The Randomized Golub-Klema-Stewart Algorithm

Robin Armstrong

PhD Student, Cornell University

Center for Applied Mathematics

BASED ON WORK SUPPORTED BY THE NATIONAL SCIENCE FOUNDATION.

Joint work with…

Alex Buzali
MS Student, Harvard University School of Engineering and Applied Sciences

Anil Damle
Professor, Cornell University Department of Computer Science

PART 1: BACKGROUND

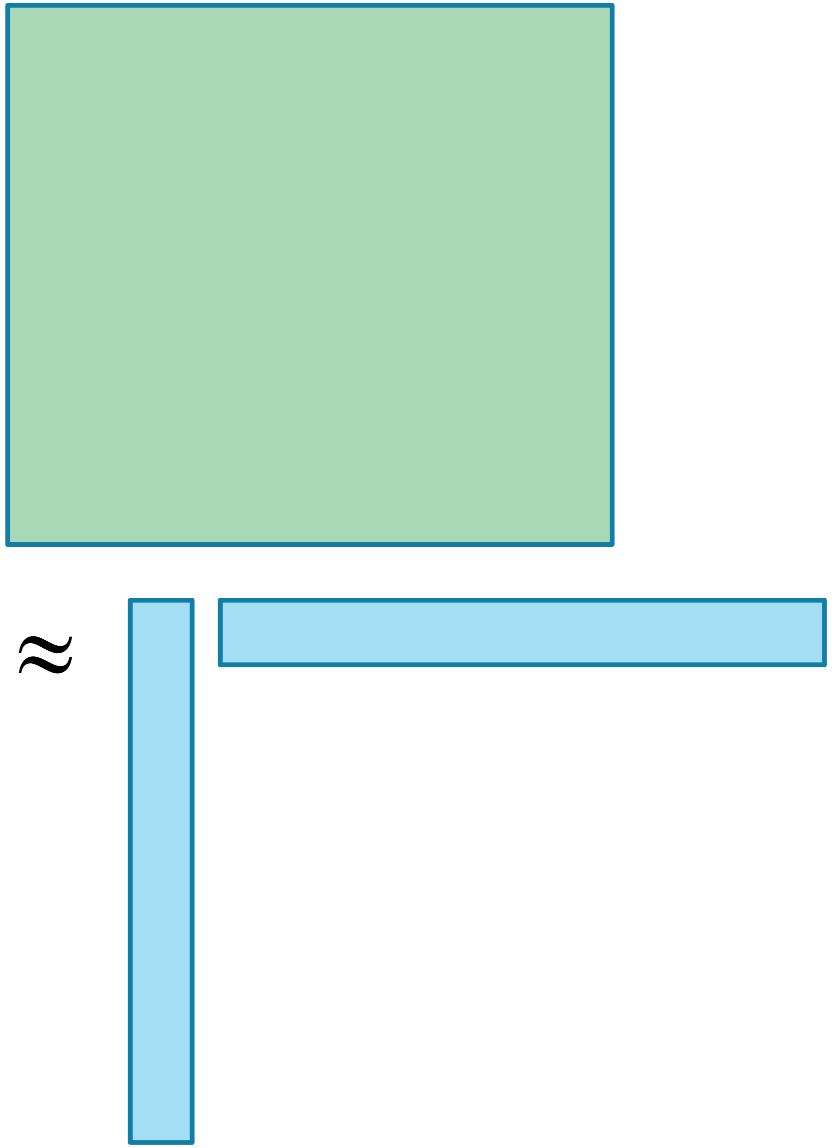
Low-Rank Approximations

- How to compress a large matrix into one of lower rank?
- Gold standard: the singular value decomposition,

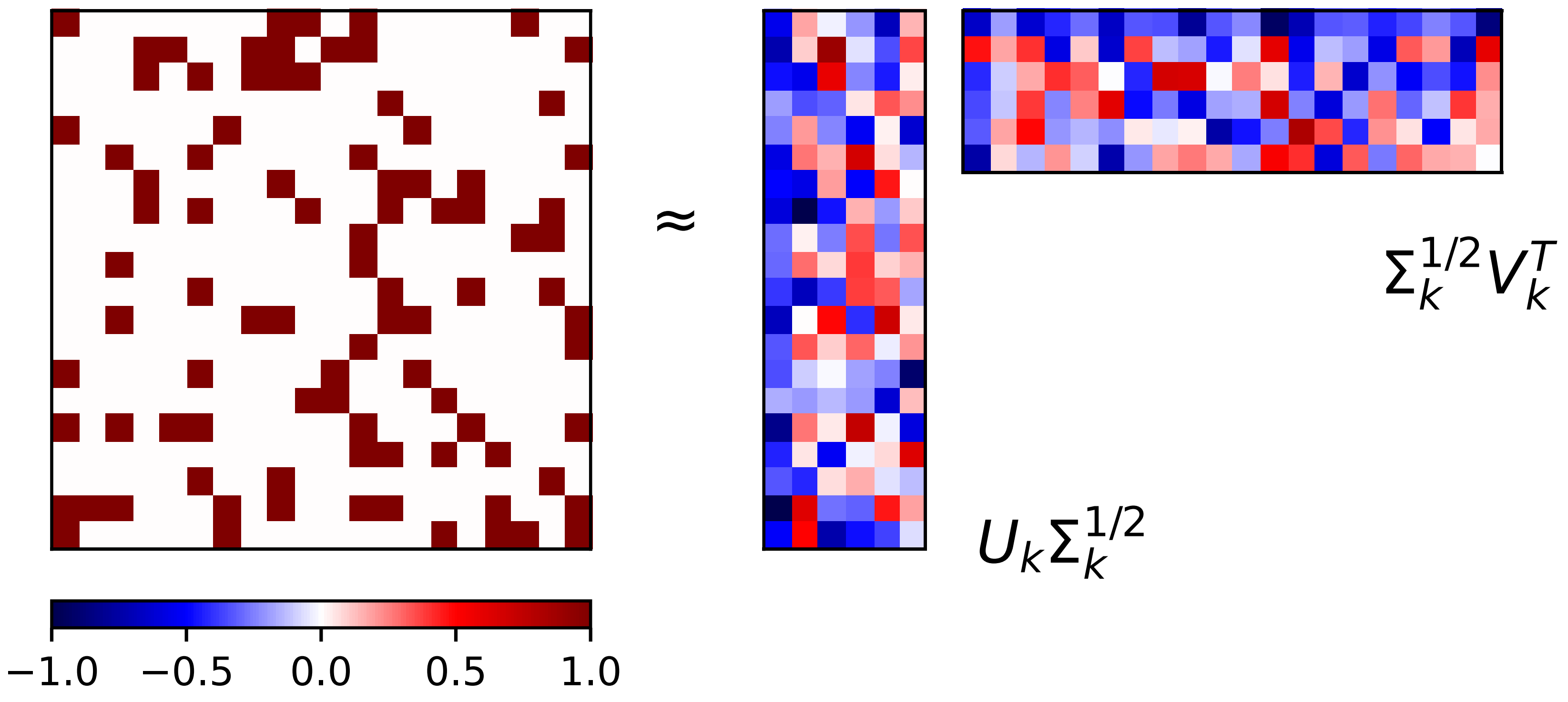
$$A = U\Sigma V^T = \begin{bmatrix} U_k & U_\perp \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_\perp \end{bmatrix} \begin{bmatrix} V_k^T \\ V_\perp^T \end{bmatrix}.$$

- **Optimal rank-k approximation is $U_k \Sigma_k V_k^T$.**
- **Optimal error is $\|\Sigma_\perp\|$.**
- What really matters are the **leading singular subspaces**:

$$\mathcal{U}_k = \mathrm{range}(U_k), \qquad \mathcal{V}_k = \mathrm{range}(V_k).$$



SVD: Optimal, But Very Unstructured



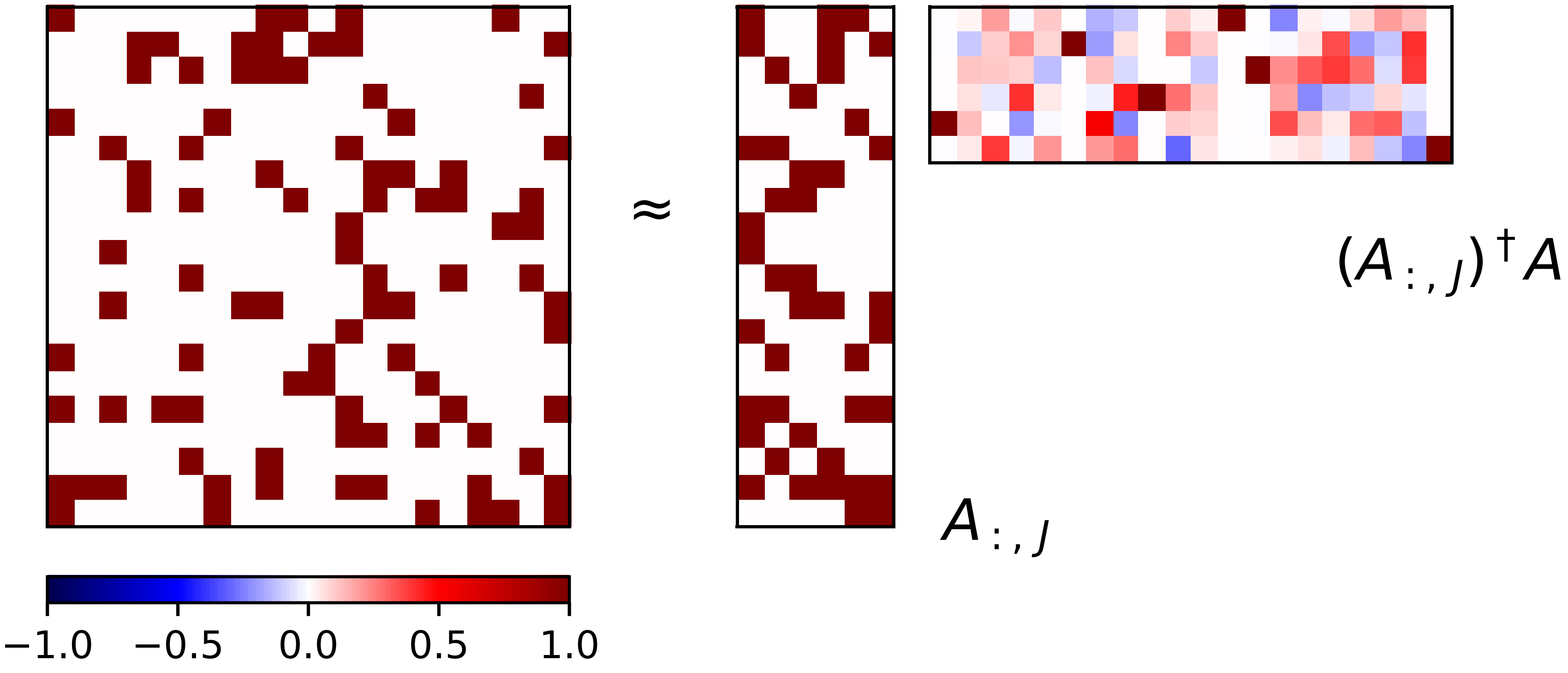
Interpolative Decompositions

- We might prefer to **approximate *A*'s columns in terms of other columns**.
- We want ***J*, a set of *k* column indices**, such that

$$\|A - A_{:,\,J}(A_{:,\,J})^{\dagger}A\| \approx \|\Sigma_{\perp}\|.$$

- This is called an **interpolative decomposition**. Applications in:
  - Gaussian process regression,
  - Neural network pruning,
  - Electronic structure theory,
  - And more…
- Also possible to interpolate using rows, or rows and columns (CUR decomposition).

Much More Structured!



How To Choose The Basis Columns?

- Projecting into $\mathcal{U}_k$ is optimal, so **try to make range$(\boldsymbol{A}_{:,J}) \approx \boldsymbol{\mathcal{U}_k}$.**
- Let $\phi_{\max}(J) =$ largest principal angle between $\mathrm{range}(A_{:,J})$ and $\mathcal{U}_k$.

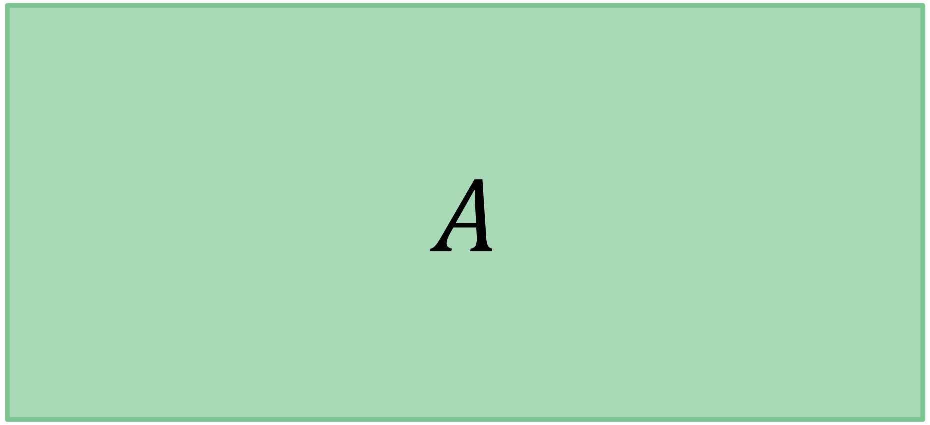
Theorem (Golub, Klema, and Stewart, 1976)[1].

$$\sin \phi_{\max}(J) \leq \frac{\sigma_{k+1}(A)}{\sigma_k(A)\sigma_{\min}(V_{J,\,1:k})}.$$

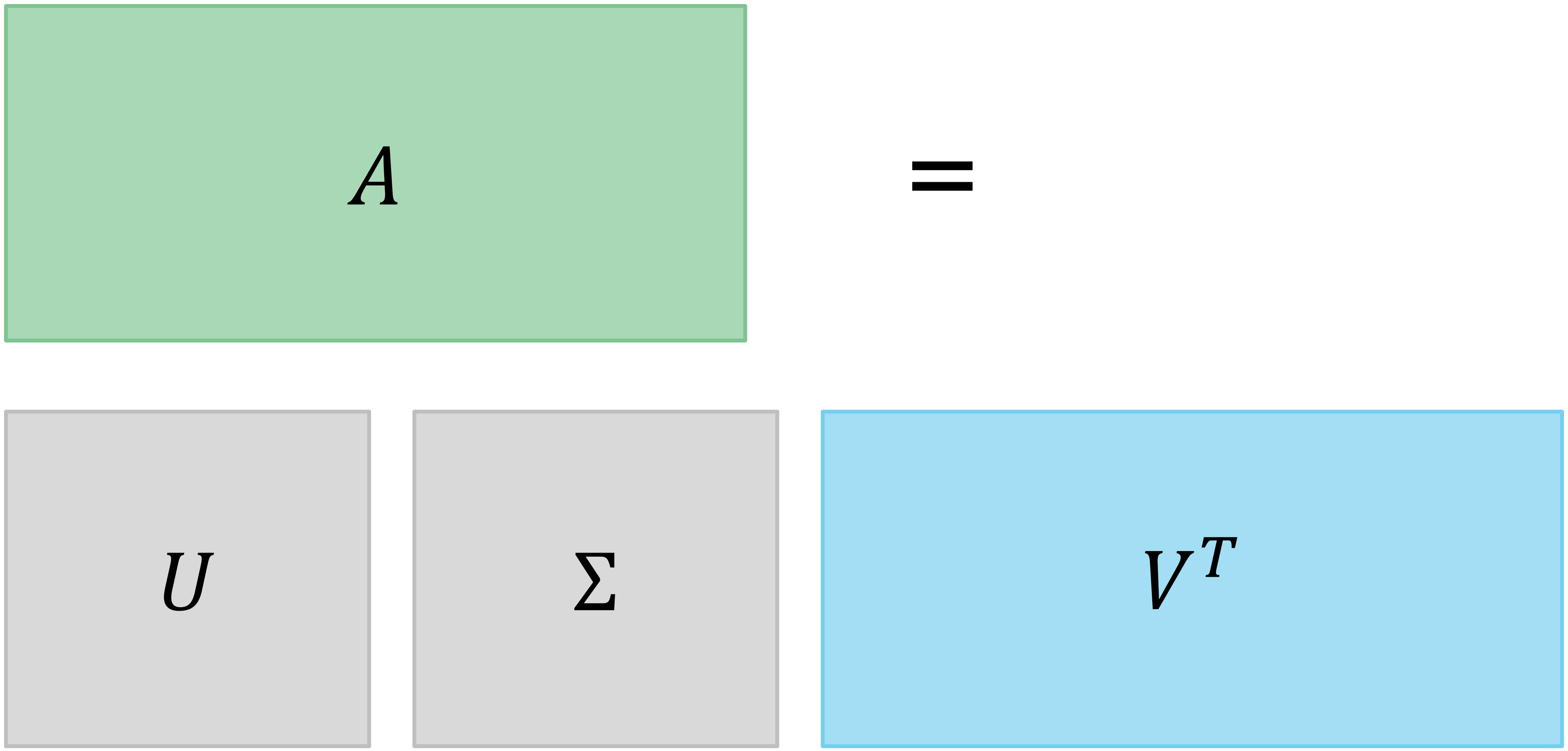
- Leads to the **Golub-Klema-Stewart (GKS) Algorithm[1].**

1. G.H. Golub, V. Klema, and G. Stewart, *Rank Degeneracy and Least Squares Problems*, Tech. Report STAN-CS-76-559, Stanford University, 1976.

The GKS Algorithm



The GKS Algorithm

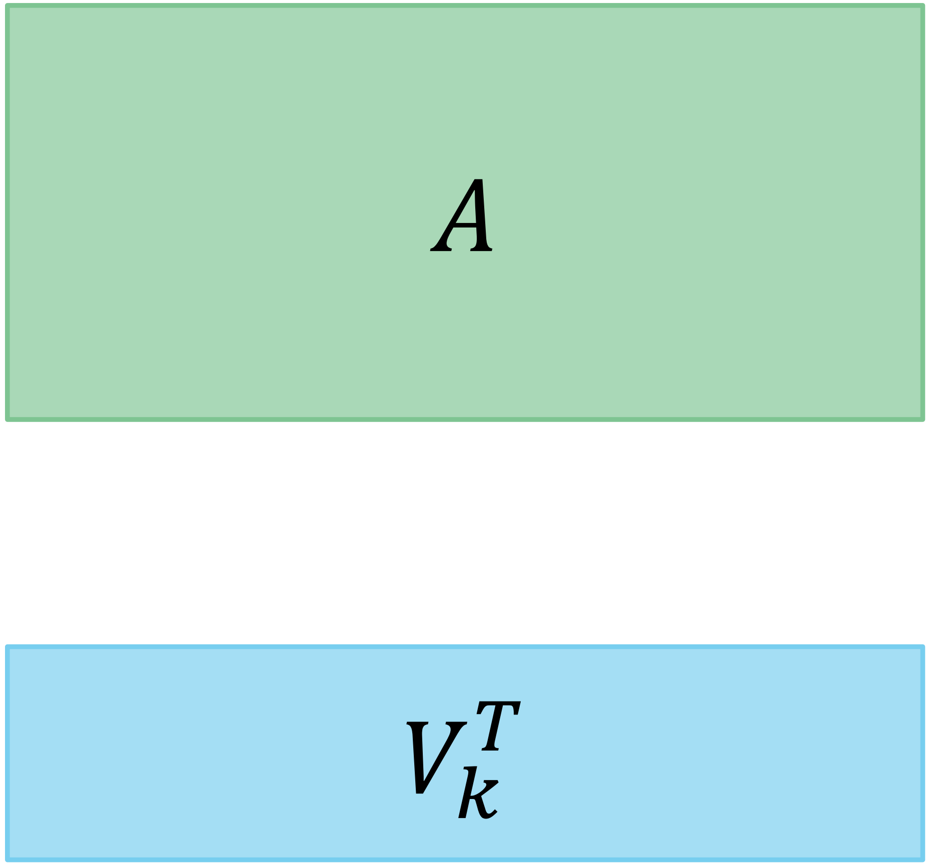


The GKS Algorithm

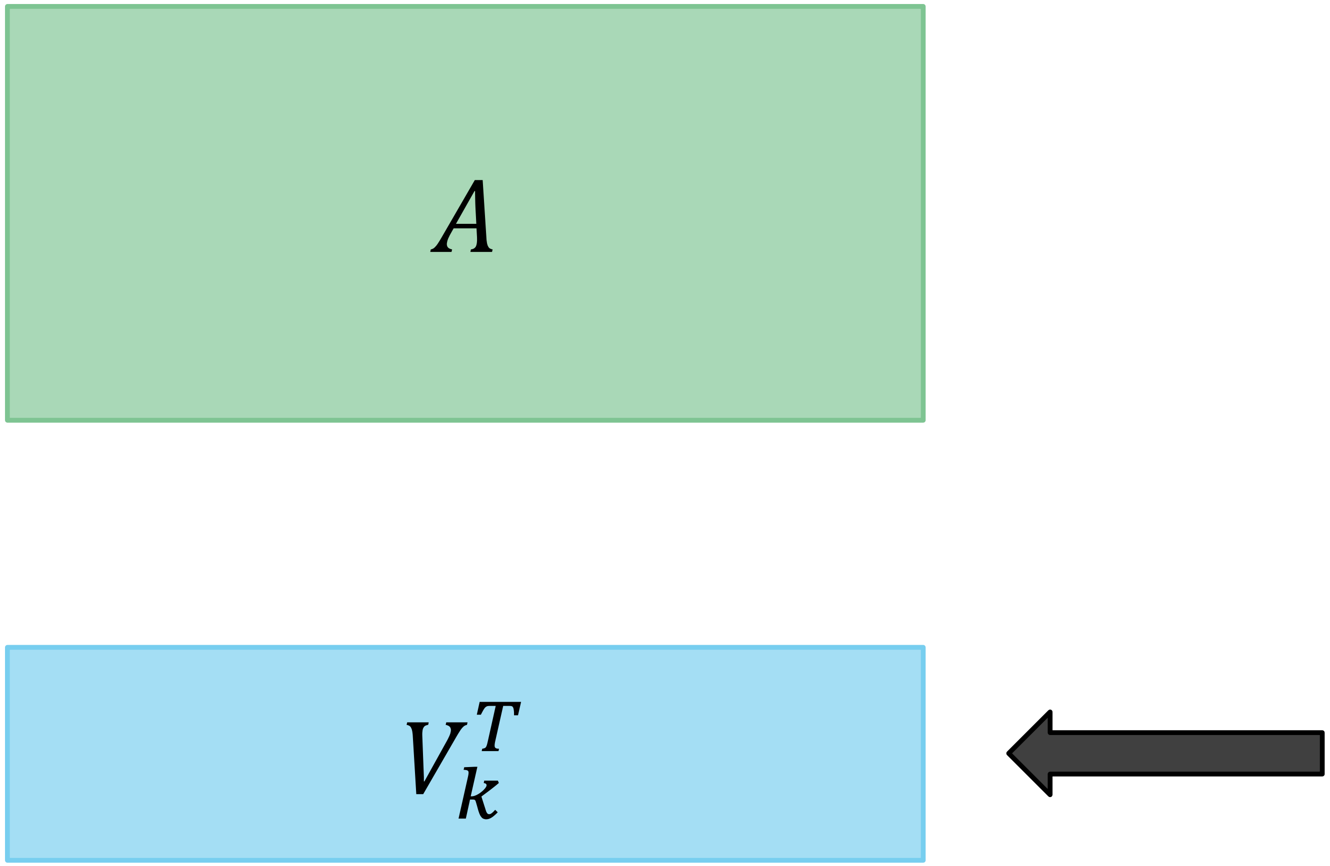
$$A = \begin{bmatrix} U_k & U_\perp \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_\perp \end{bmatrix} \begin{bmatrix} V_k^T \\ V_\perp^T \end{bmatrix}$$

The GKS Algorithm

The GKS Algorithm



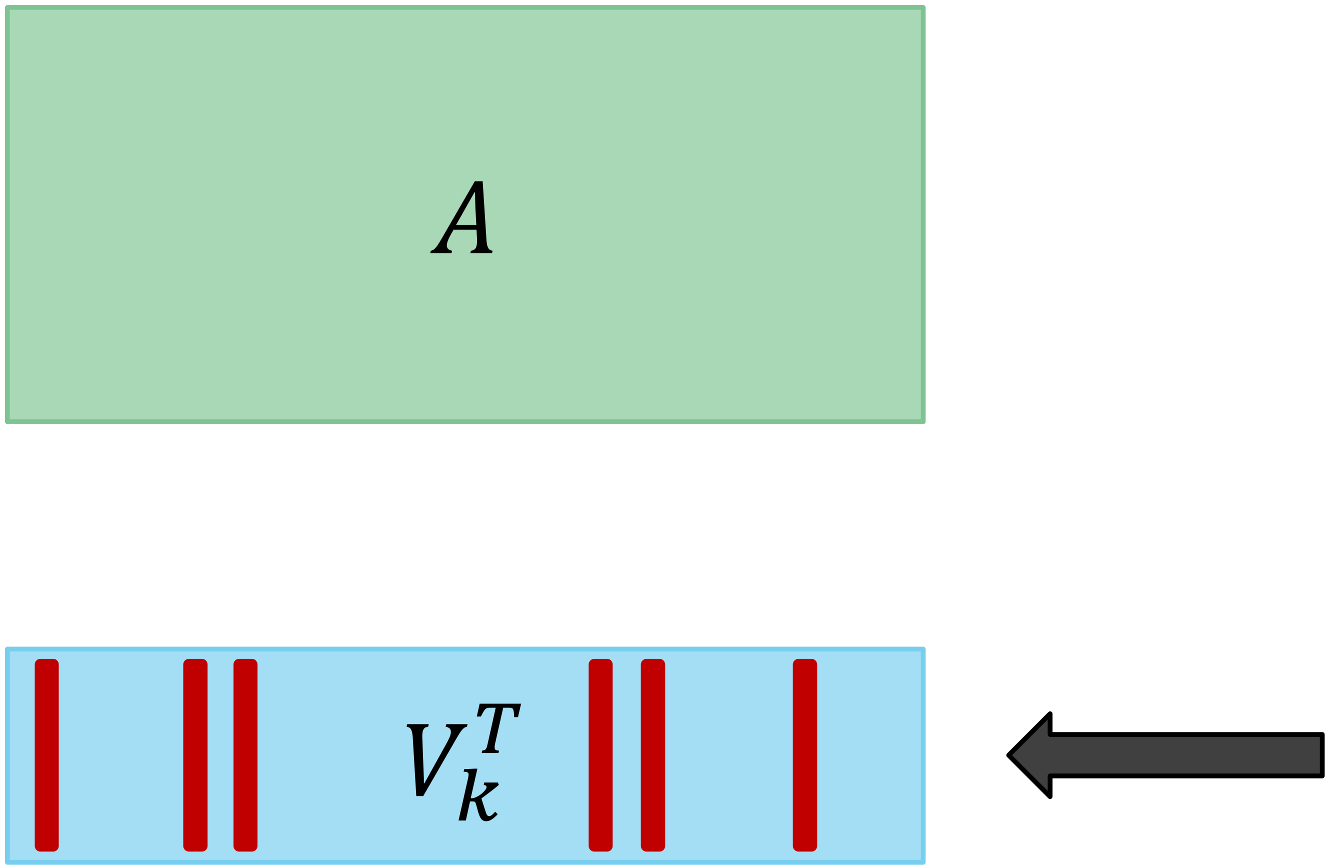
The GKS Algorithm



rank-revealing QR factorization:

$$V_k^T[\Pi_1\ \Pi_2] = Q[R_1\ \ R_2]$$

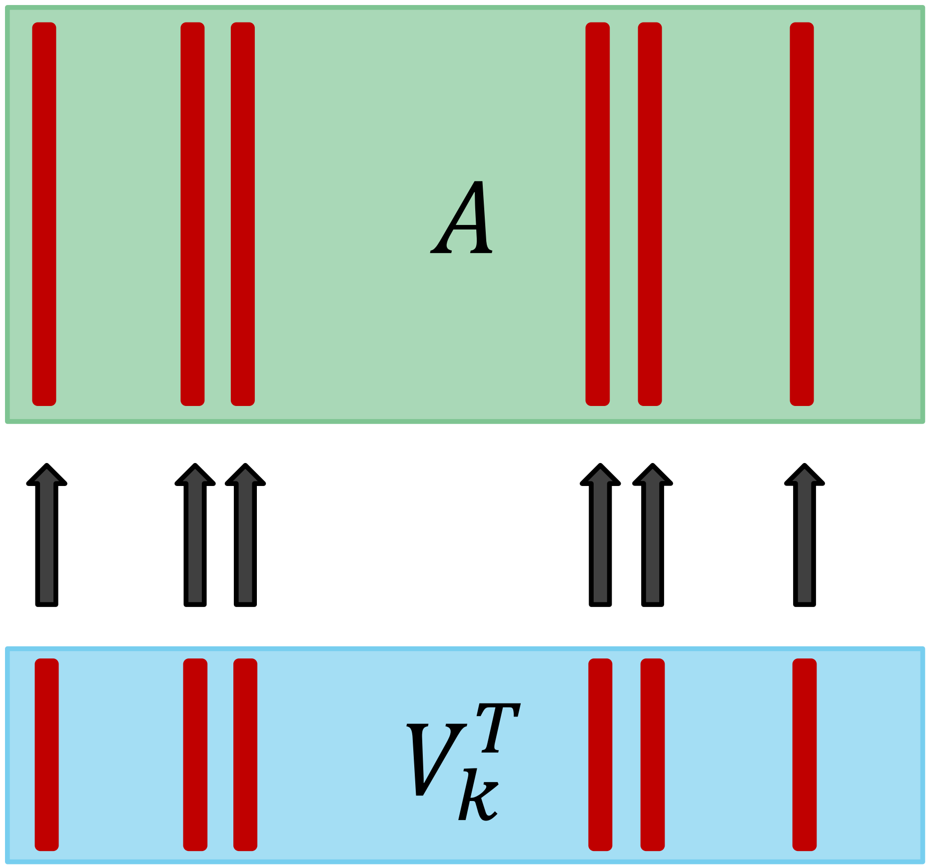
The GKS Algorithm



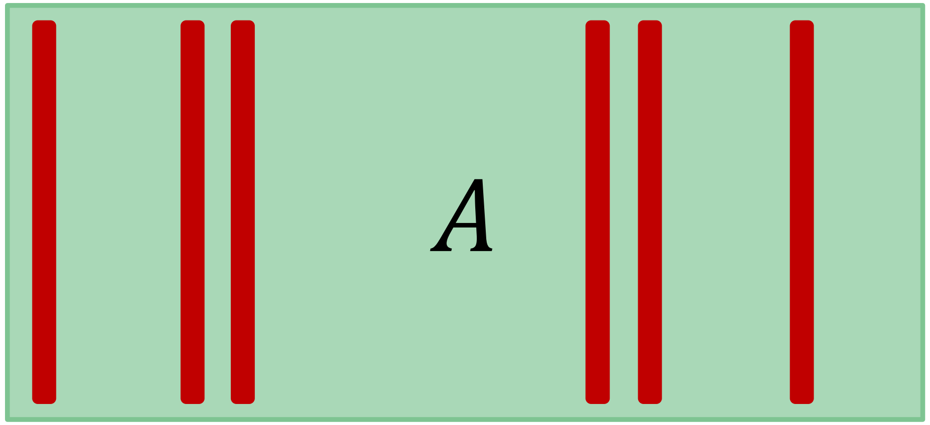
rank-revealing QR factorization:

$$V_k^T[\Pi_1 \; \Pi_2] = Q[R_1 \;\; R_2]$$

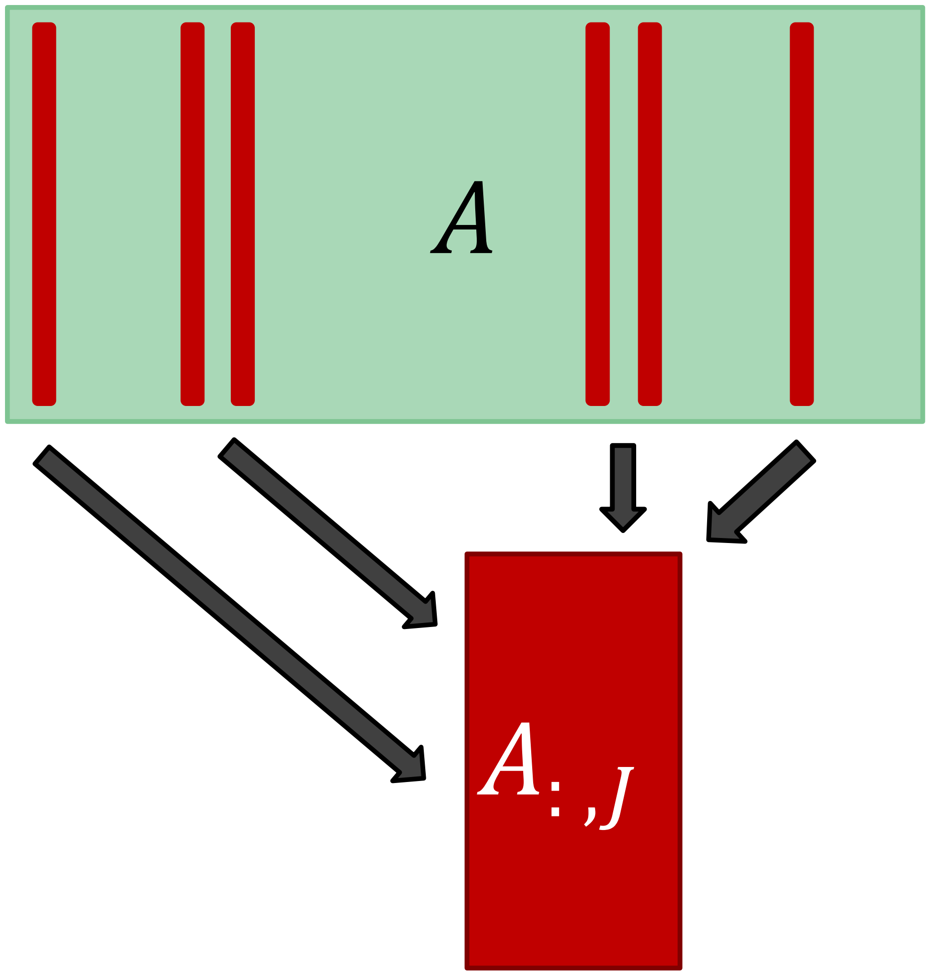
The GKS Algorithm



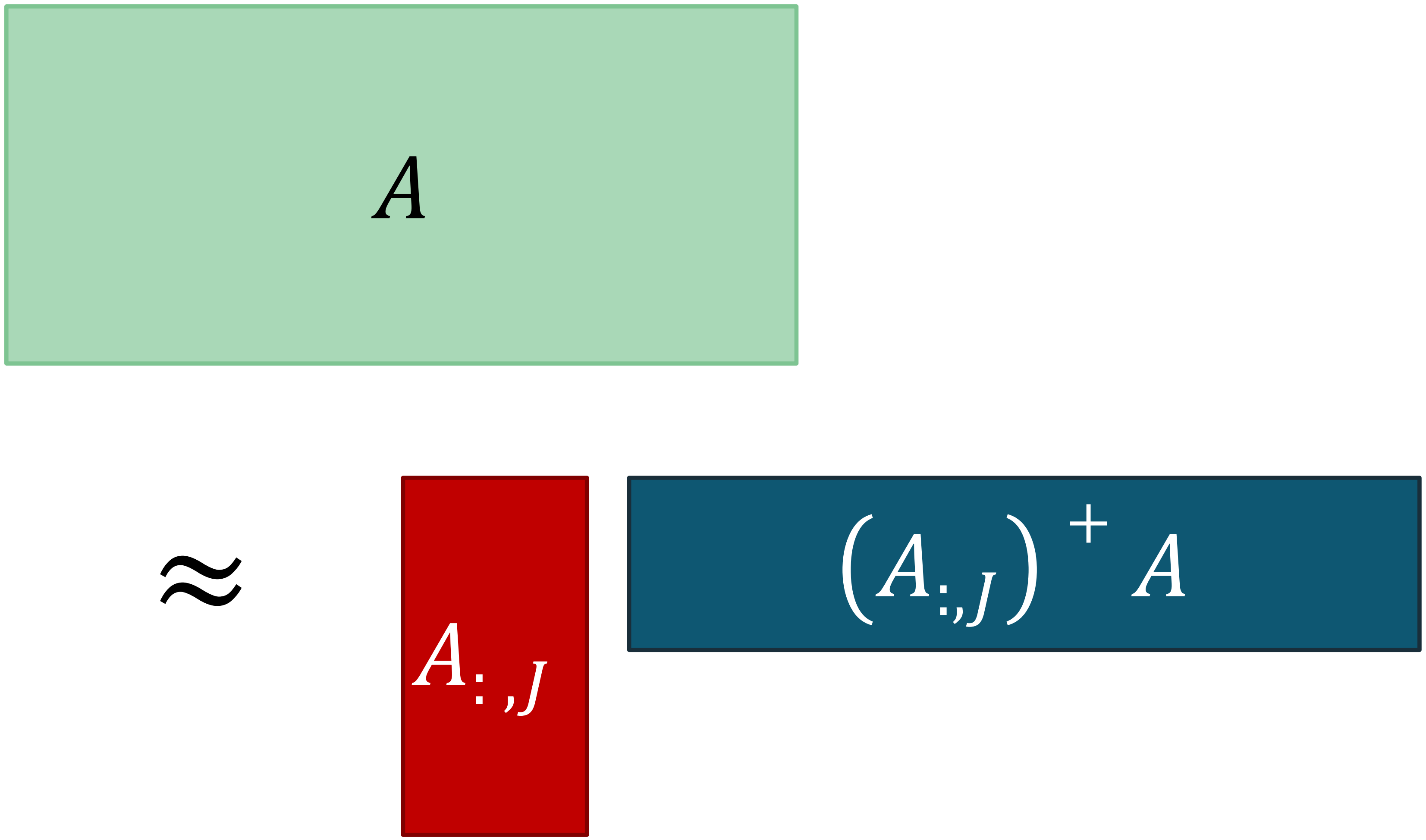
The GKS Algorithm



The GKS Algorithm



The GKS Algorithm



The GKS Algorithm

A

Pseudocode:

1. $U_k, \Sigma_k, V_k \leftarrow \text{svd}(A)$.
2. $\Pi, Q, R \leftarrow \text{rrqr}(V_k^T, k)$.
3. return $A_{:,J} = A\Pi_{:,1:k}$.

$$\approx A_{:,J} \; \left(A_{:,J}\right)^+ A$$

PART 2: THE RGKS ALGORITHM

GKS Algorithm

1. $U_k, \Sigma_k, V_k \leftarrow \text{svd}(A)$.
2. $\Pi, Q, R \leftarrow \text{rrqr}(V_k^T, k)$.
3. **return** $A_{:,J} = A\Pi_{:,1:k}$.

- Replace the SVD with a **randomized SVD**[2].

RGKS Algorithm

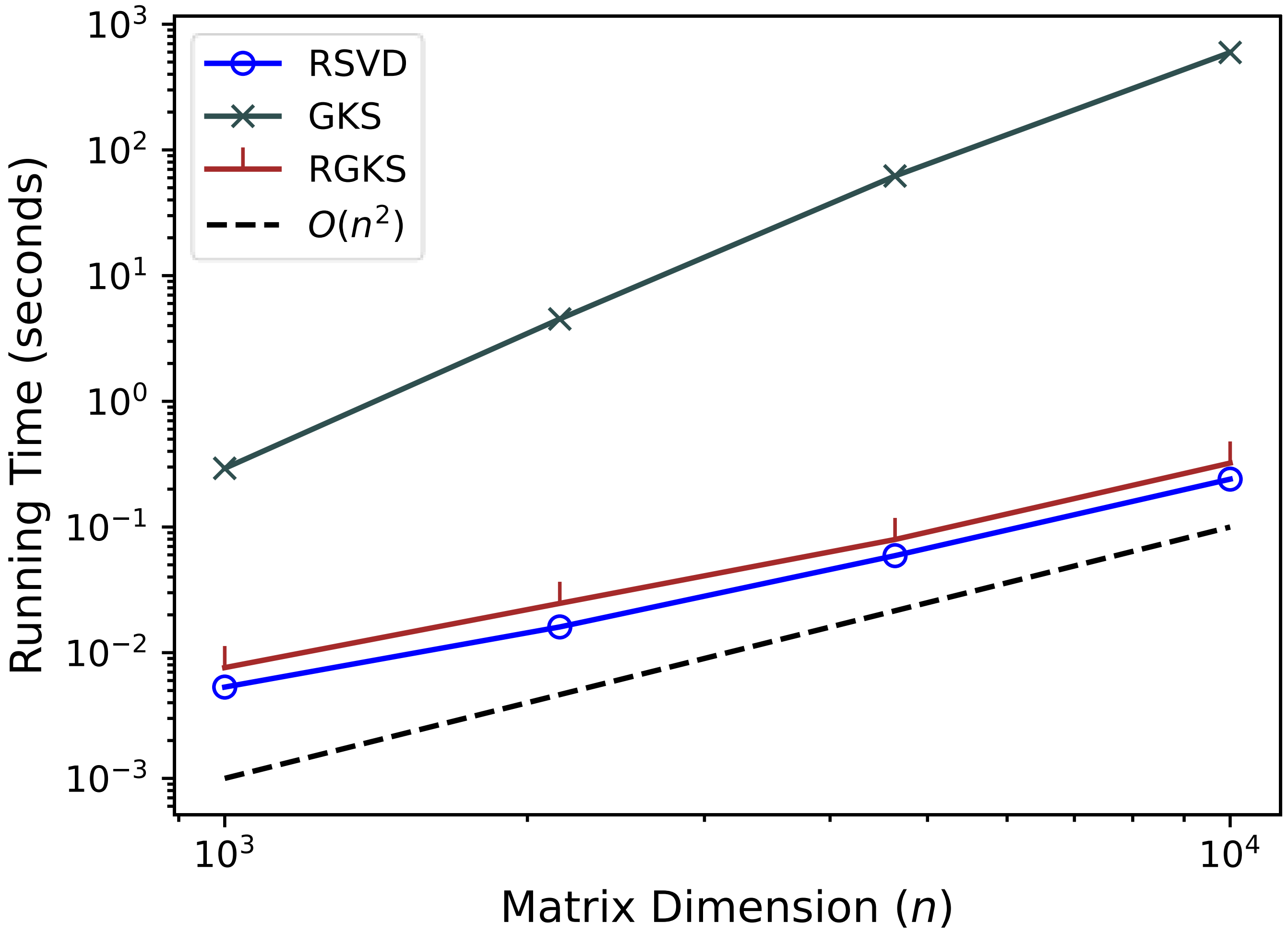
1. $\hat{U}_k, \hat{\Sigma}_k, \hat{V}_k \leftarrow \textbf{randomized\_svd}(A, k, p, q)$
2. $\Pi, Q, R \leftarrow \text{rrqr}(\hat{V}_k^T, k)$.
3. **return** $A_{:,J} = A\Pi_{:,1:k}$.

- p = oversampling, q = power iterations.
- Like GKS, but with approximated subspaces:

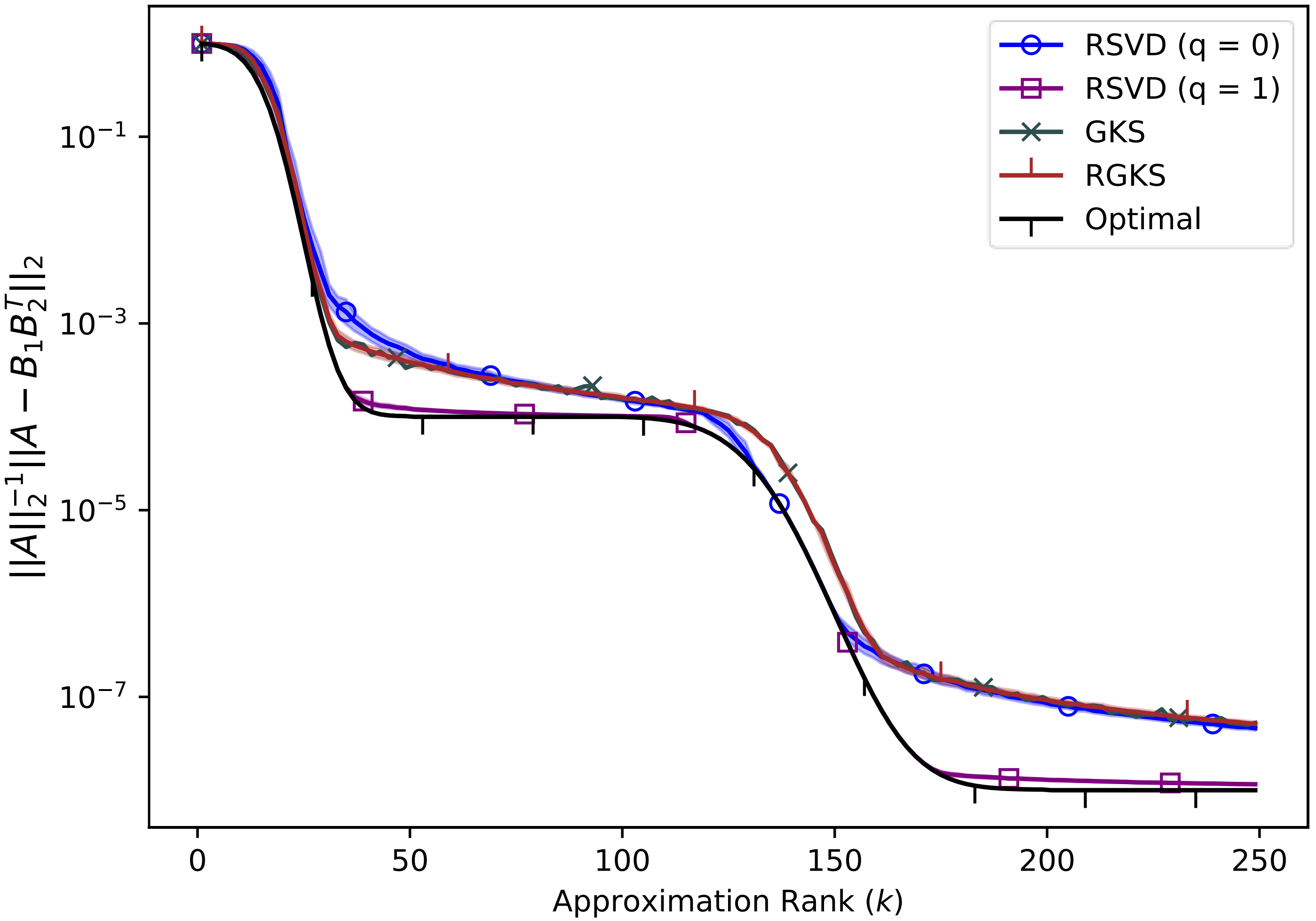
$$\widehat{\mathcal{U}}_k = \text{range}(\widehat{U}_k), \qquad \widehat{\mathcal{V}}_k = \text{range}(\widehat{V}_k).$$

2. N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding Structure With Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Review, 53 (2011).

RGKS Is Efficient



RGKS Is Accurate



PART 3: ERROR ANALYSIS OF RGKS

Reframing RGKS

- RGKS uses A's columns to approximate $\mathcal{U}_k$.
- Does so by making $\sigma_{\min}(V_{J,1:k})$ large. **Why does this work?**

$$\mathrm{range}(A_{:,\,J}) \xleftarrow{\;A\;} \mathrm{span}\{e_j : j \in J\}$$

$$\mathcal{U}_k \xleftarrow{\;A\;} \mathcal{V}_k$$

- Let $\varphi_i(J)$ = principal angles between $\mathrm{span}\{e_j : j \in J\}$ and $\mathcal{V}_k$.
- $\boldsymbol{\varphi_{\max}(J) = \max_i \varphi_i = \arccos(\sigma_{\min}(V_{J,1:k}))}$.

Theorem. Provided that $k \leq n/2$ and $\varphi_{\max}(J) < \pi/2$,

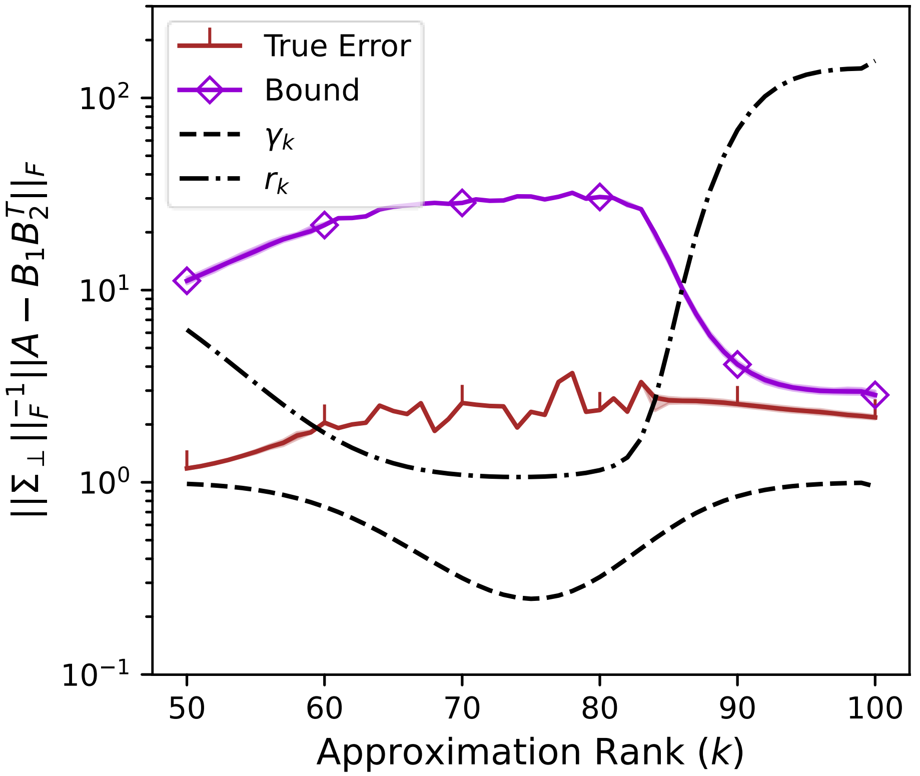
$$\|A - A_{:,J}(A_{:,J})^{\dagger}A\|_{\mathrm{F}} \leq \|\Sigma_{\perp}\|_{\mathrm{F}}\sqrt{1 + \frac{1}{r_k}\sum_{i=1}^{k}\tan^2\varphi_i(J)},$$

where $r_k = \sigma_{k+1}(A)^{-2}\sum_{i\geq k+1}\sigma_i(A)^2$ is the **residual stable rank.**

- GKS guarantees $\tan\varphi_{\max}(J) \leq \sqrt{q(n,k)^2 - 1}$, where q depends on the RRQR algorithm.

$$\widehat{\mathcal{V}}_k \xleftrightarrow{\;\widehat{\varphi}_{\max}\;} \mathrm{span}\{e_j : j \in J\} \xleftrightarrow{\;\varphi_{\max}\;} \mathcal{V}_k$$

- RGKS guarantees $\tan\widehat{\boldsymbol{\varphi}}_{\mathbf{max}} \leq \sqrt{q(n,k)^2 - 1}$, where **$\widehat{\boldsymbol{\varphi}}_{\mathbf{max}}$ = largest angle between $\widehat{\boldsymbol{\mathcal{V}}}_k$ and $\mathrm{span}\{\boldsymbol{e_j} : \boldsymbol{j} \in \boldsymbol{J}\}$.**



Relating $\widehat{\varphi}_{\max}(J)$ And $\varphi_{\max}(J)$

- Let $\theta_{\max}$ = largest principal angle between $\hat{\mathcal{V}}_k$ and $\mathcal{V}_k$.

Theorem. $\varphi_{\max}(J) \leq \hat{\varphi}_{\max}(J) + \theta_{\max}$.

- We get a bound for RGKS:

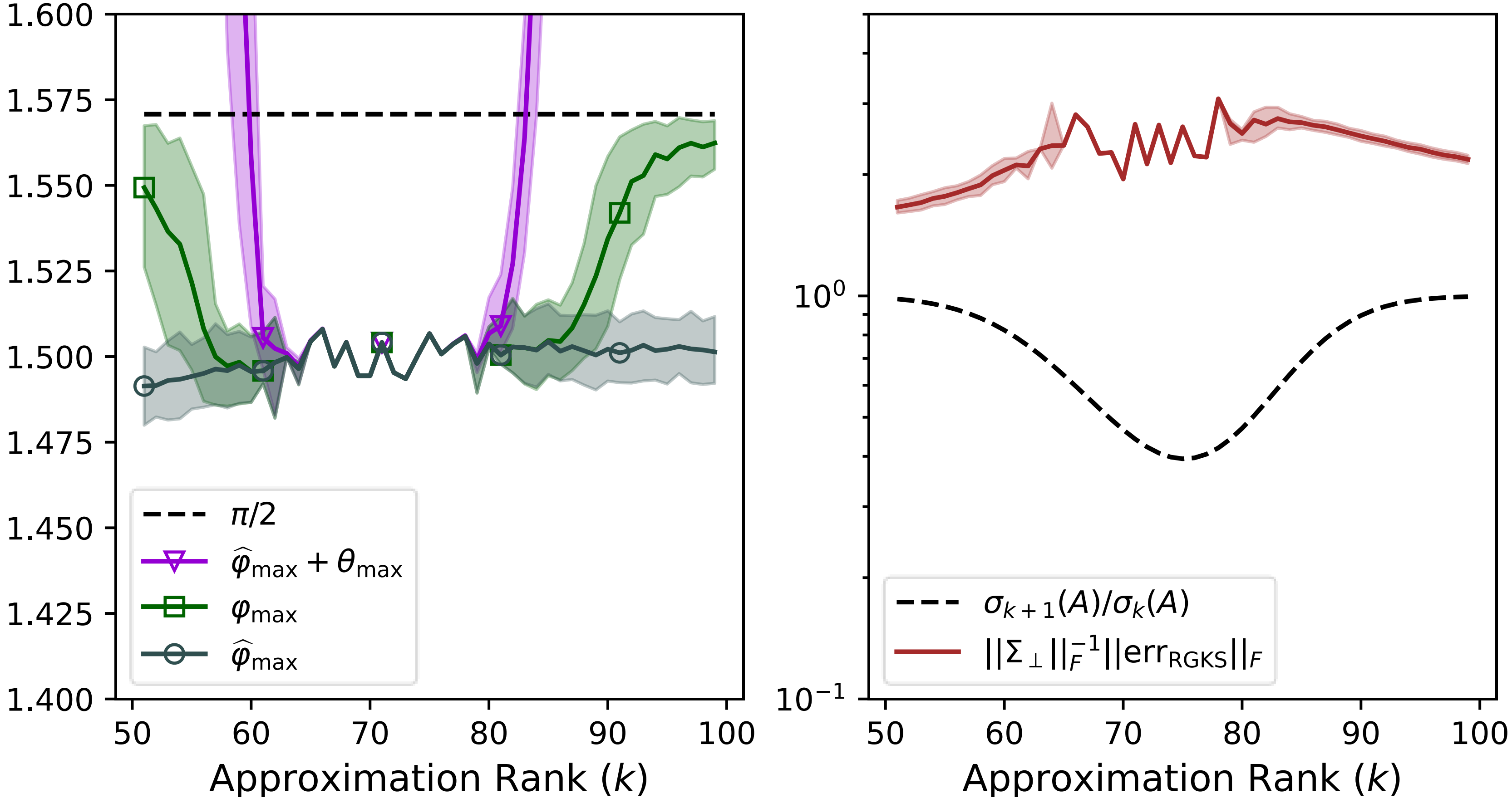
$$\|A - A_{:,J}(A_{:,J})^{\dagger}A\|_{\mathrm{F}} \leq \|\Sigma_{\perp}\|_{\mathrm{F}}\sqrt{1 + \frac{k}{r_k}\tan^2(\widehat{\varphi}_{\max}(J) + \theta_{\max})},$$

...where $\theta_{\max}$ captures randomization errors.

- It's known[3] that $\theta_{\max}$ is small when $\sigma_{k+1}(A) \ll \sigma_k(A)$.

3. A. K. Saibaba, Randomized Subspace Iteration: Analysis of Canonical Angles and Unitarily Invariant Norms, *SIAM Journal on Matrix Analysis and Applications*, 40 (2019).

Tight When $\sigma_{k+1}(A) \ll \sigma_k(A)$



What About When $\sigma_{k+1}(A) \approx \sigma_k(A)$?

- $\theta_{\max} \approx \pi/2$, but **RGKS still performs well.**
- $\theta_{\max}$ measures **aggregate** errors:

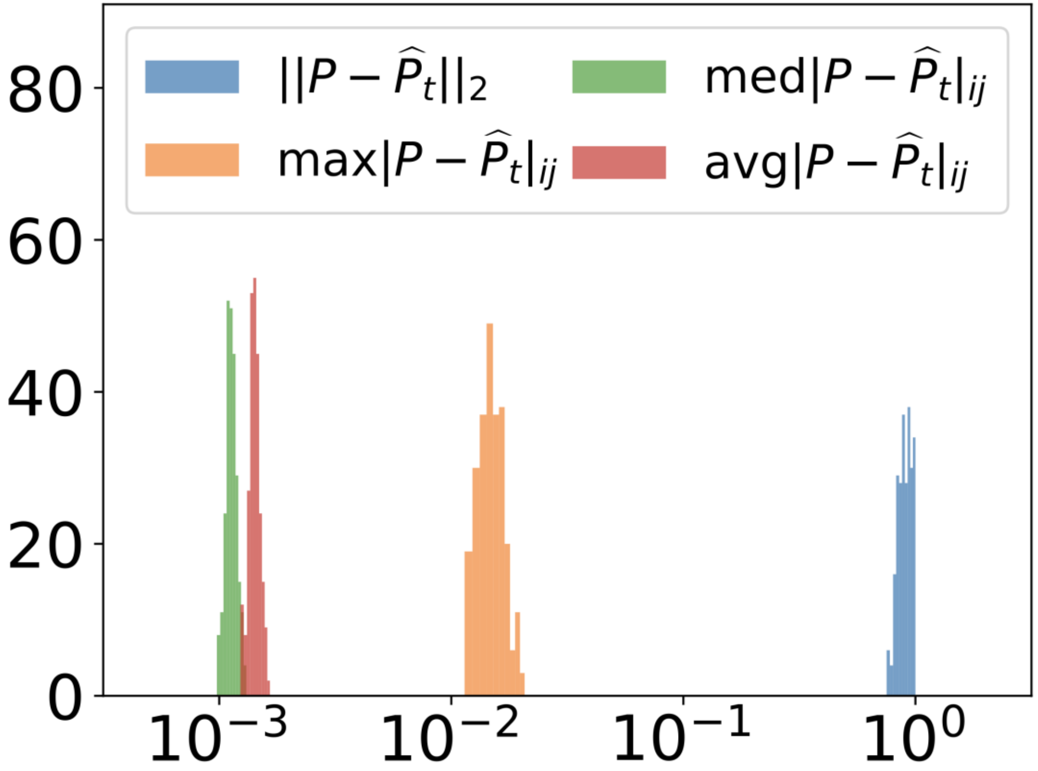
$$\sin\theta_{\max} = \|P - \widehat{P}\|_2,$$

...where $= V_k V_k^T$, $\hat{P} = \hat{V}_k \hat{V}_k^T$.

- Better to consider **component-wise** errors, like

$$\eta = \max_{i,j} |P_{i,j} - \widehat{P}_{i,j}|, \quad \text{or}$$

$$\mu = \min_{Q \in \mathbb{O}(k)} \|V_k - \widehat{V}_k Q\|_{2\to\infty}.$$



Theorem. If $\varphi_{\max}(J) < \pi/2$, then

$$\cos\varphi_{\max}(J) \geq \cos\widehat{\varphi}_{\max}(J) - \frac{2kc_k\mu}{\cos\widehat{\varphi}_{\max}(J)} + \mathcal{O}(\mu^2).$$

Want to learn more?

Preprint at https://doi.org/10.48550/arXiv.2310.09452.

THANK YOU!