Near-Optimal Streaming PCA: Matching Matrix Bernstein with Oja's Algorithm

Prateek Jain, Chi Jin, Sham M. Kakade,
Praneeth Netrapalli, Aaron Sidford

Microsoft Research, UC Berkeley, Univ of Washington

Problem Definition: Streaming PCA

- Given: $x_1, \ldots, x_n \in \mathbb{R}^d$
- $x_i \sim \mathcal{D}(\mathbb{R}^d)$ i.i.d.
- $\mathbb{E}\left[x_i x_i^\top\right] = \Sigma$

Task: Find top eigenvector v^* of Σ

Formally: Find $\hat{v}$ to $\min_{\hat{v}} \theta(\hat{v}, v^*)^2$

Streaming: Cannot store n vectors

Problem Definition: Streaming PCA

- Given: $x_1, \ldots, x_n \in \mathbb{R}^d$
- $x_i \sim \mathcal{D}(\mathbb{R}^d)$ i.i.d.
- $\mathbb{E}\left[x_i x_i^\top\right] = \Sigma$

Parameters

- $\|x_i\|^2 \leq \mathcal{M}$ w.p. 1
- 2^{nd} moment: $\mathcal{V} \triangleq \left\|\mathbb{E}[\|x_i\|^2 x_i x_i^\top]\right\|$
  - Natural notion of variance for matrices
- λ_1 and λ_2 top two eigenvalues of Σ
- $\text{gap} = \lambda_1 - \lambda_2$

Task: Find top eigenvector v^* of Σ

Formally: Find $\hat{v}$ to $\min_{\hat{v}} \theta(\hat{v}, v^*)^2$

Streaming: Cannot store n vectors

Problem Definition: Streaming PCA

More precise notions:

- $\left\| x_i x_i^\top - \Sigma \right\| \leq \mathcal{M}$
- $\mathcal{V} \triangleq \left\| \mathbb{E}[(x_i x_i^\top - \Sigma)^2] \right\|$

Considered in the paper but not for this talk.

Parameters

- $\|x_i\|^2 \leq \mathcal{M}$ w.p. 1
- 2^{nd} moment: $\mathcal{V} \triangleq \left\| \mathbb{E}[\|x_i\|^2 x_i x_i^\top] \right\|$
  - ➢ Natural notion of variance for matrices
- λ_1 and λ_2 top two eigenvalues of Σ
- $\text{gap} = \lambda_1 - \lambda_2$

Task: Find top eigenvector v^* of Σ

Formally: Find $\hat{v}$ to $\min_{\hat{v}} \theta(\hat{v}, v^*)^2$

Streaming: Cannot store n vectors

Standard Approach: Empirical Covariance Matrix

Compute top eigenvector $\hat{v}$ of $\hat{\Sigma} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$

Standard Approach: Empirical Covariance Matrix

Compute top eigenvector $\hat{v}$ of $\hat{\Sigma} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$

Matrix Bernstein + Wedin's theorem (With constant probability)

Standard Approach: Empirical Covariance Matrix

Compute top eigenvector $\hat{v}$ of $\hat{\Sigma} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$

- $\|x_i\| \leq \mathcal{M}$
- $\mathcal{V} \triangleq \|\mathbb{E}[\|x_i\|^2 x_i x_i^\top]\|$
- λ_1 and λ_2 top two e.v. of Σ

Matrix Bernstein + Wedin's theorem (With constant probability)

Standard Approach: Empirical Covariance Matrix

Compute top eigenvector $\hat{v}$ of $\hat{\Sigma} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$

- $\|x_i\| \le \mathcal{M}$
- $\mathcal{V} \triangleq \|\mathbb{E}[\|x_i\|^2 x_i x_i^\top]\|$
- λ_1 and λ_2 top two e.v. of Σ

Matrix Bernstein + Wedin's theorem (With constant probability)

$$\theta(\hat{v}, v^*)^2 \le O\left(\frac{\mathcal{V}\log d}{\text{gap}^2}\cdot\frac{1}{n} + \left(\frac{\mathcal{M}\log d}{\text{gap}}\right)^2 \cdot\frac{1}{n^2}\right)$$

Standard Approach: Empirical Covariance Matrix

Compute top eigenvector $\hat{v}$ of $\hat{\Sigma} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$

- $\|x_i\| \le \mathcal{M}$
- $\mathcal{V} \triangleq \|\mathbb{E}[\|x_i\|^2 x_i x_i^\top]\|$
- λ_1 and λ_2 top two e.v. of Σ

Matrix Bernstein + Wedin's theorem (With constant probability)

$$\theta(\hat{v}, v^*)^2 \le O\left(\frac{\mathcal{V}\log d}{\text{gap}^2}\cdot\frac{1}{n} + \left(\frac{\mathcal{M}\log d}{\text{gap}}\right)^2\cdot\frac{1}{n^2}\right)$$

Asymptotic error

Lower order error

Standard Approach: Empirical Covariance Matrix

Com

Mat

of Σ

Can we achieve the same accuracy using $O(d)$ space?

$$\theta(\hat{v}, v^*)^2 \leq O\left(\frac{\mathcal{V}\log d}{\text{gap}^2}\cdot\frac{1}{n} + \left(\frac{\mathcal{M}\log d}{\text{gap}}\right)^2\cdot\frac{1}{n^2}\right)$$

Asymptotic error

Lower order error

Existing Results

- Existing results are suboptimal by $O(d)$ or $O\left(\frac{\lambda_1}{\text{gap}}\right)$

Algorithm	Asymptotic Error
Matrix Bernstein	$O\left(\frac{\mathcal{V}\log d}{\text{gap}^2}\cdot\frac{1}{n}\right)$
Alecton (Sa et al. 2015)	$O\left(\frac{\mathcal{V}d\log d}{\text{gap}^2}\cdot\frac{1}{n}\right)$
Block power method (Mitliagkas et al. 2013, Hardt and Price 2014)	$O\left(\frac{\mathcal{V}\log d}{\text{gap}^2}\cdot\frac{\lambda_1}{\text{gap}}\cdot\frac{1}{n}\right)$

Oja's Algorithm (1982)

1. Choose v_0 uniformly random from unit sphere
2. For $i = 1, \ldots, n$
   a. $v_i \leftarrow v_{i-1} + \eta_i \left(x_i^\top v_{i-1}\right) x_i$
   b. $v_i \leftarrow \frac{v_i}{\|v_i\|}$
3. Output v_n

Oja's Algorithm (1982)

1. Choose v_0 uniformly random from unit sphere
2. For $i = 1, \ldots, n$
   a. $v_i \leftarrow v_{i-1} + \eta_i \left(x_i^\top v_{i-1}\right) x_i$
   b. $v_i \leftarrow \frac{v_i}{\|v_i\|}$
3. Output v_n

- Asymptotic convergence well known
- Best known finite sample guarantees (Balasubramani et al. 2013) are suboptimal by $\mathrm{poly}(d)$ factors

Oja's Algorithm (1982)

1. Choose v_0 uniformly random from unit sphere
2. For $i = 1, \ldots, n$
   a. $v_i \leftarrow v_{i-1} + \eta_i \left(x_i^\top v_{i-1}\right) x_i$
   b. $v_i \leftarrow \frac{v_i}{\|v_i\|}$
3. Output v_n

- Asymptotic convergence well known
- Best known finite sample guarantees (Balasubramani et al. 2013) are suboptimal by $\mathrm{poly}(d)$ factors

Our contribution

Near optimal finite sample guarantees for Oja's algorithm

Our Result

For a particular choice of η_i, once $n > \frac{\mathcal{V} \log^2 d}{\text{gap}^2}$

Algorithm	Error
Our Result ($O(d)$ space)	$O\left(\frac{\mathcal{V} \log d}{\text{gap}^2} \cdot \frac{1}{n} + \left(\frac{\mathcal{M} \log d}{\text{gap}}\right)^2 \cdot \frac{1}{n^2}\right)$
Matrix Bernstein ($O(d^2)$ space)	$O\left(\frac{\mathcal{V} \log d}{\text{gap}^2} \cdot \frac{1}{n} + \left(\frac{\mathcal{M} \log d}{\text{gap}}\right)^2 \cdot \frac{1}{n^2}\right)$

$\mathcal{M}$ is upper bound, $\mathcal{V}$ is second moment

Our Result

For a different step size sequence η_i, asymptotically

Algorithm	Asymptotic Error
Our Result	$O\left(\frac{\mathcal{V}}{\text{gap}^2} \cdot \frac{1}{n}\right)$
Matrix Bernstein	$O\left(\frac{\mathcal{V} \log d}{\text{gap}^2} \cdot \frac{1}{n}\right)$

$\mathcal{V}$ is second moment

Difficulty with Earlier Approaches

- Earlier approaches: Try to track $\theta(v_i, v^*)$ and show descent

- Main challenge:

  - ❑ $\theta(v_i, v^*)^2$ need not monotonically decrease

  - ❑ Even worse, hard to rule out $v_i \perp v^*$

  - ❑ Bottomline: Large noise in initial iterations very hard to tackle

Proof Idea

Essentially, $v_n = \frac{1}{Z} B_n v_0$, where $B_n \triangleq \prod_{i=n}^{1} \left(\mathbb{I} + \eta_i x_i x_i^\top \right)$

Proof Idea

Essentially, $v_n = \frac{1}{Z} B_n v_0$, where $B_n \triangleq \prod_{i=n}^{1} \left(\mathbb{I} + \eta_i x_i x_i^\top \right)$

Lemma 1

If $\hat{v} = \frac{B v_0}{\|B v_0\|}$ for v_0 uniformly random on unit sphere, then $\forall\, v^*$:

$$\theta(\hat{v}, v^*)^2 \leq \frac{\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)}{{v^*}^\top B B^\top v^*} \text{ (with prob > 3/4)}$$

Proof Idea (Ctd.)

$$\theta(\hat{v}, v^*)^2 \leq \frac{\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)}{v^{*\top} B B^\top v^*}; \quad B = \prod_{i=n}^{1} \left(\mathbb{I} + \eta_i x_i x_i^\top\right)$$

Proof Idea (Ctd.)

$$\theta(\hat{v}, v^*)^2 \leq \frac{\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)}{v^{*\top} B B^\top v^*}; \quad B = \prod_{i=n}^{1} \left(\mathbb{I} + \eta_i x_i x_i^\top\right)$$

Prove upper bound on $\mathbb{E}\left[\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)\right]$ ⟶ Use Markov's inequality

Proof Idea (Ctd.)

$$\theta(\hat{v}, v^*)^2 \leq \frac{\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)}{{v^*}^\top B B^\top v^*}; \quad B = \prod_{i=n}^{1}\left(\mathbb{I} + \eta_i x_i x_i^\top\right)$$

Prove upper bound on $\mathbb{E}\left[\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)\right]$ ⟶ Use Markov's inequality

Prove lower bound on $\mathbb{E}\left[{v^*}^\top B B^\top v^*\right]$
and upper bound on $\mathbb{E}\left[\left({v^*}^\top B B^\top v^*\right)^2\right]$ ⟶ Use Chebyshev inequality

Proof Idea (Ctd.)

$$\theta(\hat{v}, v^*)^2 \leq \frac{\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)}{{v^*}^\top B B^\top v^*}; \quad B = \prod_{i=n}^{1}\left(\mathbb{I} + \eta_i x_i x_i^\top\right)$$

Prove upper bound on $\mathbb{E}\left[\mathrm{Tr}\left(V_\perp^\top B B^\top V_\perp\right)\right]$ ⟶ Use Markov's inequality

Prove lower bound on $\mathbb{E}\left[{v^*}^\top B B^\top v^*\right]$
and upper bound on $\mathbb{E}\left[\left({v^*}^\top B B^\top v^*\right)^2\right]$ ⟶ Use Chebyshev inequality

- Characterizes error in terms of step sizes η_i
- Particular choice of η_i yields our results claimed earlier

Conclusion

- Near optimal guarantees for Oja's algorithm
- Error bound as an explicit function of step size sequence
  - ❑ Helps in choosing step-size sequence for any given regime
- New and simple analysis; could be useful in a wider context

Open problems

- Gap independent results
- Extension to top-k PCA
- High probability results