \( \varphi_1(T^{(n+1)} - T^{(n)} \mid \mathbb{F}_n) = \eta_{n+1} H_o + R^{(n)}_o \), where \( H_o = \varphi_1 \left( \mathbb{Y}^{(n+1)}(T^{(n+1)} - 1) \right) \).
   (a) \( H_o \leq 16\psi^4 H \), where \( H = \mathbb{Y}^{(n+1)}(d-p) \times p \) with \( \psi_{ij} = \lambda_{p+1} - \lambda_j \) for \( i = 1, \ldots, d \), \( j = 1, \ldots, p \);
   (b) \( \| R^{(n)}_o \|_2 \leq 2 v_1 \lambda_1 - \rho 2 \mu \eta_{n+1}^2 \left( 1 + \left[ 1 + v_1 \lambda_1 \right] \tau \right) \tau + \tau^2 + \frac{1}{2} \tau^3 + 8 \lambda_1 - \rho \mu \eta_{n+1}^2 (1 + \tau^2)^{\frac{3}{2}} + 1 \).

3.2 Whole iteration process with a good initial guess

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

\[ T^{(n)} = E \left\{ T^{(n)} \mid \mathbb{F}_n \right\} = 0, \quad E \left\{ D^{(n+1)} \mid \mathbb{F}_n \right\} = 0, \]

\[ E \left\{ D^{(n+1)} \mid \mathbb{F}_n \right\} = \varphi_1(T^{(n+1)} - T^{(n)} \mid \mathbb{F}_n). \]

By Item 3 of Lemma 3, we have

\[ T^{(n+1)} = D^{(n+1)} + T^{(n+1)} + E \left\{ T^{(n+1)} - T^{(n)} \mid \mathbb{F}_n \right\} \]

\[ = D^{(n+1)} + T^{(n+1)} + \eta_n(\Delta T^{(n)} - T^{(n)} \bar{A}) + R^{(n)}_E \]

\[ = \mathcal{L}_{n+1} T^{(n+1)} + D^{(n+1)} + R^{(n+1)} \]

where \( \mathcal{L}_{n+1} : T \mapsto T + \eta_{n+1} \Delta T - \eta_{n+1} T \bar{A} \) is a bounded linear operator. It can be verified that \( \mathcal{L}_{n+1} T = L_{n+1} \circ T \), the Hadamard product of \( L_n \) and \( T \), where \( L_{n+1} = \lambda_{ij}^{(n+1)}(d-p) \times p \) with \( \lambda_{ij}^{(n+1)} = 1 + \eta_{n+1} (\lambda_{p+1} - \lambda_j) \). Clearly \( \mathcal{L}_{n+1} \mathcal{L}_{n+2} = \mathcal{L}_{n+2} \mathcal{L}_{n+1} \) for any \( n_1, n_2 \). Moreover, it can be shown that \( \sup \{ T \} = \rho(\mathcal{L}) \), where \( \mathcal{L}_{n+1} \mathcal{L}_{n+1} \mathcal{L} \) is an operator norm induced by the matrix norm \( \| \| \).

Define events

\[ M_n(\kappa, \mu) = \left\{ \| T^{(n)} - T^{(n)} \|_2 \leq \frac{1}{2} (1 + \kappa^2) \mu \eta_{n+1} \Gamma \right\}, \]

\[ T_n(\kappa) = \left\{ \| T^{(n)} \|_2 \leq \kappa \right\}, \quad Q_n(\mu) = \left\{ n < N_{\text{qb}} \{ \mu \} \right\}. \]

Suppose \( F_{\eta, \mu}(n, \kappa, \nu) \leq C_{\eta, \mu}(n, \kappa, \nu) \) for any \( n \), where \( C_{\eta, \mu}(n, \kappa, \nu) \) is an absolute constant, which can be easily examined in any specific strategy to choose \( \eta_{n+1} \).

For \( s > 0 \) and \( \eta_s \gamma < 1 \), define

\[ N_{\gamma} = \min \left\{ n \in \mathbb{N} : F_{\eta, \mu}(n, \kappa, \nu) \leq (\eta_s \gamma)^s \right\}, \]

which implies \( F_{\eta, \mu}(n, \kappa, \nu) \leq (\eta_s \gamma)^s < F_{\eta, \mu}(n, \kappa, \nu)^{-1} \). Define

\[ T^{(n+1)} = \left( \prod_{r=1}^{n} \mathcal{L}_r \right) T^{(0)} + \sum_{s=1}^{n} \left( \prod_{r=s+1}^{n} \mathcal{L}_r \right) D^{(s)} \]

\[ + \sum_{s=1}^{n} \left( \prod_{r=s+1}^{n} \mathcal{L}_r \right) R^{(s-1)} \]

Define events

\[ M_n(\kappa, \mu) = \left\{ \| T^{(n)} - T^{(n)} \|_2 \leq \frac{1}{2} (1 + \kappa^2) \mu \eta_{n+1} \Gamma \right\}, \]

\[ T_n(\kappa) = \left\{ \| T^{(n)} \|_2 \leq \kappa \right\}, \quad Q_n(\mu) = \left\{ n < N_{\text{qb}} \{ \mu \} \right\}. \]

where \( \nu = 8 C_{D,2,1} \lambda_1 \lambda_1 - (1 + \tau^2) \gamma^{-2} \).

It can be shown that under some conditions, if the initial \( \nu^{(0)} \) is not too bad, then with high probability \( \| T^{(n)} \|_2 \) will never become too large and eventually become as small as expected. The formal statement is given in Lemma 4.

\[ \rho(\mathcal{L}) = 1 - \eta_{n+1} (\lambda_1 - (\lambda_p - \lambda_{p+1})). \]

Thus for any \( T \),

\[ \| LT \|_u = \| T(I - \eta \bar{A}) + \eta \Delta T \|_u \]

\[ \leq \| T - \eta \bar{A} \Delta \|_u + \| \Delta T \|_u \]

\[ = (1 - \eta \lambda_p + \eta \lambda_{p+1}) \| T \|_u = \rho(\mathcal{L}) \| T \|_u, \]

which means \( \| T \|_u \leq \rho(\mathcal{L}) \). This ensures \( \| T \|_u = \rho(\mathcal{L}) \).

Footnote 1 continued
Lemma 4 Let \( N_1 := N^{(1)}_{\ln(e-\ln \kappa)/\ln(\eta_2 \gamma)} \). Suppose that (6), and the following hold,

\[
\begin{align}
&\nu(1 + \kappa^2)\mu^2 n_0^2 \leq \kappa, \\
&\kappa \rho_n^{-1} \leq \epsilon \quad \text{for } n \geq N_1.
\end{align}
\]

(7a)

If \( V^{(0)} \in S(\kappa/2) \), then for any \( n \geq N_1 \),

\[
P \left( \left( \mathbb{H}_n \cap \mathbb{H}_0 \right)^c \right) \leq 2nd \exp(-C_M \lambda^2 n^{-1} \gamma^{-2} \mu) + n(ed + p + 1) \exp \left( -C_\psi \psi^{-1}(1 \wedge \psi^{-1}) \mu \right),
\]

where \( \mathbb{H}_n := \bigcap_{r \leq n} T_r(\kappa), \mathbb{H}_0 := \{ N_{\ln(\epsilon/\ln(\eta_2 \gamma))} \leq N_1 \} \).

3.3 Estimation with a random initial guess

In order to compare our result with the previous results, we will estimate \( |\tan \Theta(\text{span}(U^{(n)}), U_n)| = \| T^{(n)} \| \).

First we give the estimation with a good initial guess in Lemma 5.

Lemma 5 Suppose that (6) and (7) hold, and let \( N_1 = N_{\ln(e/\ln(\eta_2 \gamma))} \). If \( V^{(0)} \in S(1) \), then there exists a high-probability event \( \mathbb{H} \) with \( P(\mathbb{H}) \geq 1 - \delta_1 \), such that for any \( n \geq N_1 \),

\[
E \left( \| T^{(n)} \| \right) 
\leq \frac{n^2}{\eta_2^2} \sum_{r=s+1}^{n} \mathcal{L}_r^2 H_\sigma + R, \quad (8)
\]

where \( R \leq C_R \frac{\epsilon^2}{\ln \frac{1}{\delta_1}} \) with \( C_R \) an absolute constant, and

\( H_\sigma \leq 16 \psi^4 H \) is as in Item 4(a) of Lemma 3.

Note that in (8) the inequality holds for each entry in the matrices. In other words, the inequality in (8) represents \( (d-p) \times p \) scalar inequalities. Hence, in this sense, we can say that the iteration process is decoupled. This observation is very useful for the gap-free consideration.

To deal with a random initial guess, a theorem by Huang et al. (2021) is adopted.

Lemma 6 (Huang et al. 2021, Theorem 2.4) If \( \sup_{F} \| F^T (X X^T A) F \|_F \leq B \), and \( \tilde{V}_0 \) has i.i.d. standard Gaussian entries, writing

\[
N_0 = C_o \frac{pB^2}{\delta^2 \gamma^2} \left( \ln \frac{d B}{\delta \gamma} \right)^{4}, \quad \eta_n = \eta_o = C_o \frac{\ln \frac{d}{\delta}}{\gamma N_o},
\]

then with probability at least \( 1 - \delta, \| T^{(N_o)} \|_2 \leq 1 \).

The whole iteration process can be split into two parts: first the iteration goes from the initial guess into \( S(1) \), whose probability is estimated by Lemma 6; then the iteration goes from an approximation in \( S(1) \) to any precision we would like, whose probability is estimated by Lemma 5. Note that Lemma 6 is only valid for a bounded distribution. Thus we have to use it on the event \( Q_n(\mu) \) whose probability is bounded in Lemma 7.

Lemma 7 For any \( n \geq 1 \),

\[
P \left( \{ N_{\phi} | \mu | > n \} \right) \geq 1 - n(ed + p + 1) \exp \left( -C_\psi \psi^{-1}(1 \wedge \psi^{-1}) \mu \right),
\]

where \( C_\psi \) is an absolute constant.

To sum up, we have Theorem 8.

Theorem 8 If \( U^{(0)} \in \mathbb{R}^{d \times p} \) has i.i.d standard Gaussian entries, and \( \eta_n = \begin{cases} C_o \frac{\ln \frac{d}{\delta}}{\gamma N_o} \quad & \text{if } n \leq N_o, \\ C_o \frac{1}{\gamma n} \quad & \text{if } n > N_o, \end{cases} \) then there exists a high-probability event \( \mathbb{H}_n \) with \( P(\mathbb{H}_n) \geq 1 - \delta_n \), such that for any \( n \geq N_o \),

\[
E \left( \| \tan \Theta (U^{(n)}, U_n) \|_F \right) \| \mathbb{H}_n \| \leq \frac{64C_o \psi^4 \psi(\Lambda)}{(1 - \delta)n^2},
\]

where \( C \) is an absolute constant, and \( \psi \) is \( X \)'s Orlicz \( \psi_2 \) norm. Here

\[
\psi(\Lambda) = \frac{1}{\psi} \sum_{i=1}^{d-p} \sum_{j=1}^{p} \frac{\lambda_{p+i} \lambda_j}{\lambda_{p+i} - \lambda_j} \leq \frac{p(d-p)\lambda_p \lambda_{p+1}}{(\lambda_p - \lambda_{p+1})^2}. \quad (9)
\]

Proof The probability of the whole process is guaranteed by Lemmas 5 to 7. In the following, we will show the upper bound.

Introduce sum(\( A \)) for the sum of all the entries of \( A \). In particular, \( \sum(\mathcal{A}(A) = \| A \|_F^2 \). Write \( F_{p+1,n}(\gamma) = \sum_{s=1}^{n} \sum_{r=s+1}^{n} (1 - \eta_{r} \gamma)^2 \). By Lemma 5 with \( \epsilon = \eta_{n}/n, \eta_n = C_o \frac{\ln \frac{d}{\delta}}{\gamma n} \), with high probability we have

\[
E \left( \| T^{(n)} \|_2^2 ; \mathbb{H}_n \right) \leq \frac{64C_o \psi^4 \psi(\Lambda)}{(1 - \delta)n^2},
\]
\[ \leq (F^{(N_o+1,n)}_*)^2 \| T(\Lambda) \|_F^2 + 2 \operatorname{sum} (G \circ H_o) + \operatorname{sum} (R). \]

where \( G = [\gamma_{ij}]_{(d-p) \times p} \) with \( \gamma_{ij} = F^{(N_o+1,n)}_{D,2,2}(\lambda_j - \lambda_{p+i}). \) Then,

\[ F^{(N_o+1,n)}_{D,2,2}(\lambda_j - \lambda_{p+i}) \leq \left( 1 - C_\eta \frac{\ln n - \ln N_o}{n - N_o} \right)^{n-N_o} \]

\[ \leq \left( 1 - C_\eta \frac{\ln n - \ln N_o}{n - N_o} \right)^{n-N_o} \]

\[ \leq e^{-C_\eta (\ln n - \ln N_o)} = \left( \frac{N_o}{n} \right)^{C_\eta}, \]

and

\[ F^{(N_o+1,n)}_{D,2,2}(\lambda_j - \lambda_{p+i}) \]

\[ = C_n^2 \sum_{s=1}^{n-N_o} \frac{(n-\nu_{ij})!}{(n-1)!(n-s-1)!} \]

\[ \leq C_n^2 \sum_{s=1}^{n-N_o} \frac{(n-s+1)}{n} \]

\[ \leq C_n^2 \left( \frac{1}{y^2} \sum_{s=1}^{n-N_o} \frac{n-N_o}{n} \right) \]

3.4 Lower bound and gap-free feature

Vu and Lei (2013) gives a minimax lower bound for the classical PCA, namely (3). First we point out the factor \( \frac{\lambda_{p+1}}{(p-p+1)} \) in the lower bound can be replaced by \( \frac{\lambda_{p+1}}{(p-p+1)} \). The key to the point is a lemma where, which is presented in Lemma 9.

Lemma 9 (Vu and Lei 2013, Lemma A.2) Let \( X_1, X_2 \in \mathbb{R}^{d \times p} \) are column orthonormal matrices, \( \beta \geq 0 \), and \( \Sigma_i = I_p + \beta X_i X_i^T \) for \( i = 1, 2 \). If \( \mathbb{P}_i \) are the \( n \)-fold product of the \( N(0, \Sigma_i) \) probability measure, then the Kullback–Leibler (KL) divergence is

\[ D(\mathbb{P}_1, \mathbb{P}_2) = \frac{n \beta^2}{1 + \beta} \| \sin \Theta(X_1, X_2) \|_F^2. \]

Then they use a substitution \( \frac{\lambda_{p+1}}{(p-p+1)} \geq 1 + \beta \) to obtain the lower bound. However, as a matter of fact, in their construction on the covariance matrix \( \Sigma_i \), it holds that \( \lambda_1 = \cdots = \lambda_p \), so when they proved the former bound, actually they were proving the latter bound.

Then we show the construction is also valid for the gap-free consideration.

Lemma 10 Let \( X_1, X_2 \in \mathbb{R}^{d \times p} \) are column orthonormal matrices satisfying \( X_1^T X_2 = X_2^T X_1 = 0 \). \( \beta \geq p, \beta \geq 0, \beta \geq 0, \) and \( \Sigma_1 = I_p + \beta X_1 X_1^T + \beta X_2 X_2^T \). \( \Sigma_2 = I_p + \beta X_2 X_2^T + \beta X_3 X_3^T \). If \( \mathbb{P}_i \) are the \( n \)-fold product of the \( N(0, \Sigma_i) \) probability measure, then the Kullback–Leibler (KL) divergence is

\[ D(\mathbb{P}_1, \mathbb{P}_2) = \frac{n \beta^2}{1 + \beta} \| \sin \Theta(X_1, [X_2 X_3]) \|_F^2. \]

Proof The proof is similar as that of Vu and Lei (2013, Lemma A.2). The key is to ensure

\[ \operatorname{tr} \left( \Sigma_2^{-1}(\Sigma_1 - \Sigma_2) \right) = \frac{\beta^2}{1 + \beta} \| \sin \Theta(X_1, [X_2 X_3]) \|_F^2. \]

In fact,

\[ \operatorname{tr} \left( \Sigma_2^{-1}(\Sigma_1 - \Sigma_2) \right) \]

\[ = \operatorname{tr} \left( (I_p + \beta X_2 X_2^T + \beta X_3 X_3^T)^{-1} \beta (X_1 X_1^T - X_2 X_2^T) \right) \]

\[ = \beta \operatorname{tr} \left( \left( \frac{1}{1 + \beta} (X_1 X_1^T + X_3 X_3^T) + I_p - X_2 X_2^T - X_3 X_3^T \right)^{-1} \beta (X_1 X_1^T - X_2 X_2^T) \right) \]

\[ = \frac{\beta}{1 + \beta} \operatorname{tr} \left( \left( 1 + \beta \right) I_p - \beta (X_2 X_2^T + X_3 X_3^T) (X_1 X_1^T - X_2 X_2^T) \right) \]

\[ = \frac{\beta^2}{1 + \beta} \operatorname{tr} \left( I_p - X_2 X_2^T - X_3 X_3^T \right) \left( X_1 X_1^T - X_2 X_2^T \right) \]

\[ + \frac{\beta}{1 + \beta} \operatorname{tr} \left( X_1 X_1^T - X_2 X_2^T \right) \]
\[ \frac{\beta^2}{1 + \beta} \text{tr} \left( \left[ I_p - X_2 X_2^T - X_3 X_3^T \right] X_1 X_1^T \right) + 0 \]

\[ \frac{\beta^2}{1 + \beta} \|\sin \Theta(X_1, [X_2 \ X_3])\|_F^2. \]

Following Lemma 10, we can use a substitution \( \frac{\lambda_p \lambda_{p+q}}{(\lambda_p - \lambda_{p+q})^2} \)

\[ = \frac{1 + \beta}{\beta^2} \text{ to obtain the lower bound under the gap-free consideration, namely Theorem 11.} \]

Theorem 11 (Gap-free version of Vu and Lei (2013, Theorem 3.1)) Let \( \mathcal{P}_0(\sigma^2, d) \) be the set of all \( d \)-dimensional sub-Gaussian distributions for which the eigenvalues of the covariance matrix satisfy \( \frac{\lambda_p + 1}{\lambda_p - \lambda_{q+1}} \leq \sigma^2. \) Then for \( 1 \leq p \leq q \leq n, \)

\[ \inf_{\dim \mathcal{U}_a = q} \sup_{X \in \mathcal{P}_0(\sigma^2, d)} E \left\{ \| \sin \Theta(\mathcal{U}_a, \mathcal{U}_a)\|_F^2 \right\} \]

\[ \geq c p(d - q) \sigma^2 \frac{n}{n} \geq c \frac{\lambda_p \lambda_{q+1}}{(\lambda_p - \lambda_{q+1})^2} \frac{p(d - q)}{n}, \]

where \( c > 0 \) is an absolute constant.

On the other hand, the Oja’s method can be easily proved to be gap-free, as is shown in Theorem 12, which exactly matches the lower bound in Theorem 11.

Theorem 12 If \( U^{(0)} \in \mathbb{R}^{d \times q} \) has i.i.d standard Gaussian entries, and \( \eta_n \) are chosen as in Theorem 8 with \( \gamma \) replaced by \( \gamma \), then Theorem 8 still holds, except that \( \varphi(A) \) in (9) can be replaced by

\[ \tilde{\varphi}(A) = \frac{1}{\gamma} \sum_{i=1}^{d-q} \sum_{j=1}^p \lambda_{q+i} \lambda_j - \lambda_j \leq \frac{p(d - q) \lambda_p (\lambda_p - \gamma)}{\gamma^2}. \]

Proof The proof is nearly the same as that of Theorem 8 except the bound.

Note that the iteration process produced by the Oja’s method is decoupled, as is argued after Lemma 5. Now what we need to estimate is

\[ E \left\{ \left\| \mathcal{Z}_q(U^{(n)}) \right\|_F^2 ; \mathbb{E}_a \right\} \]

\[ \leq (F^{-1}_{n, n})^2 \left\| \mathcal{Z}_q(U^{(n)}) \right\|_F^2 + 32 \psi^4 \sum (G_q \circ H_q) + \sum (R) \]

where \( G_q = \left[ \gamma_1 \right]_{d(q - d) \times p}, H_q = \left[ \eta_n \right]_{d(q - d) \times p}, R_q \in \mathbb{R}^{(d-q) \times p} \)

\[ \leq \frac{p N_0^2}{n^2} + \frac{64 C_n \psi^4}{\gamma} \sum_{i=1}^{d-q} \sum_{j=1}^{p} \frac{\lambda_{p+i} \lambda_j}{n - N_0} \left( 1 + C_n \frac{N_0}{n} \right) \]

\[ + C_R \frac{p N_0^2}{n^2 \ln \frac{md}{n}}. \]

\[ \square \]

4 Conclusion

We presented a convergence analysis of the Oja’s method for online/streaming PCA iteration with sub-Gaussian samples, by combining the idea in Li et al. (2018) and Liang et al. (2022) and the convergence result from a random guess to a good guess by Huang et al. (2021). Our results show for the first time that the Oja’s method for online PCA is optimal in the sense that with high probability the convergence rate exactly matches the minimax information lower bound for offline PCA.

Recently Amid and Warmuth (2020), Henriksen and Ward (2019) developed a method to decide the learning rate adaptively. Though our framework works for different strategies on choosing the learning rates, it remains an open but very worthwhile problem to consider how fast the adaptive learning rates would accelerate the convergence, because with a small probability to cover bad events, it is quite possible to attain a convergence rate even below the lower bound.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-023-10229-z.

Funding Supported in part by National Natural Science Foundation of China NSFC-11901340.

Declarations

Conflict of interest The author declare no competing interests.

References

Abed-Meraim, K., Attallah, S., Chkeif, A., et al.: Orthogonal Oja algorithm. IEEE Signal Process. Lett. 7(5), 116–119 (2000)

Allen-Zhu, Z., Li, Y.: First efficient convergence for streaming k-PCA: a global gap-free and near-optimal rate. In: 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pp. 487–492 (2017)

Amid, E., Warmuth, M.K.: An implicit form of Krasulina’s k-PCA update without the orthonormality constraint. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3179–3186. AAAI Press (2020)

Arora, R., Cotter, A., Livescu, K., et al.: Stochastic optimization for PCA and PLS. In: 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 861–868 (2012)

Arora, R., Cotter, A., Srebro, N.: Stochastic optics of PCA with capped MSG. In: Advances in Neural Information Processing System, pp. 1815–1823 (2013)

Balean, M.F., Du, S.S., Wang, Y., et al.: An improved gap-dependency analysis of the noisy power method. In: 29th Annual Conference on Learning Theory, vol. 49, pp. 284–309. PMLR (2016)

Björk, Å., Golub, G.: Numerical methods for computing angles between linear subspaces. Math. Comput. 27(123), 579–594 (1973)

De Sa, C., Re, C., Oluotukun, K.: Global convergence of stochastic gradient descent for some non-convex matrix problems. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 2332–2341. PMLR (2015)
Garber, D., Hazan, E., Jin, C., et al.: Faster eigenvector computation via shift-and-invert preconditioning. In: Proceedings of The 33rd International Conference on Machine Learning, vol 48, pp. 2626–2634. PMLR (2016)

Hardt, M., Price, E.: The noisy power method: a meta algorithm with applications. In: Advances in Neural Information Processing Systems, pp. 2861–2869 (2014)

Henriksen, A., Ward, R.: AdaOja: Adaptive learning rates for streaming PCA (2019). ArXiv e-prints arXiv:1905.12115 [stat.ML]

Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)

Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24(6), 417–441 (1933)

Huang, D., Niles-Weed, J., Ward, R.: Streaming k-PCA: Efficient guarantees for Oja’s algorithm, beyond rank-one updates. In: Proceedings of the 34th Conference on Learning Theory (COLT), vol 134, pp. 248–2498. PMLR (2021)

Huang, D., Niles-Weed, J., Ward, R.: Streaming k-PCA: Efficient guarantees for Oja’s algorithm, beyond rank-one updates. In: Proceedings of the 34th Conference on Learning Theory (COLT), vol 134, pp. 2463–2478. PMLR (2021)

Li, C.J., Wang, M., Liu, H., et al.: Near-optimal stochastic approximation for online principal component estimation. Math. Program. Ser. B 167(1), 75–97 (2018)

Liang, X., Guo, Z.C., Wang L., et al.: Nearly optimal stochastic approximation for online principal subspace estimation (2022). Science China Math. https://doi.org/10.1007/s11425-021-1972-5

Marinov, T.V., Mianjy, P., Arora, R.: Streaming principal component analysis in noisy setting. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 3413–3422. PMLR (2018)

Mianjy, P., Arora, R.: Stochastic PCA with \( \ell_2 \) and \( \ell_1 \) regularization. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 3531–3539. PMLR (2018)

Muirhead, R.J.: Aspects of Multivariate Statistical Theory. Wiley Series in Probability and Mathematical Statistics. Wiley, New York (1982)

Oja, E.: Simplified neuron model as a principal component analyzer. J. Math. Biol. 15(3), 267–273 (1982)

Oja, E., Karhunen, J.: On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. J. Math. Anal. Appl. 106(1), 69–84 (1985)

Pearson, K.F.R.S.: On lines and planes of closest fit to systems of points in space. Philos. Mag. 2(11), 559–572 (1901)

Shamir, O.: Fast stochastic algorithms for SVD and PCA: convergence properties and convexity. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33rd International Conference on Machine Learning, vol. 48, pp. 248–256. PMLR (2016b)

Shamir, O.: Convergence of stochastic gradient descent for PCA. In: Proceedings of The 33rd International Conference on Machine Learning, vol. 48, pp. 248–256. PMLR (2016a)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.