Provable Model-Parallel Distributed Principal Component Analysis with Parallel Deflation

Fangshuo Liao[1], Wenyi Su[1], Anastasios Kyrillidis[1,2]

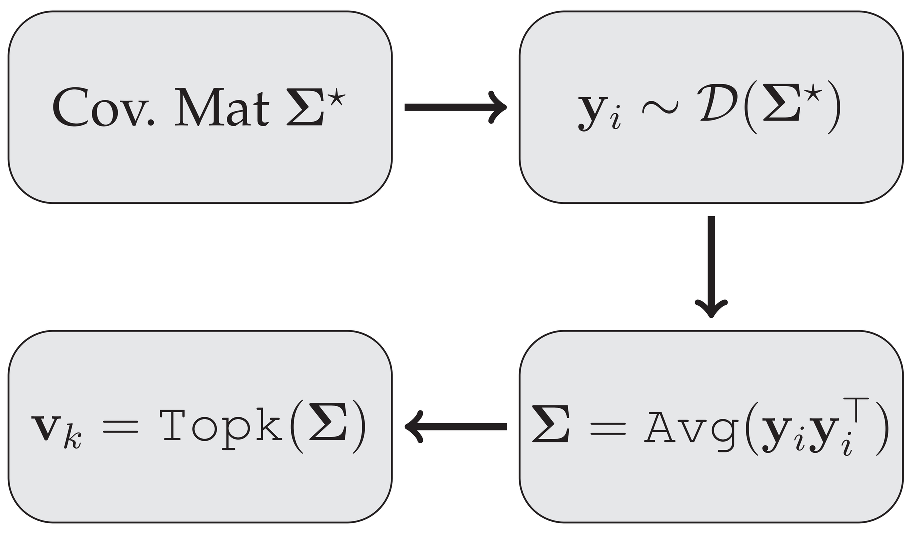
[1]Computer Science Department, Rice University

[2]Ken Kennedy Institute, Rice University

{Fangshuo.Liao,bs82,anastasios}@rice.edu

Background & Motivation

Principal Component Analysis.



Computing Top-1 Eigenvector.

$$\mathbf{u}_1^\star = \arg\max_{\mathbf{u}:\|\mathbf{u}\|_2=1} \mathbf{u}^\top\Sigma\mathbf{u}$$

This can be solved with linear convergence using e.g. power iteration, starting at initialized vector $\mathbf{v}$:

$$\mathbf{x}_0 = \mathbf{v}; \quad \hat{\mathbf{x}}_{t+1} = \Sigma\mathbf{x}_t; \quad \mathbf{x}_{t+1} = \frac{\hat{\mathbf{x}}_{t+1}}{\|\hat{\mathbf{x}}_{t+1}\|_2}.$$

Deflation Method. Gradually remove the solved eigenvector from the matrix.

$$\Sigma_1 = \Sigma; \quad \mathbf{v}_k = \texttt{Top1}\left(\Sigma_k, \hat{\mathbf{v}}_{k,\text{init}}, T\right);$$
$$\Sigma_{k+1} = \Sigma_k - \mathbf{v}_k\mathbf{v}_k^\top\Sigma_k\mathbf{v}_k\mathbf{v}_k^\top,$$

Central Question

*Can we design **model-parallel** distributed PCA methods based on deflation?*

Model-parallel: different workers are responsible for solving different eigenvectors.

Why model-parallel? Possibility of exploiting another level of parallelism that is independent of data-parallel computing [1, 2].

Why deflation? Solving $\mathbf{v}_k$ only needs the knowledge of Σ_k.

Game-Theoretical Perspective

As in the EigenGame[1] paper, we also study our algorithm under game-theoretic setting by viewing the solver of each eigenvector as a player. In parallel deflation, the kth solver maximizes the utility given by

$$\mathcal{V}_k\left(\mathbf{v} \mid \{\mathbf{v}_{k'}\}_{k'=1}^{k-1}\right) = \mathbf{v}^\top\Sigma\mathbf{v} - \sum_{k'=1}^{k-1}\mathbf{v}_{k'}^\top\Sigma\mathbf{v}_{k'}\cdot\left(\mathbf{v}_{k'}^\top\mathbf{v}\right)^2$$

Let $\mathbf{u}_k^\star$ be the kth eigenvector of Σ. (1)

Theorem 1. *Assume that the covariance matrix Σ has positive and strictly decreasing eigenvalues $\lambda_1^\star > \dots > \lambda_K^\star > 0$. Then, $\{\mathbf{u}_k^\star\}_{k=1}^K$ is the unique strict Nash Equilibrium defined by the utilities in (1) up to sign perturbation, i.e., replacing $\mathbf{u}_k^\star$ with $-\mathbf{u}_k^\star$.*

Under the condition that$\mathbf{v}_{k'} = \mathbf{u}_{k'}^\star$, $\mathcal{V}_k$ above is equivalent to the utility of EigenGame.

Reference

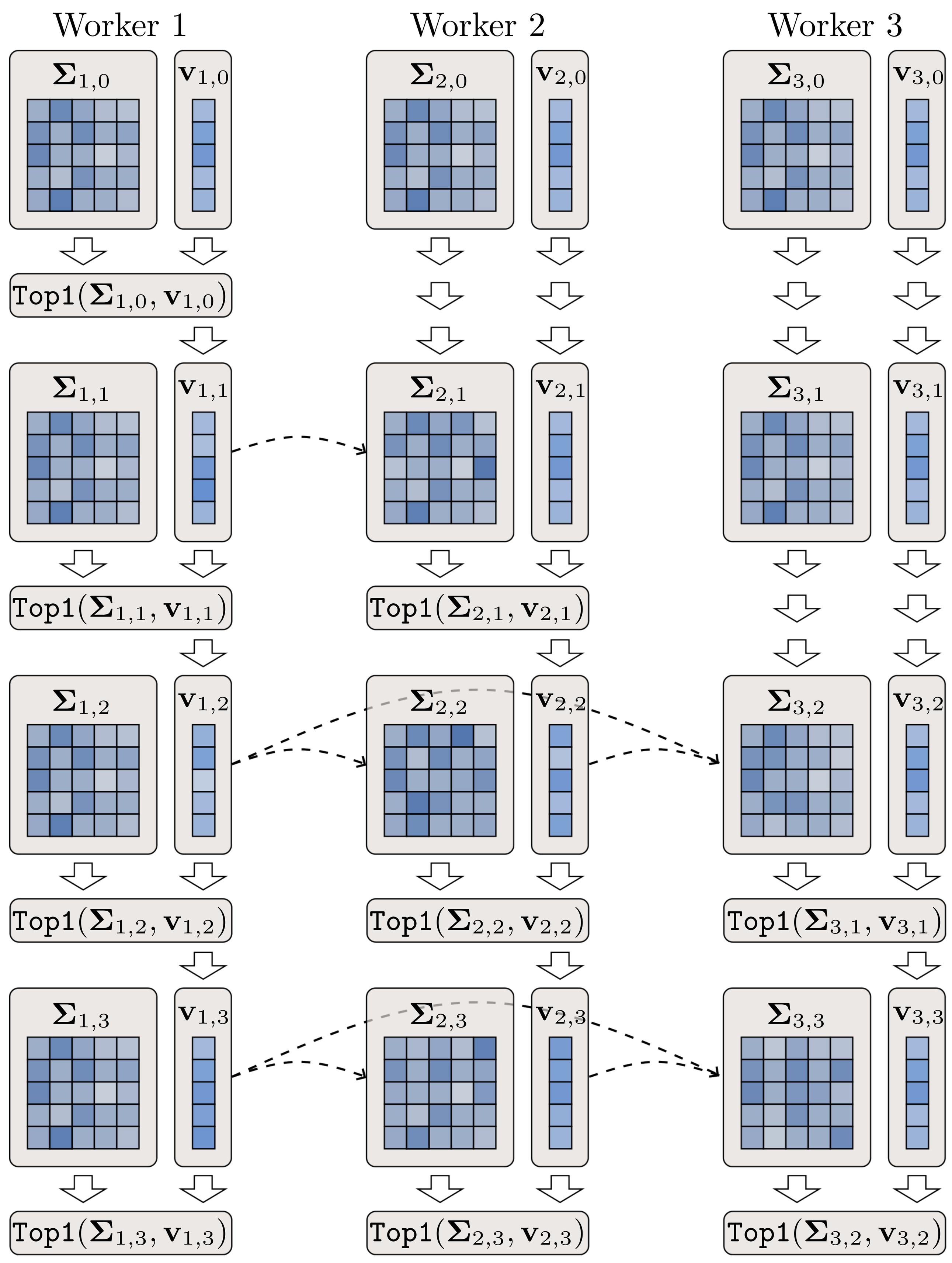
[1] Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. Eigengame: Pca as a nash equilibrium, 10 2020.

[2] Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. Eigengame unloaded: When playing games is better than optimizing, 2022.

[3] Fangshuo Liao, Junhyung Lyle Kim, Cruz Barnum, and Anastasios Kyrillidis. On the error-propagation of inexact deflation for principal component analysis, 2023.

Our Algorithm: Parallel Deflation

Breaking Sequential Dependency. Sequential dependency is important in previous algorithm because solving $\mathbf{v}_k$ need $\mathbf{v}_{k-1}$ to be *completely* solved. Our method provides a *rough estimation* of $\mathbf{v}_{k-1}$ to the solver of $\mathbf{v}_k$, and continuously provide improved versions of $\mathbf{v}_{k-1}$ later.



Mathematical Description. In the ℓ-th communication round, Worker k executes:

- Updating deflated matrix

$$\lambda_{k',\ell} = \mathbf{v}_{k',\ell-1}^\top\Sigma\mathbf{v}_{k',\ell-1}; \quad \forall k' \le k$$
$$\Sigma_{k,\ell} = \Sigma - \sum_{k'=1}^{k-1}\lambda_{k',\ell}\mathbf{v}_{k',\ell-1}\mathbf{v}_{k',\ell-1}^\top$$

- Updating eigenvector estimate

$$\mathbf{v}_{k,\ell} = \texttt{Top1}\left(\Sigma_{k,\ell}, \mathbf{v}_{k,\ell-1}\right); \quad \forall\ell \ge k.$$

Extension to Stochastic Setting. Let $\hat{\mathbf{Y}} \in \mathbb{R}^{n\times d}$ be a *mini-batch* of (properly scaled) data. Then $\Sigma \approx \hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}$. The algorithm can be rewritten as

- Compute eigenvalue estimations

$$\hat{\lambda}_{k',\ell} = \|\hat{\mathbf{Y}}\mathbf{v}_{k',\ell-1}\|_2^2; \quad \forall k' \in [k-1]$$

- Computation of matrix-vector product

$$\Sigma_{k,\ell}\mathbf{x} = \hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}\mathbf{x}_t - \sum_{k'=1}^{k-1}\hat{\lambda}_{k'}\left(\mathbf{v}_{k',\ell-1}^\top\mathbf{x}_t\right)\cdot\mathbf{v}_{k',\ell-1}.$$

- Apply the matrix-vector product computations to the `Top1` algorithm to obtain $\mathbf{v}_{k,\ell}$.

Convergence Analysis

Assumption 1. *[3] We assume that there exists a real value $\mathcal{F}\left(\hat{\Sigma}\right) \in (0,1)$ that depends on $\hat{\Sigma}$ such that for any $\mathbf{x}_0 \in \mathbb{R}^d$, $\texttt{Top1}(\cdot)$ satisfies $\|\texttt{Top1}(\hat{\Sigma},\mathbf{x}_0) - \mathbf{u}^\star\|_2 \le \mathcal{F}(\hat{\Sigma})\|\mathbf{x}_0 - \mathbf{u}^\star\|_2$.*

Theorem 2. *Assume that Assumption 1 holds, and let $\mathcal{F}_k = \max_{\ell\ge k}\mathcal{F}\left(\hat{\Sigma}_{k,\ell}\right)$. Let $\{m_k\}_{k=0}^K$ be defined recursively by $m_k = \max\left\{\mathcal{F}_k, \frac{1}{k} + \frac{k-1}{k}m_{k-1}\right\}$ and $m_0 = \mathcal{F}_1$. Let $\{s_k\}_{k=1}^n$ be defined as $s_1 = 1$ and for all $k \in [K-1]$ and $k' \in [k]$:*

$$s_{k+1} \ge s_k + O\left(\max\left\{\left(\log\frac{1}{m_k}\right)^{-1}\left(1 + \log\frac{k\lambda_k^\star}{\lambda_{k+1}^\star - \lambda_{k+2}^\star}\right), \frac{km_k+1}{1-m_k}\right\}\right). \quad (2)$$

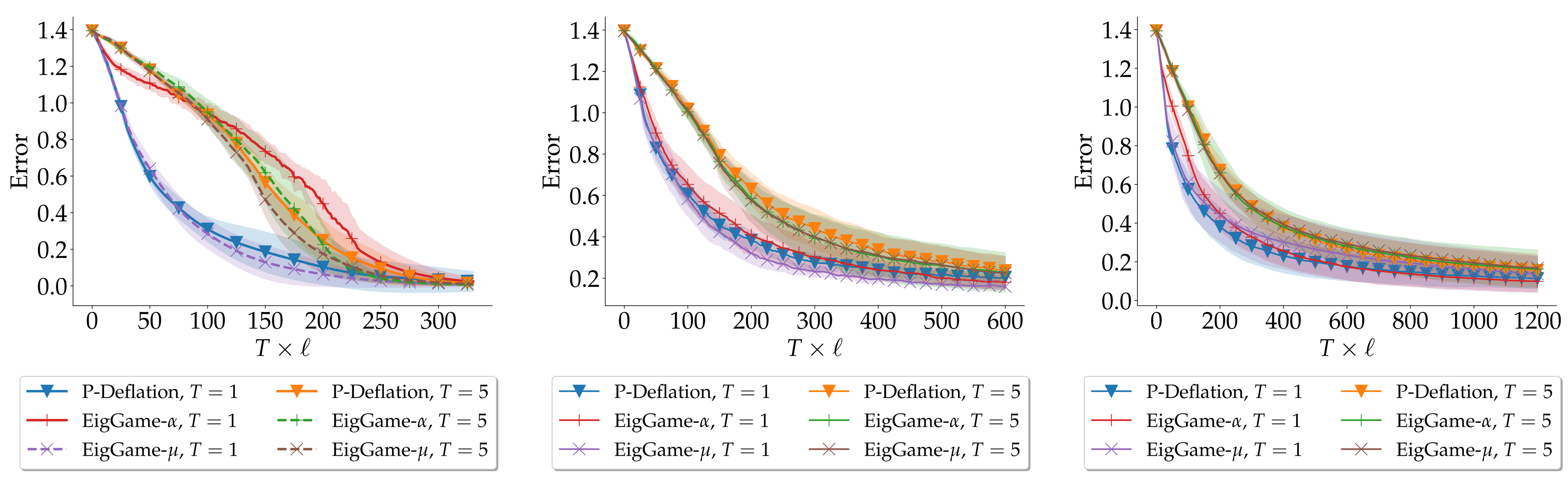
Then, we have that the following holds for all $k \in [K]$

$$\|\mathbf{v}_{k,\ell} - \mathbf{u}_k^\star\|_2 \le O\left((\ell - s_k)\, m_k^{\ell - s_k}\right); \quad \forall\ell \ge s_k - 1. \quad (3)$$

Interpretation of the theorem.

- (3) shows convergence of the kth eigenvector with rate $m_k \in (0,1)$, starting at round s_k.
- (2) characterize the gap between the two consecutive convergence starting point, s_k and s_{k+1}.

Experimental Result



Comparison of the convergence behavior of parallel deflation, EigenGame-α, and EigenGame-μ (left). in deterministic setting on synthetic dataset with power-law decaying eigenvalues, (middle)in stochastic setting on synthetic dataset with power-law decaying eigenvalues, and (right). in stochastic setting on MNIST dataset.

Acknowledgement

This work is supported by NSF CMMI no. 2037545, NSF CAREER award no. 2145629, Welch Foundation Grant no. A22-0307, a Microsoft Research Award, an Amazon Research Award, and a Ken Kennedy Research Cluster award (Rice K2I). AK was supported in part through seed funding from the Ken Kennedy Institute at Rice University.