58th Annual IEEE Symposium on Foundations of Computer Science

First Efficient Convergence for Streaming k-PCA: a Global, Gap-Free, and Near-Optimal Rate

Zeyuan Allen-Zhu
Microsoft Research
zeyuan@csail.mit.edu

Yuanzhi Li
Princeton University
yuanzhil@cs.princeton.edu

***Abstract*—We study streaming principal component analysis (PCA), that is to find, in $O(dk)$ space, the top k eigenvectors of a $d \times d$ hidden matrix $\mathbf{\Sigma}$ with online vectors drawn from covariance matrix $\mathbf{\Sigma}$.**

We provide *global* convergence for Oja's algorithm which is popularly used in practice but lacks theoretical understanding for $k > 1$. We also provide a modified variant Oja^{++} that runs *even faster* than Oja's. Our results match the information theoretic lower bound in terms of dependency on error, on eigengap, on rank k, and on dimension d, up to poly-log factors. In addition, our convergence rate can be made gap-free, that is proportional to the approximation error and independent of the eigengap.

In contrast, for general rank k, before our work (1) it was open to design any algorithm with efficient global convergence rate; and (2) it was open to design any algorithm with (even local) gap-free convergence rate in $O(dk)$ space.

***Keywords*-principal component analysis, streaming algorithm, online algorithm, global convergence, stochastic optimization, convergence, optimal algorithm, nonconvex optimization**

I. Introduction

Principle component analysis (PCA) is the problem of finding the subspace of largest variance in a dataset consisting of vectors, and is a fundamental tool used to analyze and visualize data in machine learning, computer vision, statistics, and operations research. In the big-data scenario, since it can be unrealistic to store the entire dataset, it is interesting and more challenging to study the streaming model (a.k.a. the stochastic online model) of PCA.

Suppose the data vectors $x \in \mathbb{R}^d$ are drawn i.i.d. from an unknown distribution with covariance matrix $\mathbf{\Sigma} = \mathbb{E}[xx^\top] \in \mathbb{R}^{d\times d}$, and the vectors are presented to the algorithm in an online fashion. Following (1; 2), we assume the Euclidean norm $\|x\|_2 \le 1$ with probability 1 (therefore $\mathbf{Tr}(\mathbf{\Sigma}) \le 1$) and we are interested in approximately computing the top k eigenvectors of $\mathbf{\Sigma}$. We are interested in algorithms with memory storage $O(dk)$, the same as the memory needed to store any k vectors in d dimensions. We call this the *streaming k-PCA problem*.

For streaming k-PCA, the popular and natural extension of Oja's algorithm originally designed for the $k = 1$ case works as follows. Beginning with a random Gaussian matrix $\mathbf{Q}_0 \in \mathbb{R}^{d\times k}$ (each entry i.i.d $\sim \mathcal{N}(0,1)$), it repeatedly applies

$$\text{rank-}k \text{ Oja's algorithm:} \quad \mathbf{Q}_t \leftarrow (\mathbf{I} + \eta_t x_t x_t^\top)\mathbf{Q}_{t-1}, \quad \mathbf{Q}_t = \mathsf{QR}(\mathbf{Q}_t) \tag{I.1}$$

where $\eta_t > 0$ is some learning rate that may depend on t, vector x_t is the random vector in iteration t, and $\mathsf{QR}(\mathbf{Q}_t)$ is the Gram-Schmidt decomposition that orthonormalizes the columns of $\mathbf{Q}_t$.

Although Oja's algorithm works reasonably well in practice, very limited theoretical results are known for its convergence in the $k > 1$ case. Even worse, little is known for *any* algorithm that solves streaming PCA in the $k > 1$. Specifically, there are three major challenges for this problem:

1) Provide an *efficient* convergence rate that only logarithmically depends on the dimension d.
2) Provide a *gap-free* convergence rate that is independent of the eigengap.
3) Provide a *global* convergence rate so the algorithm can start from a random initial point.

In the case of $k > 1$, to the best of our knowledge, only Shamir (7) successfully analyzed the original Oja's algorithm. His convergence result is only local and not gap-free.[1]

Other groups of researchers (2; 6; 8) studied a *block variant* of Oja's, that is to sample multiple vectors x in each round t, and then use their empirical covariance to replace the use of $x_t x_t^\top$. This algorithm is more

The full and future version of this paper can be found at https://arxiv.org/abs/1607.07837.

Most of the work was done when Z. Allen-Zhu was a research member at Princeton University and the Institute for Advanced Study.

We thank Jieming Mao for discussing our lower bound Theorem 6, and thank Dan Garber and Elad Hazan for useful conversations. Z. Allen-Zhu is partially supported by a Microsoft research award, no. 0518584, and an NSF grant, no. CCF-1412958.

[1] A local convergence rate means that the algorithm needs a warm start that is sufficiently close to the solution. However, the complexity to reach such a warm start is not clear.

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.51

	Paper	Global Convergence	Is It "Efficient"?	Local Convergence
$k=1$ gap-dependent	Shamir (3)	$\widetilde{O}\big(\frac{d}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ ♭	no	$\widetilde{O}\big(\frac{1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ ♭
	Sa et al. (4)	$\widetilde{O}\big(\frac{d}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ ♭	no	$\widetilde{O}\big(\frac{d}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ ♭
	Li et al. (5) [a]	$\widetilde{O}\big(\frac{d\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ ♭	no	$\widetilde{O}\big(\frac{d\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ ♭
	Jain et al. (1)	$\widetilde{O}\big(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$	yes	$\widetilde{O}\big(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$
	Theorem 1 (Oja)	$\widetilde{O}\big(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$	yes	$\widetilde{O}\big(\frac{\lambda_1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$
$k=1$ gap-free	Shamir (3) (Remark I.3)	$\widetilde{O}\big(\frac{d}{\rho^2}\cdot\frac{1}{\varepsilon^2}\big)$ ♭	no	$\widetilde{O}\big(\frac{1}{\rho^2}\cdot\frac{1}{\varepsilon^2}\big)$ ♭
	Theorem 2 (Oja)	$\widetilde{O}\big(\frac{\lambda_{1\sim(1+m)}}{\rho^2}\cdot\frac{1}{\varepsilon}\big)$	yes	$\widetilde{O}\big(\frac{\lambda_{1\sim(1+m)}}{\rho^2}\cdot\frac{1}{\varepsilon}\big)$
$k\geq 1$ gap-dependent	Hardt-Price (6) [b]	$\widetilde{O}\big(\frac{d\lambda_k}{\mathsf{gap}^3}\cdot\frac{1}{\varepsilon}\big)$ ♭	no	$\widetilde{O}\big(\frac{d\lambda_k}{\mathsf{gap}^3}\cdot\frac{1}{\varepsilon}\big)$ ♭
	Li et al. (2) [b]	$\widetilde{O}\big(\frac{k\lambda_k}{\mathsf{gap}^3}\cdot\big(kd+\frac{1}{\varepsilon}\big)\big)$ ♭	no	$\widetilde{O}\big(\frac{k\lambda_k}{\mathsf{gap}^3}\cdot\frac{1}{\varepsilon}\big)$ ♭
	Shamir (7)	unknown ♭	no	$O\big(\frac{1}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ ♭
	Balcan et al. (8) [b]	$\widetilde{O}\big(\frac{d(\lambda_{1\sim k})^2\lambda_k}{\mathsf{gap}^3}\cdot\frac{1}{\varepsilon}\big)$ ♭ (when $\lambda_{1\sim k}\geq k/d$) [c]	no	$\widetilde{O}\big(\frac{d(\lambda_{1\sim k})^2\lambda_k}{\mathsf{gap}^3}\cdot\frac{1}{\varepsilon}\big)$ ♭ (when $\lambda_{1\sim k}\geq k/d$)
	Theorem 1 (Oja)	$\widetilde{O}\big(\frac{\lambda_{1\sim k}}{\mathsf{gap}^2}\cdot\big(\frac{1}{\varepsilon}+k\big)\big)$	yes	$\widetilde{O}\big(\frac{\lambda_{1\sim k}}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$
	Theorem 4 (Oja^{++})	$\widetilde{O}\big(\frac{\lambda_{1\sim k}}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$	yes	$\widetilde{O}\big(\frac{\lambda_{1\sim k}}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$
	Theorem 6 (LB)	$\Omega\big(\frac{k\lambda_k}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon}\big)$ (lower bound)		
$k\geq 1$ gap-free	**Theorem 2** (Oja)	$\widetilde{O}\big(\frac{\min\{1,\,(\lambda_{1\sim k}+k\cdot\lambda_{(k+1)\sim(k+m)})\}}{\rho^2}\cdot k\big)$ $+\widetilde{O}\big(\frac{\lambda_{1\sim k+m}}{\rho^2}\cdot\frac{1}{\varepsilon}\big)$	yes	$\widetilde{O}\big(\frac{\lambda_{1\sim k+m}}{\rho^2}\cdot\frac{1}{\varepsilon}\big)$
	Theorem 5 (Oja^{++})	$\widetilde{O}\big(\frac{\lambda_{1\sim k+m}}{\rho^2}\cdot\frac{1}{\varepsilon}\big)$	yes	$\widetilde{O}\big(\frac{\lambda_{1\sim k+m}}{\rho^2}\cdot\frac{1}{\varepsilon}\big)$
	Theorem 6 (LB)	$\Omega\big(\frac{k\lambda_k}{\rho^2}\cdot\frac{1}{\varepsilon}\big)$ (lower bound)		

Table I: Comparison of known results. For $\mathsf{gap}\stackrel{\text{def}}{=}\lambda_k-\lambda_{k+1}$, every $\varepsilon\in(0,1)$ and $\rho\in(0,1)$:

- "gap-dependent convergence" means $\|\mathbf{Q}_T^\top\mathbf{Z}\|_F^2\leq\varepsilon$ where $\mathbf{Z}$ consists of the last $d-k$ eigenvectors.
- "gap-free convergence" means $\|\mathbf{Q}_T^\top\mathbf{W}\|_F^2\leq\varepsilon$ where $\mathbf{W}$ consists of all eigenvectors with eigenvalues $\leq\lambda_k-\rho$.
- a global convergence is "efficient" if it only (poly-)logarithmically depend on the dimension d.
- k is the target rank; in gap-free settings m be the largest index so that $\lambda_{k+m}>\lambda_k-\rho$.
- we denote by $\lambda_{a\sim b}\stackrel{\text{def}}{=}\sum_{i=a}^b\lambda_i$ in this table. Since $\|x\|_2\leq 1$ for each sample vector, we have
$$\mathsf{gap}\in[0,1/k],\quad \lambda_i\in[0,1/i],\quad k\mathsf{gap}\leq k\lambda_k\leq\lambda_{1\sim k}\leq\lambda_{1\sim k+m}\leq 1\ .$$
- we use ♭ to indicate the result is outperformed.
- some results in this table (both ours and prior work) depend on $\lambda_{1\sim k}$. In principle, this requires the algorithm to know a constant approximation of $\lambda_{1\sim k}$ upfront. In practice, however, since one always tunes the learning rate η (for any algorithm in the table), we do not need additional knowledge on $\lambda_{1\sim k}$.

[a] The result of (5) is in fact $\widetilde{O}(\frac{d\lambda_1^2}{\mathsf{gap}^2}\cdot\frac{1}{\varepsilon})$ by under a stronger 4-th moment assumption. It slows down at least by a factor $1/\lambda_1$ if the 4-th moment assumption is removed.

[b] These results give guarantees on spectral norm $\|\mathbf{Q}_T^\top\mathbf{W}\|_2^2$, so we increased them by a factor k for a fair comparison.

[c] If $\|x_t\|_2$ is always 1 then $\lambda_{1\sim k}\geq k/d$ always holds. Otherwise, even in the rare case of $\lambda_{1\sim k}<k/d$, their complexity becomes $\widetilde{O}(\frac{k^2\lambda_k}{d\cdot\mathsf{gap}^3})$ and is still worse than ours.

stable and easier to analyze, but only leads to suboptimal convergence.

We discuss them more formally below (and see Table I):

- Shamir (7) implicitly provided a *local* but efficient convergence result for Oja's algorithm,[2] which requires a very accurate starting matrix $\mathbf{Q}_0$: his theorem relies on $\mathbf{Q}_0$ being correlated with the top k eigenvectors by a correlation value at least $k-1/2$. If using random initialization, this event happens with probability at most $2^{-\Omega(d)}$.
- Hardt and Price (6) analyzed the block variant of Oja's,[3] and obtained a global convergence that linearly scales with the dimension d. Their result also

[2] The original method of Shamir (7) is an offline one. One can translate his result into a streaming setting and this requires a lot of extra work including the martingale techniques we introduce in this paper.

[3] They are in fact only able to output $2k$ vectors, guaranteed to approximately include the top k eigenvectors.

has a cubic dependency on the gap between the k-th and $(k+1)$-th eigenvalue which is not optimal. They raised an open question regarding how to provide any convergence result that is gap-free.
- Balcan et al. (8) analyzed the block variant of Oja's. Their results are also not efficient and cubically scale with the eigengap. In the gap-free setting, their algorithm runs in space more than $O(kd)$, and also outputs more than k vectors.[4] For such reason, we do not include their gap-free result in Table I, and shall discuss it more in the full version.
- Li et al. (2) also analyzed the block variant of Oja's. Their result also cubically scales with the eigengap, and their global convergence is not efficient.
- In practice, researchers observed that it is advantageous to choose the learning rate η_t to be high at the beginning, and then gradually decreasing (c.f. (9)). To the best of our knowledge, there is no theoretical support behind this learning rate scheme for general k.

In sum, it remains open before our work to obtain (1) any gap-free convergence rate in space $O(kd)$, (2) any global convergence rate that is efficient, or (3) any global convergence rate that has the optimal quadratic dependence on eigengap.

Over Sampling. Let us emphasize that it is often desirable to directly output a $d \times k$ matrix $\mathbf{Q}_T$. Some of the previous results, such as Hardt and Price (6), or the gap-free case of Balcan et al. (8), are only capable of finding an over-sampled matrix $d \times k'$ for some $k' > k$, with the guarantee that these k' columns approximately contain the top k eigenvectors of $\mathbf{\Sigma}$. However, it is not clear how to find "the best k vectors" out of this k'-dimensional subspace.

Special Case of $k = 1$. Jain (1) obtained the first convergence result that is both efficient and global for streaming 1-PCA. Shamir (3) obtained the first gap-free result for streaming 1-PCA, but his result is not efficient. Both these results are based on Oja's algorithm, and it remains open before our work to obtain a gap-free result that is also efficient even when $k = 1$.

Other Related Results. Mitliagkas et al. (10) obtained a streaming PCA result but in the restricted spiked covariance model. Balsubramani et al. (11) analyzed a modified variant of Oja's algorithm and needed an extra $O(d^5)$ factor in the complexity.

The offline problem of PCA (and SVD) can be solved via iterative algorithms that are based on variance-reduction techniques on top of stochastic gradient methods (7; 12) (see also (13; 14) for the $k = 1$ case); these methods do multiple passes on the input data so are not relevant to the streaming model. Offline PCA can also be solved via power method or block Krylov method (15), but since each iteration of these methods relies on one full pass on the dataset, they are not suitable for streaming setting either. Other offline problems and efficient algorithms relevant to PCA include canonical correlation analysis and generalized eigenvector decomposition (16–18).

Offline PCA is *significantly easier* to solve because one can (although non-trivially) reduce a general k-PCA problem to k times of 1-PCA using the techniques of (12). However, this is *not the case* in streaming PCA because one can lose a large polynomial factor in the sampling complexity.

A. Results on Oja's Algorithm

We denote by $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ the eigenvalues of $\mathbf{\Sigma}$, and it satisfies $\lambda_1 + \cdots + \lambda_d = \mathbf{Tr}(\mathbf{\Sigma}) \leq 1$. We present convergence results on Oja's algorithm that are *global, efficient and gap-free*.

Our first theorem works when there is a eigengap between λ_k and λ_{k+1}:

Theorem 1 (Oja, gap-dependent)**.** *Letting* $\mathsf{gap} \stackrel{\text{def}}{=} \lambda_k - \lambda_{k+1} \in \left(0, \frac{1}{k}\right]$ *and* $\Lambda \stackrel{\text{def}}{=} \sum_{i=1}^{k} \lambda_i \in \left(0, 1\right]$*, for every* $\varepsilon, p \in (0,1)$ *define learning rates*

$$T_0 = \widetilde{\Theta}\left(\frac{k\Lambda}{\mathsf{gap}^2 p^2}\right), \quad T_1 = \widetilde{\Theta}\left(\frac{\Lambda}{\mathsf{gap}^2}\right),$$

$$\eta_t = \begin{cases} \widetilde{\Theta}\left(\frac{1}{\mathsf{gap} \cdot T_0}\right) & 1 \leq t \leq T_0; \\ \widetilde{\Theta}\left(\frac{1}{\mathsf{gap} \cdot T_1}\right) & T_0 < t \leq T_0 + T_1; \text{ [5]} \\ \widetilde{\Theta}\left(\frac{1}{\mathsf{gap} \cdot (t - T_0)}\right) & t > T_0 + T_1. \end{cases}$$

Let $\mathbf{Z}$ *be the column orthonormal matrix consisting of all eigenvectors of* $\mathbf{\Sigma}$ *with values no more than* λ_{k+1}*. Then, the output* $\mathbf{Q}_T \in \mathbb{R}^{d \times k}$ *of Oja's algorithm satisfies with probability at least* $1 - p$*:*

$$\textit{for every}^{[6]} \quad T = T_0 + T_1 + \widetilde{\Theta}\left(\tfrac{T_1}{\varepsilon}\right)$$

it satisfies

$$\|\mathbf{Z}^\top \mathbf{Q}_T\|_F^2 \leq \varepsilon \ .$$

Above, $\widetilde{\Theta}$ *hides poly-log factors in* $\frac{1}{p}, \frac{1}{\mathsf{gap}}$ *and* d*.*

In other words, after a warm up phase of length T_0, we obtain a $\frac{\lambda_1 + \cdots + \lambda_k}{\mathsf{gap}^2} \cdot \frac{1}{T}$ convergence rate for the quantity $\|\mathbf{Z}^\top \mathbf{Q}_T\|_F^2$. We make several observations (see also Table I):

[4] They require space $O((k+m)d)$ where $k+m$ is the number of eigenvalues in the interval $[\lambda_k - \rho, 1]$ for some "virtual gap" parameter ρ. See our Theorem 2 for a definition. This may be as large as $O(d^2)$. Also, they output $k+m$ vectors which are only guaranteed to approximately "contain" the top k eigenvectors.

[5] The intermediate stage $[T_0, T_0 + T_1]$ is in fact unnecessary, but we add this phase to simplify proofs.

[6] Theorem also applies to every $T \geq T_0 + T_1 + \widetilde{\Omega}(T_1/\varepsilon)$ by making η_t poly-logarithmically dependent on T.

- In the $k = 1$ case, Theorem 1 matches the best known result of Jain et al. (1).
- In the $k > 1$ case, Theorem 1 gives the first efficient global convergence rate.
- In the $k > 1$ case, even in terms of local convergence rate, Theorem 1 is faster than the best known result of Shamir (7) by a factor $\lambda_1 + \cdots + \lambda_k \in (0, 1)$.

Remark I.1. The quantity $\|\mathbf{Z}^\top \mathbf{Q}_T\|_F^2$ captures the correlation between the resulting matrix $\mathbf{Q}_T \in \mathbb{R}^{d\times k}$ and the smallest $d - k$ eigenvectors of $\mathbf{\Sigma}$. It is a natural generalization of the sin-square quantity widely used in the $k = 1$ case, because if $k = 1$ then $\|\mathbf{Z}^\top \mathbf{Q}_T\|_F^2 = \sin^2(q, \nu_1)$ where q is the only column of $\mathbf{Q}$ and ν_1 is the leading eigenvector of $\mathbf{\Sigma}$.

Some literatures instead adopt the spectral-norm guarantee (i.e., bounds on $\|\mathbf{Z}^\top \mathbf{Q}_T\|_2^2$) as opposed to the Frobenius-norm one. The two guarantees are only up to a factor k away. We choose to prove Frobenius-norm results because: (1) it makes the analysis significantly simpler, and (2) k is usually small comparing to d, so if one can design an efficient (i.e., dimension free) convergence rate for the Frobenius norm that also implies an efficient convergence rate for the spectral norm.

Remark I.2. Our lower bound later (i.e. Theorem 6) implies, at least when λ_1 and λ_k are within a constant factor of each other, the local convergence rate in Theorem 1 is optimal up to log factors.

Gap-Free Streaming k-PCA. When the eigengap is small which is usually true in practice, it is desirable to obtain *gap-free* convergence (3; 15). We have the following theorem which answers the open question of Hardt and Price (6) regarding gap-free convergence rate for streaming k-PCA.

Theorem 2 (Oja, gap-free)**.** *For every $\rho, \varepsilon, p \in (0, 1)$, let $\lambda_1, \ldots, \lambda_m$ be all eigenvalues of $\mathbf{\Sigma}$ that are $> \lambda_k - \rho$, let $\Lambda_1 \stackrel{\text{def}}{=} \sum_{i=1}^{k} \lambda_i \in \big(0, 1\big]$, $\Lambda_2 \stackrel{\text{def}}{=} \sum_{j=k+1}^{k+m} \lambda_j \in \big(0, 1\big]$, define learning rates*

$$T_0 = \widetilde{\Theta}\left(\frac{k \cdot \min\{1,\ \Lambda_1 + \frac{k\Lambda_2}{p^2}\}}{\rho^2 \cdot p^2}\right)$$

$$T_1 = \widetilde{\Theta}\left(\frac{\Lambda_1 + \Lambda_2}{\rho^2}\right),$$

$$\eta_t = \begin{cases} \widetilde{\Theta}\big(\frac{1}{\rho \cdot T_0}\big) & t \le T_0; \\ \widetilde{\Theta}\big(\frac{1}{\rho \cdot T_1}\big) & t \in (T_0, T_0 + T_1]; \\ \widetilde{\Theta}\big(\frac{1}{\rho \cdot (t - T_0)}\big) & t > T_0 + T_1. \end{cases}$$

Let $\mathbf{W}$ be the column orthonormal matrix consisting of all eigenvectors of $\mathbf{\Sigma}$ with values no more than $\lambda_k - \rho$. Then, the output $\mathbf{Q}_T \in \mathbb{R}^{d\times k}$ of Oja's algorithm satisfies with prob. at least $1 - p$:

$$\text{for every}^{7} \quad T = T_0 + T_1 + \widetilde{\Theta}\left(\tfrac{T_1}{\varepsilon}\right) \quad \text{it satisfies} \quad \|\mathbf{W}^\top \mathbf{Q}_T\|_F^2 \le \varepsilon \ .$$

Above, $\widetilde{\Theta}$ hides poly-log factors in $\frac{1}{p}, \frac{1}{\rho}$ and d.

Note that the above theorem is a *double approximation*. The number of iterations depend both on ρ and ε, where ε is an upper bound on the correlation between $\mathbf{Q}_T$ and all eigenvectors in $\mathbf{W}$ (which depends on ρ). This is the first known gap-free result for the $k > 1$ case using $O(kd)$ space.

One may also be interested in single-approximation guarantees, such as the rayleigh-quotient guarantee. Note that a single-approximation guarantee by definition loses information about the ε-ρ tradeoff; furthermore, (good) single-approximation guarantees are not easy to obtain.[8]

We show in this paper the following theorem regarding the rayleigh-quotient guarantee:

Theorem 3 (Oja, rayleigh quotient, informal)**.** *There exist learning rate choices so that, for every $T = \widetilde{\Theta}\big(\frac{k}{\rho^2 \cdot p^2}\big)$, letting q_i be the i-th column of the output matrix $\mathbf{Q}_T$, then*

$$\mathbf{Pr}\left[\forall i \in [k], \quad q_i^\top \mathbf{\Sigma} q_i \ge \lambda_i - \widetilde{\Theta}(\rho)\right] \ge 1 - p \ .$$

Again, $\widetilde{\Theta}$ hides poly-log factors in $\frac{1}{p}, \frac{1}{\rho}$ and d.

Remark I.3. Before our work, the only gap-free result with space $O(kd)$ is Shamir (3) — but it is not efficient and only for $k = 1$. His result is in Rayleigh quotient but not double-approximation. If the initialization phase is ignored, Shamir's local convergence rate matches our *global* one in Theorem 3. However, if one translates his result into double approximation, the running time loses a factor ε. This is why in Table I Shamir's result is in terms of $1/\varepsilon^2$ as opposed to $1/\varepsilon$.

B. Results on Our New Oja^{++} *Algorithm*

Oja's algorithm has a slow initialization phase (which is also observed in practice (9)). For example, in the gap-dependent case, Oja's running time $\widetilde{O}\big(\frac{\lambda_1 + \cdots + \lambda_k}{\rho^2} \cdot \big(k + \frac{1}{\varepsilon}\big)\big)$ is dominated by its initialization when $\varepsilon > 1/k$. We propose in this paper a modified variant of Oja's that initializes *gradually*.

Our Oja^{++} Algorithm. At iteration 0, instead putting all the dk random Gaussians into $\mathbf{Q}_0$ like Oja's, our Oja^{++} only fills the first $k/2$ columns of $\mathbf{Q}_0$ with random Gaussians, and sets the remaining columns be zeros. It applies the same iterative rule as Oja's to go

[7] Theorem also applies to every $T \ge T_0 + \widetilde{\Omega}\big(T_0/\varepsilon\big)$ by making η_t poly-logarithmically dependent on T.

[8] Pointed out by (1), a direct translation from double approximation to a rayleigh-quotient type of convergence loses a factor on the approximation error. They raised it as an open question regarding how to design a direct proof without sacrificing this loss. Our next theorem answers this open question (at least in the gap-free case).

from $\mathbf{Q}_t$ to $\mathbf{Q}_{t+1}$, but after every T_0 iterations for some $T_0 \in \mathbb{N}^*$, it replaces the zeros in the next $k/4, k/8, \ldots$ columns with random Gaussians and continues.[9] This gradual initialization ends when all the k columns become nonzero, and the remaining algorithm of Oja^{++} works exactly the same as Oja's.

We provide pseudocode of Oja^{++} in the full version, and state below its main theorems:

Theorem 4 (Oja^{++}, gap-dependent, informal)**.** *Letting* $\mathsf{gap} \stackrel{\text{def}}{=} \lambda_k - \lambda_{k+1} \in \left(0, \frac{1}{k}\right]$*, our* Oja^{++} *outputs a column-orthonormal* $\mathbf{Q}_T \in \mathbb{R}^{d\times k}$ *with* $\|\mathbf{Z}^\top \mathbf{Q}_T\|_F^2 \leq \varepsilon$ *in* $T = \widetilde{\Theta}\left(\frac{\lambda_1+\cdots+\lambda_k}{\mathsf{gap}^2\varepsilon}\right)$ *iterations.*

Theorem 5 (Oja^{++}, gap-free, informal)**.** *Given* $\rho \in (0,1)$*, our* Oja^{++} *outputs a column-orthonormal* $\mathbf{Q}_T \in \mathbb{R}^{d\times k}$ *with* $\|\mathbf{W}^\top \mathbf{Q}_T\|_F^2 \leq \varepsilon$ *in* $T = \widetilde{\Theta}\left(\frac{\lambda_1+\cdots+\lambda_{k+m}}{\rho^2\varepsilon}\right)$ *iterations.*

C. Result on Lower Bound

We have the following information-theoretical lower bound for any (possibly offline) algorithm:

Theorem 6 (lower bound, informal)**.** *For every integer* $k \geq 1$*, integer* $m \geq 0$*, every* $0 < \rho < \lambda < 1/k$*, every (possibly randomized) algorithm* $\mathcal{A}$*, we can construct a distribution* μ *over unit vectors with* $\lambda_{k+m+1}(\mathbb{E}_\mu[xx^\top]) \leq \lambda - \rho$ *and* $\lambda_k(\mathbb{E}_\mu[xx^\top]) \geq \lambda$*. The output* $\mathbf{Q}_T$ *of* $\mathcal{A}$ *with samples* $x_1, ..., x_T$ *i.i.d. drawn from* μ *satisfies*

$$\mathop{\mathbb{E}}_{x_1,\ldots,x_T,\mathcal{A}}\left[\|\mathbf{W}^\top \mathbf{Q}_T\|_F^2\right] = \Omega\left(\frac{k\lambda}{\rho^2 \cdot T}\right) \quad .$$

($\mathbf{W}$ *consists of the last* $d-(k+m)$ *eigenvectors of* $\mathbb{E}_\mu[xx^\top]$*.)*

Our Theorem 6 (with $m = 0$ and $\rho = \mathsf{gap}$) implies that, in the gap-dependent setting, the global convergence rate of Oja^{++} is optimal up to log factors, at least when $\lambda_1 = O(\lambda_k)$. Our gap-free result does not match this lower bound. We explain in the full version that if one increases the space from $O(kd)$ to $O((k+m)d)$ in the gap-free case, our Oja^{++} can also match this lower bound.

[9] Zeros columns will remain zero according to the usage of Gram-Schmidt in Oja's algorithm.

REFERENCES

[1] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford, "Streaming PCA: Matching Matrix Bernstein and Near-Optimal Finite Sample Guarantees for Oja's Algorithm," in *COLT*, 2016.

[2] C.-L. Li, H.-T. Lin, and C.-J. Lu, "Rivalry of two families of algorithms for memory-restricted streaming pca," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016, pp. 473–481.

[3] O. Shamir, "Convergence of stochastic gradient descent for pca," in *ICML*, 2016.

[4] C. D. Sa, C. Re, and K. Olukotun, "Global convergence of stochastic gradient descent for some non-convex matrix problems," in *ICML*, 2015, pp. 2332–2341.

[5] C. J. Li, M. Wang, H. Liu, and T. Zhang, "Near-Optimal Stochastic Approximation for Online Principal Component Estimation," *ArXiv e-prints*, vol. abs/1603.05305, Mar. 2016.

[6] M. Hardt and E. Price, "The noisy power method: A meta algorithm with applications," in *NIPS*, 2014, pp. 2861–2869.

[7] O. Shamir, "Fast stochastic algorithms for svd and pca: Convergence properties and convexity," in *ICML*, 2016.

[8] M.-F. Balcan, S. S. Du, Y. Wang, and A. W. Yu, "An improved gap-dependency analysis of the noisy power method," in *COLT*, 2016, pp. 284–309.

[9] B. Xie, Y. Liang, and L. Song, "Scale up nonlinear component analysis with doubly stochastic gradients," in *NIPS*, 2015, pp. 2341–2349.

[10] I. Mitliagkas, C. Caramanis, and P. Jain, "Memory limited, streaming pca," in *NIPS*, 2013, pp. 2886–2894.

[11] A. Balsubramani, S. Dasgupta, and Y. Freund, "The fast convergence of incremental pca," in *NIPS*, 2013, pp. 3174–3182.

[12] Z. Allen-Zhu and Y. Li, "LazySVD: Even Faster SVD Decomposition Yet Without Agonizing Pain," in *NIPS*, 2016.

[13] D. Garber and E. Hazan, "Fast and simple PCA via convex optimization," *ArXiv e-prints*, Sep. 2015.

[14] D. Garber, E. Hazan, C. Jin, S. M. Kakade, C. Musco, P. Netrapalli, and A. Sidford, "Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation," in *ICML*, 2016.

[15] C. Musco and C. Musco, "Randomized block krylov methods for stronger and faster approximate singular value decomposition," in *NIPS*, 2015, pp. 1396–1404.

[16] R. Ge, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford, "Efficient Algorithms for Large-scale Generalized Eigenvector Computation and Canonical Correlation Analysis," in *ICML*, 2016.

[17] W. Wang, J. Wang, D. Garber, and N. Srebro, "Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis," in *NIPS*, 2016.

[18] Z. Allen-Zhu and Y. Li, "Doubly Accelerated Methods for Faster CCA and Generalized Eigendecomposition," in *Proceedings of the 34th International Conference on Machine Learning*, ser. ICML '17, 2017.

[19] M. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, 2015, ch. Chapter 2: Basic tail and concentration bounds.

[20] S. J. Szarek, "Condition numbers of random matrices," *Journal of Complexity*, vol. 7, no. 2, pp. 131–149, 1991.

[21] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.

[22] M. Rudelson and R. Vershynin, "Smallest singular value of a random rectangular matrix," *Communications on Pure and Applied Mathematics*, vol. 62, no. 12, pp. 1707–1739, 2009.