Efficient Convex Relaxations for Streaming PCA

Teodor V. Marinov

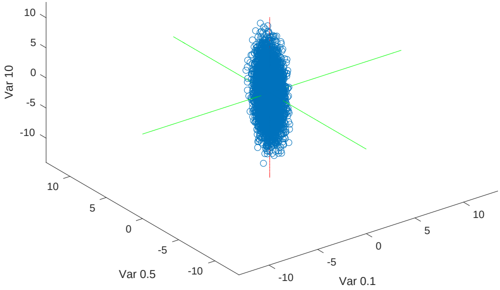
Johns Hopkins University

Joint work with Dr. Raman Arora
To appear at NeurIPS 2019

Outline

PCA as a geometric problem

- Given a data matrix $\mathrm{X} \in \mathbb{R}^{d \times T}$
- Output a $\mathrm{U} \in \mathbb{R}^{d \times k}, \mathrm{U}^\top \mathrm{U} = \mathrm{I}_k$ to minimize reconstruction error: $\frac{1}{T}\|\mathrm{X} - \mathrm{U}\mathrm{U}^\top \mathrm{X}\|_F^2$



PCA as a geometric problem

Optimization problem

$$\begin{aligned} \underset{\mathrm{U}\in\mathbb{R}^{d\times k}}{\text{minimize:}} \quad & \frac{1}{T}\|\mathrm{X}-\mathrm{U}\mathrm{U}^\top\mathrm{X}\|_F^2 \\ \text{subject to:} \quad & \mathrm{U}^\top\mathrm{U}=\mathrm{I}_k \end{aligned}$$

PCA as a geometric problem

Optimization problem

$$\begin{aligned} \underset{\mathrm{U}\in\mathbb{R}^{d\times k}}{\text{minimize:}} \quad & \frac{1}{T}\|\mathrm{X}-\mathrm{U}\mathrm{U}^\top\mathrm{X}\|_F^2 \\ \text{subject to:} \quad & \mathrm{U}^\top\mathrm{U}=\mathrm{I}_k \end{aligned}$$

Quickly rewriting the objective:

$$\begin{aligned} \frac{1}{T}\|\mathrm{X}-\mathrm{U}\mathrm{U}^\top\mathrm{X}\|_F^2 &= \frac{1}{T}\left\{\|\mathrm{X}\|_F^2+\|\mathrm{U}\mathrm{U}^\top\mathrm{X}\|_F^2-2\mathrm{Tr}\left(\mathrm{X}^\top\mathrm{U}\mathrm{U}^\top\mathrm{X}\right)\right\} \\ &= \frac{1}{T}\left\{\|\mathrm{X}\|_F^2+\mathrm{Tr}\left((\mathrm{U}\mathrm{U}^\top\mathrm{X})^\top\mathrm{U}\mathrm{U}^\top\mathrm{X}\right)-2\mathrm{Tr}\left(\mathrm{X}^\top\mathrm{U}\mathrm{U}^\top\mathrm{X}\right)\right\} \\ &= \frac{1}{T}\left\{\|\mathrm{X}\|_F^2+\mathrm{Tr}\left(\mathrm{X}^\top\mathrm{U}\mathrm{U}^\top\mathrm{U}\mathrm{U}^\top\mathrm{X}\right)-2\mathrm{Tr}\left(\mathrm{X}^\top\mathrm{U}\mathrm{U}^\top\mathrm{X}\right)\right\} \\ &= \frac{1}{T}\{\|\mathrm{X}\|_F^2-\mathrm{Tr}\left(\mathrm{U}^\top\mathrm{X}\mathrm{X}^\top\mathrm{U}\right)\} \end{aligned}$$

PCA as variance maximization

Equivalent optimization problem

$$\begin{aligned} \underset{\mathrm{U}\in\mathbb{R}^{d\times k}}{\text{maximize}} \quad & \frac{1}{T}\operatorname{Tr}\left(\mathrm{U}^\top \mathrm{X}\mathrm{X}^\top \mathrm{U}\right) \\ \text{subject to} \quad & \mathrm{U}^\top \mathrm{U} = \mathrm{I}_k \end{aligned}$$

PCA as variance maximization

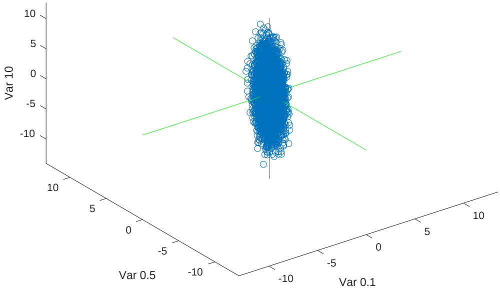
Equivalent optimization problem

$$\begin{aligned} \underset{\mathrm{U} \in \mathbb{R}^{d \times k}}{\text{maximize}} \quad & \frac{1}{T} \mathrm{Tr}\left(\mathrm{U}^\top \mathrm{X} \mathrm{X}^\top \mathrm{U}\right) \\ \text{subject to} \quad & \mathrm{U}^\top \mathrm{U} = \mathrm{I}_k \end{aligned}$$

Solution to optimization problem

Optimal solution is given by a set of k eigenvectors associated with the top k eigenvalues of $\frac{1}{T}\mathrm{X}\mathrm{X}^\top$

The stochastic optimization perspective



$$\text{Each } \mathrm{x}_t \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 10 \end{pmatrix}\right) \text{ independently.}$$

The stochastic optimization perspective

- Assume each column x_t of X is i.i.d as some probability law $\mathcal{D}$.
- Empirical problem is just a proxy for the stochastic optimization problem:

Stochastic optimization problem

$$\begin{aligned} \underset{\mathrm{U}\in\mathbb{R}^{d\times k}}{\text{minimize:}} \quad & \mathbb{E}_{\mathrm{x}\sim\mathcal{D}}\|\mathrm{x}-\mathrm{U}\mathrm{U}^\top\mathrm{x}\|_F^2 \\ \text{subject to:} \quad & \mathrm{U}^\top\mathrm{U}=\mathrm{I}_k \end{aligned}$$

The stochastic optimization perspective

Minimizing reconstruction

$$\begin{aligned} \underset{\mathrm{U}\in\mathbb{R}^{d\times k}}{\text{minimize:}} \quad & \mathbb{E}_{\mathrm{x}\sim\mathcal{D}}\|\mathrm{x}-\mathrm{U}\mathrm{U}^\top\mathrm{x}\|_F^2 \\ \text{subject to:} \quad & \mathrm{U}^\top\mathrm{U}=\mathrm{I}_k \end{aligned}$$

$$\Longleftrightarrow$$

Variance maximization

$$\begin{aligned} \underset{\mathrm{U}\in\mathbb{R}^{d\times k}}{\text{maximize}} \quad & \mathbb{E}_{\mathrm{x}\sim\mathcal{D}}\mathrm{Tr}\left(\mathrm{U}^\top\mathrm{x}\mathrm{x}^\top\mathrm{U}\right) \\ \text{subject to} \quad & \mathrm{U}^\top\mathrm{U}=\mathrm{I}_k \end{aligned} \tag{1}$$

Oja's algorithm

- One possible way to solve Problem 1 is to use Stochastic Gradient Descent (SGD)
- Gradient for objective in Problem 1 is $2\mathbb{E}_{\mathrm{x}\sim\mathcal{D}}[\mathrm{x}\mathrm{x}^\top]\mathrm{U}$
- Since we do not have direct access to distribution $\mathcal{D}$, we use an unbiased estimator of the gradient based on a sample $\mathrm{x}_t \sim \mathcal{D}$ given by $\mathrm{x}_t\mathrm{x}_t^\top\mathrm{U}$

Oja's algorithm

SGD on Problem 1 is also known as Stochastic Power Method (Oja's algorithm) Allen-Zhu and Li [2017]

$$\mathrm{U}_t \leftarrow (\mathrm{I} + \eta_t \mathrm{x}_t \mathrm{x}_t^\top)\mathrm{U}_{t-1}, \mathrm{U}_t = \mathsf{Orth}(\mathrm{U}_t); \tag{2}$$

Convergence guarantee (informal)

After T iterations of Oja's algorithm w.p. $1-\delta$ it holds that

$$\|\mathrm{U}_*^\top \mathrm{U}_T\|_F^2 \leq k - \tilde{O}\left(\frac{1}{\Delta(\mathrm{C})^2 T}\right),$$

where $\mathrm{C} = \mathbb{E}_{\mathrm{x}\sim\mathcal{D}}[\mathrm{x}\mathrm{x}^\top]$ and $\Delta(\mathrm{C})$ is the eigengap at the k-th eigenvalue.

$$\Delta(\mathrm{C}) := \lambda_k(\mathrm{C}) - \lambda_{k+1}(\mathrm{C})$$

Convex relaxations to PCA

Maximizing variance formulation

$$\begin{aligned} \underset{\mathrm{U}\in\mathbb{R}^{d\times k}}{\text{maximize}} \quad & \mathbb{E}_{\mathrm{x}\sim\mathcal{D}}\mathrm{Tr}\left(\mathrm{U}\mathrm{U}^\top \mathrm{x}\mathrm{x}^\top\right) \\ \text{subject to} \quad & \mathrm{U}^\top\mathrm{U} = \mathrm{I}_k \end{aligned}$$

The above is equivalent to:

Maximizing variance formulation

$$\begin{aligned} \underset{\mathrm{P}\in\mathbb{R}^{d\times d}}{\text{maximize}} \quad & \mathbb{E}_{\mathrm{x}\sim\mathcal{D}}\mathrm{Tr}\left(\mathrm{P}\mathrm{x}\mathrm{x}^\top\right) \\ \text{subject to} \quad & \mathrm{P}^2 = \mathrm{P}, \mathrm{P}^\top = \mathrm{P}, \mathsf{rank}(\mathrm{P}) = k \end{aligned}$$

Convex relaxations to PCA (continued)

The convex hull of $\{\mathrm{P} \in \mathbb{R}^{d \times d} : \mathrm{P}^2 = \mathrm{P}, \mathrm{P}^\top = \mathrm{P}, \operatorname{rank}(\mathrm{P}) = k\}$ is $\{\mathrm{P} \in \mathbb{R}^{d \times d} : \operatorname{Tr}(\mathrm{P}) \leq k, 0 \preceq \mathrm{P} \preceq \mathrm{I}, \mathrm{P}^\top = \mathrm{P}\}$

Convex relaxation [Arora et al., 2013]

$$\begin{aligned} \underset{\mathrm{P} \in \mathbb{R}^{d \times d}}{\text{maximize}} \quad & \operatorname{Tr}(\mathrm{PC}) \\ \text{subject to} \quad & \operatorname{Tr}(\mathrm{P}) \leq k, 0 \preceq \mathrm{P} \preceq \mathrm{I}, \mathrm{P}^\top = \mathrm{P} \end{aligned}, \tag{3}$$

Convex relaxation with regularization [Mianjy and Arora, 2018]

$$\begin{aligned} \underset{\mathrm{P} \in \mathbb{R}^{d \times d}}{\text{maximize}} \quad & \operatorname{Tr}(\mathrm{PC}) - \frac{\lambda}{2}\|\mathrm{P}\|_F^2 \\ \text{subject to} \quad & \operatorname{Tr}(\mathrm{P}) \leq k, 0 \preceq \mathrm{P} \preceq \mathrm{I}, \mathrm{P}^\top = \mathrm{P} \end{aligned}. \tag{4}$$

Convex relaxations to PCA (continued)

Projected SGD for Problem 3 (MSG)

$$\mathrm{P}_t \leftarrow \mathcal{P}\left(\mathrm{P}_{t-1} + \eta_t \mathrm{x}_t \mathrm{x}_t^\top\right)$$

Projected SGD for Problem 4 (RMSG)

$$\mathrm{P}_t \leftarrow \mathcal{P}\left((1 - \lambda\eta_t)\mathrm{P}_{t-1} + \eta_t \mathrm{x}_t \mathrm{x}_t^\top\right)$$

In the above $\mathcal{P}(\cdot)$ is the projection onto the convex set of constraints $\{\mathrm{P} : \mathrm{Tr}\,(\mathrm{P}) \leq k, 0 \preceq \mathrm{P} \preceq \mathrm{I}, \mathrm{P}^\top = \mathrm{P}\}$.

Convex relaxations to PCA (continued)

Running projected SGD on the above problems, comes with the following guarantees:

Convergence guarantee for MSG (informal)

After T iterations of MSG, it holds that

$$\mathbb{E}[\langle \mathrm{P}_* - \mathrm{P}_T, \mathrm{C}\rangle] \leq \tilde{O}\left(\frac{1}{\sqrt{T}}\right).$$

Convergence guarantee for RMSG (informal)

After T iterations of RMSG, it holds that

$$\mathbb{E}[\langle \mathrm{P}_* - \mathrm{P}_T, \mathrm{C}\rangle] \leq \tilde{O}\left(\frac{1}{\Delta(\mathrm{C})^2 T}\right).$$

Angle between subspaces and suboptimality in objective

- Oja's guarantee is of the form $k - \|\mathrm{U}_*^\top \mathrm{U}_T\|_F^2 \leq \epsilon$
- MSG and RMSG guarantees are of the form $\langle \mathrm{U}_*\mathrm{U}_*^\top - \mathrm{U}_T\mathrm{U}_T^\top, \mathrm{C}\rangle \leq \epsilon$
- We have [Mianjy and Arora, 2018]

$$\langle \mathrm{U}_*\mathrm{U}_*^\top - \mathrm{U}_T\mathrm{U}_T^\top, \mathrm{C}\rangle \leq \lambda_1(\mathrm{C})(k - \|\mathrm{U}_*^\top \mathrm{U}_T\|_F^2)$$

- No known relation in opposite direction

Per iteration complexity

- The computational complexity of Oja's algorithm per iteration is $O(dk^2)$ (can be reduced to $O(dk)$ if we do not call the $(Orth)$ procedure)
- The computational complexity of MSG and RMSG per iteration is $O(d\text{rank}(\mathrm{P}_t)^2)$
- Worst case for MSG and RMSG is $O(d^3)$

Per iteration complexity in practice

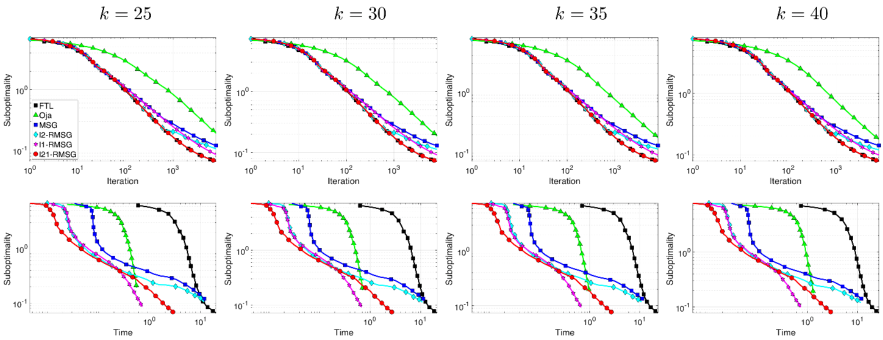


Figure 3: Experiment on MNIST by Mianjy and Arora [2018]

Controlling the rank of P_t

- If distribution is well-behaved rank of P_t stays in $O(k)$
- We get a similar computational complexity to Oja's algorithm but experiments suggest MSG and RMSG have better convergence properties
- Can we formalize the above statements through theoretical results?

Controlling the rank of P_t

- If distribution is well-behaved rank of P_t stays in $O(k)$
- We get a similar computational complexity to Oja's algorithm but experiments suggest MSG and RMSG have better convergence properties
- Can we formalize the above statements through theoretical results?

Yes, if we make slight tweaks to MSG and RMSG!

MB-RMSG

Input: Stream of data $\{\mathrm{x}_{t_n}\}$, parameters $\Delta(C)$, probability of failure δ, number of components k

Output: P_T

$$n = \log\left(d/\delta\right) \frac{128k\log(1/\delta)}{\Delta(\mathrm{C})^5}$$

$$\mathrm{P}_1 = \text{Top-k}\left(\frac{1}{n}\sum_{l=1}^{n} \mathrm{x}_{0_l}\mathrm{x}_{0_l}^{\top}\right)$$

%% $\{\mathrm{x}_{0_l}\}_{l=1}^{n}$ is the warm-start mini-batch

$$n = \log\left(\frac{Td}{\delta}\right)\frac{8(k+1)}{\Delta(\mathrm{C})^2}$$

for $t = 1, \ldots, T-1$ **do**

$$\eta_t = \frac{1}{\frac{\Delta(\mathrm{C})}{2}\left(t + \frac{128\log\left(\frac{1}{\delta}\right)}{\Delta(\mathrm{C})^3}\right)}$$

$$\mathrm{C}_t \leftarrow \frac{1}{n}\sum_{l=1}^{n} \mathrm{x}_{t_l}\mathrm{x}_{t_l}^{\top}$$

%% $\{\mathrm{x}_{t_l}\}_{l=1}^{n}$ is the mini-batch for the t^{th} epoch

$$\mathrm{P}_{t+1/2} \leftarrow \left(1 - \frac{\Delta(\mathrm{C})}{2}\eta_t\right)\mathrm{P}_t + \eta_t\mathrm{C}_t$$

$$\mathrm{P}_{t+1} = \mathcal{P}(\mathrm{P}_{t+1/2})$$

end for

Formal guarantee 1

Theorem

There exists an algorithm, solving Problem 4, which after T iterations, with probability at least $1 - 3e\delta$, returns a sequence of iterates $\{\mathrm{P}_t\}_{t=1}^{T}$, such that for all $t \leq T$

$$\langle \mathrm{P}^* - \mathrm{P}_t, \mathrm{C} \rangle \leq \frac{32 \log(1/\delta)}{\Delta(\mathrm{C})^2 \left(t + \frac{1}{\gamma} - 1\right)},$$

where $\gamma = \frac{\Delta(\mathrm{C})^3}{128 \log(1/\delta)}$. Further, for all $t \leq T$ it holds that P_t is a rank-k projection matrix and the per-iteration computational complexity of the algorithm is bounded by $\tilde{O}\left(\frac{dk^2}{\Delta(\mathrm{C})^2} + dk^2\right)$.

Formal guarantee 1 (continued)

- Total computational complexity for ϵ-suboptimality of MB-RMSG is $\tilde{O}(\frac{dk^2}{\Delta(\mathrm{C})^4\epsilon})$
- Total computational complexity for Oja's algorithm to reach ϵ-suboptimality is $\tilde{O}(\frac{dk}{\Delta(\mathrm{C})^2\epsilon})$

We can do better in expectation!

Formal guarantee 2

Theorem

Let $\mathcal{A}$ be the event that for all $t \in [T]$ it holds that $\|\mathrm{C}_t - \mathrm{C}\| \leq \frac{\Delta(C)}{8(k+1)}$ and P_t is a rank-k projection matrix. Then Algorithm 22 guarantees that $\mathcal{A}$ occurs with probability at least $1 - \delta$ and that

$$\mathbb{E}\left[\langle \mathrm{P}^* - \mathrm{P}_T, \mathrm{C}\rangle | \mathcal{A}\right] \leq \tilde{O}\left(\frac{\Delta(\mathrm{C})}{T} + \min(\Delta(\mathrm{C}) \times d, 1)\frac{1}{kT}\right).$$

Above theorem implies that the total computational complexity for achieving ϵ-suboptimality is $\tilde{O}\left(\frac{dk^2}{\epsilon\Delta(\mathrm{C})^2} \times \min(d\Delta(\mathrm{C})), 1)\right)$, which is only a factor of k away from Oja's algorithm whenever the gap is large, and actually improves by a factor of $1/\Delta(\mathrm{C})$ over Oja's in the case when $\Delta(\mathrm{C}) \in o(1/kd)$!

Key lemma

Lemma

Let P_t *be rank* k *and suppose* $\|\mathrm{C} - \mathrm{C}_t\| \leq \beta$. *Then, a sufficient condition for* P_{t+1} *to be rank* k *is*

$$\langle \mathrm{P}_t, \mathrm{C} \rangle \geq \langle \mathrm{P}^*, \mathrm{C} \rangle - \frac{\Delta(\mathrm{C})}{2} + \frac{\lambda}{2} + \beta(k+1). \tag{5}$$

- Projection works by shifting all eigenvalues of $\mathrm{P}_t + \eta_t \mathrm{C}_t$ and then clipping them between 0 and 1
- Let $\lambda_k(\mathrm{P}_t + \eta_t \mathrm{C}_t) = 1 + \lambda_k$ and $\lambda_{k+1}(\mathrm{P}_t + \eta_t \mathrm{C}_t) = \lambda_{k+1}$ for some λ_k and λ_{k+1}
- If $\lambda_k > \lambda_{k+1}$, then the shift from the projection is larger than λ_{k+1} and thus projection will clip λ_{k+1} to 0.
- We can guarantee that this happens with high probability if P_t is close enough to P^* and C_t is close enough to C.

Rest of analysis

- The rest of the analysis requires last iterate SGD guarantees with high probability
- For results in expectation we need to adapt the analysis for smooth SGD
- Both of these are non-trivial to do!

But does it work?

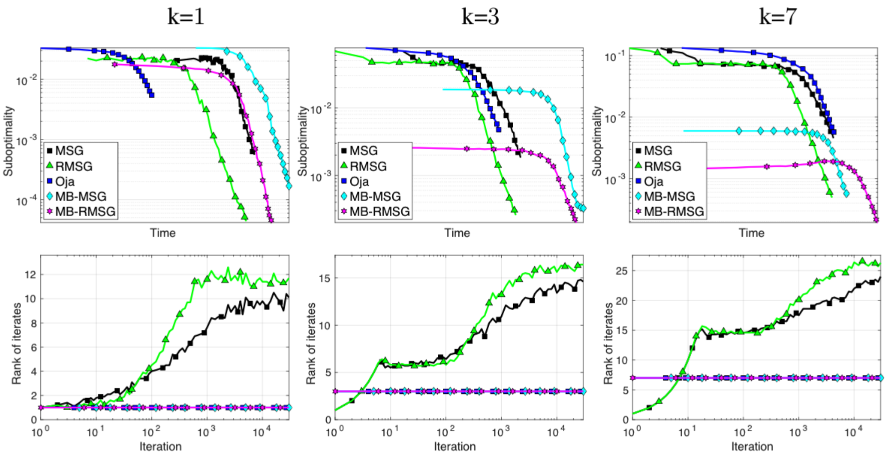


Figure 4: Experiments on MNIST

Open problems

- We gave an algorithm based on the convex relaxation to PCA which achieves (almost) optimal rates
- But goal was to study MSG and RMSG directly
- We need better tools to control rank of MSG and RMSG iterates

Other related work

- There are many other works on Streaming PCA, mainly focused on studying Oja's algorithm [De Sa et al., 2014, Hardt and Price, 2014, Balcan et al., 2016, Jain et al., 2016, Shamir, 2016a,b, Allen-Zhu and Li, 2017, Li et al., 2018]
- Other important work in the Online Learning setting is by Warmuth and Kuzmin [2008], Grabowska and Kotłowski [2018], Garber [2018]

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

Raman Arora, Andy Cotter, and Nati Srebro. Stochastic optimization of pca with capped msg. In *Advances in Neural Information Processing Systems*, pages 1815–1823, 2013.

Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309, 2016.

Christopher De Sa, Kunle Olukotun, and Christopher Ré. Global convergence of stochastic gradient descent for some non-convex matrix problems. *arXiv preprint arXiv:1411.1134*, 2014.

Dan Garber. On the regret minimization of nonconvex online gradient ascent for online pca. *arXiv preprint arXiv:1809.10491*, 2018.

Monika Grabowska and Wojciech Kotłowski. Online principal component analysis for evolving data streams. In *International Symposium on Computer and Information Sciences*, pages 130–137. Springer, 2018.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Conference on Learning Theory*, pages 1147–1164, 2016.

Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1):75–97, 2018.

Poorya Mianjy and Raman Arora. Stochastic pca with ℓ_2 and ℓ_1 regularization. In *International Conference on Machine Learning*, pages 3531–3539, 2018.

Ohad Shamir. Convergence of stochastic gradient descent for pca. In *International Conference on Machine Learning*, pages 257–265, 2016a.

Ohad Shamir. Fast stochastic algorithms for svd and pca: Convergence properties and convexity. In *International Conference on Machine Learning*, pages 248–256, 2016b.

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