Dimensionality Reduction for Stationary Time Series via Stochastic Nonconvex Optimization

Minshuo Chen*, Lin F. Yang◇, Mengdi Wang◇, Tuo Zhao*
*Georgia Tech ◇Princeton University

Background

Consider the following stochastic optimization problem,

$$\min_u \mathbb{E}_{Z\sim\mathcal{D}}[f(u,Z)] \quad \text{subject to } u \in \mathcal{U},$$

- f is a loss function (possibly **nonconvex**);
- Z is the random sample;
- $\mathcal{D}$ is the underlying data distribution;
- $\mathcal{U}$ is a feasible set (possibly **nonconvex**).

Consider n samples $\{z_1, ..., z_n\}$ from $\mathcal{D}$, we have

$$\mathbb{E}[f(u,z)] = \frac{1}{n}\sum_{i=1}^{n} f(u, z_i).$$

For differentiable f, stochastic gradient descent (SGD) takes

$$u_{k+1} = \Pi_{\mathcal{U}}[u_k - \eta\nabla_u f(u_k, z_k)],$$

- η is the step size parameter;
- $\nabla_u f(u_k, z_k)$ is an **unbiased** stochastic gradient for approximating $\nabla_u \mathbb{E}_{Z\sim\mathcal{D}} f(u_k, Z)$, i.e.,

$$\mathbb{E}_{z_k}\nabla_u f(u_k, z_k) = \nabla_u \mathbb{E}_{Z\sim\mathcal{D}} f(u_k, Z);$$

- $\Pi_{\mathcal{U}}$ is a projection operator onto the feasible set $\mathcal{U}$.

Challenges:
- Data dependency $\Longrightarrow$ Biased stochastic gradient;
- Nonconvex $f, \mathcal{U} \Longrightarrow$ Complicated landscapes.

Streaming PCA Problem

A simple but fundamental problem for time series data:

$$U^* \in \underset{U}{\mathrm{argmin}} \; -\mathrm{Trace}(U^\top \Sigma U) \quad \text{subject to } U^\top U = I_r,$$

- $U \in \mathbb{R}^{m\times r}$ aims to recover r leading eigenvectors;
- Σ is the covariance matrix of the stationary distribution.

Time series ► Biased estimation due to data dependency:

$$\mathbb{E}[z_k z_k^\top U_k | U_k] \neq \Sigma U_k;$$

Nonconvexity ► Solution space is rotational-invariant:

$$U \Longleftrightarrow QU \text{ for any orthogonal matrix } Q \in \mathbb{R}^{r\times r}.$$

Our Approaches:
▷ Downsampling $\Longrightarrow$ ~~Data dependency~~;
▷ Principle Angle $\Longrightarrow$ ~~Rotational invariance~~.

Downsampled Oja's Algorithm

Lemma 1. *For time series $\{z_k\}$ with covariance matrix Σ,*

- *$\{z_k\}_{k=1}^\infty$ is Markov, geometrically ergodic with parameter ρ, and sub-Gaussian;*
- *The stationary distribution has zero mean.*

$\Longrightarrow$ *Given a pre-specified accuracy τ, there exists $h = O\left(\kappa_\rho \log\frac{1}{\tau}\right)$ such that*

$$\mathbb{E}\left[z_{h+k}z_{h+k}^\top \middle| z_k\right] = \Sigma + E\Sigma \quad \textit{with } \|E\|_2 \le \tau.$$

► Motivate us to chunk up the time series:

$\boxed{z_1, z_2, \ldots, z_h}, \boxed{z_{h+1}, \ldots, z_{2h}}, \cdots, \boxed{z_{2bh+1}, \ldots, z_{2(b+1)h}}$

Downsampled Oja's Algorithm for Streaming PCA

Input: data points z_k, block size h, step size η.
Init: set U_1 with orthonormal columns;
  set $s \leftarrow 1$.
Repeat:
  Take sample z_{sh}, and set $X_s \leftarrow z_{sh}z_{sh}^\top$;
  $U_{s+1} \leftarrow \Pi_{\mathrm{Orth}}(U_s + \eta X_s U_s)$;
  $s \leftarrow s+1$;
Until Convergence.
Output: U_s.

Principle Angle Based Landscape

Principle Angle: Given two matrices $U \in \mathbb{R}^{m\times r_1}$ and $V \in \mathbb{R}^{m\times r_2}$ with orthonormal columns, where $1 \le r_1 \le r_2 \le m$, the principle angle between U and V is defined as,

$$\Theta(U,V) = \mathrm{diag}\left[\cos^{-1}\left(\sigma_1(U^\top V)\right), \ldots, \cos^{-1}\left(\sigma_{r_1}(U^\top V)\right)\right].$$

Landscape of Steaming PCA: Eigenvalue decomposition

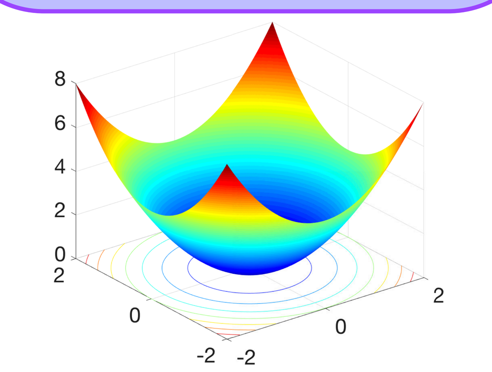
$$\Sigma = \sum_{i=1}^{m} \lambda_i v_i v_i^\top$$

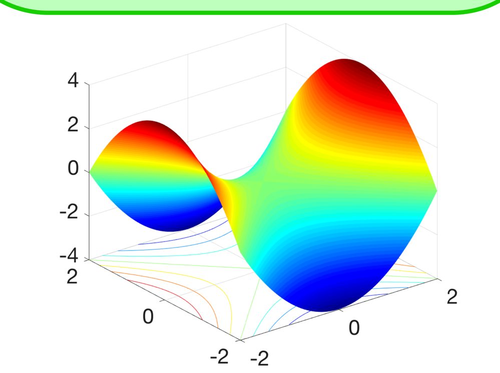
with λ_i and v_i being eigenvalue and eigenvector, respectively.

Stationary Point
$\mathrm{span}(U) = \mathrm{span}(v_{a_1}, \ldots, v_{a_r})$ with $\mathcal{A}_r = \{a_1, \ldots, a_r\} \subset \{1, \ldots, m\}$
$\Longleftrightarrow \|\cos\Theta(U, V_r)\|_{\mathrm{F}} \in \{0, \ldots, r\}$ with $V_r = [v_1, \ldots, v_r]$

Global Optimum
$\mathrm{span}(U) = \mathrm{span}(V_{\mathcal{A}_r})$ with $\mathcal{A}_r = \{1, \ldots, r\}$
$\Longleftrightarrow \|\cos\Theta(U, V_r)\|_{\mathrm{F}} = r$

Saddle Point (Global Maximum)
$\mathrm{span}(U) = \mathrm{span}(V_{\mathcal{A}_r})$ with $\mathcal{A}_r \neq \{1, \ldots, r\}$
$\Longleftrightarrow \|\cos\Theta(U, V_r)\|_{\mathrm{F}} < r$





Convergence Analysis — Intuition

Consider Taylor expansion of downsampled Oja's algorithm:

$$U_{s+1} = U_s + \eta\left(I - U_s U_s^\top\right) X_s U_s + \eta^2 W_s.$$

Define principle angle $\gamma_{i,s}^2 = \|U_s v_i\|_2^2$ for $i = 1, \ldots, m$.

- **ODE** Approximation:

$$\text{Discrete: } \frac{\gamma_{i,s+1}^2 - \gamma_{i,s}^2}{\eta} = \mathcal{F}_{i,s}\gamma_{i,s}^2 + O(\eta).$$

$$\text{weakly} \Downarrow \eta \to 0$$

$$\text{Continuous: } d\gamma_i^2 = b_i\gamma_i^2 dt.$$

Analogous to Law of Large Number, not reliable!

- **SDE** Approximation ($\gamma_{i,s}^2 = O(\eta)$ for some $i \in \{1, \ldots, r\}$):

Decompose principle angle as $\gamma_{i,s}^2 = \eta\sum_{j=1}^{r}\zeta_{ij,s}^2$.

$$\text{Discrete: } \frac{\zeta_{ij,s+1} - \zeta_{ij,s}}{\sqrt{\eta}} = \mathcal{F}_{ij,s}\zeta_{ij,s} + O(\eta).$$

$$\text{weakly} \Downarrow \eta \to 0$$

$$\text{Continuous: } d\zeta_{ij} = K_{ij}\zeta_{ij}dt + G_{ij}dB_t.$$

Randomness Returns.
Analogous to Central Limit Theorem!

Convergence Analysis — Three Stages

Assumption 1. *There exists an eigengap in the covariance matrix Σ, i.e., $\lambda_1 \ge \cdots \ge \lambda_r > \lambda_{r+1} \ge \cdots \ge \lambda_m > 0$.*

Stage 1. Escaping from Saddle Points: We need asymptotically,

$$S_1 \asymp \frac{\log(K+1)}{\eta(\lambda_r - \lambda_{r+1})}$$

iterations to escape from a saddle point.

Stage 2. Traverse between Stationary Points: We need asymptotically,

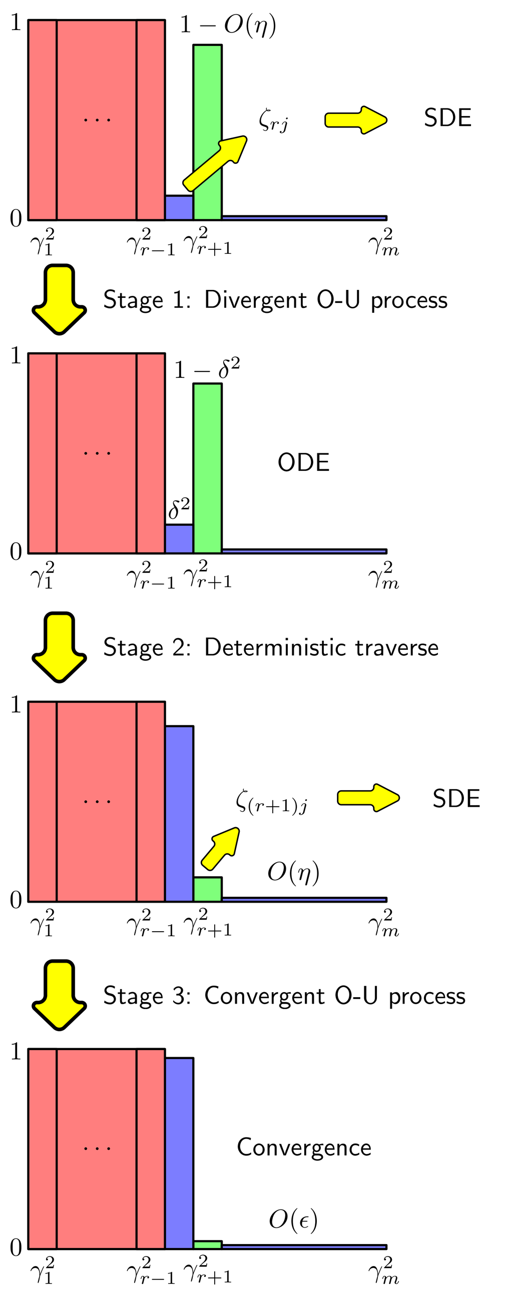
$$S_2 \asymp \frac{1}{\eta(\lambda_r - \lambda_{r+1})}\log\frac{1}{\delta^2}$$

iterations to reach the neighborhood of the global optima.

Stage 3. Convergence to Global Optima: We need asymptotically,

$$S_3 \asymp \frac{\log K'}{\eta(\lambda_r - \lambda_{r+1})}$$

iterations to converge to an ϵ-optimal solution.



Choosing $\eta \asymp \frac{(\lambda_r - \lambda_{r+1})\epsilon}{5rG_m}$, the total sample complexity is

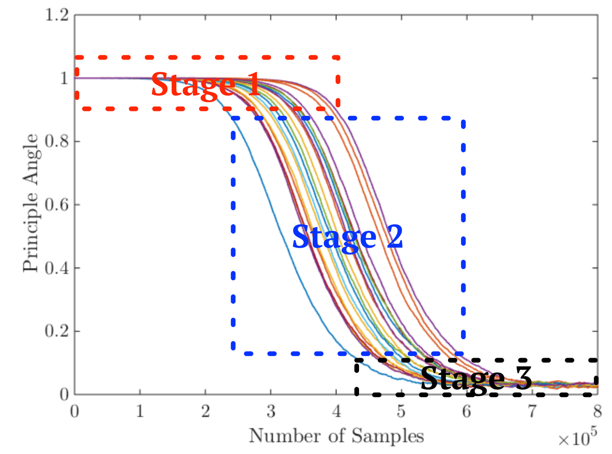
$$N = (S_1 + S_2 + S_3)h \asymp \frac{rG_m}{\epsilon(\lambda_r - \lambda_{r+1})^2}\log^2\frac{rG_m}{\epsilon(\lambda_r - \lambda_{r+1})}.$$

Experiments

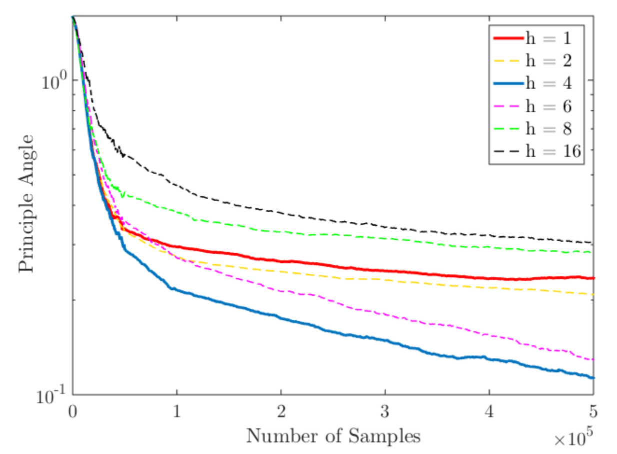
Simulated Data. Gaussian VAR model,

$$z_{k+1} = Az_k + \epsilon_k,$$

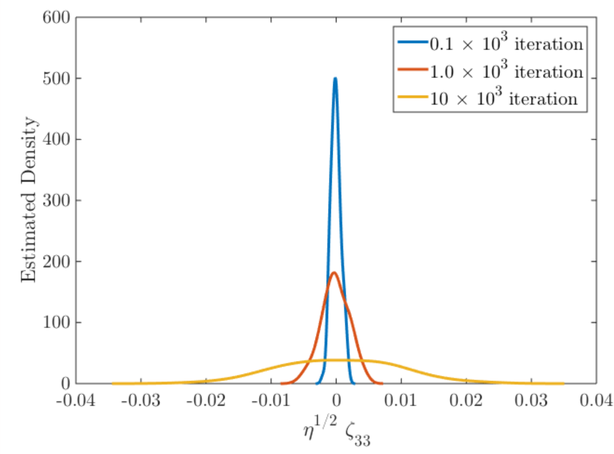
with $z_k \in \mathbb{R}^{16}$ and $\|A\|_2 < 1$ to guarantee stationarity.
▷ We aim to recover the first 3 largest eigenvalues of the covariance matrix Σ of the stationary distribution.



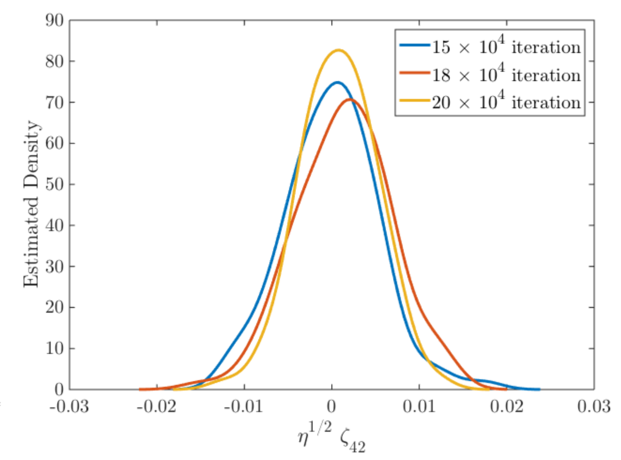
Solution trajectories



Different block sizes



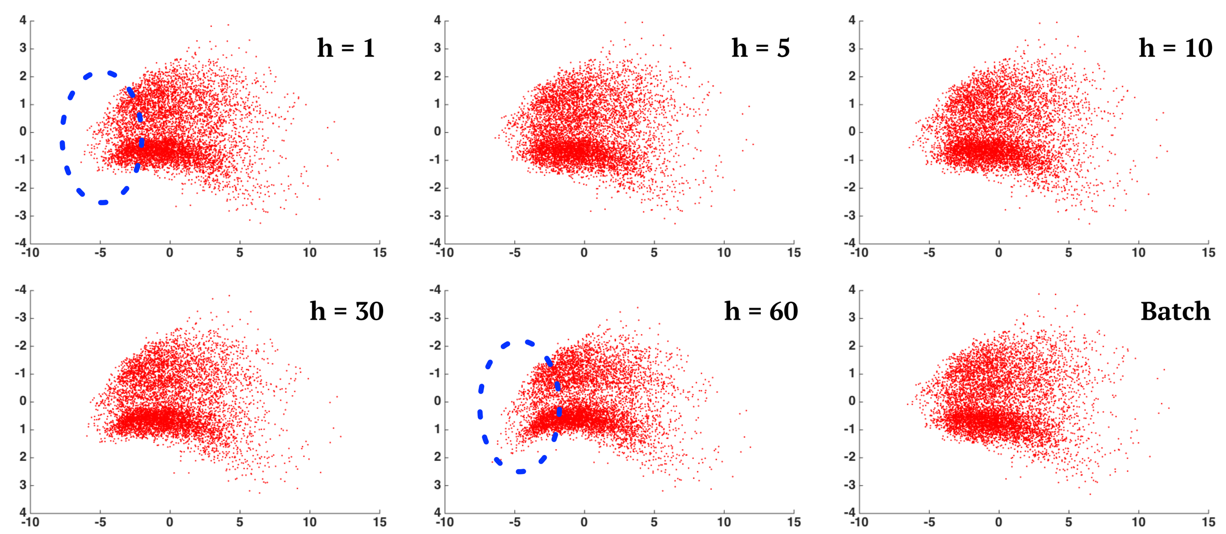
Distribution of ζ_{33}



Distribution of ζ_{42}

- We can clearly distinguish three stages;
- Trade-off between sample efficiency and convergence property;
- Estimated distributions of ζ_{33} and ζ_{42} over 100 runs roughly follow Gaussian distributions.

Real Data. Air Quality dataset with 9358 instances of concentrations of 9 different gases in a heavily polluted area.
▷ We aim to estimate the first 2 principle components.
▷ We project each data point onto the leading and the second principle components.



- $h = 1 \Longrightarrow$ some distortion in the circled area;
- $h = 3, h = 5 \Longrightarrow$ quite similar to Batch;
- As h increases to 30 or 60 $\Longrightarrow$ obvious distortion in the circled area again.