# Average Performance Analysis of the Stochastic Gradient Method for Online PCA

Stephane Chretien, Christophe Guyeux, Zhen-Wai Olivier Ho

# Average performance analysis of the stochastic gradient method for online PCA

Stéphane Chrétien[1], Christophe Guyeux[2], and Zhen-Wai Olivier Ho[3]

[1] National Physical Laboratory, Teddington, UK, stephane.chretien@npl.co.uk
https://sites.google.com/view/stephanechretien
[2] FEMTO-ST Institute, UMR 6174 CNRS, France
christophe.guyeux@univ-fcomte.fr
[3] LMB Université de Bourgogne Franche-Comté, 16 route de Gray, 25030, Besançon, France. zhen-wai-olivier@univ-fcomte.fr

**Abstract.** This paper studies the complexity of the stochastic gradient algorithm for PCA when the data are observed in a streaming setting. We also propose an online approach for selecting the learning rate. Simulation experiments confirm the practical relevance of the plain stochastic gradient approach and that drastic improvements can be achieved by learning the learning rate.

**Keywords:** Stochastic gradient · online PCA · non-convex optimisation· average case analysis.

## 1 Introduction

### 1.1 Background

Principal Component Analysis (PCA) is a paramount tool in an amazingly wide scope of applications. PCA belongs to the small list of algorithms which are extensively used in data science, medicine, finance, machine learning, etc. and the list is almost infinite. PCA is one of the basic blocks in Data Analytics. Computing singular/eigenvectors also appears key to discovering nonlinear embeddings of the data such as Laplacian eigenmaps [2].

In the era of Big Data, computing a set of singular vectors might turn to be a computationally difficult task to achieve. In practice the data matrix itself cannot be imported into the RAM and the data can only be accessed in small samples. In face of such hard memory management problems, Online Convex Optimisation often provides efficient alternatives to standard computations in machine learning [6,13,10]. On the other hand, computing eigen/singular vectors is not a convex optimisation problem. Instead, PCA can be seen as an optimisation problem over the sphere and as such, requires a different type of analysis. Online or stochastic versions of PCA have been extensively studied lately; see in particular the review [3]. On the theoretical side, [11] proposed a very clear analysis of the stochastic gradient algorithm for PCA which does not require information about the gap between successive eigenvalues. Better convergence

rates were subsequently obtained in [12], [7], [1] using more advanced algorithms. All these previous works rely on the assumption that the data arrive sequentially and are i.i.d., and their objective is to compute their common covariance matrix.

Our contribution explores a different set up. In the present work, we assume that the entrees of the covariance matrix are revealed one at a time in a sequential fashion. In such a set up, only some correlations between certain components of the data vectors, supposed to be chosen uniformly at random, are assumed to be available at each round, and not the data themselves. Therefore, our set up pertains to the activity around the important problem of Positive Semi-Definite matrix completion [5], [8], [9].

Our first main contribution is a mathematical proof that the method of [11] extends to the online matrix completion problem. Our theoretical findings also include a formula for the learning rate which can be optimised depending on the problem at hand. Practical optimisation of the learning rate is our second contribution. Our tuning algorithm is an adaptation of Freund and Shapire's online Hedge algorithm and is shown to provide substantial improvement of the practical convergence speed of the online gradient scheme for PCA.

### 1.2 Organisation of the paper

Our main results are presented in Section 2 where the algorithm is described and our main theorem is given. The proof of our main theorem is exposed in Section 3. Implementation and numerical experiments are given in Section 4. In particular, a simple method for choosing the learning rate is described in Section 4.1. The technical lemmæ which are used in the proof of Section 3 are gathered in Section A at the end of the paper.

## 2 Main results

### 2.1 Presentation of the problem and prior result

We use bold-faced letters to denote vectors, and capital letters to denote matrices unless specified otherwise. Given a matrix $A$, we denote by $A^\top$ its transpose matrix, $\|A\|$ its spectral norm and $\|A\|_{1\to 2} = \max \|\mathbf{A}_j\|_2$ the maximum $\ell_2$ norm of its column. For a vector $\mathbf{v}$, we denote by $\mathbf{v}^\top$ its transpose. Moreover $(\mathbf{e}_i)_i$ denote the canonical basis of $\mathbb{R}^d$. The optimisation problem can be written

$$\min_{\mathbf{w}:\|\mathbf{w}\|=1} -\mathbf{w}^\top A\mathbf{w}, \tag{1}$$

where $d > 1$ and $A$ is a symmetric positive semi-definite matrix supposed unknown. We suppose that we have access to a stream of i.i.d. matrices $A_t$ defined as

$$A_t = d^2\ A_{i_t,j_t}\ \mathbf{e}_{i_t}\mathbf{e}_{j_t}^\top \tag{2}$$

and $(i_t, j_t)$ is drawn uniformly at random from $\{1,\dots,n\}^2$. It is easily seen that $\mathbb{E}[A_t] = A$, therefore each matrix $A_t$ can be seen as a properly rescaled noiseless

random component of $A$. It can be readily seen that any leading eigenvector of $A$ is a solution of the optimisation problem.

### 2.2 The stochastic projected gradient algorithm

Given a symmetric matrix $A \in \mathbb{R}^{d\times d}$, the projected gradient algorithm writes

$$\mathbf{w}_{t+1} = (I + \eta A)\mathbf{w}_t / \|(I + \eta A)\mathbf{w}_t\|_2 \tag{3}$$

where $\eta$ is a step-size parameter and $\mathbf{w}_0$ is the initial estimate for a leading eigenvector of $A$. This algorithm correspond to initialising at $\mathbf{w}_0$ then make a gradient step at each iteration followed by a projection into the unit sphere. However, since $A$ is unknown, the stochastic gradient we will study in this paper is simply defined as

$$\mathbf{w}_{t+1} = (I + \eta A_t)\mathbf{w}_t / \|(I + \eta A_t)\mathbf{w}_t\|_2 \tag{4}$$

obtained by replacing $A$ with the random matrix $A_t$. Since the projection on the unit sphere is a rescaling operation which is commutative with respect to the matrix product, we can leave the projection operation to the end. That is, for our analysis, it is enough to consider the equivalent algorithm which only performs projection at the end:

- Initialise $\mathbf{w}_0$ on a unit sphere,
- Perform $T > 0$ stochastic gradient step : $\mathbf{w}_{t+1} = (I + \eta A_t)\mathbf{w}_t$
- Return $\mathbf{w}_T / \|\mathbf{w}_T\|_2$.

In [12], the stream of i.i.d. matrices $A_t$ are also assumed positive semidefinite. The main result in [12] is the following theorem.

**Theorem 1.** *Suppose that the matrices $(A_t)_{t\in\mathbb{N}}$ are positive semi-definite, real i.i.d for some leading eigenvector $\mathbf{v}$ of $A$, $\frac{1}{p} < \langle \mathbf{w}_0, \mathbf{v}\rangle^2$ for some $p > 0$ and that for some $b \geq 1$, both $\|A_t\|/\|A\|$ and $\|A_t - A\|/\|A\|$ are at most $b$ with probability 1. Then, after $T$ iterations of (4) with $\eta = \frac{1}{b\sqrt{pT}}$, then with probability at least $\frac{1}{cp}$, the return $\mathbf{w}_T$ satisfies*

$$1 - \frac{\mathbf{w}_T^\top A \mathbf{w}_T}{\|A\|} \leq c' \frac{\log(T) b\sqrt{p}}{\sqrt{T}}, \tag{5}$$

*where $c$ and $c'$ are positive constants.*

Note that our online positive semidefinite matrix completion framework is not compatible with the assumptions required for Theorem 2.2 to apply. In our problem, the matrices $A_t$ are not themself positive semidefinite.

### 2.3 Main theorem

Without loss of generality, we will throughout assume that $\|A\| = 1$. Our goal is to show that, for $\varepsilon > 0$, the vector $\mathbf{w}_T$ obtained after $T$ iterations of the stochastic gradient method, satisfies

$$1 - \mathbf{w}_T A \mathbf{w}_T \leq \varepsilon \tag{6}$$

in expectation for $T$ a sufficiently large integer and $\eta$ tuned accordingly. Since $\|\mathbf{w}_T\|_2 = 1$, this is equivalent to showing that

$$\mathbf{w}_T^\top ((1-\varepsilon)I - A)\mathbf{w}_T \leq 0. \tag{7}$$

The next theorem summarizes our main findings.

**Theorem 2.** *Let $\varepsilon > 0$ and assume that $0 < \frac{1}{p} < \langle \mathbf{w}_0, \mathbf{v} \rangle^2$ for a leading eigenvector $\mathbf{v}$ of $A$. Define*

$$V_T = \mathbf{w}_0^\top \prod_{i=T}^{1} (I + \eta A_i)^\top ((1-\epsilon)I - A) \prod_{i=1}^{T} (I + \eta A_i)\mathbf{w}_0. \tag{8}$$

*Then for $T$ satisfying*

$$T > \max\left( \frac{4p^2 d^2}{\varepsilon}, \frac{\log 4p\varepsilon^{-1}}{\log\left(1 + \frac{\varepsilon}{pd^2}\right)} \right), \tag{9}$$

*and $\eta = \frac{\varepsilon}{4pd^2}$, it holds that*

$$\mathbb{E}[V_T] \leq -\frac{\varepsilon}{4p}(1 + 2\eta)^T. \tag{10}$$

Since $V_T = \|\mathbf{w}_T\|_2^2 \mathbf{w}_T^\top ((1-\varepsilon)I - A)\mathbf{w}_T$, the theorem implies the desired result.

## 3 Proof of the Theorem 2

In this section, we prove our main result, namely Theorem 2. Define

$$\mathbf{B}_T = \prod_{i=T}^{1} (I + \eta A_i)^\top ((1-\epsilon)I - A) \prod_{i=1}^{T} (I + \eta A_i) \tag{11}$$

so that $V_T = \mathbf{w}_0^\top B_T \mathbf{w}_0$.

**Lemma 1.** *We have that*

$$\begin{aligned} \mathbb{E}[B_T] = \mathbb{E}[B_{T-1}] + \eta \left( A^\top \mathbb{E}[B_{T-1}] + \mathbb{E}[B_{T-1}]A \right) \\ + \eta^2 d^2 \mathrm{diag}\left( A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A \right). \end{aligned} \tag{12}$$

*Proof.* Expand the recurrence relationship and take the expectation. Finally use Lemma 2 to obtain the last term of the inequality.

Expanding the recurrence in Lemma 1, we have

$$\mathbb{E}[V_T] \leq \mathbf{w}_0^\top (I + 2\eta A)^\top ((1-\varepsilon)I - A)\mathbf{w}_0 + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\| \|\mathbf{w}_0\|_2^2. \tag{13}$$

where the last term was obtained by using inequality (28) and $\|A\|_{1\to 2} \leq 1$. Using an eigendecomposition of $A$ and $\|\mathbf{w}_0\|_2^2 = 1$ gives

$$\mathbb{E}[V_T] \leq \sum_{j=1}^{d} (1+2\eta s_j)^T (1-\varepsilon-s_j) w_{0,j}^2 + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\|. \tag{14}$$

where $s_1 > \cdots > s_d$ denote the eigenvalues of $A$ and $w_{0,j} = \langle \mathbf{w}_0, \mathbf{v}_j \rangle$ denotes the $j-th$ component of $\mathbf{w}_0$ in the basis of the eigenvectors of $A$. Since $s_1 = 1$, this inequality rewrites

$$\mathbb{E}[V_T] \leq -\varepsilon(1+2\eta)^T w_{0,1}^2 + \sum_{j=2}^{d} (1+2\eta s_j)^T (1-\varepsilon-s_j) w_{0,j}^2 + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\|. \tag{15}$$

In the remainder of the proof, we prove that the negative term $-\varepsilon(1+2\eta)^T w_{0,1}^2$ dominates the positive terms. The terms $w_{0,j}^2$ sum to $1 - w_{0,1}^2$. Therefore the sum $\sum_{j=2}^{d} (1+2\eta s_j)^T (1-\varepsilon-s_j) w_{0,j}^2$ is less than $\max_{s\in[0,1]} (1+2\eta s)^T (1-\varepsilon-s)$, which can be bounded from above using Lemma 7. Therefore, we get the following inequality

$$\mathbb{E}[V_T] \leq -\varepsilon(1+2\eta)^T w_{0,1}^2 + (1 + \frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}) + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\|. \tag{16}$$

Factoring out $(1+2\eta)^T$, the inequality now writes

$$\mathbb{E}[V_T] \leq (1+2\eta)^T \Big( -\varepsilon w_{0,1}^2 + \frac{1}{(1+2\eta)^T} + \frac{(1+2\eta(1-\varepsilon))^T}{(1+2\eta)^T \eta(T+1)} + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\| \Big) \tag{17}$$

For the sake of simplifying the analysis, we will use a uniform bound on the spectral norm of $\mathrm{diag}(\mathbb{E}[B_k])$. More precisely, Lemma 6 implies that

$$\begin{aligned}\|\mathrm{diag}(\mathbb{E}[B_k])\| &\leq 2\frac{\eta}{\eta d^2+1}\left(\frac{1}{1-\eta(\eta d^2+2)}-\frac{1}{1-\eta}\right)(1-\varepsilon)\\ &\quad+2\frac{\eta}{\eta d^2+1}\Big(\eta d^2\frac{1}{1-\eta(\eta d^2+2)}\\ &\qquad+\frac{1}{1-\eta}\Big)(2-\varepsilon)+\left(1+\frac{\eta^2 d^2}{1-\eta(\eta d^2+2)}\right)(1-\varepsilon) \qquad (18)\\ &\leq 2\frac{\eta}{\eta d^2+1}\left(\frac{1-\varepsilon+(2-\varepsilon)\eta d^2}{1-\eta(\eta d^2+2)}+\frac{1}{1-\eta}\right)+\left(1+\frac{\eta^2 d^2}{1-\eta(\eta d^2+2)}\right)(1-\varepsilon)\\ &\leq 2\frac{\eta}{\eta d^2+1}\frac{2-\varepsilon+(2-\varepsilon)\eta d^2}{1-\eta(\eta d^2+2)}+1+\frac{\eta^2 d^2}{1-\eta(\eta d^2+2)} \qquad (19)\end{aligned}$$

for all $k$. This simplifies into

$$\|\mathrm{diag}(\mathbb{E}[B_k])\| \leq 1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}. \qquad (20)$$

Thus we obtain

$$\begin{aligned}\mathbb{E}[V_T] &\leq (1+2\eta)^T\Big(-\varepsilon w_{0,1}^2+\frac{1}{(1+2\eta)^T}+\frac{(1+2\eta(1-\varepsilon))^T}{(1+2\eta)^T\eta(T+1)}\\ &\qquad+\eta^2 d^2\left(1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}\right)\sum_{i=1}^{T}(1+2\eta)^{-i}\Big) \qquad (21)\end{aligned}$$

Bounding $\sum_{i=1}^{T}(1+2\eta)^{-i}$ by its infinite series $\sum_{i=1}^{\infty}(1+2\eta)^{-i}=(2\eta)^{-1}$ yields

$$\begin{aligned}\mathbb{E}[V_T] &\leq (1+2\eta)^T\Bigg(-\varepsilon w_{0,1}^2+\frac{1}{(1+2\eta)^T}+\frac{(1+2\eta(1-\varepsilon))^T}{(1+2\eta)^T\eta(T+1)} \qquad (22)\\ &\qquad+\eta/2d^2\left(1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}\right)\Bigg). \qquad (23)\end{aligned}$$

We can show that, for well chosen values of $\eta$ and $T$, the term between parenthesis can be made to be less that $-\varepsilon/4p$. Taking for example $\eta=\frac{\varepsilon}{4Cpd^2}$ for some constant $C$ such that $\left(1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}\right)\leq 2$ and $T>\max(4p^2d^2C/\varepsilon,\log(4p\varepsilon^{-1})/\log(1+\varepsilon/(Cpd^2)))$ is consistent with the constraints. Notice further that for $\varepsilon$ sufficiently small, this can be simplified further by taking $C=1$. One of the benefits of using this approach over standard methods from the literature, is that it is all at the same time elementary, intuitive and it can easily be checked to enjoy the same theoretical guarantees as the original method devised in [4]. Full details will be provided in a longer version of the paper.

# 4 Implementation

## 4.1 Choosing the learning rate

In this section, we address the question of choosing the learning rate, i.e. the step-size $\eta$ in iterations (4). Tuning the learning rate is essential in practice as it is well known to have a huge impact on the convergence speed of the method. Our idea to tune the learning rate is as follows:

- Choose the tolerance $\epsilon \in (0,1)$, and the algorithm's parameters $R,\ K \in \mathbb{N}_*$, $\rho \in (0,1)$ and $\beta > 0$.
- *Burn-in period:*
  - For $\eta \in \{\rho^k\}_{k=1:K}$, run $R$ gradient iterations in parallel whose iterates are denoted by $\mathbf{w}_t^{(k,r)}$, $t = 1, \ldots, B$.
  - Define $\pi_0^{(k)} = 1/K$, $k = 1, \ldots, K$. For $t = 1, \ldots, B$, let

$$L_t^{(k)} = \frac{2}{R(R-2)} \sum_{r<r'=2,\ldots,R} \langle \mathbf{w}_t^{(k,r)}, \mathbf{w}_t^{(k,r')} \rangle, \tag{24}$$

    and for $k = 1, \ldots, K$, define $\pi_{t+1}^{(k)} = \pi_t^{(k)}\ \exp\left(\beta\ L_t^{(k)}\right)$.
  - Stop when $\max_{k=1,\ldots,K}\ L_t^{(k)} \geq 1 - 10\ \epsilon$.
- *After burn-in*:
  - Reset $R$ to 1 and $K$ to 1.
  - Normalise $\pi$.
  - At each step $t = B+1, \ldots$, choose the stepsize with probability $\pi_B$.
  - Stop when $L_t^{(1)} \geq 1 - \epsilon$.

Choosing the parameter $\beta$ is more robust than choosing the learning rate. Moreover, a reasonably effective value for $\beta$ is given by (see [4]):

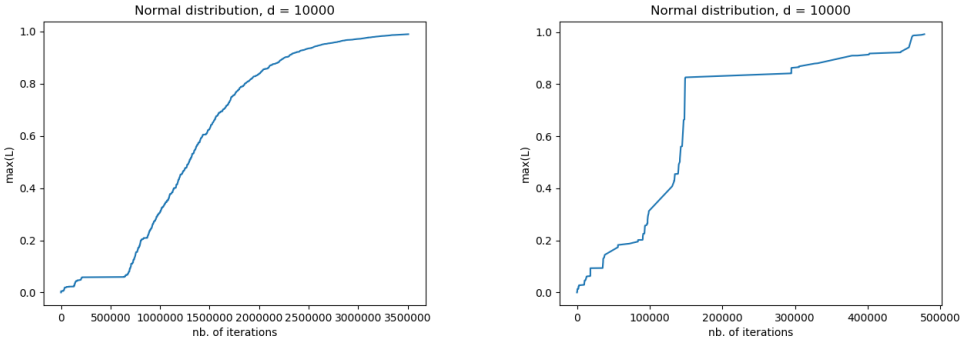
$$\beta = \sqrt{\frac{\log(K)}{B}}. \tag{25}$$

### 4.2 Numerical experiment

In this section, we present a simple numerical experiment which shows that

- The stochastic gradient method actually works in practice
- The adaptive selection of the learning rate/step-size described in the previous subsection actually accelerates the method's convergence drastically.

We run a simple experiment on a random i.i.d. Gaussian matrix of size $10000 \times 10000$. The convergence of $(L_t^{(1)})_{t\in\mathbb{N}}$ to 1 of the plain stochastic gradient method is shown in Figure 1a below. The accelerated version's convergence for the same experiment is shown in Figure 1b below. These results show that the method of the previous Section actually provides a substantial acceleration. We carefully checked that the selected learning rate is not equal to the smallest nor the largest value on the proposed grid of values between $2^{-3}$, $2^{-2}$, $\dots 2^{17}$. The observed gain in convergence speed was by a factor of 8.75. Extensive numerical experiment demonstrating this behaviour at larger scales will be included in an expanded version of this work.




(a) $K = 1$ (b) $K = 20$

Fig. 1: Convergence of $(L_t^{(1)})_{t\in\mathbb{N}}$ as a function of the iteration index: (a) is for the case of the arbitrary choice of learning rate equal to $2^-4$ and (b) shows the behaviour of the method using the learning procedure of Section 4.1 for values of the learning rate equal to $2^{-3}$, $2^{-2}$, $2^{-1}$, 1, 2, $\dots$, $2^{17}$.

## 5 Conclusion

In the present paper, we have studied the average behaviour of the stochastic gradient for the computation of the principal eigen-vector of positive semi-definite matrices, in the setting where the entrees are revealed one at a time. The analysis provides the first complexity analysis in this online setting. A preliminary

computer experiment integrating a novel learning rate optimisation procedure is included.

## A Technical lemmæ

Recall that

$$B_T = \prod_{t=T}^{1} (I + \eta A_t)^\top ((1-\varepsilon)I - A) \prod_{t=1}^{\top} (I + \eta A_t). \tag{26}$$

**Lemma 2.** *In the case of matrix completion, given a matrix $X$, we have*

$$\mathbb{E}[A_t^\top X A_t] = d^2 \ \mathrm{diag}(A \ \mathrm{diag}(X) A).$$

*Proof.* The resulting matrix writes

$$\begin{aligned} A_t^\top X A_t &= d^4 A_{ij} A_{ji} \mathbf{e}_{j_t} \mathbf{e}_{i_t}^\top X \mathbf{e}_{i_t} \mathbf{e}_{j_t}^\top \\ &= d^4 A_{ij} A_{ji} X_{ii} \mathbf{e}_{j_t} \mathbf{e}_{j_t}^\top. \end{aligned}$$

Therefore the expected matrix writes

$$\mathbb{E}[A_t^\top X A_t] = d^2 \sum_{i,j}^{d} A_{ij} A_{ji} X_{ii} \mathbf{e}_j \mathbf{e}_j^\top$$

Using the symmetry of $A$ gives the result.

Now our next goal is to see how $\mathrm{diag}\left(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A\right)$ evolves with the iterations. For this purpose, take the diagonal of (12), multiply from the left by $A^\top$ and from the right by $A$ and take the diagonal of the resulting expression.

**Lemma 3.** *We have that*

$$\|\mathrm{diag}\left(\mathbb{E}[B_T]\right)\| \leq 2\eta \ \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + (1+\eta^2 d^2) \ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\| \tag{27}$$

*Proof.* Expanding the recurrence relationship (12) gives

$$\begin{aligned} \mathrm{diag}(\mathbb{E}[B_T]) = \mathrm{diag}(\mathbb{E}[B_{T-1}]) &+ \eta \left(\mathrm{diag}\left(A^\top \mathbb{E}[B_{T-1}] + \mathbb{E}[B_{T-1}]A\right)\right) \\ &+ \eta^2 d^2 \mathrm{diag}\left(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A\right). \end{aligned}$$

For any diagonal matrix $\Delta$ and symmetric matrix $A$, we have

$$\|\mathrm{diag}(A^\top \Delta A)\| \leq \|A\|_{1\to 2}^2 \|\Delta\|. \tag{28}$$

Therefore, by taking the operator norm on both sides of the equality, we have

$$\|\mathrm{diag}(\mathbb{E}[B_T])\| \leq (1+\eta^2 d^2 \|A\|_{1\to 2}^2)\|\mathrm{diag}(\mathbb{E}[B_{T-1}])\| + 2\eta \|\mathrm{diag}(A^\top \mathbb{E}[B_{T-1}])\| \tag{29}$$

We conclude using $\|\mathrm{diag}(A^\top E[B_{T-1}])\| \leq \|A\|_{1\to 2}\|\mathbb{E}[B_{T-1}]\|_{1\to 2}$ and $\|A\|_{1\to 2} \leq 1$.

We also have to understand how the $\ell_{1\to 2}$ norm evolves.

**Lemma 4.** *We have*

$$\|\mathbb{E}[B_T]\|_{1\to 2} \leq \eta \ \|\mathbb{E}[B_{T-1}]\| + (1+\eta) \ \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + \eta^2 d^2 \ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\|. \tag{30}$$

*Proof.* Expanding the recurrence relationship gives

$$\begin{aligned}\|\mathbb{E}[B_T]\|_{1\to 2} = \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + \eta \ \left(\|A^\top \mathbb{E}[B_{T-1}]\|_{1\to 2} + \|\mathbb{E}[B_{T-1}]^\top A\|_{1\to 2}\right) \\ + \eta^2 d^2 \|\mathrm{diag}(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A)\|_{1\to 2}.\end{aligned}$$

For a diagonal matrix $\Delta$, we have $\|\Delta\|_{1\to 2} = \|\Delta\|$. This leads to

$$\begin{aligned}\|\mathbb{E}[B_T]\|_{1\to 2} = \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + \eta \ (\|A\|\|E[B_{T-1}]\|_{1\to 2} + \|\mathbb{E}[B_{T-1}]\|\|A\|_{1\to 2}) \\ + \eta^2 d^2 \|A\|_{1\to 2}^2 \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\|.\end{aligned}$$

Finally, using $\|A\|_{1\to 2} \leq 1$ concludes the proof.

We then have to understand how the operator norm of $\mathbb{E}[B_T]$ evolves

**Lemma 5.** *We have*

$$\|\mathbb{E}[B_T]\| \leq (1+2\eta)\|\mathbb{E}[B_{T-1}]\| + \eta^2 d^2 \ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\|. \tag{31}$$

*Proof.* Expanding the recurrence relationship (12) return

$$\|\mathbb{E}[B_T]\| = \mathbb{E}[B_{T-1}] + \eta(\|A^\top \mathbb{E}[B_{T-1}]\| + \|\mathbb{E}[B_{T-1}]A\|) + \eta^2 d^2 \|\mathrm{diag}(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A)\|.$$

Then using similar inequalities as in the proof of the lemmas above, we have the result.

**Lemma 6.** *Let $\|A\| = 1$, then we have*

$$\|\mathrm{diag}(\mathbb{E}[B_T])\| \leq \alpha \max_j (1-\varepsilon-s_j) + \beta \|(1-\varepsilon)I - A\|_{1\to 2} + \gamma \max_j (1-\varepsilon-A_{jj}) \tag{32}$$

*where*

$$\begin{aligned}\alpha &= 2\frac{\eta}{\eta d^2+1}\left(\frac{1-\eta^{T-2}(\eta d^2+2)^{T-2}}{1-\eta(\eta d^2+2)} - \frac{1-\eta^{T-2}}{1-\eta}\right) \\ \beta &= 2\frac{\eta}{\eta d^2+1}\left(\eta d^2 \frac{1-\eta^{T-2}(\eta d^2+2)^{T-2}}{1-\eta(\eta d^2+2)} + \frac{1-\eta^{T-2}}{1-\eta}\right) \\ \gamma &= 1 + \eta^2 d^2 \frac{1-\eta^{T-2}(\eta d^2+2)^{T-2}}{1-\eta(\eta d^2+2)}\end{aligned}$$

*Proof.* Expanding the recurrence and using equations (27), (30), and (31) yields the following system

$$\begin{bmatrix} \|\mathbb{E}[B_T]\| \\ \|\mathbb{E}[B_T]\|_{1\to 2} \\ \|\mathrm{diag}(\mathbb{E}[B_T])\| \end{bmatrix} \leq \left( I + \eta \begin{bmatrix} 2 & 0 & \eta d^2 \\ 1 & 1 & \eta d^2 \\ 0 & 2 & \eta d^2 \end{bmatrix} \right) \begin{bmatrix} \|\mathbb{E}[B_{T-1}]\| \\ \|\mathbb{E}[B_{T-1}]\|_{1\to 2} \\ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\| \end{bmatrix} \tag{33}$$

To obtain the result, we expand the inequality by recurrence. Therefore, we are interested in computing the $T$-th power of the matrix in inequality (33). We have

$$\left( I + \eta \begin{bmatrix} 2 & 0 & \eta d^2 \\ 1 & 1 & \eta d^2 \\ 0 & 2 & \eta d^2 \end{bmatrix} \right)^T = I + \sum_{i=1}^{T} \eta^i \begin{bmatrix} 2 & 0 & \eta d^2 \\ 1 & 1 & \eta d^2 \\ 0 & 2 & \eta d^2 \end{bmatrix}^i . \tag{34}$$

After computing the power matrices, it result that

$$\begin{aligned} \|\mathrm{diag}(\mathbb{E}[B_T])\| \leq & \sum_{i=1}^{T} \left( \eta^i \frac{2(\eta d^2 + 2)^{i-1} - 1}{\eta d^2 + 1} \right) \|\mathbb{E}[B_0]\| \\ & + \sum_{i=1}^{T} \left( \eta^i \frac{2\eta d^2 (\eta d^2 + 2)^{i-1} + 1}{\eta d^2 + 1} \right) \|\mathbb{E}[B_0]\|_{1\to 2} \\ & + \left( 1 + \eta^2 d^2 \sum_{i=1}^{T} (\eta^2 d^2 + 2\eta)^{i-1} \right) \|\mathrm{diag}(\mathbb{E}[B_0])\|. \end{aligned} \tag{35}$$

We conclude after computing the sums and bounding from above $\|\mathbb{E}[B_0]\|$ by $\max_j(1 - \varepsilon - s_j)$.

**Lemma 7.** *For $\eta < 1$ and $\varepsilon > 0$, we have*

$$\max_{s\in[0,1]} (1 + 2\eta\ s)^T (1 - \varepsilon - s) \leq 1 + \frac{(1 + 2\eta(1 - \varepsilon))^T}{\eta(T + 1)} \tag{36}$$

*Proof.* Denote $f(s) = (1 + 2\eta\ s)^T (1 - \varepsilon - s)$. Differentiating $f$ and setting to zero, we obtain

$$\begin{aligned} & 2\eta T (1 + 2\eta\ s)^{T-1} (1 - \varepsilon - s) - (1 + 2\eta\ s)^T = 0 \\ & \iff 2\eta T (1 - \varepsilon - s) - (1 + 2\eta\ s) = 0 \\ & \iff \frac{T(1 - \varepsilon) - 1/2\eta}{T + 1} = s \end{aligned}$$

Let $s_c = \frac{T - \varepsilon - 1/2\eta}{T+1}$ denote this critical point. Consider the two following cases :

- if $s_c \notin [0, 1]$, then $f$ has no critical point in the domain and therefore is maximised at either domain endpoint, i.e.

$$\max_{s\in[0,1]} f(s) = \max\{f(0) = 1 - \varepsilon, f(1) = -\varepsilon(1 + 2\eta)^T\} \leq 1$$

- if $s_c \in [0,1]$, then $f$ is maximised at $s_c$ and the value of $f$ at $s_c$ is

$$\left(1+2\eta\frac{T(1-\varepsilon)-1/2\eta}{T+1}\right)^T\left(1-\varepsilon-\frac{T(1-\varepsilon)-1/2\eta}{T+1}\right)$$
$$=\left(1+\frac{2\eta T(1-\varepsilon)-1}{T+1}\right)^T\left(\frac{1-\varepsilon+1/2\eta}{T+1}\right)$$
$$\leq (1+2\eta(1-\varepsilon))^T\left(\frac{1+1/2\eta}{T+1}\right)\leq\frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}.$$

This analysis proves that the maximum value $f$ can achieve is less than $\max\{1,\frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}\}\leq 1+\frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}\}$. Hence the result.

## References


1. Zeyuan Allen-Zhu and Yuanzhi Li, *Lazysvd: Even faster svd decomposition yet without agonizing pain*, Advances in Neural Information Processing Systems, 2016, pp. 974–982.
2. Afonso S Bandeira, *Ten lectures and forty-two open problems in the mathematics of data science*, 2015.
3. Hervé Cardot and David Degras, *Online principal component analysis in high dimension: Which algorithm to choose?*, arXiv preprint arXiv:1511.03688 (2015).
4. Yoav Freund and Robert E Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of computer and system sciences **55** (1997), no. 1, 119–139.
5. Robert Grone, Charles R Johnson, Eduardo M Sá, and Henry Wolkowicz, *Positive definite completions of partial hermitian matrices*, Linear algebra and its applications **58** (1984), 109–124.
6. Elad Hazan et al., *Introduction to online convex optimization*, Foundations and Trends® in Optimization **2** (2016), no. 3-4, 157–325.
7. Chi Jin, Sham M Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford, *Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation*, arXiv preprint arXiv:1510.08896 (2015).
8. Monique Laurent, *A tour dâĂŹhorizon on positive semidefinite and euclidean distance matrix completion problems*, Topics in Semidefinite and Interior-Point Methods **18** (1998), 51–76.
9. ———, *Matrix completion problemsmatrix completion problems*, Encyclopedia of Optimization, Springer, 2001, pp. 1311–1319.
10. Shai Shalev-Shwartz et al., *Online learning and online convex optimization*, Foundations and Trends® in Machine Learning **4** (2012), no. 2, 107–194.
11. Ohad Shamir, *A stochastic pca and svd algorithm with an exponential convergence rate.*, ICML, 2015, pp. 144–152.
12. ———, *Convergence of stochastic gradient descent for pca*, Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, JMLR.org, 2016, pp. 257–265.
13. Suvrit Sra, Sebastian Nowozin, and Stephen J Wright, *Optimization for machine learning*, Mit Press, 2012.