Convergence of Stochastic Gradient Descent for PCA

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science, Israel

Abstract

We consider the problem of principal component analysis (PCA) in a streaming stochastic setting, where our goal is to find a direction of approximate maximal variance, based on a stream of i.i.d. data points in $\mathbb{R}^d$. A simple and computationally cheap algorithm for this is stochastic gradient descent (SGD), which incrementally updates its estimate based on each new data point. However, due to the non-convex nature of the problem, analyzing its performance has been a challenge. In particular, existing guarantees rely on a non-trivial eigengap assumption on the covariance matrix, which is intuitively unnecessary. In this paper, we provide (to the best of our knowledge) the first eigengap-free convergence guarantees for SGD in the context of PCA. This also partially resolves an open problem posed in (Hardt & Price, 2014). Moreover, under an eigengap assumption, we show that the same techniques lead to new SGD convergence guarantees with better dependence on the eigengap.

1. Introduction

Principal component analysis (PCA) (Pearson, 1901; Hotelling, 1933) is a fundamental tool in data analysis and visualization, designed to find the subspace of largest variance in a given dataset (a set of points in Euclidean space). We focus on a simple stochastic setting, where the data $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathbb{R}^d$ is assumed to be drawn i.i.d. from an unknown underlying distribution, and our goal is to find a direction of approximately maximal variance. This can be written as the optimization problem

$$\min_{\mathbf{w}:\|\mathbf{w}\|=1} -\mathbf{w}^\top \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{w}, \tag{1}$$

or equivalently, finding an approximate leading eigenvector of the covariance matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

The conceptually simplest method for this task, given m sampled points $\mathbf{x}_1, \ldots, \mathbf{x}_m$, is to construct the empirical covariance matrix $\frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top$, and compute its leading eigenvector by an eigendecomposition. Based on concentration of measure arguments, it is not difficult to show that this would result in an $\mathcal{O}(\sqrt{1/m})$-optimal solution to Eq. (1). Unfortunately, the runtime of this method is $\mathcal{O}(md^2 + d^3)$. In large-scale applications, both m and d might be huge, and even forming the $d \times d$ covariance matrix, let alone performing an eigendecomposition, can be computationally prohibitive. A standard alternative to exact eigendecomposition is iterative methods, such as power iterations or the Lanczos method, which require performing multiple products of a vector with the empirical covariance matrix. Although this doesn't require computing and storing the matrix explicitly, it still requires multiple passes over the data, whose number may scale with eigengap parameters of the matrix or the target accuracy (Kuczynski & Wozniakowski, 1992; Musco & Musco, 2015). Recently, new randomized algorithms for this problem were able to significantly reduce the required number of passes, while maintaining the ability to compute high-accuracy solutions (Shamir, 2015b;a; Garber & Hazan, 2015; Jin et al., 2015).

In this work, we consider the efficacy of algorithms which perform a *single pass* over the data, and in particular, stochastic gradient descent (SGD). For solving Eq. (1), SGD corresponds to initializing at some unit vector $\mathbf{w}_0$, and then at each iteration t perform a stochastic gradient step with respect to $\mathbf{x}_t\mathbf{x}_t^\top$ (which is an unbiased estimate of $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$), followed by a projection to the unit sphere:

$$\mathbf{w}_t := (I + \eta\mathbf{x}_t\mathbf{x}_t^\top)\mathbf{w}_{t-1} \quad , \quad \mathbf{w}_t := \mathbf{w}_t/\|\mathbf{w}_t\|.$$

Here, η is a step size parameter. In the context of PCA, this is also known as Oja's method (Oja, 1982; Oja & Karhunen, 1985). The algorithm is highly efficient in terms of memory and runtime per iteration, requiring storage of a single d-dimensional vector, and performing only vector-vector and a vector-scalar products in each iteration.

In the world of convex stochastic optimization and learning, SGD has another remarkable property: Despite it being a simple, one-pass algorithm, it is essentially (worst-case) statistically optimal, attaining the same statistical estima-

tion error rate as exact empirical risk minimization (Bousquet & Bottou, 2008; Shalev-Shwartz et al., 2009; Shalev-Shwartz & Ben-David, 2014). Thus, it is quite natural to ask whether SGD also performs well for the PCA problem in Eq. (1), compared to statistically optimal but computationally heavier methods.

The study of SGD (or variants thereof) for PCA has gained interest in recent years, with some notable examples including (Arora et al., 2012; Balsubramani et al., 2013; Arora et al., 2013; Mitliagkas et al., 2013; Hardt & Price, 2014; De Sa et al., 2015; Jin et al., 2015; Jain et al., 2016). While experimentally SGD appears to perform quite well, its theoretical analysis has proven difficult, due to the non-convex nature of the objective function in Eq. (1). Remarkably, despite this non-convexity, finite-time convergence guarantees have been obtained under an eigengap assumption – namely, that the difference between the largest and 2nd-largest eigenvalues of $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ are separated by some fixed value $\lambda > 0$. For example, (De Sa et al., 2015) requires $\mathcal{O}(d/\lambda^2\epsilon)$ iterations to ensure with high probability that one of the iterates is ϵ-optimal. Very recently, (Jain et al., 2016) improved this to $\mathcal{O}(1/\lambda^2\epsilon)$ with constant probability. (Jin et al., 2015) requires $\mathcal{O}(1/\lambda^2 + 1/\lambda\epsilon)$ iterations, provided we begin close enough to an optimal solution.

Nevertheless, one may ask whether such eigengap dependencies are indeed necessary, if our goal is simply to find an approximately optimal solution of Eq. (1). Intuitively, if $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ has two equal (or near equal) top eigenvalues, then we may still expect to get a solution which lies close to the subspace of these two top eigenvalues, and approximately minimizes Eq. (1), with the runtime not dependent on any eigengap. Unfortunately, existing results tell us nothing about this regime, and not just for minor technical reasons: These results are based on tracking the geometric convergence of the SGD iterates $\mathbf{w}_t$ to a leading eigenvector of the covariance matrix. When there is no eigengap, there is also no single eigenvector to converge to, and such a geometric approach does not seem to work. Getting an eigengap-free analysis has also been posed as an open problem in (Hardt & Price, 2014). We note that while there are quite a few other single-pass, eigengap-free methods for this problem, such as (Warmuth & Kuzmin, 2006; 2008; Nie et al., 2013; Boutsidis et al., 2015; Garber et al., 2015; Kotłowski & Warmuth, 2015), their memory and runtime-per iteration requirements are much higher than SGD, often $\mathcal{O}(d^2)$ or worse.

In this work, we study the convergence of SGD for PCA, using a different technique than those employed in previous works, with the following main results:

- We provide the first (to the best of our knowledge) SGD convergence guarantee which does not pose an eigengap assumption. Roughly speaking, we prove that if the step size is chosen appropriately, then after T iterations starting from random initialization, with positive probability, SGD returns an $\tilde{\mathcal{O}}(\sqrt{p/T})$-optimal[1] solution of Eq. (1), where p is a parameter depending on how the algorithm is initialized:
  - If the algorithm is initialized from a warm-start point $\mathbf{w}_0$ such that $\frac{1}{\langle \mathbf{v}, \mathbf{w}_0 \rangle^2} \leq \mathcal{O}(1)$ for some leading eigenvector $\mathbf{v}$ of the covariance matrix, then $p = \mathcal{O}(1)$.
  - Under uniform random initialization on the unit Euclidean sphere, $p = \mathcal{O}(d)$, where d is the dimension.
  - Using a more sophisticated initialization (requiring the usage of the first $\mathcal{O}(d)$ iterations, but no warm-start point), $p = \tilde{\mathcal{O}}(n_A)$, where n_A is the *numerical rank* of the covariance matrix. The numerical rank is a relaxation of the standard notion of rank, is always at most d and can be considered a constant under some mild assumptions.
- In the scenario of a positive eigengap $\lambda > 0$, and using a similar proof technique, we prove an SGD convergence guarantee of $\mathcal{O}(p/\lambda T)$ (where p is as above) with positive probability. This guarantee is optimal in terms of dependence on T, λ, and in particular, has better dependence on λ compared to all previous works on SGD-like methods we are aware of ($1/\lambda$ as opposed to $1/\lambda^2$).

Unfortunately, a drawback of our guarantees is that they only hold with rather low probability: $\Omega(1/p)$, which can be small if p is large. Formally, this can be overcome by repeating the algorithm $\tilde{\mathcal{O}}(p)$ times, which ensures that with high probability, at least one of the outputs will be close to optimal. However, we suspect that these low probabilities are an artifact of our proof technique, and resolving it is left to future work.

2. Setting

We use bold-faced letters to denote vectors, and capital letters to denote matrices. Given a matrix M, we let $\|M\|$ denote its spectral norm, and $\|M\|_F$ its Frobenius norm.

We now present the formal problem setting, in a somewhat more general way than the PCA problem considered earlier. Specifically, we study the problem of solving

$$\min_{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| = 1} -\mathbf{w}^\top A \mathbf{w}, \tag{2}$$

where $d > 1$ and A is a positive semidefinite matrix, given access to a stream of i.i.d. positive semidefinite matrices $\tilde{A}_t$

[1] Throughout, we use $\mathcal{O}, \Omega$ to hide constants, and $\tilde{\mathcal{O}}, \tilde{\Omega}$ to hide constants and logarithmic factors.

where $\mathbb{E}[\tilde{A}_t] = A$ (e.g. $\mathbf{x}_t\mathbf{x}_t^\top$ in the PCA case). Notice that the gradient of Eq. (2) at a point $\mathbf{w}$ equals $2A\mathbf{w}$, with an unbiased stochastic estimate being $2\tilde{A}_t\mathbf{w}$. Therefore, applying SGD to Eq. (2) reduces to the following: Initialize at some unit-norm vector $\mathbf{w}_0$, and for $t = 1, \ldots, T$, perform $\mathbf{w}_t = (I + \eta\tilde{A}_t)\mathbf{w}_{t-1}, \mathbf{w}_t = \mathbf{w}_t/\|\mathbf{w}_t\|$, returning $\mathbf{w}_T$. In fact, for the purpose of the analysis, it is sufficient to consider a formally equivalent algorithm, which only performs the projection to the unit sphere at the end:

- Initialize by picking a unit norm vector $\mathbf{w}_0$
- For $t = 1, \ldots, T$, perform $\mathbf{w}_t = (I + \eta\tilde{A}_t)\mathbf{w}_{t-1}$
- Return $\frac{\mathbf{w}_T}{\|\mathbf{w}_T\|}$

It is easy to verify that the output of this algorithm is mathematically equivalent to the original SGD algorithm, since the stochastic gradient step amounts to multiplying $\mathbf{w}_{t-1}$ by a matrix independent of $\mathbf{w}_{t-1}$, and the projection just amounts to re-scaling. In both cases, we can write the algorithm's output in closed form as

$$\frac{\left(\prod_{t=T}^{1}(I + \eta\tilde{A}_t)\right)\mathbf{w}_0}{\left\|\left(\prod_{t=T}^{1}(I + \eta\tilde{A}_t)\right)\mathbf{w}_0\right\|}.$$

3. Convergence Without an Eigengap Assumption

Our main result is the following theorem, which analyzes the performance of SGD for solving Eq. (2).

Theorem 1. *Suppose that*

- *For some leading eigenvector* $\mathbf{v}$ *of* A, $\frac{1}{\langle \mathbf{v}, \mathbf{w}_0\rangle^2} \leq p$ *for some* p *(assumed to be* ≥ 8 *for simplicity).*
- *For some* $b \geq 1$, *both* $\frac{\|\tilde{A}_t\|}{\|A\|}$ *and* $\frac{\|\tilde{A}_t - A\|}{\|A\|}$ *are at most* b *with probability* 1.

If we run the algorithm above for T *iterations with* $\eta = \frac{1}{b\sqrt{pT}}$ *(assumed to be* ≤ 1*), then with probability at least* $\frac{1}{cp}$, *the returned* $\mathbf{w}$ *satisfies*

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

where c, c' *are positive numerical constants.*

The proof appears in Subsection 5.1. Note that this is a *multiplicative* guarantee on the suboptimality of Eq. (2), since we normalize by $\|A\|$, which is the largest magnitude Eq. (2) can attain. By multiplying both sides by $\|A\|$, we can convert this to an additive bound of the form

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

where b' is a bound on $\max\left\{\|\tilde{A}_t\|, \|\tilde{A}_t - A\|\right\}$. Also, note that the choice of η in the theorem is not crucial, and similar bounds (with different c, c') can be shown for other $\eta = \Theta(1/b\sqrt{pT})$.

The value of p in the theorem depends on how the initial point $\mathbf{w}_0$ is chosen. One possibility, of course, is if we can initialize the algorithm from a "warm-start" point $\mathbf{w}_0$ such that $\frac{1}{\langle \mathbf{v}, \mathbf{w}_0\rangle^2} \leq \mathcal{O}(1)$, in which case the bound in the theorem becomes $\mathcal{O}(\log(T)/\sqrt{T})$ with probability $\Omega(1)$. Such a $\mathbf{w}_0$ may be given by some other algorithm, or alternatively, if we are interested in analyzing SGD in the regime where it is close to one of the leading eigenvectors.

Of course, such an assumption is not always relevant, so let us turn to consider the performance without such a "warm-start". For example, the simplest and most common way to initialize $\mathbf{w}_0$ is by picking it uniformly at random from the unit sphere. In that case, for any $\mathbf{v}$, $\langle \mathbf{v}, \mathbf{w}_0\rangle^2 = \Theta(1/d)$ with high constant probability[2], so the theorem above applies with $p = \mathcal{O}(d)$:

Corollary 1. *If* $\mathbf{w}_0$ *is chosen uniformly at random from the unit sphere in* $\mathbb{R}^d$*, then Thm. 1 applies with* $p = \mathcal{O}(d)$*, and the returned* $\mathbf{w}$ *satisfies, with probability at least* $\Omega(1/d)$*,*

$$1 - \frac{\mathbf{w}^\top A\mathbf{w}}{\|A\|} \;\leq\; \mathcal{O}\left(\frac{\log(T)b\sqrt{d}}{\sqrt{T}}\right),$$

While providing some convergence guarantee, note that the probability of success is low, scaling down linearly with d. One way to formally solve this is to repeat the algorithm $\Omega(d)$ times, which ensures that with high probability, at least one output will succeed (and finding it can be done empirically by testing the outputs on a validation set). However, it turns out that by picking $\mathbf{w}_0$ in a smarter way, we can substantially improve those factors (at least in the analysis).

Specifically, we consider the following method, parameterized by number of iterations T_0, which are implemented before the main algorithm above:

- Sample $\mathbf{w}$ from a standard Gaussian distribution on $\mathbb{R}^d$
- Let $\mathbf{w}_0 = 0$.
- For $t = 1, \ldots, T_0$, let $\mathbf{w}_0 := \mathbf{w}_0 + \frac{1}{T_0}\tilde{A}_t\mathbf{w}$
- Return $\mathbf{w}_0 := \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|}$.

[2] One way to see this is by assuming w.l.o.g. that $\mathbf{v} = \mathbf{e}_1$ and noting that the distribution of $\mathbf{w}_0$ is the same as $\mathbf{w}/\|\mathbf{w}\|$ where $\mathbf{w}$ has a standard Gaussian distribution, hence $\langle \mathbf{v}, \mathbf{w}_0\rangle^2 = w_1^2/\sum_j w_j^2$, and by using standard concentration tools it can be shown that the numerator is $\Theta(1)$ and the denominator is $\Theta(d)$ with high probability.

Essentially, instead of initializing from a random point $\mathbf{w}$, we initialize from

$$\frac{\tilde{A}\mathbf{w}}{\|\tilde{A}\mathbf{w}\|} \quad , \quad \text{where} \quad \tilde{A} = \frac{1}{T_0}\sum_{t=1}^{T_0}\tilde{A}_t.$$

Since $\tilde{A}$ is a mean of T_0 random matrices with mean A, this amounts to performing a single approximate power iteration. Recently, it was shown that a single exact power iteration can improve the starting point of stochastic methods for PCA (Shamir, 2015a). The method above extends this idea to a purely streaming setting, where we only have access to stochastic approximations of A.

The improved theoretical properties of $\mathbf{w}_0$ with this initialization is formalized in the following lemma (where $\|A\|_F$ denotes the Frobenius norm of A):

Lemma 1. *The following holds for some numerical constants $c, c' > 0$: For $\mathbf{w}_0$ as defined above, if $T_0 \geq cdb^2\log(d)$, then with probability at least $\frac{7}{10} - \frac{2}{d} - \exp(-d/8)$,*

$$\frac{1}{\langle \mathbf{v}, \mathbf{w}_0\rangle^2} \leq c'\log(d)n_A,$$

where $n_A = \frac{\|A\|_F^2}{\|A\|^2}$ is the numerical rank of A.

The proof is provided in Subsection 5.2. Combining this with Thm. 1, we immediately get the following corollary:

Corollary 2. *If $\mathbf{w}_0$ is initialized as described above, then Thm. 1 applies with $p = \mathcal{O}(\log(d)n_A)$, and the returned $\mathbf{w}$ satisfies, with probability at least $\Omega(1/n_A\log(d))$,*

$$1 - \frac{\mathbf{w}^\top A\mathbf{w}}{\|A\|} \leq \mathcal{O}\left(\frac{\log(T)b\sqrt{\log(d)n_A}}{\sqrt{T}}\right),$$

The improvement of Corollary 2 compared to Corollary 1 depends on how much smaller is n_A, the numerical rank of A, compared to d. We argue that in most cases, n_A is much smaller, and often can be thought of as a moderate constant, in which case Corollary 2 provides an $\tilde{\mathcal{O}}\left(\frac{b}{\sqrt{T}}\right)$ error bound with probability $\tilde{\Omega}(1)$, at the cost of $\tilde{\mathcal{O}}(db^2)$ additional iterations at the beginning. Specifically:

- n_A is always in $[1, d]$, and in particular, can never be larger than d.
- n_A is always upper bounded by the rank of A, and is small even if A is only approximately low rank. For example, if the spectrum of A has polynomial decay $i^{-\alpha}$ where $\alpha > 1$, then n_A will be a constant independent of d. Moreover, to begin with, PCA is usually applied in situations where we hope A is close to being low rank.
- When $\tilde{A}_t$ is of rank 1 (which is the case, for instance, in PCA, where $\tilde{A}_t$ equals the outer product of the t-th datapoint $\mathbf{x}_t$), we have $n_A \leq b^2$, where we recall that b upper bounds the scaled spectral norm of $\tilde{A}_t$. In machine learning application, the data norm is often assumed to be bounded, hence b is not too large. To see why this holds, note that for rank 1 matrices, the spectral and Frobenius norms coincide, hence

$$n_A = \left(\frac{\|A\|_F}{\|A\|}\right)^2 = \left(\frac{\|\mathbb{E}[\tilde{A}_1]\|_F}{\|A\|}\right)^2$$
$$\leq \left(\mathbb{E}\left[\frac{\|\tilde{A}_1\|_F}{\|A\|}\right]\right)^2 = \left(\mathbb{E}\left[\frac{\|\tilde{A}_1\|}{\|A\|}\right]\right)^2 \leq b^2,$$

  where we used Jensen's inequality.

Similar to Corollary 1, we can also convert the bound of Corollary 2 into a high-probability bound, by repeating the algorithm $\tilde{\mathcal{O}}(n_A)$ times.

4. Convergence under an Eigengap Assumption

Although our main interest so far has been the convergence of SGD without any eigengap assumptions, we show in this section that our techniques also imply new bounds for PCA with an eigengap assumptions, which in certain aspects are stronger than what was previously known.

Specifically, we consider the same setting as before, but where the ratio $\frac{s_1 - s_2}{s_1}$, where s_1, s_2 are the leading singular values of the covariance matrix A is assumed to be strictly positive and lower bounded by some fixed $\lambda > 0$. Using this assumption and a proof largely similar to that of Thm. 1, we have the following theorem:

Theorem 2. *Under the same conditions as Thm. 1, suppose furthermore that*

- *The top two eigenvalues of A have a gap $\lambda\|A\| > 0$*
- $\frac{\log^2(T)b^2p}{\lambda T} \leq \frac{\log(T)b\sqrt{p}}{\sqrt{T}}$

If we run the algorithm above for $T > 1$ iterations with $\eta = \frac{\log(T)}{\lambda T}$ (assumed to be ≤ 1), then with probability at least $\frac{1}{cp}$, the returned $\mathbf{w}$ satisfies

$$1 - \frac{\mathbf{w}^\top A\mathbf{w}}{\|A\|} \leq c'\frac{\log^2(T)b^2p}{\lambda T},$$

where c, c' are positive numerical constants.

The proof (which uses techniques similar to the proof of Thm. 1) appears in the supplementary material. Considering first the technical conditions of the theorem, we note

that assuming $\frac{\log^2(T)b^2p}{\lambda T} \leq \frac{\log(T)b\sqrt{p}}{\sqrt{T}}$ simply amounts to saying that T is sufficiently large so that the $\mathcal{O}\left(\frac{\log^2(T)b^2p}{\lambda T}\right)$ bound provided by Thm. 2 is better than the $\mathcal{O}\left(\frac{\log(T)b\sqrt{p}}{\sqrt{T}}\right)$ bound provided by Thm. 1, by more than a constant. This is the interesting regime, since otherwise we might as well choose η as in Thm. 1 and get a better bound without any eigengap assumptions. Moreover, as in Thm. 1, a similar proof would hold if the step size is replaced by $c\log(T)/\lambda T$ for some constant $c \geq 1$.

As in Thm. 1, we note that p can be as large as d under random initialization, but this can be improved to the numerical rank of A using an approximate power iteration, or by analyzing the algorithm starting from a warm-start point $\mathbf{w}_0$ for which $\frac{1}{\langle \mathbf{v}, \mathbf{w}_0\rangle^2} \leq \mathcal{O}(1)$ for a leading eigenvector $\mathbf{v}$ of A. Also, note that under an eigengap assumption, if $1 - \frac{\mathbf{w}^\top A\mathbf{w}}{\|A\|}$ goes to 0 with the number of iterations T, it must hold that $\langle \mathbf{v}, \mathbf{w}\rangle^2$ goes to 1 for a leading eigenvector of A, so the analysis with $p = \mathcal{O}(1)$ is also relevant for analyzing SGD for sufficiently large T, once we're sufficiently close to the optimum.

Comparing the bound to previous bounds in the literature for SGD-like methods (which all assume an eigengap, e.g. (Balsubramani et al., 2013; Hardt & Price, 2014; De Sa et al., 2015; Jin et al., 2015)), an interesting difference is that the dependence on the eigengap λ is only $1/\lambda$, as opposed to $1/\lambda^2$ or worse. Intuitively, we are able to improve this dependence since we track the suboptimality directly, as opposed to tracking how $\mathbf{w}_T$ converges to a leading eigenvector, say in terms of the Euclidean norm. This has an interesting parallel in the analysis of SGD for λ-strongly convex functions, where the suboptimality of $\mathbf{w}_T$ decays as $\tilde{\mathcal{O}}(1/\lambda T)$, although $\mathbb{E}[\|\mathbf{w}_T - \mathbf{w}^*\|^2]$ can only be bounded by $\mathcal{O}(1/\lambda^2 T)$ (compare for instance Lemma 1 in (Rakhlin et al., 2012) and Theorem 1 in (Shamir & Zhang, 2013)). Quite recently, Jin et al. ((Jin et al., 2015)) proposed another streaming algorithm which does have only $1/\lambda$ dependence (at least for sufficiently large T), and a high probability convergence rate which is even asymptotically optimal in some cases. However, their formal analysis is from a warm-start point (which implies $p = \mathcal{O}(1)$ in our notation), whereas the analysis here applies to any starting point. Moreover, the algorithm in (Jin et al., 2015) is different and more complex, whereas our focus here is on the simple and practical SGD algorithm. Finally, we remark that although an $\mathcal{O}(1/\lambda T)$ convergence rate is generally optimal (using any algorithm), we do not know whether the dependence on b and p in the convergence bound of Thm. 2 for SGD is optimal, or whether it can be improved.

5. Proofs

5.1. Proof of Thm. 1

The proof is based on a combination of several lemmas, whose proofs are rather technical. Due to lack of space, these proofs are deferred to the supplementary material.

To simplify things, we will assume that we work in a coordinate system where A is diagonal, $A = \mathrm{diag}(s_1, \ldots, s_d)$, where $s_1 \geq s_2 \geq \ldots \geq s_d \geq 0$, and s_1 is the eigenvalue corresponding to $\mathbf{v}$. This is without loss of generality, since the algorithm and the theorem conditions are invariant to the choice of coordinate system. Moreover, since the objective function in the theorem is invariant to $\|A\|$, we shall assume that $\|A\| = s_1 = 1$. Under these assumptions, the theorem's conditions reduce to:

- $\frac{1}{w_{0,1}^2} \leq p$, for some $p \geq 8$
- $b \geq 1$ is an upper bound on $\|\tilde{A}_t\|, \|\tilde{A}_t - A\|$

Let $\epsilon \in (0, 1)$ be a parameter to be determined later. The proof works by lower bounding the probability of the objective function (which under the assumption $\|A\| = 1$, equals $1 - \mathbf{w}^\top A\mathbf{w}$) being suboptimal by at most ϵ. This can be written as

$$\Pr\left(\frac{\mathbf{w}_T^\top(I - A)\mathbf{w}_T}{\|\mathbf{w}_T\|^2} \leq \epsilon\right),$$

or equivalently,

$$\Pr\left(\mathbf{w}_T^\top((1-\epsilon)I - A)\mathbf{w}_T \leq 0\right).$$

Letting

$$V_T = \mathbf{w}_T^\top((1-\epsilon)I - A)\mathbf{w}_T,$$

we need to lower bound $\Pr(V_T \leq 0)$.

In analyzing the convergence of stochastic gradient descent, a standard technique to bound such probabilities is via a martingale analysis, showing that after every iteration, the objective function decreases by a certain amount. Unfortunately, due to the non-convexity of the objective function here, the amount of decrease at iteration t critically depends on the current iterate $\mathbf{w}_t$, and in the worst case may even be 0 (e.g. if $\mathbf{w}_t$ is orthogonal to the leading eigenvector, and there is no noise). Moreover, analyzing the evolution of $\mathbf{w}_t$ is difficult, especially without eigengap assumptions, where there isn't necessarily some fixed direction which $\mathbf{w}_t$ converges to. Hence, we are forced to take a more circuitous route.

In a nutshell, the proof is composed of three parts. First, we prove that if ϵ and the step size η are chosen appropriately, then $\mathbb{E}[V_T] \leq -\tilde{\Omega}\left((1+\eta)^{2T}\frac{\epsilon}{p}\right)$. If we could also prove

a concentration result, namely that V_T is not much larger than its expectation, this would imply that $\Pr(V_T \leq 0)$ is indeed large. Unfortunately, we do not know how to prove such concentration. However, it turns out that it is possible to prove that V_T is not much *smaller* than its expected value: More precisely, that $V_T \geq -\tilde{\mathcal{O}}\left((1+\eta)^{2T}\epsilon\right)$ with high probability. We then show that given such a high-probability *lower bound* on V_T, and a bound on its expectation, we can produce an *upper bound* on V_T which holds with probability $\tilde{\Omega}(1/p)$, hence leading to the result stated in the theorem.

We begin with a preliminary technical lemma:

Lemma 2. *For any $\epsilon, \eta \in (0,1)$, and integer $k \geq 0$,*

$$\max_{s\in[0,1]} (1+\eta s)^k (1-\epsilon-s) \;\leq\; 1 + 2\frac{(1+\eta(1-\epsilon))^k}{\eta(k+1)}.$$

Using this lemma, we now prove that $V_T = \mathbf{w}_T^\top((1-\epsilon)I - A)\mathbf{w}_T$ has a large negative expected value. To explain the intuition, note that if we could have used the exact A instead of the stochastic approximations $\tilde{A}_t$ in deriving $\mathbf{w}_T$, then we would have

$$\begin{aligned}
&\mathbf{w}_T^\top((1-\epsilon)I - A)\mathbf{w}_T \\
&\quad= \mathbf{w}_0^\top(I+\eta A)^T((1-\epsilon)I - A)(I+\eta A)^T\mathbf{w}_0 \\
&\quad= \sum_{j=1}^d (1+\eta s_j)^{2T}(1-\epsilon-s_j)w_{0,j}^2 \\
&\quad\leq \frac{1}{p}(1+\eta s_1)^{2T}(1-\epsilon-s_1) \\
&\qquad\quad + \sum_{j=2}^d (1+\eta s_j)^{2T}(1-\epsilon-s_j)w_{0,j}^2 \\
&\quad\leq \frac{1}{p}(1+\eta s_1)^{2T}(1-\epsilon-s_1) \\
&\qquad\quad + \left(\sum_{j=2}^d w_{0,j}^2\right)\max_{s\in[0,1]}(1+\eta s)^{2T}(1-\epsilon-s),
\end{aligned}$$

which by the assumptions $s_1 = 1$ and $1 = \|\mathbf{w}_0\|^2 = \sum_{j=1}^d w_{0,j}^2$ is at most

$$-\frac{\epsilon}{p}(1+\eta)^{2T} + \max_{s\in[0,1]}(1+\eta s)^{2T}(1-\epsilon-s).$$

Applying Lemma 2 and picking η, ϵ appropriately, it can be shown that the above is at most $-\Omega\left(\frac{\epsilon}{p}(1+\eta)^{2T}\right)$.

Unfortunately, this calculation doesn't apply in practice, since we use the stochastic approximations $\tilde{A}_t$ instead of A. However, using more involved calculations, we prove in the lemma below that the expectation is still essentially the same, provided ϵ, η are chosen appropriately.

Lemma 3. *If $\eta = \frac{1}{b}\sqrt{\frac{1}{pT}} \leq 1$ and $\epsilon = c\frac{\log(T)b\sqrt{p}}{\sqrt{T}} \leq 1$ for some sufficiently large constant c, then it holds that*

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

Having proved an upper bound on $\mathbb{E}[V_T]$, we now turn to prove a high-probability lower bound on V_T. The proof is based on relating V_T to $\|\mathbf{w}_T\|^2$, and then performing a rather straightforward martingale analysis of $\log(\|\mathbf{w}_T\|^2)$.

Lemma 4. *Suppose that $\tilde{A}_t$ is positive semidefinite for all t, and $\Pr(\|\tilde{A}_t\| \leq b) = 1$. Then for any $\delta \in (0,1)$, we have with probability at least $1-\delta$ that*

$$V_T > -\exp\left(\eta b\sqrt{T\log(1/\delta)} + (b^2+3)T\eta^2\right)(1+\eta)^{2T}\epsilon.$$

We now have most of the required components to prove Thm. 1. First, we showed in Lemma 3 that if $\eta = \frac{1}{b}\sqrt{\frac{1}{pT}}$, then

$$\mathbb{E}[V_T] \leq -(1+\eta)^{2T}\frac{\epsilon}{4p}. \tag{3}$$

for $\epsilon = \mathcal{O}(b\log(T)\sqrt{p/T})$. Using the same step size η, Lemma 4 implies that

$$\Pr\left(V_T \leq -\exp\left(\sqrt{\frac{\log(1/\delta)}{p}} + \frac{1+3/b^2}{p}\right)(1+\eta)^{2T}\epsilon\right)$$

is at most δ, and since we assume $b \geq 1$ (hence $1+3/b^2 \leq 4$), this implies that

$$\Pr\left(-\frac{V_T}{\exp(4/p)(1+\eta)^{2T}\epsilon} \geq \exp\left(\sqrt{\frac{\log(1/\delta)}{p}}\right)\right) \leq \delta. \tag{4}$$

Now, define the non-negative random variable

$$R_T = \max\left\{0, -\frac{V_T}{\exp(4/p)(1+\eta)^{2T}\epsilon}\right\},$$

and note that by its definition,

$$\mathbb{E}[R_T] \geq \mathbb{E}\left[-\frac{V_T}{\exp(4/p)(1+\eta)^{2T}\epsilon}\right]$$

and

$$\begin{aligned}
&\Pr\left(R_T \geq \exp\left(\sqrt{\frac{\log(1/\delta)}{p}}\right)\right) \\
&= \Pr\left(-\frac{V_T}{\exp(4/p)(1+\eta)^{2T}\epsilon} \geq \exp\left(\sqrt{\frac{\log(1/\delta)}{p}}\right)\right).
\end{aligned}$$

Using Eq. (3) and Eq. (4), this implies that

$$\mathbb{E}[R_T] \geq 1/\left(4p\exp(4/p)\right)$$

and

$$\Pr\left(R_T \geq \exp\left(\sqrt{\frac{\log(1/\delta)}{p}}\right)\right) \leq \delta.$$

To summarize the development so far, we defined a non-negative random variable R_T, which is bounded with high probability, yet its expectation is at least $\Omega(1/p)$. The following lemma shows that for a bounded non-negative random variable with "large" expectation, the probability of it being on the same order as its expectation cannot be too small:

Lemma 5. *Let X be a non-negative random variable such that for some $\alpha, \beta \in [0,1]$, we have $\mathbb{E}[X] \geq \alpha$, and for any $\delta \in (0,1]$,*

$$\Pr\left(X \geq \exp\left(\beta\sqrt{\log(1/\delta)}\right)\right) \leq \delta.$$

Then

$$\Pr\left(X > \frac{\alpha}{2}\right) \;\geq\; \frac{\alpha - \exp\left(-\frac{2}{\beta^2}\right)}{15}.$$

Before sketching the proof of the lemma, let us show to use it to prove Thm. 1. Applying it on the random variable R_T, which satisfies the lemma conditions with $\alpha = \frac{1}{4p\exp(4/p)}$, $\beta = \sqrt{\frac{1}{p}}$, we have

$$\begin{aligned}
&\frac{1}{15}\left(\frac{1}{4p\exp(4/p)} - \exp\left(-2p\right)\right) \\
&\leq \Pr\left(R_T > \frac{1}{8p\exp(4/p)}\right) \\
&= \Pr\left(\max\left\{0, \frac{-V_T}{\exp(4/p)(1+\eta)^{2T}\epsilon}\right\} > \frac{1}{8p\exp(4/p)}\right) \\
&= \Pr\left(-\frac{V_T}{\exp(4/p)(1+\eta)^{2T}\epsilon} > \frac{1}{8p\exp(4/p)}\right) \\
&= \Pr\left(V_T \leq -\frac{(1+\eta)^{2T}\epsilon}{8p}\right) \;\leq\; \Pr\left(V_T \leq 0\right)
\end{aligned}$$

The term $\frac{1}{15}\left(\frac{1}{4p\exp(4/p)} - \exp\left(-2p\right)\right)$ can be verified to be at least $\frac{1}{100p}$ for any $p \geq 8$, hence we obtained

$$\Pr(V_T \leq 0) \geq \frac{1}{100p}.$$

As discussed at the beginning of the proof, $V_T \leq 0$ implies that $\frac{\mathbf{w}_T(I-A)\mathbf{w}_T}{\|\mathbf{w}_T\|^2} \leq \epsilon$, where $\epsilon = c\frac{\log(T)b\sqrt{p}}{\sqrt{T}}$ is the value chosen in Lemma 3, and the theorem is established.

We now turn to sketch the proof of Lemma 5 (like all lemmas in this subsection, the formal proof appears in the supplementary material). To explain the intuition, suppose that X in the lemma was actually at most 1 with probability 1, rather than just bounded with high probability. Then we would have

$$\begin{aligned}
\alpha &\leq \mathbb{E}[X] \\
&= \Pr\left(X \geq \frac{\alpha}{2}\right)\mathbb{E}\left[X | X \geq \frac{\alpha}{2}\right] \\
&\quad + \Pr\left(X < \frac{\alpha}{2}\right)\mathbb{E}\left[X | X \leq \frac{\alpha}{2}\right] \\
&\leq \Pr\left(X \geq \frac{\alpha}{2}\right)\cdot 1 + \Pr\left(X < \frac{\alpha}{2}\right)\cdot\frac{\alpha}{2} \\
&= \Pr\left(X \geq \frac{\alpha}{2}\right) + \left(1 - \Pr\left(X \geq \frac{\alpha}{2}\right)\right)\frac{\alpha}{2},
\end{aligned}$$

which implies that $\alpha \leq \left(1-\frac{\alpha}{2}\right)\Pr\left(X \geq \frac{\alpha}{2}\right) + \frac{\alpha}{2}$, and therefore $\Pr\left(X \geq \frac{\alpha}{2}\right) \geq \frac{\alpha/2}{1-\alpha/2} \geq \frac{\alpha}{2}$. As a result, X is at least one-half its expectation lower bound (α) with probability at least $\alpha/2$. The proof of Lemma 5, presented below, follows the same intuition, but uses a more delicate analysis since X is actually only upper bounded with high probability.

5.2. Proof of Lemma 1

Define $\Delta = \|\tilde{A} - A\|$. Also, let $s_1 \geq s_2 \geq \ldots \geq s_d \geq 0$ be the d eigenvalues of A, with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$, where we assume that $\mathbf{v} = \mathbf{v}_1$. Using the facts $(x+y)^2 \leq 2x^2 + 2y^2$ and $\|\mathbf{v}_1\| = 1$, we have

$$\begin{aligned}
\frac{1}{\langle \mathbf{v}_1, \mathbf{w}_0\rangle^2} &= \frac{\|\tilde{A}\mathbf{w}\|^2}{\langle \mathbf{v}_1, \tilde{A}\mathbf{w}\rangle^2} \\
&= \frac{\|A\mathbf{w} + (\tilde{A}-A)\mathbf{w}\|^2}{\left(\langle \mathbf{v}_1, A\mathbf{w}\rangle + \langle \mathbf{v}_1, (\tilde{A}-A)\mathbf{w}\rangle\right)^2} \\
&\leq \frac{2\|A\mathbf{w}\|^2 + 2\|(\tilde{A}-A)\mathbf{w}\|^2}{\langle \mathbf{v}_1, A\mathbf{w}\rangle^2 + 2\langle \mathbf{v}_1, A\mathbf{w}\rangle\langle \mathbf{v}_1, (\tilde{A}-A)\mathbf{w}\rangle} \\
&\leq \frac{2\|A\mathbf{w}\|^2 + 2\|\mathbf{w}\|^2\Delta^2}{\langle \mathbf{v}_1, A\mathbf{w}\rangle^2 - 2|\langle \mathbf{v}_1, A\mathbf{w}\rangle|\|\mathbf{w}\|\Delta},
\end{aligned}$$

where we implicitly assume that Δ is sufficiently small for the denominator to be positive (eventually, we will pick T_0 large enough to ensure this).

Recall that $\mathbf{v}_1, \ldots, \mathbf{v}_d$ forms an orthonormal basis for $\mathbb{R}^d$, so $\mathbf{w} = \sum_{i=1}^d \mathbf{v}_i\langle \mathbf{v}_i, \mathbf{w}\rangle$. Therefore, we can write the above as

$$\begin{aligned}
&\frac{2\left(\sum_{i=1}^d s_i\mathbf{v}_i\langle \mathbf{v}_i, \mathbf{w}\rangle\right)^2 + 2\|\mathbf{w}\|^2\Delta^2}{\left(s_1\langle \mathbf{v}_1, \mathbf{w}\rangle\right)^2 - 2|s_1\langle \mathbf{v}_1, \mathbf{w}\rangle|\|\mathbf{w}\|\Delta} \\
&= \frac{2\sum_{i=1}^d s_i^2\langle \mathbf{v}_i, \mathbf{w}\rangle^2 + 2\|\mathbf{w}\|^2\Delta^2}{s_1^2\langle \mathbf{v}_1, \mathbf{w}\rangle^2 - 2|s_1\langle \mathbf{v}_1, \mathbf{w}\rangle|\|\mathbf{w}\|\Delta} \\
&\leq \frac{2\left(\sum_{i=1}^d s_i^2\right)\left(\max_i\langle \mathbf{v}_i, \mathbf{w}\rangle^2\right) + 2\|\mathbf{w}\|^2\Delta^2}{s_1^2\langle \mathbf{v}_1, \mathbf{w}\rangle^2 - 2|s_1\langle \mathbf{v}_1, \mathbf{w}\rangle|\|\mathbf{w}\|\Delta}.
\end{aligned}$$

To simplify notation, since $\mathbf{w}$ is drawn from a standard Gaussian distribution, which is rotationally invariant, we can assume without loss of generality that $(\mathbf{v}_1,\ldots,\mathbf{v}_d) = (\mathbf{e}_1,\ldots,\mathbf{e}_d)$, the standard basis, so the above reduces to

$$\frac{2\left(\sum_{i=1}^d s_i^2\right)\max_i w_i^2 + 2\|\mathbf{w}\|^2\Delta^2}{s_1^2 w_1^2 - 2|s_1 w_1|\|\mathbf{w}\|\Delta}.$$

Recall that $\Delta = \|\tilde{A} - A\|$, where $\tilde{A}$ is the average of T_0 independent random matrices with mean A, and spectral norm at most $\|A\|b$. Using a Hoeffding matrix bound (e.g. (Tropp, 2012)), and the fact that $\|A\| = s_1$, it follows that with probability at least $1-\delta$,

$$\Delta \leq \|A\|b\sqrt{\frac{8\log(d/\delta)}{T_0}} \;=\; s_1\sqrt{\frac{8b^2\log(d/\delta)}{T_0}}.$$

Plugging into the above, we get an upper bound of

$$\frac{2\left(\sum_{i=1}^d s_i^2\right)\max_i w_i^2 + \|\mathbf{w}\|^2 s_1^2\frac{16b^2\log(d/\delta)}{T_0}}{s_1^2 w_1^2 - 2s_1^2|w_1|\|\mathbf{w}\|\sqrt{\frac{8b^2\log(d/\delta)}{T_0}}},$$

holding with probability at least $1-\delta$. Dividing both numerator and denominator by s_1^2, and recalling that $n_A = \frac{\|A\|_F^2}{\|A\|^2} = \frac{\sum_{i=1}^d s_i^2}{s_1^2}$, the above equals

$$\begin{aligned}&\frac{2n_A\max_i w_i^2 + \|\mathbf{w}\|^2\frac{16b^2\log(d/\delta)}{T_0}}{w_1^2 - 2|w_1|\|\mathbf{w}\|\sqrt{\frac{8b^2\log(d/\delta)}{T_0}}}\\ &= \frac{2n_A\max_i w_i^2 + \|\mathbf{w}\|^2\frac{16b^2\log(d/\delta)}{T_0}}{|w_1|\left(|w_1| - 2\|\mathbf{w}\|\sqrt{\frac{8b^2\log(d/\delta)}{T_0}}\right)}. \qquad (5)\end{aligned}$$

Based on standard Gaussian concentration arguments, it holds that $\Pr\left(w_1^2 \leq \frac{1}{8}\right) \leq \frac{3}{10}$, $\Pr\left(\max_i w_i^2 \geq 18\log(d)\right) \leq \frac{1}{d}$, and $\Pr\left(\|\mathbf{w}\| \geq \sqrt{2d}\right) \leq \exp\left(-\frac{d}{8}\right)$ (see for instance the proof of Lemma 1 in (Shamir, 2015a), and Corollary 2.3 in (Barvinok, 2005)). Combining the above with a union bound, it holds that with probability at least $1 - \delta - \frac{3}{10} - \frac{1}{d} - \exp(-d/8)$, Eq. (5) is at most

$$\frac{36\log(d)n_A + \frac{32db^2\log(d/\delta)}{T_0}}{\frac{1}{8}\left(\frac{1}{8} - 2\sqrt{2}\sqrt{\frac{8db^2\log(d/\delta)}{T_0}}\right)}.$$

Recalling that this is an upper bound on $\frac{1}{\langle\mathbf{v}_1,\mathbf{w}_0\rangle^2}$, picking $\delta = 1/d$ for simplicity, and slightly simplifying, we showed that with probability at least $\frac{7}{10} - \frac{2}{d} - \exp(-d/8)$,

$$\frac{1}{\langle\mathbf{v}_1,\mathbf{w}_0\rangle^2} \;\leq\; \frac{36\log(d)n_A + \frac{64b^2\log(d)}{T_0}}{\frac{1}{8}\left(\frac{1}{8} - 8\sqrt{2}\sqrt{\frac{db^2\log(d)}{T_0}}\right)}.$$

Since $n_A \geq 1$, then by picking $T_0 \geq cdb^2\log(d)$ for a sufficiently large constant c, we get that $\frac{1}{\langle\mathbf{v}_1,\mathbf{w}_0\rangle^2} \leq c'\log(d)n_A$ for a numerical constant c', as required.

6. Discussion

In this paper, we studied the convergence of plain-vanilla stochastic gradient descent (SGD) for the PCA optimization problem (Eq. (1)) in $\mathbb{R}^d$. We proved an $\mathcal{O}(\sqrt{p/T})$ convergence rate after T iterations, with probability $\Omega(1/p)$, where p is a parameter depending on the quality of the initialization point. To the best of our knowledge, this is the first eigengap-free bound for SGD in the context of PCA. Moreover, the proof technique is quite different than standard SGD convergence proofs, and instead directly analyses the probability of convergence after a given number of iterations. In the case of a fixed eigengap $\lambda > 0$, the same techniques yield a new $\mathcal{O}(p/\lambda T)$ convergence bound, which has better dependence on λ compared to previous works. Although p can be as large as $\mathcal{O}(d)$ with random initialization, we showed how a more sophisticated initialization strategy can improve it to $\mathcal{O}(n_A)$ where n_A is the numerical rank of the covariance matrix.

The paper leaves open several questions. First, the probabilities at which the bounds above hold are rather small, and depend on the initialization point. We conjecture that it is possible to improve these to at least a universal constant (without the need to repeat the algorithm several times), possibly using a tighter large deviation analysis. A second question is whether SGD is worst-case optimal for PCA in a streaming setting, or whether there is an inherent price to pay for using such an $\mathcal{O}(d)$-memory method (e.g. the p factors in our bounds). In a very recent and elegant work, (Jain et al., 2016) showed that standard SGD is indeed essentially optimal, but for a different performance metric than ours, and under an eigengap assumption. Whether this can be extended to our setting remains to be seen. Finally, it would be interesting to extend the analysis here to block methods, where $k > 1$ eigenvectors are to be extracted simultaneously.

Acknowledgments: This research is supported in part by an FP7 Marie Curie CIG grant, the Intel ICRI-CI Institute, and Israel Science Foundation grant 425/13. We thank Ofer Zeitouni for several illuminating discussions.

References

Arora, R., Cotter, A., Livescu, K., and Srebro, N. Stochastic optimization for PCA and PLS. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

Arora, R., Cotter, A., and Srebro, N. Stochastic optimiza-

tion of PCA with capped MSG. In *NIPS*, 2013.

Balsubramani, A., Dasgupta, S., and Freund, Y. The fast convergence of incremental PCA. In *NIPS*, 2013.

Barvinok, A. Measure concentration lecture notes. http://www.math.lsa.umich.edu/~barvinok/total710.pdf, 2005.

Bousquet, Olivier and Bottou, Léon. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pp. 161–168, 2008.

Boutsidis, C., Garber, D., Karnin, Z., and Liberty, E. Online principal components analysis. In *SODA*, 2015.

De Sa, C., Olukotun, K., and Ré, C. Global convergence of stochastic gradient descent for some nonconvex matrix problems. In *ICML*, 2015.

Garber, D. and Hazan, E. Fast and simple pca via convex optimization. *arXiv preprint arXiv:1509.05647*, 2015.

Garber, D., Hazan, E., and Ma, T. Online learning of eigenvectors. In *ICML*, 2015.

Hardt, M. and Price, E. The noisy power method: A meta algorithm with applications. In *NIPS*, 2014.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Jain, Prateek, Jin, Chi, Kakade, Sham M, Netrapalli, Praneeth, and Sidford, Aaron. Matching matrix bernstein with little memory: Near-optimal finite sample guarantees for oja's algorithm. *arXiv preprint arXiv:1602.06929*, 2016.

Jin, C., Kakade, S., Musco, C., Netrapalli, P., and Sidford, A. Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation. *arXiv preprint arXiv:1510.08896*, 2015.

Kotłowski, W. and Warmuth, M. Pca with gaussian perturbations. *arXiv preprint arXiv:1506.04855*, 2015.

Kuczynski, J. and Wozniakowski, H. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13(4):1094–1122, 1992.

Mitliagkas, I., Caramanis, C., and Jain, P. Memory limited, streaming PCA. In *NIPS*, 2013.

Musco, C. and Musco, C. Stronger approximate singular value decomposition via the block lanczos and power methods. *arXiv preprint arXiv:1504.05477*, 2015.

Nie, J., Kotłowski, W., and Warmuth, M. Online pca with optimal regrets. In *Algorithmic Learning Theory*, pp. 98–112. Springer, 2013.

Oja, E. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

Oja, E. and Karhunen, J. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.

Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Stochastic convex optimization. In *COLT*, 2009.

Shamir, O. Fast stochastic algorithms for svd and pca: Convergence properties and convexity. *arXiv preprint arXiv:1507.08788*, 2015a.

Shamir, O. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *ICML*, 2015b.

Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*, 2013.

Tropp, J. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Warmuth, M. and Kuzmin, D. Online variance minimization. In *Learning theory*, pp. 514–528. Springer, 2006.

Warmuth, M. and Kuzmin, D. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9 (10):2287–2320, 2008.