Near-optimal stochastic approximation for online principal component estimation

Chris Junchi Li¹ · Mengdi Wang¹ · Han Liu¹ · Tong Zhang²

Received: 5 December 2015 / Accepted: 9 July 2017 / Published online: 19 August 2017
© Springer-Verlag GmbH Germany and Mathematical Optimization Society 2017

Abstract Principal component analysis (PCA) has been a prominent tool for high-dimensional data analysis. Online algorithms that estimate the principal component by processing streaming data are of tremendous practical and theoretical interests. Despite its rich applications, theoretical convergence analysis remains largely open. In this paper, we cast online PCA into a stochastic nonconvex optimization problem, and we analyze the online PCA algorithm as a stochastic approximation iteration. The stochastic approximation iteration processes data points incrementally and maintains a running estimate of the principal component. We prove for the first time a nearly optimal finite-sample error bound for the online PCA algorithm. Under the subgaussian assumption, we show that the finite-sample error bound closely matches the minimax information lower bound.

Electronic supplementary material The online version of this article (doi:10.1007/s10107-017-1182-z) contains supplementary material, which is available to authorized users.

✉ Mengdi Wang
mengdiw@princeton.edu

Chris Junchi Li
junchil@princeton.edu

Han Liu
hanliu@princeton.edu

Tong Zhang
tzhang@stat.rutgers.edu

¹ Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

² Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA
Keywords Principal component analysis · Stochastic approximation · Nonconvex optimization · Stochastic gradient method · High-dimensional data · Online algorithm · Finite-sample analysis

Mathematics Subject Classification 62H25 · 68W27 · 60H10 · 60H15

1 Introduction

Principal component analysis (PCA) [20,36] is one of the most popular dimension reduction methods for high-dimensional data analysis. It has wide applications in bioinformatics, healthcare, imaging, computer vision, artificial intelligence, social science, finance and economy. Let $X$ be a random vector in $\mathbb{R}^d$ with mean zero and unknown covariance matrix $\Sigma = \mathbb{E} [XX^\top] \in \mathbb{R}^{d \times d}$, where the eigenvalues of $\Sigma$ are $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Principal component analysis aims to find the principal eigenvector of $\Sigma$ that corresponds to the largest eigenvalue $\lambda_1$, based on independent and identically distributed sample realizations $X^{(1)}, \ldots, X^{(n)}$. This can be casted into a nonconvex stochastic optimization problem, given by

$$\begin{align*}
\text{maximize} & \quad u^\top \mathbb{E} [XX^\top] u, \\
\text{subject to} & \quad \|u\| = 1, \\
& \quad u \in \mathbb{R}^d,
\end{align*}$$

where $\| \cdot \|$ denotes the Euclidean norm. We assume throughout this paper that the covariance matrix $\Sigma$ has a unique principal component, which we denote by $u^*$. Accordingly, the principal component $u^*$ is the unique solution to problem (1.1).

The classical PCA method estimates $u^*$ using the principal component of the empirical covariance matrix, i.e.,

$$\hat{u}^{(n)} = \text{argmax}_{\|u\|=1} u^\top \hat{\Sigma}^{(n)} u,$$

where $\hat{\Sigma}^{(n)}$ is the empirical covariance matrix based on $n$ samples

$$\hat{\Sigma}^{(n)} = \frac{1}{n} \sum_{i=1}^n X^{(i)} (X^{(i)})^\top.$$

This can be viewed as a sample average approximation method for problem (1.1). According to both literatures of PCA and sample average approximation, it is well-known that this method produces a non-improvable solution $\hat{u}^{(n)}$, in the sense that the estimation error $\hat{u}^{(n)} - u^*$ achieves the information lower bound. More precisely, we are interested in the angle between the two unit vectors $\hat{u}^{(n)}$ and $u^*$, given by
\[
\angle (\hat{u}^{(n)}, u^*) = \arccos \left( \hat{u}^{(n)} \top u^* \right),
\]

which takes value in \([0, \pi]\). It has been shown that the angle between any \(n\)-sample estimator \(\hat{u}^{(n)}\) and the true principal component \(u^*\) satisfies

\[
\inf_{\hat{u}^{(n)} \in M(\sigma_*^2, d)} \sup_{X \in \mathcal{M}(\sigma_*^2, d)} \mathbb{E} \left[ \sin^2 \angle (\hat{u}^{(n)}, u^*) \right] \geq c \cdot \sigma_*^2 \cdot \frac{d - 1}{n},
\]

where \(c\) is some positive constant (Theorem 3.1 of Vu and Lei [46]). In Eq. (2.18), the infimum of \(\hat{u}^{(n)}\) is taken over all \(n\)-sample estimators, and \(\mathcal{M}(\sigma_*^2, d)\) is the collection of all \(d\)-dimensional subgaussian distributions with mean zero and eigenvalues satisfying \(\lambda_1 \lambda_2 / (\lambda_1 - \lambda_2)^2 \leq \sigma_*^2\). Classical PCA method has time complexity \(O(nd^2)\) and space complexity \(O(d^2)\). When the raw data are high-dimensional, storing and computing a large empirical covariance matrix can be expensive.

In this paper, we focus on online methods for principal component analysis. These methods update the iterates incrementally by processing data points one by one. The practical goal is to be able to learn the principal eigenvector “on the fly”, without explicitly computing and storing the empirical covariance matrix \(\hat{\Sigma}^{-1}\). In particular, we focus on an iteration for online PCA that was first proposed by Oja [34], which is given by

\[
u^{(n)} = \Pi \left\{ u^{(n-1)} + \beta X^{(n)} (X^{(n)} \top u^{(n-1)}) \right\},
\]

where \(\beta\) is some positive stepsize, and \(\Pi\) denotes the Euclidean projection operator onto the unit sphere \(S^{d-1} = \{ u \in \mathbb{R}^d \mid \|u\| = 1 \}\), i.e., \(\Pi u = \|u\|^{-1}u\) for all \(u \neq 0\). The iteration (1.4) only requires vector product operations. It has time complexity \(O(d)\) per iteration, and has space complexity \(O(d)\). Iteration (1.4) is very easy to implement in practice and has been used as a heuristic method for fast principal component analysis.

In contrast to classical PCA which is a sample average approximation method, the online PCA iteration (1.4) is essentially a stochastic approximation iteration for the optimization problem (1.1). Surprisingly, theoretical analysis of its convergence remains largely open (except in special cases). The theoretical challenge is due to our attempt to maximize a convex function over a unit sphere, on which there are infinitely many unstable stationary solutions. In this paper, we aim to take an important step in analyzing the convergence of online PCA methods. We are interested in finite-sample analysis of the online PCA iteration, and we aim to match the convergence rate with the information lower bound. An intriguing open question is: Can online principal component analysis be optimal? An answer “yes” would imply that principal component analysis is as simple as estimating the mean of a distribution, for which an iterative online process is sufficient to provide non-improvable estimates. Our results provide a partial answer to this question: almost yes. Our analysis involves the weak convergence theory for Markov processes [15]. Such analysis has a potential for a broader class of stochastic algorithms for nonconvex optimization.
1.1 Related literatures

Although the online PCA iteration (1.4) was proposed over thirty years ago [34,35], its convergence rate analysis remains somewhat limited. It was not until recently, due to the need to handle massive amounts of data, did online PCA gain attention. Two recent works Balsubramani et al. [6] and Shamir [39] study the convergence of online PCA from different perspectives, and obtain some useful rate results. We provide a detailed comparison between their results and ours in Sect. 2.3.

In the mathematical programming and statistics communities, the computational and statistical aspects of PCA are often studied separately. From the statistical perspective, recent developments have focused on estimating principal components for very high-dimensional data. When the data dimension is much larger than the sample size, i.e., $d \gg n$, classical methods using decomposition of the empirical covariance matrix produce inconsistent estimates [22,29]. Sparsity-based methods have been studied, such as the truncated power method studied by Yuan and Zhang [51] and Wang et al. [50]. Other sparsity regularization methods for high dimensional PCA has been studied in Johnstone and Lu [22], Vu and Lei [45,46], Zou [52], d’Aspremont et al. [12], Amini and Wainwright [2], Ma [25], Cai et al. [11], etc. Note that in this paper we do not consider the high-dimensional regime and sparsity regularization.

From the computational perspective, power iterations or the Lanczos method are well studied. These iterative methods require performing multiple products between vectors and empirical covariance matrices. Such operation usually involves multiple passes over the data, whose complexity may scale with the eigengap and dimensions [17,23,28]. Recently, randomized algorithms have been developed to reduce the computation complexity [16,40,41]. A critical trend today is to combine the computational and statistical aspects and to develop algorithmic estimator that admits fast computation as well as good estimation properties. Related literatures include [3,4,18,26,38].

The idea of using stochastic approximation for PCA can be traced back to Oja and Karhunen [35]. Stochastic approximation (SA) was first studied for the root-finding problem and later extended to stochastic optimization and stochastic variational inequalities; see e.g., the textbooks by Kushner and Yin [24], by Benveniste et al. [7], by Borkar [10], by Bertsekas and Tsitsiklis [9]. The idea of processing one sample at a time is also related to the class of incremental methods, which are developed for minimizing the sum of a large number of component functions. These methods update incrementally by making use of one component at a time, through a gradient-type or proximal-type iteration (see for example, [8,30–32,47–49]). However, existing convergence rate analysis of incremental algorithms do not apply to the optimization over the nonconvex sphere constraint.

In the context of machine learning and signal processing, stochastic approximation (or more commonly referred to as stochastic gradient descent) has been extensively studied for stochastic convex optimization. In machine learning applications, the objective is usually the expectation of a convex loss function parameterized by a random variable. It has been shown that after $n$ samples/iterations, the average of the iterates has $\mathcal{O}(1/n)$ optimization error for strongly convex objective, and $\mathcal{O}(1/\sqrt{n})$ error for general convex objective (see [37,42]). For nonsmooth problems with noisy gradients, there are $\Theta(1/n)$ and $\Theta(1/\sqrt{n})$ minimax information-theoretic lower bounds.
for convex and strongly convex problems, respectively (see e.g., [1], and see the book by Nemirovsky and Yudin [33] for a comprehensive study on optimization complexities). The matching between the lower bounds and the upper bounds suggests that stochastic approximation (or equivalently, stochastic gradient) is optimal for convex optimization under the stochastic first-order oracle. In contrast, to the best knowledge of the authors, there has been no such work for stochastic nonconvex optimization problem (1.1) with sphere constraint.

1.2 Our contributions

The contributions of this work is summarized as follows.

- We provide the first convergence rate result for online PCA that nearly matches the information minimax lower bound (2.18) under the subgaussian distributional assumption. We show that, when the initial iterate \( u(0) \) is randomly chosen according to a uniform distribution and the stepsize \( \beta \) is chosen in accordance with the sample size \( n \), there is a high-probability event \( A^* \) with \( \mathbb{P}(A^*) \geq 1 - \delta \) such that

\[
\mathbb{E}\left[ \sin^2 \angle(u(n), u^*) | A^* \right] \leq C(d, n, \delta) \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{(d - 1) \log n}{n},
\]

where \( \delta \in [0, 1) \) and \( C(d, n, \delta) \) is some factor that can be approximately treated as a constant; see details in Eq. (2.16). We also show that, when both the data dimension \( d \) and data size \( n \) scale up with \( d/n^{1-\varepsilon} \to 0 \) for some constant \( \varepsilon \in (0, 1) \), the factor \( C(d, n, \delta) \) approaches to some absolute constant. Moreover, we show that

\[
\mathbb{E}\left[ \sin^2 \angle(u(n), u^*) ; A^* \right] \leq C'(d, n, \delta) \cdot \frac{\lambda_1}{\lambda_1 - \lambda_2} \cdot \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \cdot \frac{\log n}{n},
\]

where the factor \( C'(d, n, \delta) \) converges to some absolute constant as long as \( d^{-1} \sum_{k=1}^{d} \frac{\lambda_k}{\lambda_1} \) is bounded away from 0. To the authors’ best knowledge, this is the tightest convergence result known for the online PCA iteration under the near-optimal scaling condition, i.e., \( d/n^{1-\varepsilon} \to 0 \). A detailed comparison between our result and concurrent ones is given in Sect. 2.3.

- Our convergence rate results are nearly global. In particular, convergence rate results hold as long as the initial iterate satisfies

\[
| \sin \angle(u(0), u^*) | \leq 1 - \frac{c}{d},
\]

for some constant \( c > 0 \). Here the error tolerance scales inverse-linearly as the dimension \( d \) increases. This is critical because, when \( d \) is large, a uniformly distributed initial iterate is nearly perpendicular to the principal component with high probability. Our initial condition allows one to randomly sample \( u(0) \) according to a uniform distribution over the sphere, while preserving the near-optimal conver-
gence rate. In contrast, most existing results on PCA, for instance [51], require that the initial condition be $|\sin \angle(u^{(0)}, u^*)| \leq 1 - c_0$ for some $c_0 \in (0, 1)$. Such initial condition becomes increasingly stringent as the dimension $d$ increases [5]. Our choice of initial iterate does not require any prior knowledge about the principal component. Therefore, our convergence results are nearly global in the sense that a randomly selected initial point achieves near-optimal convergence rate with high probability.

**Organization.** Section 2 states the assumptions on subgassian PCA and give the main results on convergence rate and finite-sample error bound of the online PCA iteration. Section 3 analyzes the convergence of the stochastic iteration and gives proofs of the main results. Section 4 summarizes the results of this paper and presents directions for future works. Some technical analyses are deferred to the appendix.

**Notations.** For a sequence of $\{x_k\}$ and positive $\{y_k\}$, we write $x_k = O(y_k)$ if there exists constant $M < \infty$ such that $|x_k| \leq M y_k$. Also let $\lfloor x \rfloor$ denotes the largest integer $\leq x$, and $\lceil x \rceil$ the smallest integer $\geq x$. Let $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. Let $\|A\|_2$ be the spectral norm of square matrix $A$. Lastly, let $\|X\|_\infty$ be the $L^\infty$-norm of random variable $X$, and also $E[X;A]$ denotes the expectation of $X$ over event $A$, i.e. we have $E[X;A] = E[X 1_A]$.

**2 Main results**

In this section, we present the main convergence results for the online PCA algorithm. The algorithm maintains a running estimate $u^{(n)}$ of the true principal component $u^*$, and updates it while interacting with a streaming data source. Due to the nature of PCA, we are interested in the convergence of the angle process $\{\angle(u^{(n)}, u^*)\}$. A detailed description of the algorithm is given by Algorithm 1.

**Algorithm 1** Online PCA Algorithm

1. Initialize $u^{(0)}$ and choose the stepsize $\beta$ under some condition (to be specified)
2. for $n = 1, 2, \ldots, N$ do
3. Draw one sample $X^{(n)}$ from the (streaming) data source
4. Update the iterate $u^{(n)}$ by

$$u^{(n)} = \Pi \left\{ u^{(n-1)} + \beta X^{(n)} (X^{(n)})^\top u^{(n-1)} \right\}.$$ 

end for

**2.1 Distributional assumptions**

In this paper we focus on the setting where $X$ has a subgaussian distribution. We first introduce the Orlicz $\psi_2$-norm of a random variable.

**Definition 1** (Orlicz $\psi_2$-norm) Let $\psi_2 : \mathbb{R} \mapsto \mathbb{R}$ be a nondecreasing and convex function with $\psi_2(0) = 0$. For a random variable $Y \in \mathbb{R}$, its Orlicz $\psi_2$-norm is defined as

$$\|Y\|_{\psi_2} \equiv \inf \left\{ \xi > 0 : E\left(|Y|^2 / \xi^2\right) \leq 2 \right\}. \quad (2.1)$$
Near-optimal stochastic approximation for online principal...

When a zero-mean random variable $Y$ has a finite $\psi_2$ norm, its distribution is known to have a subgaussian tail. For instance if $Y \sim N(0, \sigma^2)$, we can verify that $\|Y\|_{\psi_2} \leq C\sigma$ for some constant $C$.

**Definition 2** (Subgaussian) We say that a random vector $Z \in \mathbb{R}^d$ follows a subgaussian distribution if $\sup_{v \in S^{d-1}} \|v^T Z\|_{\psi_2} < \infty$.

We refer the readers to §B and Van Der Vaart and Wellner [43], Vershynin [44] for more discussions on subgaussian distributions. Let $\Sigma^{1/2}$ be the unique symmetric positive semidefinite matrix that satisfies $\Sigma^{1/2} \Sigma^{1/2} = \Sigma$. We state our first assumption as follows.

**Assumption 1** (Subgaussian Distribution) Let $Z \in \mathbb{R}^d$ be a random vector satisfying

$$
\mathbb{E}[Z] = 0, \quad \mathbb{E}[ZZ^T] = I_d, \quad \sup_{v \in S^{d-1}} \|v^T Z\|_{\psi_2} \leq 1.
$$

Let $X^{(1)}, X^{(2)}, \ldots \in \mathbb{R}^d$ be independent and identically distributed realizations of $X \equiv \Sigma^{1/2} Z$.

Assumption 1 essentially requires that streaming data $X^{(1)}, X^{(2)}, \ldots$ follow a subgaussian distribution. It follows immediately from Assumption 1 that $\mathbb{E}[X] = 0$ and $\mathbb{E}[XX^T] = \Sigma$.

**Assumption 2** (Positive Eigenvalue Gap) Let the eigenvalues of $\Sigma$ be $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_d \geq 0$.

The positive value $\lambda_1 - \lambda_2$ is often referred to as the eigengap. The nonzero eigengap implies that the principal component $u^*$ is uniquely identified. We will show later that the eigengap plays a key role in the convergence analysis of the online PCA iteration.

### 2.2 Convergence results and finite sample analysis

For easiness of presentation, we introduce a rescaling of the iteration/sample index $n$. We define the rescaled time $N^s_{\beta,s}$ as

$$
N^s_{\beta,s} = \left\lceil s \log\left(\frac{\lambda_1^{-2}(\lambda_1 - \lambda_2)\beta^{-1}}{-\log(1 - \beta(\lambda_1 - \lambda_2))}\right) \right\rceil,
$$

where $s > 0$ is a tuning parameter. Here the fixed time $N^*_{\beta,s}$ increases to infinity as the stepsize $\beta$ decreases to 0. In later analysis, we will choose the stepsize $\beta$ to be inversely proportional to the budget sample size; see Eq. (2.11). For any $c^* > 0$, we also define the rescaled time $N^0_{\beta}(c^*)$ as

$$
N^0_{\beta}(c^*) = \left\lceil \frac{\log(4c^* d)}{-\log(1 - \beta(\lambda_1 - \lambda_2))} \right\rceil.
$$
We state our first result as follows. It provides a foundation to our finite-sample convergence rate analysis.

**Theorem 1** (Convergence result with deterministic initialization) Let Assumptions 1 and 2 hold, and let there be some constant $c^* > 1$ such that $\tan^2 \angle(u^{(0)}, u^*) \leq c^* d$. Then for any $\varepsilon \in (0, 1/8)$ and $\beta > 0$ satisfying

$$
d[\lambda_1^2(\lambda_1 - \lambda_2)^{-1}\beta]^{1-2\varepsilon} \leq b_1/c^*, \quad (2.5)
$$

and any $t > 1$, there exists an event $\mathcal{H}_s$ with

$$
P(\mathcal{H}_s) \geq 1 - 2(d + 2)N_0^\beta(c^*) \exp \left(-C_0[\lambda_1^2(\lambda_1 - \lambda_2)^{-1}\beta]^{-2\varepsilon}\right)
- 4dN_0^\beta_{*,t} \exp \left(-C_1[\lambda_1^2(\lambda_1 - \lambda_2)^{-1}\beta]^{-2\varepsilon}\right), \quad (2.6)
$$

such that the iterates generated by Algorithm 1 satisfy the following for $n \in [N^\beta_{*,1} + N^\beta_{*,t}, N^\beta_{*,t}]$:

$$
\mathbb{E} \left[\tan^2 \angle(u^{(n)}, u^*); \mathcal{H}_s\right] \leq (1 - \beta(\lambda_1 - \lambda_2)^2) \frac{d}{N_0^\beta(c^*)} + C_2 \sum_{k=2}^d \frac{\lambda_1 \lambda_2 + \lambda_1^2[\lambda_1^2(\lambda_1 - \lambda_2)^{-1}\beta]^{0.5-4\varepsilon}}{\lambda_1 - \lambda_k} \cdot \beta. \quad (2.7)
$$

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

Theorem 1 gives a convergence rate estimate for the online PCA algorithm when the initial condition is nearly global, i.e., $\tan^2 \angle(u^{(0)}, u^*) \leq c^* d$ for some $c^* > 0$. It characterizes the time/sample size needed for the angles of iterates to become sufficiently small. Note that the error bound (2.7) is satisfied on a set $\mathcal{H}_s$ with probability close to 1 even if $d \to \infty$, $\beta \to 0$, as long as $d\beta^{1-2\varepsilon} \to 0$ (assuming all other parameters fixed). This means that the convergence rate result is useful in the regime of high-dimensional data analysis. The proof of Theorem 1 is provided in Sect. 3.1.

Next consider the case where the initial solution $u^{(0)}$ is sampled from the unit sphere according to a uniform distribution. In this case, we would expect that $\tan^2 \angle(u^{(0)}, u^*) = \Theta(d)$ with high probability. This implies the initial condition required by Theorem 1 is indeed satisfied. Our second result is given below.

**Theorem 2** (Convergence result with uniformly randomized initialization) Let Assumptions 1 and 2 hold, let $u^{(0)}$ be uniformly sampled from the unit sphere. Then for any $\varepsilon \in (0, 1/8)$, $\beta > 0$, $\delta > 0$ satisfying

$$
d[\lambda_1^2(\lambda_1 - \lambda_2)^{-1}\beta]^{1-2\varepsilon} \leq b_2\delta^2, \quad (2.8)
$$

and

$$
4dN_{*,2}^\beta \exp \left(-C_3[\lambda_1^2(\lambda_1 - \lambda_2)^{-1}\beta]^{-2\varepsilon}\right) \leq \delta, \quad (2.9)
$$
there exists an event $A_\ast$ with $\mathbb{P}(A_\ast) \geq 1 - 2\delta$ such that the iterates generated by Algorithm 1 satisfy

$$
\mathbb{E} \left[ \tan^2 \angle(u^{(n)}, u^*) : A_\ast \right] \leq C_4 \delta^{-4} d^2 (1 - \beta(\lambda_1 - \lambda_2))^{2n} + C_4 \sum_{k=2}^{d} \frac{\lambda_1 \lambda_k + \lambda_1^2 \left[ \frac{\lambda_1^2 (\lambda_1 - \lambda_2)^{-1} \beta}{\lambda_1 - \lambda_k} \right]^{0.5 - 4\varepsilon}}{\lambda_1 - \lambda_k} \cdot \beta.
$$

(2.10)

for $n \in [N_{\beta, 2}^\ast, N_{\beta, 3}^\ast]$. Here $b_2, C_3$ and $C_4$ are positive absolute constants.

Now let us consider the choice of stepsize $\beta$. Suppose that the eigengap $\lambda_1 - \lambda_2$ is known in advance. Also suppose that we are given a budget of sample size $N$. Our goal is to choose an appropriate constant stepsize $\beta$ and minimize the finite-sample error bound (2.10). We will pick the stepsize to be

$$
\bar{\beta}(N) = \frac{2 \log N}{(\lambda_1 - \lambda_2) N}
$$

(2.11)

which is asymptotically the minimizer of the right hand of Eq. (2.10). Then we obtain the main result of this paper: a finite-sample error bound for online PCA under sub-gaussian assumption. It is stated as follows.

**Theorem 3** (Main Result: Finite-sample error bound with uniformly randomized initialization) Let Assumptions 1 and 2 hold, let $u^{(0)}$ be uniformly sampled from the unit sphere, and let $\beta = \bar{\beta}(N)$ be given by Eq. (2.11). Then for any $\varepsilon_0 \in (0, 1/8)$, $N \geq 1$, $\delta > 0$ satisfying

$$
d \left[ \frac{\lambda_1^2 (\lambda_1 - \lambda_2)}{(\lambda_1 - \lambda_2)^2} \cdot \frac{2 \log N}{N} \right]^{1 - 2\varepsilon_0} \leq b_3 \delta^2,
$$

(2.12)

and

$$
4d N_{\bar{\beta}(N), 2}^\ast \exp \left( - C_6 [\lambda_1^2 (\lambda_1 - \lambda_2) \bar{\beta}(N)]^{-2\varepsilon_0} \right) \leq \delta,
$$

(2.13)

there exists an event $A_\ast$ with $\mathbb{P}(A_\ast) \geq 1 - 2\delta$ such that the iterates generated by Algorithm 1 satisfy the following:

(a) There is a factor $C^{(1)}(d, N, \delta)$ such that

$$
\mathbb{E} \left[ \tan^2 \angle(u^{(N)}, u^*) : A_\ast \right] \leq C^{(1)}(d, N, \delta) \cdot \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \cdot \frac{\log N}{N}.
$$

(2.14)

The factor $C^{(1)}(d, N, \delta)$ approaches to some absolute positive constant $C_5$ as $d \to \infty$, $N \to \infty$ when Eqs. (2.12), (2.13) and the following scaling condition hold:

$$
d^{-1} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \text{ is bounded away from } 0.
$$

(2.15)
(b) There is a factor $C^{(2)}(d, N, \delta)$ such that
\[
\mathbb{E} \left[ \tan^2 \angle(\mathbf{u}^{(N)}, \mathbf{u}^*) \mid A_* \right] \leq C^{(2)}(d, N, \delta) \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{(d - 1) \log N}{N}.
\]
(2.16)

The factor $C^{(2)}(d, N, \delta)$ approaches $C_5$ as $d \to \infty$, $N \to \infty$ when Eq. (2.12) holds and $\lambda_2 / \lambda_1$ is bounded away from 0.

In the above, $b_3, C_5, C_6$ are positive absolute constants.

Remarks. Theorem 3 gives an explicit estimate of the convergence rate when a fixed sample size $N$ is known in advance. We note the following:

(i) The conditions (2.12) and (2.13) are satisfied when $d / N^{1-2\varepsilon}$ is sufficiently small. They essentially require that the sample size $N$ be sufficiently large when the dimension $d$ is high.

(ii) The eigenvalue scaling condition Eq. (2.15) requires that the eigenvalues do not decay too fast. For example, Eq. (2.15) is satisfied if $\lambda_2 / \lambda_1$ is bounded away from 0. For another example, Eq. (2.15) is satisfied if $d^{-1} \sum_{k=1}^{d} \lambda_k / \lambda_1$ is bounded away from 0.

(iii) The choice of stepsize $\beta = \bar{\beta}(N)$ does not involve the dimension $d$. Both constant factors in Eqs. (2.14) and (2.16) satisfy $C^{(i)}(d, N, \delta)$ for sufficiently large $d$ and $N$, as long as both the aforementioned scaling conditions hold. From an asymptotic point of view, this implies that as long as $d \to \infty$ and $N \to \infty$ with $d / N^{1-\varepsilon} \to 0$ for some $\varepsilon > 0$, we have $C^{(i)}(d, N, \delta) \to C_5$, and hence $C^{(i)}(d, N, \delta), i = 1, 2$ can be approximately treated as universal constants.

Proofs of Theorems 2 and 3 are deferred to Sect. 3.2.

Matching the minimax information lower bound. For an arbitrary constant $\sigma^2 > 0$, let $\mathcal{M}(\sigma^2, d)$ be the collection of all distributions of $X = \Sigma^{1/2} Z$ satisfying Eq. (2.2) and the following effective noise variance condition
\[
\frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \leq \sigma^2 ;
\]
(2.17)
see [46]. For the collection of subgaussian distributions $\mathcal{M}(\sigma^2, d)$, Theorem 3.1 of [46] established the following minimax information lower bound for estimating the corresponding principal component, given by
\[
\inf_{\overline{u}^{(N)}} \sup_{X \in \mathcal{M}(\sigma^2, d)} \mathbb{E} \left[ \sin^2 \angle(\overline{u}^{(N)}, \mathbf{u}^*) \right] \geq c \cdot \sigma^2 \cdot \frac{d - 1}{N},
\]
(2.18)
for some $c > 0$, where the infimum of $\overline{u}^{(N)}$ is taken over all principal eigenvector estimator using the first $N$ data samples $X^{(1)}, \ldots, X^{(N)}$.

According to Theorem 3(b), we have
\[
\sup_{X \in \mathcal{M}(\sigma^2, d)} \mathbb{E} \left[ \sin^2 \angle(\mathbf{u}^{(N)}, \mathbf{u}^*) \mid A_* \right] \leq C^{(2)}(d, N, \delta) \cdot \sigma^2 \cdot \frac{(d - 1) \log N}{N}.
\]
Near-optimal stochastic approximation for online principal...

where \( C^{(2)}(d, N, \delta) \) can be approximately treated as a constant. In comparison with Eq. (2.18), Theorem 3 suggests that the online PCA is nearly optimal, in the sense that the finite-sample error matches the minimax information lower bound up to a \( \log N \) factor with high probability. In addition, online PCA achieves similar accuracy as that of the classical PCA in the batch setting (up to \( O(\log N) \)). More importantly, the finite-sample bound in Theorem 3 is nearly global in the sense that it does not require a warm initialization, and hence the near-optimal convergence rate is achieved even using the uniformly random initialization. In summary, our convergence rate result is both nearly optimal and nearly global.

2.3 Comparison with existing and concurrent works

We give a detailed comparison between our finite-sample error bound, i.e., Theorem 3, and related recent results. We emphasize that all comparable results assume that the samples are uniformly bounded with probability 1, i.e., \( \|X\|_2^2 \leq B \) for some constant \( B > 0 \). As a result, their results do not generalize to high-dimensional problems with unbounded distribution. In comparison, we assume that the distribution of \( X \) has subgaussian tails, which is much more general than all existing and concurrent works. A detailed list is given below. For ease of comparison, we rephrase the related results into our setting of uniformly randomized initialization.

(a) The main results of Balsubramani et al. [6] (Theorem 1.1) can be summarized as follows. Under some technical assumptions, let \( u(0) \) be uniformly sampled from \( S_d^{d-1} \), let \( n_1 \) be some starting time, and let the step sizes be \( \beta_n = 2(\lambda_1 - \lambda_2)^{-1}(n + n_1)^{-1} \). Then with probability at least 3/4, the iterates satisfy

\[
\sin^2 \angle(u^{(n)}, u^*) \leq C \cdot \frac{B^2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n + n_1} + C \cdot d^4 \cdot \left( \frac{n_1}{n + n_1} \right)^2,
\]

where \( C > 0 \) is an absolute constant. This is one of the earliest convergence rate results for the online PCA iteration. Eq. (2.19) does not fully characterize the error’s dependence on \( \lambda_1, \lambda_2 \). It does not match the minimax lower bound (2.18).

(b) The work by De Sa et al. [13] studies a different but closely related problem on minimizing the spectral error using a stochastic gradient algorithm. The algorithm’s angular part is equivalent to our online PCA iteration. Their theoretical guarantees are summarized as: Let some technical assumptions hold, let \( u^{(0)} \) be uniformly sampled from \( S^{d-1} \). Given the sample size \( N \), by setting \( \beta = 16(\lambda_1 - \lambda_2)^{-1}N^{-1} \), the output satisfies with probability at least 3/4 that

\[
\sin^2 \angle(u^{(N)}, u^*) \leq C \cdot \frac{B \lambda_1}{(\lambda_1 - \lambda_2)^2} \cdot \frac{d \log N}{N}.
\]

In addition, De Sa et al. [13] also proposed an online estimator for the top eigenvalue \( \lambda_1 \).

(c) Another related work is Shamir [39]. Shamir [39] proves a convergence rate result in terms of the objective values of problem (1.1). We rephrase their main
result (Theorem 2) as follows. Let $u^{(0)}$ be uniformly sampled from $S^{d-1}$, and let the stepsize be chosen according to the sample size, i.e., $\beta(N) = \lambda_1(\lambda_1 - \lambda_2)^{-1} \cdot N^{-1} \log N$. Then with probability at least $\mathcal{O}(d^{-1})$ the iterates generated by Algorithm 1 satisfy
\[
\sin^2 \angle(u^{(N)}, u^*) \leq \frac{u^*^\top \Sigma u^* - u^{(N)}_\top \Sigma u^{(N)}}{\lambda_1 - \lambda_2} \leq C \cdot \frac{B^2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{d \log^2 N}{N},
\] (2.21)
for all $N$ sufficiently large, where $C > 0$ is an absolute constant. The preceding result holds with probability $\mathcal{O}(1/d)$, which approaches to 0 as $d$ grows. In an updated version of Shamir [39], the success probability has been improved to $\mathcal{O}(\log^{-1} d)$.

(d) A very recent independent work by Jain et al. [21] (which is released after the initial submission of the current paper) obtains the following result: Let the $u^{(0)}$ be uniformly sampled from $S^{d-1}$, let $n_2$ be some appropriate starting time, and let the stepsize be $\beta_n = (\lambda_1 - \lambda_2)^{-1} (n + n_2)^{-1}$. Then the output $u^{(n)}$ of Algorithm 1 satisfies with probability at least $3/4$ that
\[
\sin^2 \angle(u^{(n)}, u^*) \leq C \cdot \frac{B \lambda_1}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n} + C \cdot d \cdot \left(\frac{n_2}{n}\right)^2.
\] (2.22)
Here $C > 0$ is an absolute constant.

We summarize all existing rate of convergence results for online PCA in Table 1. In short, the comparable results listed above all require a severely stringent assumption of uniform boundedness. Our results hold under the much more general assumption of subgaussian distributions. We provide the sharpest finite-sample error bound for the online PCA iteration. It is the first result that nearly matches the minimax information lower bound [46] up to a polylogarithmic factor of $N$.

### 3 Proofs of main results

This section analyzes the convergence of the stochastic iteration generated by Algorithm 1 and provides detailed proofs of the convergence rate results in Sect. 2. Arguments for convergence results are separately provided in Sects. 3.1 and 3.2, separately for the deterministic initialization case (Theorem 1) and for the uniform randomized initialization case (Theorem 2). Section 3.2 also provides the argument for the finite-sample result with uniform randomized initialization (Theorem 3). Proofs of all propositions in this section are provided in Appendix C.

#### 3.1 Proof of Theorem 1

This subsection aims to prove the convergence result for the deterministic initialization case, Theorem 1. To prepare for the proof, first we let the diagonal decomposition of the covariance matrix be
Table 1  Comparable results on the convergence rate of online PCA

| Algorithm                        | $\sin^2 \angle (u^{(n)}, u^*)$ | Optimality |
|----------------------------------|---------------------------------|------------|
| Matrix Bernstein inequality      | $C \cdot \frac{B\lambda_1}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n}$ | Lower bound |
| Alecton [13]                     | $C \cdot \frac{B\lambda_1 \cdot d}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n}$ | No         |
| Block power method [19,26]       | $C \cdot \frac{B\lambda_1^2}{(\lambda_1 - \lambda_2)^3} \cdot \frac{1}{n}$ | No         |
| Online PCA, Oja [6]              | $C \cdot \frac{B^2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n}$ | No         |
| Online PCA, Oja [39]             | $C \cdot \frac{B^2 \cdot d}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n}$ | No         |
| Online PCA, Oja [21]             | $C \cdot \frac{B\lambda_1}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n}$ | Yes        |
| Minimax rate [46], subgaussian   | $C \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n}$ | Lower bound |
| Online PCA, Oja (this work), sub-| $C \cdot \frac{\lambda_1 \lambda_2 \cdot d}{(\lambda_1 - \lambda_2)^2} \cdot \frac{1}{n}$ | Yes        |
| gaussian                         |                                  |            |

Note that our result matches the minimax information lower bound [46] in the case where $\lambda_2 = \cdots = \lambda_d$. Our result provides a finer estimate than the minimax lower bound in the more general case where $\lambda_2 \neq \lambda_d$. Note that the constant $C$ hides poly-logarithmic factors of $d$ and $n$.

$$\Sigma = \mathbb{E} \left[XX^\top\right] = U\Lambda U^\top,$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_d$, and $U$ is an orthogonal matrix consisting of column eigenvectors of $\Sigma$. Clearly the first column of $U$ is equal to the principal component $u^*$. Note that the diagonal decomposition might not be unique, in which case we work with an arbitrary one.

Second, we define the rescaled samples as

$$Y^{(n)} = U^\top X^{(n)}, \quad v^{(n)} = U^\top u^{(n)}, \quad v^* = U^\top u^*. \quad (3.1)$$

One can easily verify that

$$\mathbb{E}[Y] = 0, \quad \mathbb{E} \left[YY^\top\right] = \Lambda;$$

The principal component of the rescaled random variable $Y$, which we denote by $v^*$, is equal to $e_1$, where $\{e_1, \ldots, e_d\}$ is the canonical basis of $\mathbb{R}^d$. By applying the linear transformation $U^\top$ to the stochastic process $\{u^{(n)}\}$, we obtain an iterative process $\{v^{(n)} = U^\top u^{(n)}\}$ in the rescaled space:

$$v^{(n)} = \Pi \left\{ v^{(n-1)} + \beta Y^{(n)} \left( Y^{(n)} \right)^\top v^{(n-1)} \right\}. \quad (3.2)$$
Moreover, the angle processes associated with \( \{u^{(n)}\} \) and \( \{v^{(n)}\} \) are equivalent, i.e.,
\[
\angle(u^{(n)}, u^*) = \angle(v^{(n)}, v^*).
\]

Therefore it would be sufficient to study the rescaled process given by (3.2).

In the proof follows, we use the rescaled samples as in Eq. (3.1) and iterations as in Eq. (3.2). Consider a partition \( S^{d-1} = S_1 \cup S_2 \) where
\[
S_1 = \left\{ v \in S^{d-1} : |v_1| < 1/\sqrt{2} \right\}, \quad S_2 = \left\{ v \in S^{d-1} : |v_1| \geq 1/\sqrt{2} \right\}.
\] (3.3)

We refer to \( S_1 \) and \( S_2 \) as the cold region and the warm region, respectively.

We first focus on Algorithm 1 when the initial estimator lies in the warm region \( S_2 \), which we conveniently call warm start. Such analysis is crucial in obtaining the correct rate of convergence in the proof of Theorem 1. In terms of the angle \( \angle(v^{(0)}, v^*) \) this warm start condition is equivalent to
\[
\angle(v^{(0)}, v^*) \in [0, \pi/4] \cup [3\pi/4, \pi].
\]

To avoid uncontrollable variances we need its first coordinate \( v^{(0)}_1 \) to be bounded away from 0 throughout the algorithm for \( N_{\beta,t} \) iterates. We firstly define an auxiliary region
\[
S_3 = \left\{ v \in S^{d-1} : |v_1| \in [1/3, 1] \right\},
\] (3.4)
and set the stopping time
\[
\mathcal{N}_w = \inf \left\{ n \geq 0 : v^{(n)} \in S^c_3 \right\}, \quad (3.5)
\]
where \( A^c \) for a generic set \( A \) denotes its complement set. Also, for a positive quantity \( M \) to be determined later, let
\[
\mathcal{N}_M = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq d} \left( \max_{1 \leq k \leq d} |Y_k^{(n)}|, |v^{(n-1)}_1 Y^{(n)}| \right) \geq M^{1/2} \right\}.
\] (3.6)

In words, \( \mathcal{N}_M \) is the first \( n \) such that the maximal absolute coordinate of \( Y^{(n)} \) exceeds \( M^{1/2} \), or the inner product of \( v^{(n-1)} \) and \( Y^{(n)} \), in absolute value, exceeds \( M^{1/2} \), whichever occurs earlier. It is convenient to define the rescaled stepsize
\[
\hat{\beta} = \lambda_1^2 (\lambda_1 - \lambda_2)^{-1} \beta.
\] (3.7)

We consider the ratio iteration \( U_k^{(n)} \) defined as
\[
U_k^{(n)} = \frac{v_k^{(n)}}{v_1^{(n)}},
\] (3.8)
Geometrically, we observe that the ratio $U_k^{(n)}$ is the tangent of angle between $v^{(n)}$ and principal eigenvector $v^* = e_1$, after projected onto the two-dimensional subspace spanned by $e_1$ and $e_k$ which are the 1st and $k$th canonical unit vectors. Immediately from Eq. (3.8) we have

$$\tan^2 \angle (v^{(n)}, v^*) = \sum_{k=2}^d \left(U_k^{(n)}\right)^2.$$  

(3.9)

Our first goal in this subsection is to prove the following

**Proposition 1** Assume all conditions in Theorem 1 along with the warm start condition $v^{(0)} \in S_2$. Then there is a positive constant $b_4$ such that whenever $d \hat{\beta}^{1-2\epsilon} \leq b_4$, for any fixed $t > 0$ and for $k = 2, \ldots, d$ we have with probability $\geq 1 - 2N_{\beta,t}^* \exp (-C_{1,p} \hat{\beta}^{-2\epsilon})$, either $(\mathcal{N}_M \leq N_{\beta,t}^*)$ occurs, or

$$\sup_{n \leq N_{\beta,t}^* \wedge N_w} \left| U_k^{(n)} - U_k^{(0)} (1 - \beta(\lambda_1 - \lambda_k))^n \right| \leq C_{1,p} \hat{\beta}^{0.5-\epsilon}.$$  

Here $C_{1,p}$ and $C_{1,p}'$ are positive constants.

Proof of Proposition 1 is deferred to §C.2. Proposition 1 shows that from warm start, for each coordinate $k = 2, \ldots, d$, the $U_k^{(n)}$ approximately decays geometrically at rate $1 - \beta(\lambda_1 - \lambda_k)$, or bad event $(\mathcal{N}_M \leq N_{\beta,t}^*)$ occurs.

The next lemma controls the occurrence of the bad event by pinning down the choice of $M$.

**Lemma 1** Assume all conditions in Theorem 1. Suppose we choose $M$ as

$$M \equiv \lambda_1 \hat{\beta}^{-2\epsilon}.$$  

(3.11)

Then for each fixed $N \geq 1$

$$\mathbb{P} (\mathcal{N}_M \leq N) \leq 2(d + 1)N \exp \left(-\hat{\beta}^{-2\epsilon}\right).$$  

(3.12)

The proof of Lemma 1 is provided in §D.1. Using Proposition 1 and Lemma 1, one can obtain the rate of convergence under more careful second moment estimates, as follows.

**Proposition 2** Assume all conditions in Theorem 1 along with the warm start condition $v^{(0)} \in S_2$. Then there is a positive constant $b_4$ such that whenever Eq. (3.10) is satisfied, there exists a high-probability event $\mathcal{H}_0$ satisfying

$$\mathbb{P} (\mathcal{H}_0) \geq 1 - 4d N_{\beta,t}^* \exp \left(-C_{2,p} \hat{\beta}^{-2\epsilon}\right).$$  

(3.13)
such that for fixed $k = 2, \ldots, d$ and $n \in [N_{\beta,1}^*, N_{\beta,t}^*]$, 

$$
E \left[ (U_k^{(n)})^2 ; \mathcal{H}_0 \right] \leq (1 - \beta(\lambda_1 - \lambda_k))^{2n} \left( U_k^{(0)} \right)^2 + \frac{C_{2,p} \beta \beta^0.5 - 4 \epsilon}{\lambda_1 - \lambda_k} \cdot \beta
$$

(3.14)

Here $C_{2,p}$ and $C_{2,p}'$ are positive constants.

Proof of Proposition 2 is deferred to §C.3. Note from Eq. (3.13) $\mathcal{H}_0$ occurs with high probability as $\beta \to 0$. With $\hat{\beta}$ defined in Eq. (3.7) the second term on the right hand of Eq. (3.14) matches the second moment estimates as in Eq. (2.7).

Second, we try to relax the warm start condition. Define the stopping time

$$
N_c = \inf \left\{ n \geq 0 : v^{(n)} \in S_2 \right\}.
$$

(3.15)

In words, $N_c$ is the first $n$ such that the iterate enters the warm region. In the case of warm start, $N_c = 0$. We have the following Proposition 3 which upper bound the stopping time $N_c$, whose proof is deferred to §C.4.

**Proposition 3** Assume all conditions in Theorem 1. There exist a positive constant $b_5 < \ln^2 2/16$ such that if

$$
d\beta^{1-2\epsilon} \leq b_5/c^*,
$$

(3.16)

then we have

$$
P \left( N_c \leq N_{\beta}^*(c^*) \right) \geq 1 - 2(d + 2)N_{\beta}^*(c^*) \exp \left( -C_{3,p} \hat{\beta}^{2-2\epsilon} \right),
$$

(3.17)

where $N_{\beta}^*(c^*)$ is defined as in Eq. (2.4).

We are now ready prove Theorem 1. Note that under the setting of Theorem 1, Algorithm 1 starts from $v^{(0)} \in S^{d-1}$ where $\tan^2 \angle (v^{(0)}, v^*) \leq c^* d$. Running the algorithm for $N_{\beta}^*(c^*) \wedge N_c$ steps we know from Proposition 3 that the iterate $v^{(N_{\beta}^*(c^*) \wedge N_c)}$ lies in $S_2$ with high probability. By strong Markov property, the iterates generated by Algorithm 1 have the same law as the one starting from $v^{(N_{\beta}^*(c^*) \wedge N_c)}$, and hence Proposition 2 can be applied. We follow this reasoning and detail the proof as in below.

**Proof of Theorem 1** Let the events

$$
\mathcal{H}_{*,1} = \left( N_c \leq N_{\beta}^*(c^*) \right),
$$

$$
\mathcal{H}_{*,2} = \left( \sup_{n \in [N_{\beta,1}^*, N_{\beta,t}^*]} \left\{ U_k^{(n+N_c)} \right\} \leq 2\beta^{0.5 - \epsilon} \text{ for all } k = 2, \ldots, d \right),
$$

(3.18)

$$
\mathcal{H}_* = \mathcal{H}_{*,1} \cap \mathcal{H}_{*,2}.
$$
On $\mathcal{H}_* \cap (\mathcal{N}_c = n_\circ)$ where $n_\circ \leq N^\circ_\beta(c^*)$ is a fixed time, from the definition in Eq. (3.15) we have $v(\mathcal{N}_c \wedge N^\circ_\beta(c^*)) = v(n_\circ) \in S_2$ and hence from Eq. (3.9)

$$\tan^2 \angle(v(n_\circ), v^*) = \sum_{k=2}^d \left( U_k^{(n_\circ)} \right)^2 \leq 1.$$  \hspace{1cm} (3.19)

Therefore for

$$n \in [N^\circ_{\beta,1} + N^\circ_\beta(c^*), N^\circ_{\beta,t}] \subseteq [N^\circ_{\beta,1} + n_\circ, N^\circ_{\beta,t} + n_\circ],$$

and $\beta$ satisfying both Eqs. (3.10) and (3.16), we utilize the Markov property and Proposition 2 and conclude that for $\beta$ satisfying Eq. (2.5) where $b_1 = b_4 \wedge b_5$

$$\mathbb{E} \left[ \tan^2 \angle(v(n), v^*) 1_{\mathcal{H}_*} \mid \mathcal{N}_c = n_\circ \right] = \sum_{k=2}^d \mathbb{E} \left[ \left( U_k^{(n)} \right)^2 1_{\mathcal{H}_*} \mid \mathcal{N}_c = n_\circ \right]$$

$$\leq \sum_{k=2}^d \left( 1 - \beta(\lambda_1 - \lambda_k) \right)^{2(n-n_\circ)} \left( U_k^{(n_\circ)} \right)^2 + \sum_{k=2}^d C'_{2,p} \frac{(\lambda_1 \lambda_k + \lambda_1^2 \cdot \hat{\beta}^{0.5-4\epsilon})}{2(\lambda_1 - \lambda_k)} \beta$$

$$\leq (1 - \beta(\lambda_1 - \lambda_2))^{2(n-N^\circ_\beta(c^*))} + \sum_{k=2}^d C'_{2,p} \frac{(\lambda_1 \lambda_k + \lambda_1^2 \cdot \hat{\beta}^{0.5-4\epsilon})}{2(\lambda_1 - \lambda_k)} \beta.$$

Eqs. (3.9), (3.19) have been applied in the above estimate. The bound for

$$\mathbb{E} \left[ \tan^2 \angle(v(n), v^*) 1_{\mathcal{H}_*} \mid \mathcal{N}_c \right]$$

was argued for $\mathcal{N}_c = n_\circ \leq N^\circ_\beta(c^*)$ but is trivially valid on $(\mathcal{N}_c > N^\circ_\beta(c^*))$, since $1_{\mathcal{H}_*} = 0$ on the latter event. Furthermore, the bound on the right hand is deterministic and hence independent of $\mathcal{N}_c$. Taking expectation again gives

$$\mathbb{E} \left[ \tan^2 \angle(v(n), v^*) \mid \mathcal{H}_* \right] \leq (1 - \beta(\lambda_1 - \lambda_2))^{2(n-N^\circ_\beta(c^*))}$$

$$+ \sum_{k=2}^d C'_{2,p} \frac{(\lambda_1 \lambda_k + \lambda_1^2 \cdot \hat{\beta}^{0.5-4\epsilon})}{2(\lambda_1 - \lambda_k)} \beta.$$

Therefore Eq. (2.7) establishes with $C_2 = C'_{2,p}$.

To ensure that the event $\mathcal{H}_*$ defined in Eq. (3.18) satisfies the probability estimate in Eq. (2.6), note Eqs. (3.13) and (3.17) together imply

$$\mathbb{P}(\mathcal{H}_*) \geq \left[ 1 - 2(d + 2)N^\circ_\beta(c^*) \exp\left( -C_3,p \hat{\beta}^{2\epsilon} \right) \right] \left[ 1 - 4dN^\circ_{\beta,t} \exp\left( -C_2,p \hat{\beta}^{2\epsilon} \right) \right]$$

$$\geq 1 - 2(d + 2)N^\circ_\beta(c^*) \exp\left( -C_3,p \hat{\beta}^{2\epsilon} \right) - 4dN^\circ_{\beta,t} \exp\left( -C_2,p \hat{\beta}^{2\epsilon} \right),$$
Letting $C_0 = C_3, p$ and $C_1 = C_2, p$ this implies that Eq. (2.6) holds. Proof of Theorem 1 is accomplished. □

### 3.2 Proofs of Theorem 2 and 3

Next we prove the main results of the uniform randomized initialization case, Theorems 2 and 3. The following lemma basically says that the initial estimator condition in Theorem 1 is satisfied with high probability, where the coefficient $c^* = \delta^{\frac{-2}{d}}$ is quadratically inverse proportional to the error probability $\delta$.

**Lemma 2** Given any $\delta > 0$, if $u^{(0)}$ is sampled uniformly at random from $S^{d-1}$ in $\mathbb{R}^{d}$ then there exists a constant $C^* > 1$ independent of $\delta$ and $d$ such that

$$
P\left(\tan^2 \angle(u^{(0)}, u^*) \leq C^*\delta^{-2}d\right) \geq 1 - \delta.
$$

Proof of Lemma 2 is provided in §D.2. We are ready to prove Theorem 2.

**Proof of Theorem 2** Let the event

$$
A'_* = \left(\tan^2 \angle(u^{(0)}, u^*) \leq C^*\delta^{-2}d\right),
$$

and recall $\hat{\beta}$ is as defined in Eq. (3.7). Since the initial estimator $u^{(0)}$ is sampled uniformly at random from $S^{d-1}$, Lemma 2 indicates $P(A'_*) \geq 1 - \delta$, and our approach is to apply Theorem 1 with $c^* = \delta^{\frac{-2}{d}}$.

Firstly, Eq. (3.16) in Proposition 3, one has $d\hat{\beta}^{1-2\varepsilon} \leq (4C^*\delta^{-2})^{-1}$ i.e.

$$
\log\left(4C^*\delta^{-2}d\right) \leq (1 - 2\varepsilon) \log(\hat{\beta}^{-1}).
$$

Combined with the definitions in Eqs. (2.3) and (2.4), this implies

$$
N^*_\beta(C^*\delta^{-2}) \leq N^*_{\beta, 1-2\varepsilon} \leq N^*_{\beta, 1} - 1.
$$

(3.20)

Conditioning on $A'_*$ which satisfies the initial condition in Theorem 1, Eq. (3.20) guarantees the existence of the event $H_*$ such that

$$
P(H_* | A'_*) \geq 1 - 2(d + 2)N^*_\beta(c^*) \exp\left(-C_0\hat{\beta}^{-2\varepsilon}\right) - 4dN^*_{\beta, t} \exp\left(-C_1\hat{\beta}^{-2\varepsilon}\right)
$$

$$
\geq 1 - 4dN^*_{\beta, t+1} \exp\left(-C_3\hat{\beta}^{-2\varepsilon}\right) \geq 1 - \delta,
$$

(3.21)

where $C_3 \equiv C_0 \wedge C_1$. Thus immediately

$$
P(A'_* \cap H_*) = P(A'_*)P(H_* | A'_*) \geq (1 - \delta)^2 \geq 1 - 2\delta.
$$
Finally, for all $\beta$ satisfying $d\beta^{1-2\varepsilon} \leq (b_1/C^*)\delta^2$,

$$\mathbb{E}\left[\tan^2 \angle (u^{(n)}, u^*) \mid A_* \cap \mathcal{H}_*\right] \leq \mathbb{E}\left[\tan^2 \angle (u^{(n)}, u^*) \mathbf{1}_{\mathcal{H}_*} \mid A_*\right]$$

$$\leq 4C^*\delta^{-2}d (1 - \beta(\lambda_1 - \lambda_2))^{2n} + C_2 \sum_{k=2}^{d} \frac{\lambda_1 \lambda_k + \lambda_2^2 \beta^{0.5-4\varepsilon}}{\lambda_1 - \lambda_k} \cdot \beta.$$ 

where we applied from Eq. (2.4) that

$$d = (1 - \beta(\lambda_1 - \lambda_2))^{-2N^\beta_0(C^*\delta^{-2})} \leq (4C^*\delta^{-2}d)^2.$$ 

To summarize all above, setting $C_4 = (4C^*)^2 \vee C_2$, $b_2 = [b_1 \wedge (1/4)]/C^*$ and $A_* = A'_* \cap \mathcal{H}_*$ concludes Theorem 2. 

We finalize this paper by proving the finite-sample result, Theorem 3.

**Proof of Theorem 3** As Eq. (2.12) holds, choosing $\beta$ as in Eq. (2.11), then if for some positive constants $b_2 > 0$ and $\varepsilon = \varepsilon_0 \in (0, 1/8)$, $\beta$ and $\delta > 0$ satisfies Eq. (2.8) which is translated to

$$d \leq b_2\delta^2 \left[ \frac{\lambda_1^2}{(\lambda_1 - \lambda_2)^2} \frac{\lambda_2^2}{\lambda_1 - \lambda_2} \frac{\lambda_1^2}{\lambda_1 - \lambda_2} \frac{\log N}{N} \right]^{2\varepsilon-1}.$$ 

(3.22)

Hence Theorem 2 applies, and it is not hard to verify when $\beta = \bar{\beta}(N)$ defined in Eq. (2.11) one has immediately $N \in [N^*_\beta, 1, N^*_\beta, 2]$, and hence there exists an event $A_*$ with $\mathbb{P}(A_*) \geq 1 - 2\delta$, such that the iterates generated by Algorithm 1 satisfy Eq. (2.10) for $n \in [N^*_\beta, 2-\varepsilon, N^*_\beta, t]$. We consider the two regimes separately.

(a) Plugging in $\beta = \bar{\beta}(N)$, and hence to obtain Eq. (2.14)

$$\mathbb{E}\left[\tan^2 \angle (u^{(N)}, u^*) \mid A_*\right] \leq C^0(d, N, \delta) \cdot \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \cdot \frac{\log N}{N},$$

where the factor $C^0(d, N, \delta)$ takes the form

$$C^0(d, N, \delta) = \left( C_4\delta^{-2}d (1 - \bar{\beta}(N)(\lambda_1 - \lambda_2))^{2N} + C_4 \sum_{k=2}^{d} \frac{\lambda_1 \lambda_k + \lambda_2^2 \beta^{0.5-4\varepsilon}}{\lambda_1 - \lambda_k} \frac{\log N}{N} \right) \cdot \left( \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \cdot \frac{\log N}{N} \right)^{-1} = I + II.$$ 

(3.23)
From Eqs. (3.22) and (2.11)

\[ I \equiv C_4 \delta^{-2} d \left( 1 - \bar{\beta}(N)(\lambda_1 - \lambda_2) \right)^{2N} \cdot \left( \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \cdot \frac{\log N}{N} \right)^{-1} \]

\[ \leq C_4 \delta^{-2} d \exp \left( -2N \cdot \frac{\log N}{N} \right) \cdot \left( \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{\log N}{N} \right)^{-1} \]

\[ \leq C_4 b_2 \left[ \frac{\lambda_1^2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{\log N}{N} \right]^{2\varepsilon - 1} \cdot \left( \frac{\log N}{N} \right)^{-1} \cdot \left( \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \right)^{-1} \]

\[ \leq C_4 b_2 \left( \frac{\lambda_1}{\lambda_1 - \lambda_2} \right)^{4\varepsilon - 2} \left( \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \right)^{-1} \frac{1}{N^{2\varepsilon} \log^{2\varepsilon - 2\varepsilon} N}, \]

and also

\[ II \equiv C_4 \sum_{k=2}^{d} \frac{\lambda_1 \lambda_k + \lambda_1^2 \left[ \lambda_1^2 \left( \lambda_1 - \lambda_2 \right)^{-1} \bar{\beta}(N) \right]^{0.5 - 4\varepsilon}}{(\lambda_1 - \lambda_k)(\lambda_1 - \lambda_2)} \cdot \left( \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \right)^{-1} \]

\[ = C_4 \left( 1 + \left[ \lambda_1^2 \left( \lambda_1 - \lambda_2 \right)^{-1} \bar{\beta}(N) \right]^{0.5 - 4\varepsilon} \left[ d \left( \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \right)^{-1} + 1 \right] \right). \]

Both terms on the RHS of Eq. (3.24) have been estimated, and we let

\[ C^{(1)}(d, N, \delta) = C_4 b_2 \left( \frac{\lambda_1}{\lambda_1 - \lambda_2} \right)^{4\varepsilon - 2} \left( \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \right)^{-1} \frac{1}{N^{2\varepsilon} \log^{2\varepsilon - 2\varepsilon} N} \]

\[ + C_4 + C_4 \left[ \lambda_1^2 \left( \lambda_1 - \lambda_2 \right)^{-1} \bar{\beta}(N) \right]^{0.5 - 4\varepsilon} \left[ d \left( \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \right)^{-1} + 1 \right], \]

then we conclude immediately that as \( d, N \to \infty \) when both Eqs. (2.12) and (2.15) hold, the term goes to \( C_4 \). Letting \( C_5 \equiv C_4 \) completes the proof.

(b) For Eq. (2.16) we note \( \lambda_k \leq \lambda_2 \) for \( k \geq 2 \) and

\[ \mathbb{E} \left[ \tan^2 \angle(u^{(N)}, u^*) ; \ A_* \right] \leq C^{(1)}(d, N, \delta) \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{(d - 1) \log N}{N}. \]

where the factor \( C^{(1)}(d, N, \delta) \) has, via Eqs. (3.22) and (2.11),

\[ C^{(1)}(d, N, \delta) = \left( C_4 \delta^{-2} d \left( 1 - \bar{\beta}(N)(\lambda_1 - \lambda_2) \right)^{2N} \right) \]

\[ + C_4 \frac{\lambda_1 \lambda_2 + \lambda_1^2 \left[ \lambda_1^2 \left( \lambda_1 - \lambda_2 \right)^{-1} \bar{\beta}(N) \right]^{0.5 - 4\varepsilon}}{(\lambda_1 - \lambda_2)^2} \cdot \frac{(d - 1) \log N}{N}. \]
\[
\frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \left( \frac{d - 1}{\log N} \right)^{-1} \leq 2C_4 \delta^{-2} \exp \left( -2N \cdot \frac{\log N}{N} \right) \cdot \left( \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{\log N}{N} \right)^{-1} \\
+ C_4 \left( 1 + \left[ \frac{\lambda_1^2 (\lambda_1 - \lambda_2)^{-1} \tilde{B}(N)}{\lambda_2} \right]^{0.5 - 4\varepsilon} \cdot \frac{\lambda_1}{\lambda_2} \right) \\
\leq 2C_4 \delta^{-2} \left( \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \right)^{-1} \frac{1}{N \log N} \\
+ C_4 + C_4 \left[ \frac{\lambda_1^2 (\lambda_1 - \lambda_2)^{-1} \tilde{B}(N)}{\lambda_2} \right]^{0.5 - 4\varepsilon} \cdot \frac{\lambda_1}{\lambda_2}.
\]

Letting \( C^{(2)}(d, N, \delta) \) be the last line above concludes that as \( d, N \to \infty \) when Eq. (2.12) holds and \( \lambda_2/\lambda_1 \) being bounded away from 0, the term goes to \( C_4 = C_5 \). This completes our proof.

\[
\boxdot
\]

4 Summary

In this paper, we provide an explicit convergence rate analysis for the online PCA iteration with subgaussian samples. Our convergence rate result is nearly optimal in the sense that the output can be viewed as a running estimator that nearly attains the minimax information lower bound for subgaussian PCA. Furthermore, our convergence rate result is nearly global, in the sense that finite-sample error bound holds with high probability even if the initial solution is uniformly sampled from the unit sphere. One direction for future research is to develop and analyze online PCA method to estimate the top \( k \) eigenvectors. Another direction for research is to extend the analysis beyond PCA to a broader class of low-rank estimation problems.

Acknowledgements The authors would like to thank Ruoyu Sun and Zhaoran Wang for inspiring discussions on this work. The authors would also like to thank the two anonymous referees for their insightful comments.

References

1. Agarwal, A., Bartlett, P., Ravikumar, P., Wainwright, M.: Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. IEEE Trans. Inf. Theory 58(5), 3235–3249 (2012)
2. Amini, A., Wainwright, M.: High-dimensional analysis of semidefinite relaxations for sparse principal components. Ann. Stat. 37(5B), 2877–2921 (2009)
3. Arora, R., Cotter, A., Livescu, K., Srebro, N.: Stochastic optimization for PCA and PLS. In: 50th Annual Allerton Conference on Communication, Control, and Computing pp. 861–868 (2012)
4. Arora, R., Cotter, A., Srebro, N.: Stochastic optimization of PCA with capped msg. In: Advances in Neural Information Processing Systems, pp. 1815–1823 (2013)
5. Ball, K.: An elementary introduction to modern convex geometry. Flavors Geom. 31, 1–58 (1997)
6. Balsubramani, A., Dasgupta, S., Freund, Y.: The fast convergence of incremental PCA. In: Advances in Neural Information Processing Systems, pp. 3174–3182 (2013)
7. Benveniste, A., Métivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations. Springer, New York (2012)
8. Bertsekas, D.: Incremental proximal methods for large scale convex optimization. Math. Program. Ser. B 129, 163–195 (2011)
9. Bertsekas, D., Tsitsiklis, J.: Parallel and Distributed Computation: Numerical Methods. Athena Scientific, Belmont (1989)
10. Borkar, V.: Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press, Cambridge (2008)
11. Cai, T.T., Ma, Z., Wu, Y.: Sparse PCA: optimal rates and adaptive estimation. Ann. Stat. 41(6), 3074–3110 (2013)
12. d’Aspremont, A., Bach, F., El Ghaoui, L.: Optimal solutions for sparse principal component analysis. J. Mach. Learn. Res. 9, 1269–1294 (2008)
13. De Sa, C., Olukotun, K., Ré, C.: Global convergence of stochastic gradient descent for some non-convex matrix problems. In: Proceedings of The 32nd International Conference on Machine Learning, pp. 2332–2341 (2015)
14. Durrett, R.: Probability: Theory and Examples, 4th edn. Cambridge University Press, Cambridge (2010)
15. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence, vol. 282, 2nd edn. Wiley, Hoboken (2005)
16. Garber, D., Hazan, E.: Fast and simple PCA via convex optimization. (2015). arXiv preprint arXiv:1509.05647
17. Yang L, Braverman V, Zhao T, Wang M.: Dynamic factorization and partition of complex networks. arXiv:1705.07881
18. Hardt, M., Price, E.: The noisy power method: A meta algorithm with applications. In: Advances in Neural Information Processing Systems, pp. 2861–2869 (2014a)
19. Hardt, M., Price, E.: The Noisy Power Method: A Meta Algorithm with Applications. NIPS, pp. 2861–2869 (2014b)
20. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24(6), 417 (1933)
21. Jain, P., Jin, C., Kakade, S.M., Netrapalli, P., Sidford, A.: Matching matrix Bernstein with little memory: Near-optimal finite sample guarantees for Oja’s algorithm (2016). arXiv preprint arXiv:1602.06929
22. Johnstone, I.M., Lu, A.Y.: On consistency and sparsity for principal components analysis in high dimensions. J. Am. Stat. Assoc. 104(486), 682–693 (2009)
23. Kuczynski, J., Wozniakowski, H.: Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. SIAM J. Matrix Anal. Appl. 13(4), 1094–1122 (1992)
24. Kushner, H., Yin, G.: Stochastic Approximation and Recursive Algorithms and Applications. Springer, New York (2003)
25. Ma, Z.: Sparse principal component analysis and iterative thresholding. Ann. Stat. 41(2), 772–801 (2013)
26. Mitliagkas, I., Caramanis, C., Jain, P.: Memory limited, streaming PCA. In: Advances in Neural Information Processing Systems, pp. 2886–2894 (2013)
27. Muirhead, R.J.: Aspects of Multivariate Statistical Theory, vol. 197. Wiley, Hoboken (2005)
28. Musco, C., Musco, C.: Stronger approximate singular value decomposition via the block lanczos and power methods (2015). arXiv preprint arXiv:1504.05477
29. Nadler, B.: Finite sample approximation results for principal component analysis: a matrix perturbation approach. Ann. Stat. 41(2), 2791–2817 (2008)
30. Nedić, A.: Random algorithms for convex minimization problems. Math. Program. Ser. B 129, 225–253 (2011)
31. Nedić, A., Bertsekas, D.: Incremental subgradient methods for nondifferentiable optimization. SIAM J. Optim. 12, 109–138 (2001)
32. Nedić, A., Bertsekas, D., Borkar, V.: Distributed asynchronous incremental subgradient methods. Stud. Comput. Math. 8, 381–407 (2001)
33. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, Hoboken (1983)
34. Oja, E.: Simplified neuron model as a principal component analyzer. J. Math. Biol. 15(3), 267–273 (1982)
35. 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)
36. Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2(11), 559–572 (1901)
37. Rakhlin, A., Shamir, O., Sridharan, K.: Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Conference on Machine Learning, pp. 449–456 (2012)
38. Sa, C.D., Re, C., Olukotun, K.: Global convergence of stochastic gradient descent for some non-convex matrix problems. In: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 2332–2341 (2015)
39. Shamir, O.: Convergence of stochastic gradient descent for PCA (2015a). arXiv preprint arXiv:1509.09002
40. Shamir, O.: Fast stochastic algorithms for svd and PCA: Convergence properties and convexity (2015b). arXiv preprint arXiv:1507.08788
41. Shamir, O.: A stochastic PCA and svd algorithm with an exponential convergence rate. In: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 144–152 (2015c)
42. Shamir, O., Zhang, T.: Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In: Proceedings of The 30th International Conference on Machine Learning, pp. 71–79 (2013)
43. Van Der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, New York (1996)
44. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Compressed sensing. Cambridge University Press, pp. 210–268 (2012)
45. Vu, V.Q., Lei, J.: Minimax Rates of Estimation for Sparse PCA in High Dimensions. AISTATS, pp. 1278–1286 (2012)
46. Vu, V.Q., Lei, J.: Minimax sparse principal subspace estimation in high dimensions. Ann. Stat. 41(6), 2905–2947 (2013)
47. Wang, M., Bertsekas, D.: Incremental constraint projection methods for variational inequalities. Math. Program. Ser. A, 1–43 (2014a)
48. Wang, M., Bertsekas, D.P.: Stochastic first-order methods with random constraint projection. SIAM J. Optim. 26(1), 681–717 (2016)
49. Wang, M., Fang, X., Liu, H.: Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. Math. Prog. 161(1), 419–449 (2016)
50. Wang, Z., Lu, H., Liu, H.: Nonconvex statistical optimization: minimax-optimal sparse PCA in polynomial time. (2014b). arXiv preprint arXiv:1408.5352
51. Yuan, X.-T., Zhang, T.: Truncated power method for sparse eigenvalue problems. J. Mach. Learn. Res. 14(1), 899–925 (2013)
52. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101(476), 1418–1429 (2006)