Near-Optimal Stochastic Approximation for Online Principal Component Estimation

Chris Junchi Li

Tencent & Princeton University

Joint work with Mengdi Wang (Princeton), Han Liu (Northwestern), and Tong Zhang (Tencent AI Lab)

November 2018

Outline

Principal Component Analysis (PCA)

PCA (Pearson, 1901; Hotelling, 1933) is one of the most popular dimension reduction methods for high-dimensional data analysis

- PCA aims at learning principal leading eigenvector (or eigenspace) of the covariance matrix of a distribution from its IID data samples
- Rank-one PCA learns the eigenvector that captures most variance in data
- Wide applications in bioinformatics, healthcare, imaging, computer vision, artificial intelligence, social science, finance, economics, etc.

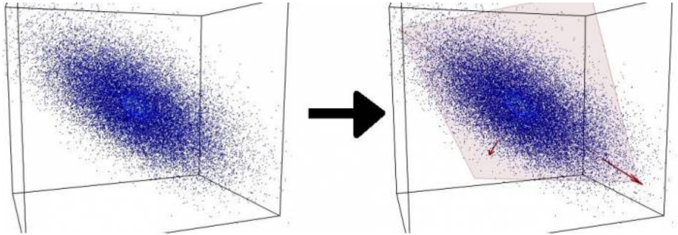


Figure 1: Illustration of PCA

PCA: Formulation

Let $\boldsymbol{X}$ be a d-dimensional random vector with mean zero and unknown covariance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] \in \mathbb{R}^{d \times d}$$

Projection of $\boldsymbol{X}$ onto unit vector $\mathbf{u}$ is $\mathbf{u}^\top \boldsymbol{X}$

- Rank-one PCA is formulated as a nonconvex stochastic optimization problem:

$$\begin{aligned} \text{minimize} \quad & -\mathbf{u}^\top \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]\mathbf{u} \\ \text{subject to} \quad & \mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\| = 1 \end{aligned} \tag{PCA}$$

- Nonconvexity due to the unit spherical constraint

Assume the eigengap of $\boldsymbol{\Sigma}$ is nonzero, so solution $\mathbf{u}^*$ to (PCA) is unique

PCA Landscape: Simplest Nonconvex Problem?

Let the covariance matrix $\mathbf{\Sigma} = \mathbf{O}\mathbf{\Lambda}\mathbf{O}^\top$ be spectral decomposition, $\mathbf{\Lambda}$ diagonal, $\mathbf{O}$ orthogonal. Let $\mathbf{e}_i = (0, \ldots, \underbrace{1}_{i^{th} \text{ coordinate}}, \ldots, 0)$ be the i^{th} coordinate vector

The stationary points of PCA landscape ($\geq 2d$ many) are of two types:

- Global minimizers: $\pm\mathbf{O}\mathbf{e}_1$;
- Global maximizers or saddle points: $\pm\mathbf{O}\mathbf{e}_2, \pm\mathbf{O}\mathbf{e}_3, \ldots, \pm\mathbf{O}\mathbf{e}_d$ and possibly more, all lying on the equator $\{\mathbf{u} : \mathbf{u}^\top\mathbf{u}^* = 0\}$,

"No spurious local minimizer" (Ge, Lee, Ma, 2016 NIPS)

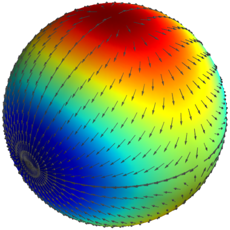


Figure 2: Quiver plot that denotes the negative-gradient directions of PCA

Classical PCA

Classical PCA estimates $\mathbf{u}^*$ using a sample average approximation method: find the top eigenvector of $\widehat{\mathbf{\Sigma}}^{(N)}$

$$\widehat{\mathbf{\Sigma}}^{(N)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}^{(i)} \left(\boldsymbol{X}^{(i)}\right)^\top$$

as an estimator of $\mathbf{u}^*$, based on i.i.d. sample realizations $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}$

- Classical PCA method produces non-improvable solution $\widehat{\mathbf{u}}^{(N)}$: estimation error achieves the minimax information lower bound (to be discussed later)
- Nevertheless, classical PCA has suboptimal (on d) time complexity $\mathcal{O}(Nd^2)$ and space complexity $\mathcal{O}(d^2)$
- When d is large, computing and storing a large empirical covariance matrix is potentially inefficient

Turn to stochastic approximation method

Online PCA

We turn to incremental or online methods for PCA, which updates the iterates incrementally by processing data points one-by-one or on-the-fly

- The gradient of objective function

$$\frac{\partial}{\partial \mathbf{u}} \left\{ -\mathbf{u}^\top \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]\mathbf{u} \right\} = -2\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]\mathbf{u}$$

- At step t, stream in data point $\boldsymbol{X}^{(t)}$ and conduct projected SGD step

$$\mathbf{u}^{(t)} = \Pi \left\{ \mathbf{u}^{(t-1)} + \eta \boldsymbol{X}^{(t)} (\boldsymbol{X}^{(t)})^\top \mathbf{u}^{(t-1)} \right\} \qquad \text{(Oja)}$$

  Here η is positive stepsize, $\Pi \mathbf{u} = \mathbf{u}/\|\mathbf{u}\|$ projects $\mathbf{u}$ onto the unit sphere

- Iteration first proposed by Oja (1982), which only gives almost sure convergence
- Known as online PCA, streaming PCA, or noisy power method (Hardt & Price, 2014 NIPS)

Online PCA

$$\mathbf{u}^{(t)} = \Pi\left\{\mathbf{u}^{(t-1)} + \eta \boldsymbol{X}^{(t)}(\boldsymbol{X}^{(t)})^\top \mathbf{u}^{(t-1)}\right\} \tag{Oja}$$

1. Essentially a stochastic approximation method for PCA but learns data on-the-fly, most applicable to both dimension d and number of samples N being large
2. Convergence rate analysis of (Oja) remains largely open until very recently. Theoretical challenge is due to the nonconvex nature

	Time complexity	Space complexity
Classical PCA	$\mathcal{O}(Nd^2)$	$\mathcal{O}(d^2)$
Oja's iteration	$\mathcal{O}(Nd)$	$\mathcal{O}(d)$

- Pros: iteration update requires only vector-vector product operation and stores only $\mathbf{u}^{(t)}$. Time complexity $\mathcal{O}(Nd)$ and space complexity $\mathcal{O}(d)$
- Cons: Choice of η, unknown convergence rate & initialization

Online PCA: New Convergence Rate Analysis

$$\mathbf{u}^{(t)} = \Pi\left\{\mathbf{u}^{(t-1)} + \eta \boldsymbol{X}^{(t)}(\boldsymbol{X}^{(t)})^\top \mathbf{u}^{(t-1)}\right\} \tag{Oja}$$

Our conclusion in one line (**L.**-Wang-Liu-Zhang, 2017 Math. Prog.):

Online PCA is *statistically optimal* and *globally convergent*

- The independent work by Jain, Jin, Kakade, Netrapalli, & Sidford (2016 COLT) also analyzes Oja's iteration and obtains an error bound that matches the matrix Bernstein's inequality under uniform initialization

Online PCA: Distributional Assumptions

Let the random samples $\boldsymbol{X} \equiv \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)} \in \mathbb{R}^d$ be i.i.d. and satisfy [1]

1. (Subgaussian) $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ where
   $\boldsymbol{Z}$ is sub-Gaussian with mean zero, covariance matrix $\mathbf{I}_d$
   Subgaussian norm $\|\boldsymbol{Z}\|_{\psi_2} = \sup_{\|\mathbf{u}\|=1} \|\mathbf{u}^\top \boldsymbol{Z}\|_{\psi_2} \leq 1$

This allows us to conclude $\mathbb{E}[\boldsymbol{X}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] = \boldsymbol{\Sigma}$

2. (Eigengap) The eigenvalues of $\boldsymbol{\Sigma}$ satisfy $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_d \geq 0$

[1] In below the matrix square root $\boldsymbol{\Sigma}^{1/2}$ satisfies $\boldsymbol{\Sigma}^{1/2} \cdot \boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$

Theorem (Convergence result, L., Wang, Liu & Zhang, 2017 Math. Prog.)

Suppose the $\mathbf{u}^{(0)}$ is uniformly sampled from the unit sphere, and scaling condition

$$d\eta^{1-\varepsilon} \text{ is sufficiently small}$$

Then for any $\delta > 0$ there is an event $\mathcal{A}$ with $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$ such that the iterates generated by (Oja) satisfy for all $\eta > 0$ sufficiently small and t sufficiently large

$$\mathbb{E}\left[\sin^2 \angle(\mathbf{u}^{(t)}, \mathbf{u}^*) \mid \mathcal{A}\right] \leq \underbrace{C \cdot \delta^{-2} d \cdot (1 - \eta(\lambda_1 - \lambda_2))^{2t}}_{\text{optimization error}} + \underbrace{C \cdot \sum_{k=2}^{d} \frac{\lambda_1 \lambda_k}{\lambda_1 - \lambda_k} \cdot \eta}_{\text{statistical error}}$$

Corollary (Finite-sample result, L., Wang, Liu & Zhang, 2017 Math. Prog.)

Suppose the $\mathbf{u}^{(0)}$ is uniformly sampled from the unit sphere, and the scaling condition

$$d/N^{1-\varepsilon} \text{ is sufficiently small}$$

Let the stepsize $\eta = \bar{\eta}(N) \asymp \dfrac{\log N}{(\lambda_1 - \lambda_2)N}$. Then for any $\delta > 0$ there exists an event $\mathcal{A}$ with $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$ such that iterates generated by (Oja) satisfy

$$\mathbb{E}\left[\sin^2 \angle(\mathbf{u}^{(N)}, \mathbf{u}^*) \mid \mathcal{A}\right] \leq C \cdot \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \cdot \frac{\log N}{N}$$

Significance of Our Result

Significance 1: statistical optimality

Oja's iteration produces estimator that nearly attains $\widetilde{\mathcal{O}}(\sqrt{d/N})$-minimax rate

- Theorem 3.1 of Vu and Lei (2013) provides the minimax information lower bound:

$$\inf_{\widetilde{\mathbf{u}}^{(N)}} \sup_{\boldsymbol{X} \in \mathcal{M}(\boldsymbol{\Sigma}, d)} \mathbb{E}\left[\sin^2 \angle(\widetilde{\mathbf{u}}^{(N)}, \mathbf{u}^*)\right] \geq c \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{d-1}{N}$$

- By choosing the stepsize carefully and under mild scaling assumptions, the output estimator nearly attains such lower bound:

$$\sup_{\boldsymbol{X} \in \mathcal{M}(\boldsymbol{\Sigma}, d)} \mathbb{E}\left[\sin^2 \angle(\mathbf{u}^{(N)}, \mathbf{u}^*) \mid \mathcal{A}\right] \leq C \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{d-1}{N} \cdot \log N$$

inf of $\widetilde{\mathbf{u}}^{(N)}$ is over all principal component estimators, and $\mathcal{M}(\boldsymbol{\Sigma}, d)$ consists of all subgaussian distributions in $\mathbb{R}^d$ with mean $\mathbf{0}$ and positive eigengap $\lambda_1 - \lambda_2$.

Significance of Our Result

Significance 2: global convergence

Finite-sample error bound of Oja's iteration holds under uniform initialization

- In contrast, most existing results requires a good initialization $\left|\sin\angle(\mathbf{u}^{(0)}, \mathbf{u}^*)\right| \le 1 - \varepsilon$. As dimension d grows, uniform initialization does not attain such good initialization with high probability, since

$$\left|\sin\angle(\mathbf{u}^{(0)}, \mathbf{u}^*)\right| \approx 1 - C/d$$

- Favorite probability question: what is the distribution of $\cos\angle(\mathbf{u}^{(0)}, \mathbf{u}^*)$?

Outline

Online PCA is rotationally invariant

- Let the diagonal decomposition of the covariance matrix be

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\top\right] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, $\mathbf{U}$ is an orthogonal matrix consisting of column eigenvectors of $\boldsymbol{\Sigma}$.

- Rescaled samples $\boldsymbol{Y}^{(t)} = \mathbf{U}^\top \boldsymbol{X}^{(t)}$, $\mathbf{v}^{(t)} = \mathbf{U}^\top \mathbf{u}^{(t)}$, $\mathbf{v}^* = \mathbf{U}^\top \mathbf{u}^*$ has

$$\mathbb{E}[\boldsymbol{Y}] = 0 \qquad \mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] = \boldsymbol{\Lambda} \qquad \angle(\mathbf{u}^{(t)}, \mathbf{u}^*) = \angle(\mathbf{v}^{(t)}, \mathbf{v}^*)$$

Study the iteration $\mathbf{v}^{(t)}$: applying the linear transformation $\mathbf{U}^\top$ to the stochastic process $\{\mathbf{u}^{(t)}\}$

$$\mathbf{v}^{(t)} \leftarrow \Pi\left\{\mathbf{v}^{(t-1)} + \eta \boldsymbol{Y}^{(t)}(\boldsymbol{Y}^{(t)})^\top \mathbf{v}^{(t-1)}\right\} \qquad \text{(Oja)}$$

Outline

Sketch of Proofs: Warm Initialization

Warm initialization: $\left|\sin\angle(\mathbf{v}^{(0)}, \mathbf{v}^*)\right| \le 1/\sqrt{2}$, "Less than 45 degree"

Let the rescaled stepsize $\widehat{\eta} = \lambda_1^2(\lambda_1 - \lambda_2)^{-1}\eta$, and rescaled time

$$N^*_{\eta,s} = \left\lceil \frac{s\log(\lambda_1^{-2}(\lambda_1-\lambda_2)\eta^{-1})}{-\log(1-\eta(\lambda_1-\lambda_2))} \right\rceil \asymp s\cdot(\lambda_1-\lambda_2)^{-1}\eta^{-1}\log(\widehat{\eta}^{-1})$$

Assume WLOG $\mathbf{\Sigma}$ is diagonal and suppose $\mathbf{v}^{(0)}$ is a warm initialization. Then each ratio iteration $v_k^{(t)}/v_1^{(t)}$ decays geometrically at rate $1-\eta(\lambda_1-\lambda_k)$:

$$v_k^{(t)}/v_1^{(t)} \approx (1-\eta(\lambda_1-\lambda_k))^t \left(v_k^{(0)}/v_1^{(0)}\right)$$

We rigorously prove via martingale concentration inequalities that with high probability

$$\sup_{t\le N^*_{\eta,s}} \left| v_k^{(t)}/v_1^{(t)} - (1-\eta(\lambda_1-\lambda_k))^t\left(v_k^{(0)}/v_1^{(0)}\right)\right| \le C\widehat{\eta}^{0.5-\varepsilon}.$$

This is a manifestation of strong convergence

Propositions

Using more careful second moment estimates in the $O(\eta^{0.5})$ neighborhood of the principal component $\mathbf{v}^*$:

Proposition 2

Assume $\mathbf{v}^{(0)}$ is a warm initialization. When $d\widehat{\eta}^{1-2\varepsilon}$ is sufficiently small, there exists a high-probability event $\mathcal{H}_0$ such that for $t \in [N^*_{\eta,1}, N^*_{\eta,s}]$

$$\mathbb{E}\left[\tan^2\angle(\mathbf{v}^{(t)}, \mathbf{v}^*)\ ; \mathcal{H}_0\right] \le (1-\eta(\lambda_1-\lambda_2))^{2t} \tan^2\angle(\mathbf{v}^{(0)}, \mathbf{v}^*) + C \cdot \sum_{k=2}^{d} \frac{\lambda_1\lambda_k + \lambda_1^2 \cdot \widehat{\eta}^{0.5-4\varepsilon}}{\lambda_1 - \lambda_k} \cdot \eta.$$

Also hold for uniform initialization? Yes! By analyzing the growth of $v_1^{(t)}/\sqrt{1-(v_1^{(t)})^2}$ via martingale concentration

Proposition 3

Assume $\mathbf{v}^{(0)}$ is a uniform initialization. When $d\widehat{\eta}^{1-2\varepsilon}$ is sufficiently small, the time required to enter the warm region $\mathcal{N}_c$ has with high probability

$$\mathcal{N}_c \le N^*_{\eta,1}$$

Putting pieces together

Lemma

Given any $\delta > 0$, if $\mathbf{u}^{(0)}$ is sampled uniformly at random from $\mathcal{S}^{d-1}$ in $\mathbb{R}^d$ then there exists a constant $C^* > 1$ independent of δ and d such that

$$\mathbb{P}\left(\tan^2\angle(\mathbf{u}^{(0)}, \mathbf{u}^*) \le C^*\delta^{-2}d\right) \ge 1 - \delta.$$

- The uniform initialization $\mathbf{v}^{(0)}$ from unit sphere has $\tan^2\angle(\mathbf{v}^{(0)}, \mathbf{v}^*) \le c^* d$
- Running the algorithm for $N_\eta^o(c^*) \wedge \mathcal{N}_c < N_{\eta,1}^*$ steps, the iterate $\mathbf{v}^{(N_\eta^o(c^*)\wedge\mathcal{N}_c)}$ is with high probability in the warm region (Proposition 2)
- By strong Markov property the iterates can be regarded as initialized from warm initialization $\mathbf{v}^{(N_\eta^o(c^*)\wedge\mathcal{N}_c)}$, then apply Proposition 1 to run for another $N_{\eta,1}^*$ steps

Theorem (Convergence result with uniform initialization)

Suppose the $\mathbf{v}^{(0)}$ is uniformly sampled from the unit sphere, and scaling condition

$$d\eta^{1-\varepsilon} \text{ is sufficiently small}$$

Then for any $\delta > 0$ there is an event $\mathcal{A}$ with $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$ such that the iterates generated by (Oja) satisfy for all $\eta > 0$ sufficiently small and $t \in [N^*_{\eta,2}, N^*_{\eta,s}]$

$$\mathbb{E}\left[\tan^2 \angle(\mathbf{v}^{(t)}, \mathbf{v}^*) \mid \mathcal{A}\right] \leq C \cdot \delta^{-2} d \cdot (1 - \eta(\lambda_1 - \lambda_2))^{2t} + C \cdot \sum_{k=2}^{d} \frac{\lambda_1 \lambda_k + \lambda_1^2 \cdot \widehat{\eta}^{0.5-4\varepsilon}}{\lambda_1 - \lambda_k} \cdot \eta$$

- Plugging in

$$\eta \equiv \bar{\eta}(N) = \frac{2 \log N}{(\lambda_1 - \lambda_2)N}$$

so $N \approx N^*_{\bar{\eta}(N),2}$ and we obtain the finite-sample error bound:

Corollary (Finite-sample result with uniform initialization)

Suppose the $\mathbf{v}^{(0)}$ is uniformly sampled from the unit sphere, and the scaling condition

$$d/N^{1-\varepsilon} \text{ is sufficiently small}$$

Let the stepsize $\eta = \bar{\eta}(N)$. Then for any $\delta > 0$ there exists an event $\mathcal{A}$ with $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$ such that iterates generated by (Oja) satisfy

$$\mathbb{E}\left[\tan^2 \angle(\mathbf{v}^{(t)}, \mathbf{v}^*) \mid \mathcal{A}\right] \leq C \cdot \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{k=2}^{d} \frac{\lambda_k}{\lambda_1 - \lambda_k} \cdot \frac{\log N}{N}$$

Outline

Differential Equation Approximation

Theorem 1

When $\eta > 0$ is small, Oja's iteration can be approximated by the solution of an ordinary differential equation (ODE), if we use an appropriate temporal scaling

The ODE is

$$\frac{\mathrm{d} V_j}{\mathrm{d} s} = V_j \sum_{k=1}^{d} (\lambda_j - \lambda_k) V_k^2, \qquad j = 1, \ldots, d \tag{ODE}$$

Solution to (ODE) is available in closed-form ($Z(s)$ be normalizing constant):

$$V_j(s) = Z(s)^{-1/2} V_j(0) \exp(\lambda_j s)$$

"Generalized logistic curves" or "Oja's flow" (Helmke & Moore, 1994)

Differential Equation Approximation

We study how Oja's iteration escapes from unstable stationary points and converges to stable stationary points:

Theorem 2

When $\eta > 0$ is small and $\mathbf{v}^{(0)} \approx \pm \boldsymbol{e}_k$, Oja's iteration can be approximated by the solution of a stochastic differential equation (SDE), if we use appropriate temporal and spatial scalings

The SDE is

$$\mathrm{d}\mathcal{V}_j = (\lambda_j - \lambda_k)\mathcal{V}_j \, \mathrm{d}s + (\lambda_j \lambda_k)^{1/2} \mathrm{d}B_j(s) \tag{SDE}$$

$B_j(s)$ is a standard Brownian motion "white noise"

"Ornstein-Uhlenbeck processes" (Uhlenbeck & Ornstein, 1930)

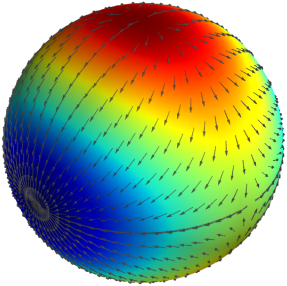
Three-Phase Analysis

Initialized near the equator $\{\mathbf{v} : \|\mathbf{v}\| = 1, v_1 = 0\}$, where all unstable stationary points lie on. Applying Theorems 1 and 2 gives the three-phase analysis:

Phase I: escaping from unstable stationary points characterized by SDE

Phase II: deterministic crossing characterized by ODE

Phase III: local converging characterized by SDE



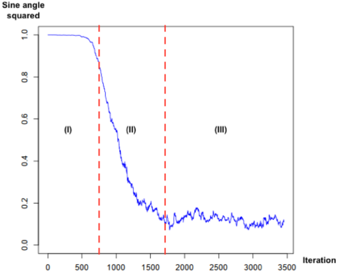


Figure 3: Animation of Oja's iteration; Illustration of three phases of diffusion processes.

Running Time Estimate and Finite-Sample Error Bound

Running time of Oja's iteration in each phase from differential equation approximation:

Phase I: iteration $T_1^\eta \asymp 0.5(\lambda_1 - \lambda_2)^{-1} \cdot \eta^{-1} \log(\eta^{-1})$

Phase II: iteration $T_2^\eta \asymp (\lambda_1 - \lambda_2)^{-1} \cdot \eta^{-1}$

Phase III: iteration $T_3^\eta \asymp 0.5(\lambda_1 - \lambda_2)^{-1} \cdot \eta^{-1} \log(\eta^{-1})$

Total running time $T^\eta = T_1^\eta + T_2^\eta + T_3^\eta \asymp (\lambda_1 - \lambda_2)^{-1} \eta^{-1} \log(\eta^{-1})$

$$\mathbb{E} \sin^2 \angle(\mathbf{v}^{(T^\eta)}, \mathbf{e}_1) \le C \cdot \sum_{k=2}^{d} \frac{\lambda_1 \lambda_k}{\lambda_1 - \lambda_k} \cdot \eta$$

Given N samples, choosing $\eta = \bar{\eta}(N) \equiv \dfrac{\log N}{(\lambda_1 - \lambda_2) N}$ we have $N \asymp T^{\bar{\eta}(N)}$ and

$$\mathbb{E} \sin^2 \angle(\mathbf{v}^{(N)}, \mathbf{e}_1) \le C \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \frac{d-1}{N} \cdot \log N$$

"Heuristically matches the statistical lower bound" (Vu & Lei, 2013 Ann. Stat.)

Summary

We provide a diffusion approximation perspective of convergence rate analysis and conclude:

Online stochastic gradient descent method for (rank-one) principal component analysis is both statistically optimal and globally convergent

Matching the statistical rate: by choosing the stepsize carefully, Oja's iteration attains $\widetilde{\mathcal{O}}(\sqrt{d/N})$-statistical rate for rank-one PCA

Global initialization: achieves optimal error bounds with no restriction on initialization, so can fastly escape from saddle points

Epilogue: Future Directions (and Non-Exhaustive Literatures)

1. Develop and analyze online PCA method for principal subspace learning that matches the statistical rate?

2. Parallelizing PCA for online data?

3. Extend the analysis beyond PCA to a broader class of nonconvex statistical estimation problems?
   - Tensor decomposition for ICA: (Ge, Huang, Jin, & Yuan, 2015 COLT) (**L.**, Wang, & Liu, 2016 NIPS) (Wang & Lu, 2017 NIPS)
   - Sparse PCA (d'Aspremont, Bach, & El Ghaoui, 2008 JMLR)
   - Partial least squares (Chen, Yang, **L.**, & Zhao, 2017 ICML)
   - Phase retrieval, Dictionary learning (Sun, Qu, & Wright, arXiv:1510.06096)
   - Matrix completion & Sensing (Sun, & Luo, 2014+ IEEE TIT) (Zheng & Lafferty, 2015 NIPS) (Zhao, Wang, & Liu, 2015 NIPS) (Ge, Jin, & Zheng, 2017 ICML)
   - Deep Learning: batch size VS generalization

Reference

L., Wang, Liu & Zhang (2017) , Near-Optimal Stochastic Approximation for Online Principal Component Estimation , Mathematical Programming

L., Wang, Liu & Zhang (2017) , Diffusion Approximation of Online PCA , NIPS

Oja, E. (1982) . Simplified neuron model as a principal component analyzer . Journal of mathematical 572 biology, 15(3), 267–273

Jain, Jin, Kakade, Netrapalli, & Sidford (2016) , Streaming PCA: Matching Matrix Bernstein and Near-Optimal Finite Sample Guarantees for Oja's Algorithm , COLT

Allen-Zhu, Li (2017) , First effcient convergence for streaming k-PCA: a global, gap-free, and near-optimal rate , FOCS