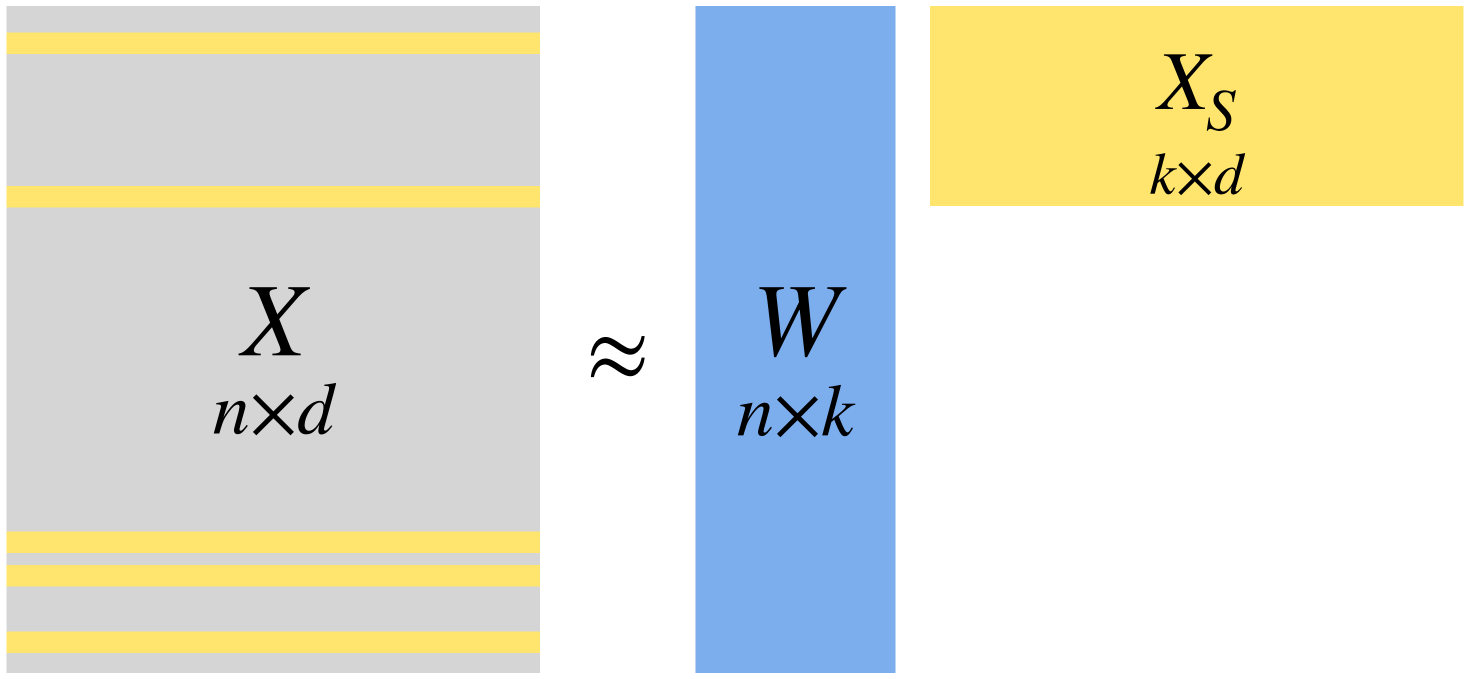
Robust Blockwise Random Pivoting (RBRP): Fast and Accurate Adaptive Interpolative Decomposition

Yijun Dong[1], Chao Chen[2], Per-Gunnar Martinsson[3], Katherine Pearce[3]

[1]Courant Institute, New York University
[2]Department of Mathematics, North Carolina State University
[3]Oden Institute, University of Texas at Austin

SIAM Conference on Parallel Processing for Scientific Computing (PP24)

Interpolative Decomposition (ID)



- Given a data matrix $X = [x_1, \cdots, x_n]^\top \in \mathbb{R}^{n \times d}$
- A target rank $1 \leq r \leq \mathrm{rank}(X)$
- A distortion constant $\epsilon > 0$
- Aim to construct a (r, ϵ)**-ID** of X —— $X \approx WX_S$ such that

$$\|X - WX_S\|_F^2 \leq (1+\epsilon)\|X - X_{\langle r \rangle}\|_F^2$$

  - $S = \{s_1, \cdots, s_k\} \subseteq [n]$ contains indices for a **skeleton subset** of size $|S| = k$ (usually $k \ll n$)
  - $X_S = [x_{s_1}, \cdots, x_{s_k}]^\top \in \mathbb{R}^{k \times d}$ is the row skeleton submatrix corresponding to S
  - $W \in \mathbb{R}^{n \times k}$ is an interpolation matrix for the given skeleton subset S
  - $X_{\langle r \rangle}$ denotes the optimal rank-r approximation of X (given by the truncated SVD)

Two Stages of ID Constructions

- **Stage I: Skeleton selection**
  - Find a good skeleton subset S:

$$\min_{S \subset [n]} \min_{W \in \mathbb{R}^{n \times |S|}} \|X - WX_S\|_F^2$$

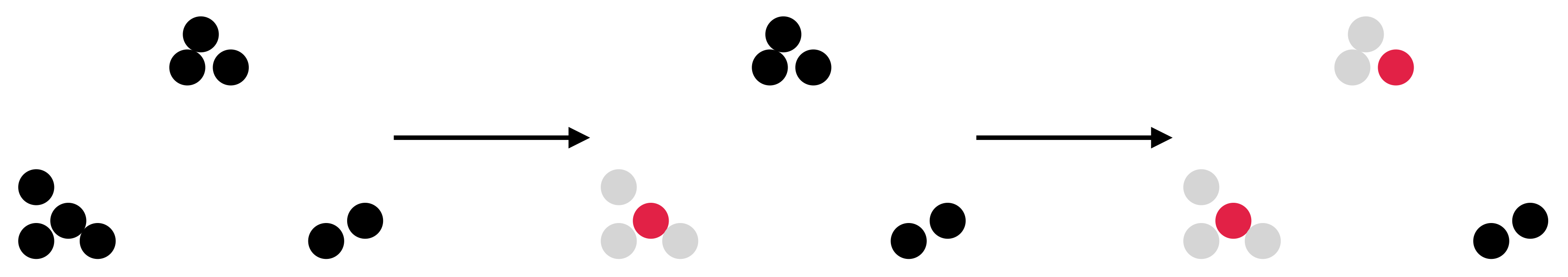
  - **Skeletonization error**: $\mathscr{E}_X(S) := \|X - XX_S^\dagger X_S\|_F^2 = \min_{W \in \mathbb{R}^{n \times |S|}} \|X - WX_S\|_F^2$
    - Naive construction of $XX_S^\dagger$ (e.g., via QR) takes $O(ndk)$ time (i.e., $k = |S|$ additional passes through X)
- **Stage II: Interpolation matrix construction**
  - For some $O(ndk)$-time selection algorithms, W can be evaluated/approximated a posteriori in $O(nk^2)$ time
  - **Interpolation error**: $\mathscr{E}_X(W \mid S) := \|X - WX_S\|_F^2$

What are Fast & Accurate ID Algorithms?

- **Skeleton complexity**: the minimum number of skeletons $k = |S|$ that an ID algorithm needs to select in order to form a (r, ϵ)-ID (in expectation), i.e., $\mathscr{E}_X(S) \leq (1+\epsilon)\|X - X_{\langle r \rangle}\|_F^2$
- **Asymptotic complexity**: the asymptotic FLOP counts of the skeleton selection stage in an ID algorithm
- **Parallelizability**: whether the dominant cost of the skeleton selection stage in an ID algorithm can be casted as matrix-matrix (fast), instead of matrix-vector (slow), multiplications with X (i.e., applicability of Level 3 BLAS)
- **Error-revealing property**: the ability of an ID algorithm to evaluate $\mathscr{E}_X(S)$ efficiently on the fly so that the target rank k does not need to be given a priori.
  - Definition: An ID algorithm is **error-revealing** if after selecting any skeleton subset S, it can evaluate the corresponding skeletonization error $\mathscr{E}_X(S)$ efficiently in at most $O(n)$ time.
- **ID-revealing property**: if the skeleton selection stage of an ID algorithm extracts sufficient information so that
  - **Exact/inexact-ID-revealing**: $W = XX_S^\dagger$ can be evaluated exactly/approximated in $O(nk^2)$ time
  - **Non-ID-revealing** otherwise

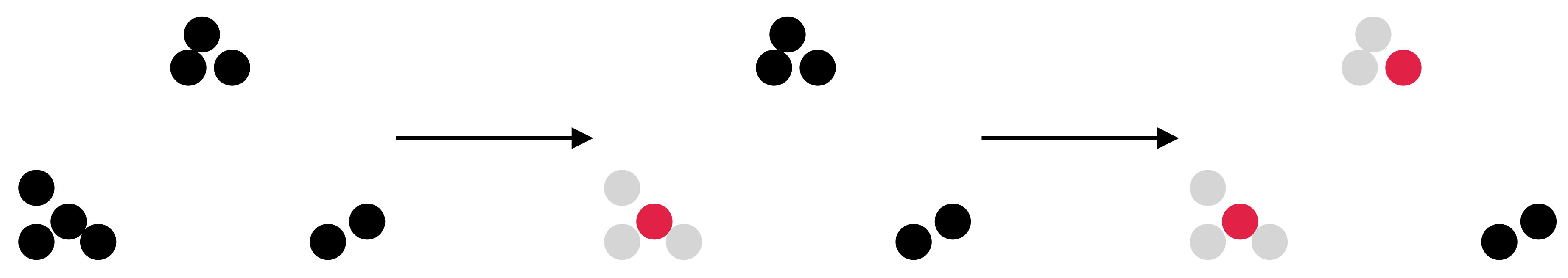
Adaptiveness & Randomness

- **Adaptiveness**
  - Each new skeleton selection is aware of the previously selected skeleton subset
  - By selecting according to the residual
  - Common adaptive residual updates:
    - Gram-Schmidt (QR)
    - Gaussian elimination (LU)

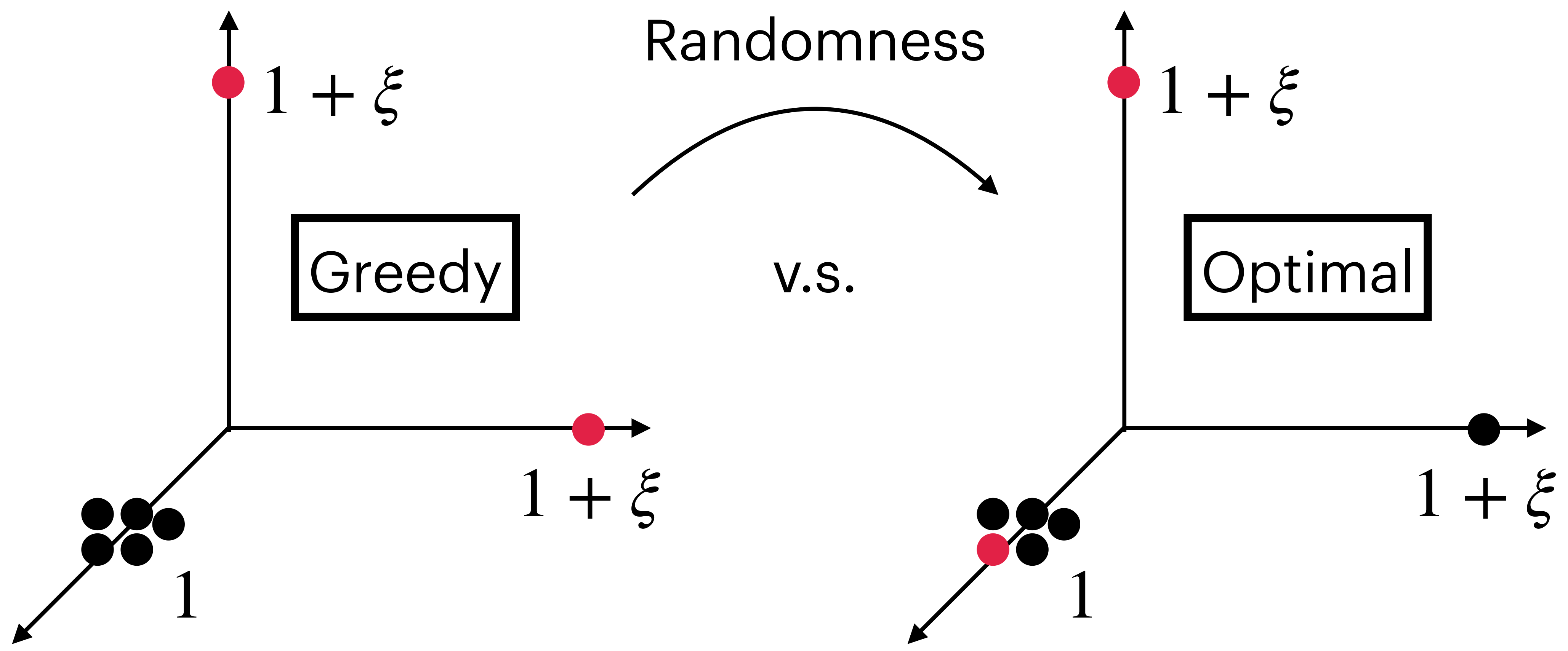


Adaptiveness & Randomness

- **Adaptiveness**
  - Each new skeleton selection is aware of the previously selected skeleton subset
  - By selecting according to the residual
  - Common adaptive residual updates:
    - Gram-Schmidt (QR)
    - Gaussian elimination (LU)



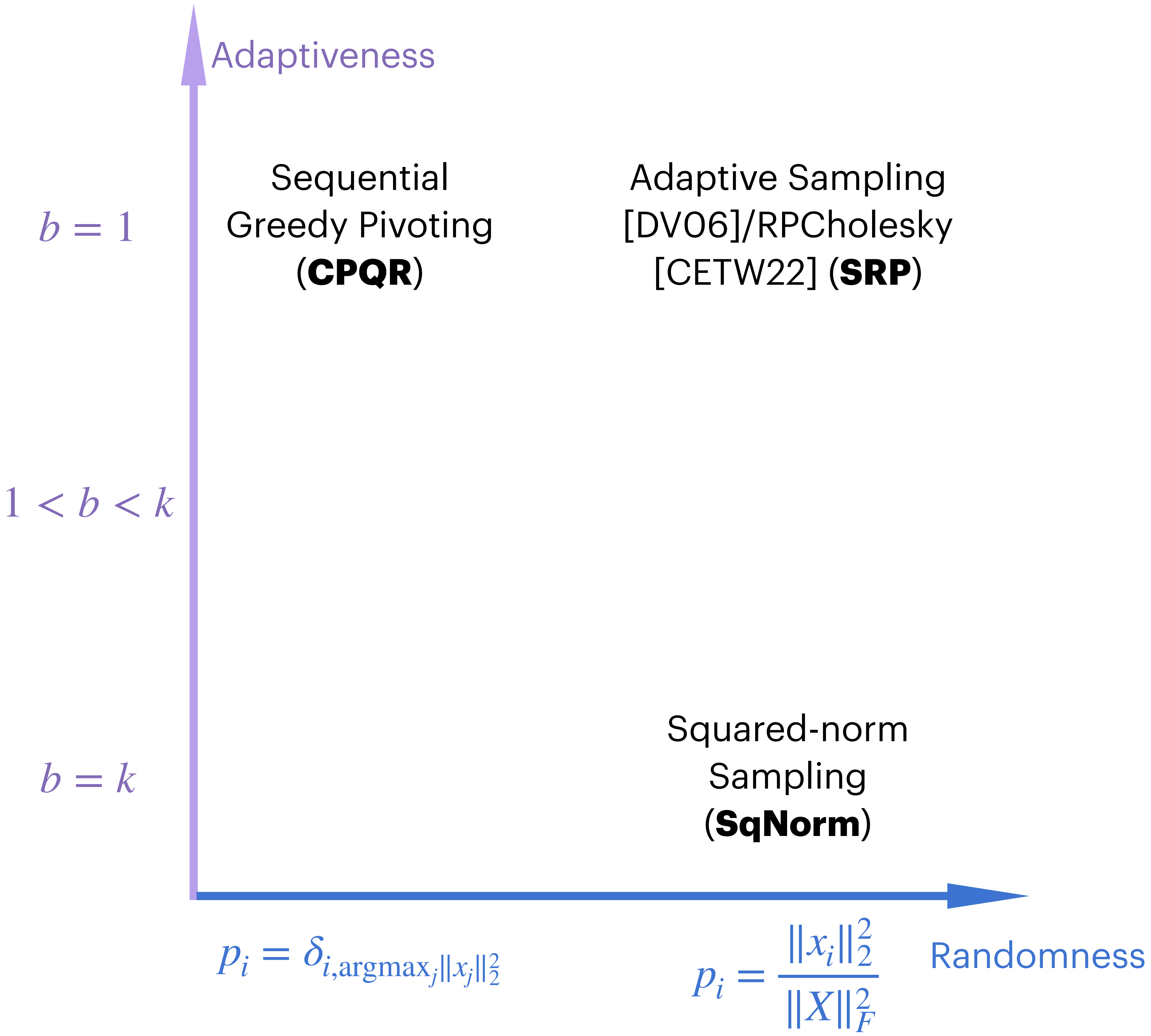
- **Randomness** (in contrast to greedy)
  - Intuition: balance exploitation with exploration
  - Effectively circumvent adversarial inputs for greedy methods
  - Achieve appealing skeleton complexities in expectation
  - Common randomness: sampling, sketching



Skeleton Selection: A General Framework

A framework for (blockwise adaptive) skeleton seletion

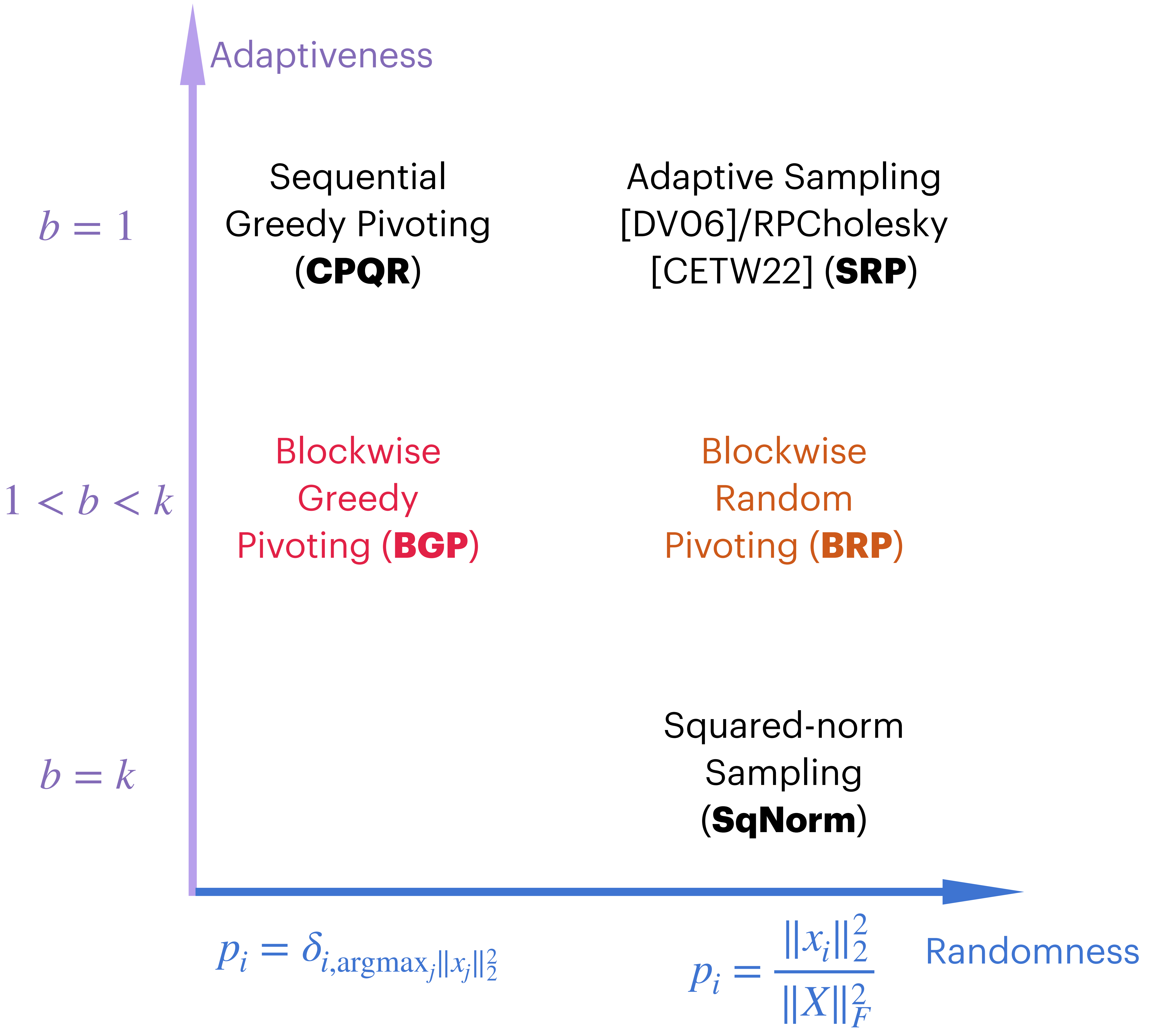
- **Inputs**: $X \in \mathbb{R}^{n \times d}$, $\tau = (1+\epsilon)\eta_r \in (0,1)$
- $X^{(0)} \leftarrow X$, $S^{(0)} \leftarrow \emptyset$, $t \leftarrow 0$
- **while** $\mathscr{E}(S^{(t)}) > \tau \|X\|_F^2$ **do**
  - $t \leftarrow t+1$
  - Select $|S_t| = b$ skeletons S_t based on $\left(p_i\left(X^{(t-1)}\right)\right)_{i\in[n]}$
  - $S^{(t)} \leftarrow S^{(t-1)} \cup S_t$
  - $X^{(t)} \leftarrow X^{(t-1)}\left(I_d - X_{S_t}^{\dagger} X_{S_t}\right)$
- $S \leftarrow S^{(t)}$, $k = |S|$



Skeleton Selection: A General Framework

A framework for (blockwise adaptive) skeleton seletion

- **Inputs**: $X \in \mathbb{R}^{n \times d}$, $\tau = (1+\epsilon)\eta_r \in (0,1)$
- $X^{(0)} \leftarrow X$, $S^{(0)} \leftarrow \emptyset$, $t \leftarrow 0$
- **while** $\mathcal{E}(S^{(t)}) > \tau \|X\|_F^2$ **do**
  - $t \leftarrow t + 1$
  - Select $|S_t| = b$ skeletons S_t based on $\left(p_i\left(X^{(t-1)}\right)\right)_{i \in [n]}$
  - $S^{(t)} \leftarrow S^{(t-1)} \cup S_t$
  - $X^{(t)} \leftarrow X^{(t-1)}\left(I_d - X_{S_t}^{\dagger} X_{S_t}\right)$
- $S \leftarrow S^{(t)}$, $k = |S|$



Skeleton Selection: Other Methods

Sampling methods

- **DPP/volume sampling** [HKPV06, BW09, DR10, KT11, GS12]
  - Pro: nearly optimal expected skeleton complexity: $k \geq \frac{r}{\epsilon} + r - 1$ is sufficient for (r, ϵ)-ID in expectation
  - Con: expensive to compute
- **Leverage score sampling** [MD09, DMMW12]
  - Pro: can be estimated efficiently for large-scale problems (e.g., tensor Khatri-Rao product)
  - Con: expensive to compute
- **Uniform sampling** [CLMMPS15]
  - Pro: linear time
  - Con: require/depend on matrix incoherence

Skeleton Selection: Other Methods

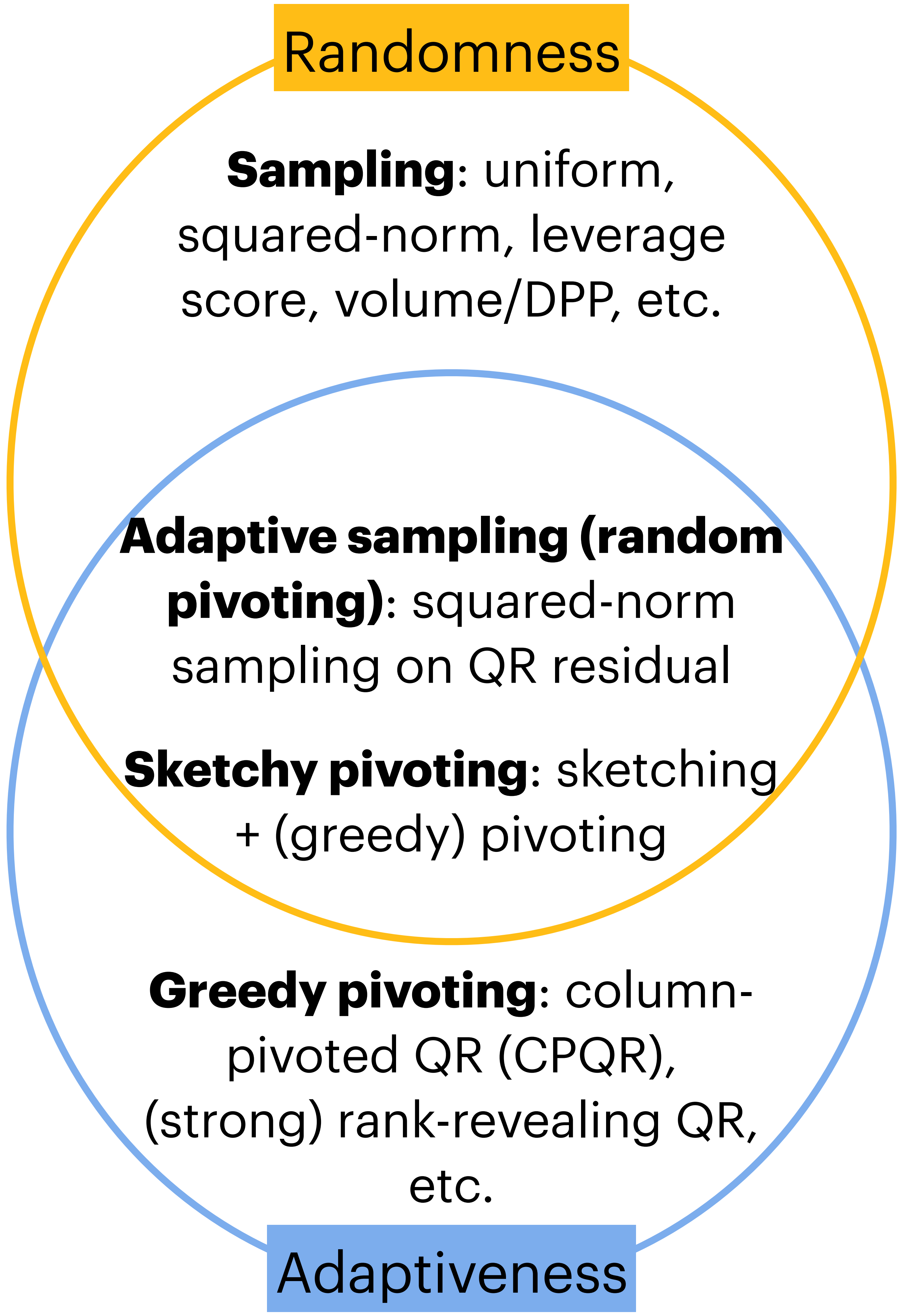
Sampling methods

- **DPP/volume sampling** [HKPV06, BW09, DR10, KT11, GS12]
  - Pro: nearly optimal expected skeleton complexity: $k \geq \frac{r}{\epsilon} + r - 1$ is sufficient for (r, ϵ)-ID in expectation
  - Con: expensive to compute
- **Leverage score sampling** [MD09, DMMW12]
  - Pro: can be estimated efficiently for large-scale problems (e.g., tensor Khatri-Rao product)
  - Con: expensive to compute
- **Uniform sampling** [CLMMPS15]
  - Pro: linear time
  - Con: require/depend on matrix incoherence

Sketchy pivoting

- **Inputs**: $X \in \mathbb{R}^{n \times d}$, $k \leq \mathrm{rank}(X)$,
- Draw JLT $\Omega \in \mathbb{R}^{d \times k}$ (e.g., $\Omega_{ij} \sim \mathcal{N}(0, 1/k)$ i.i.d.)
- Sketching $Y = X\Omega \in \mathbb{R}^{n \times k}$
- Greedy pivoting: for $t = 1, \cdots, k$
  - Column (row) pivoted QR (**CPQR**) [VM17]: $s_t \leftarrow \underset{i}{\mathrm{argmax}} \, \| Y^{(t-1)}_{i,:} \|_2^2$ + Gram-Schmidt
  - LU with partial pivoting (**LUPP**) [**D**M23]: $s_t \leftarrow \underset{i}{\mathrm{argmax}} \, | Y^{(t-1)}_{i,t} |$ + Gaussian Elimination
- Pro: fast, accurate, robust to adversarial inputs
- Con: require prior knowledge of k

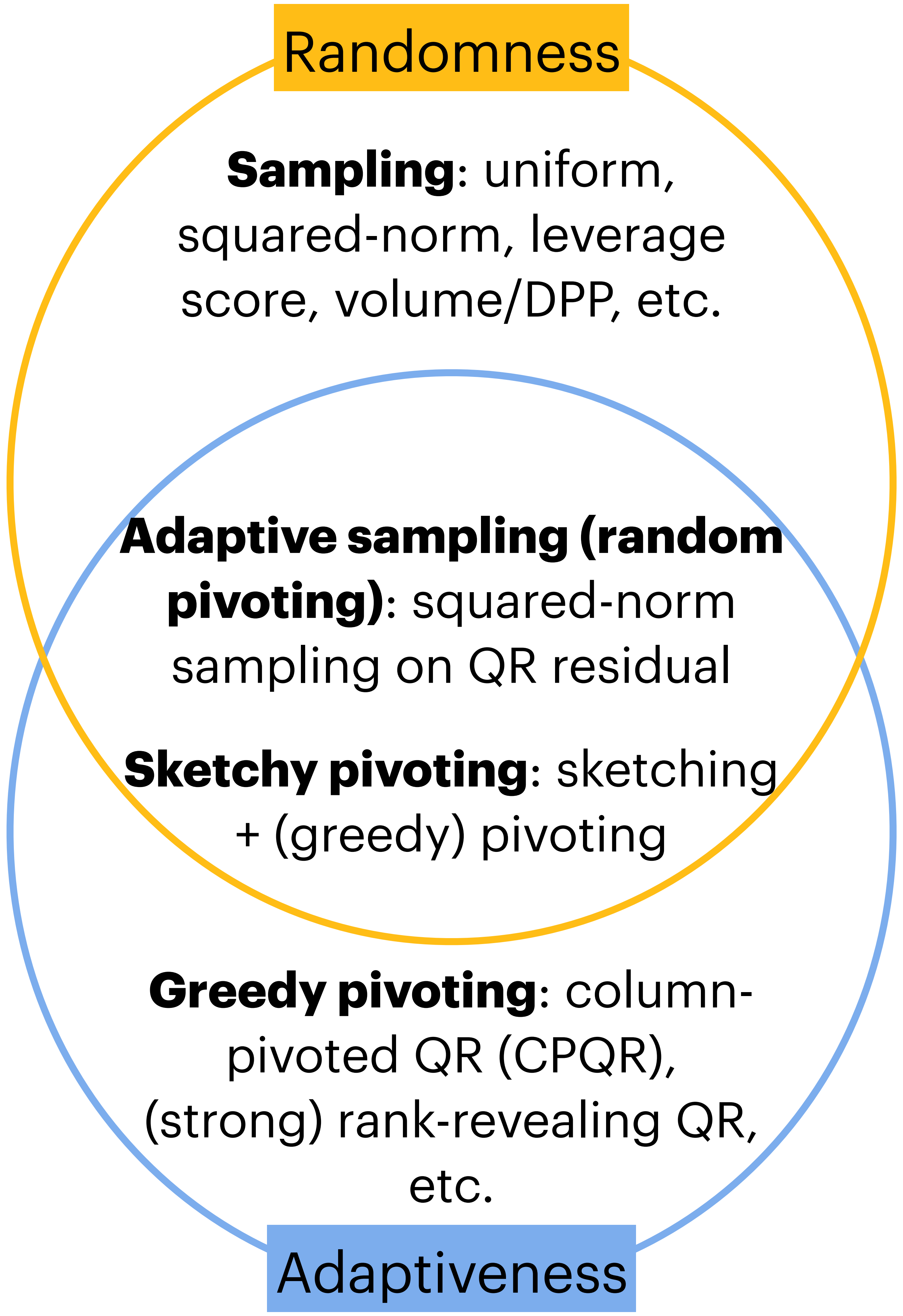
ID Algorithms with Adaptiveness & Randomness



Algorithm	Skeleton Complexity	Asymp. Cost + Parallelizability	Error-reveal	ID-reveal
Greedy Pivoting	$k \geq (1+(1+\epsilon)\eta_r)n$	$O(ndk)$ sequential	✓	**Exact**
Squared-norm Sampling	$k \geq \frac{r-1}{\epsilon\eta_r} + \frac{1}{\epsilon}$	$O(nd)$ parallel	✗	Non
Random Pivoting	$k \geq k_{RP} := \frac{r}{\epsilon} + r\log\left(\min\left\{\frac{1}{\epsilon\eta_r}, \frac{2^{r+1}}{\epsilon}\right\}\right)$	$O(ndk)$ sequential	✓	**Exact**
Sketchy Pivoting	Conjecture: $k \gtrsim k_{RP}$	$O(ndk)$ parallel	✗	Inexact
RBRP	Conjecture: $k \gtrsim k_{RP}$	$O(ndk)$ parallel	✓	**Exact**

* $\eta_r = \|X - X_{<r>}\|_F^2 / \|X\|_F^2$ quantifies the relative optimal rank-r approximation error of X

ID Algorithms with Adaptiveness & Randomness

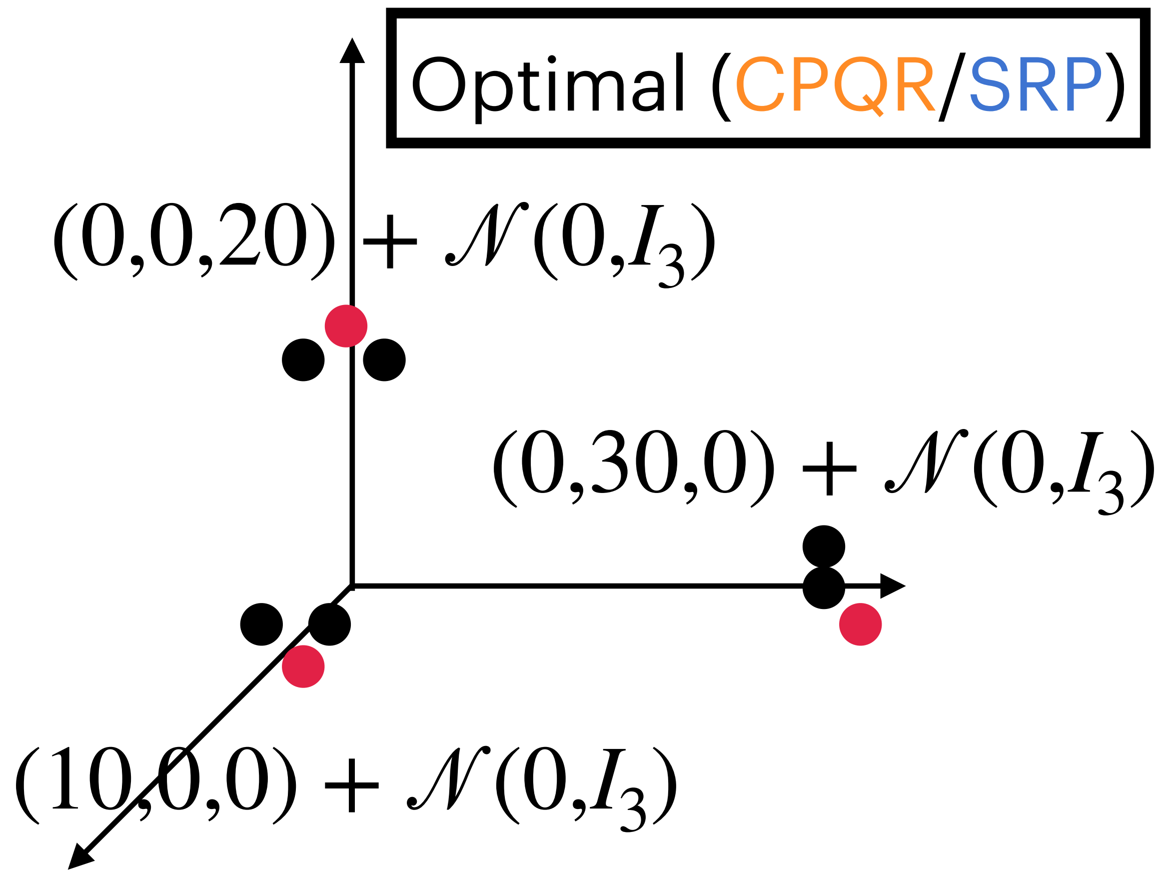


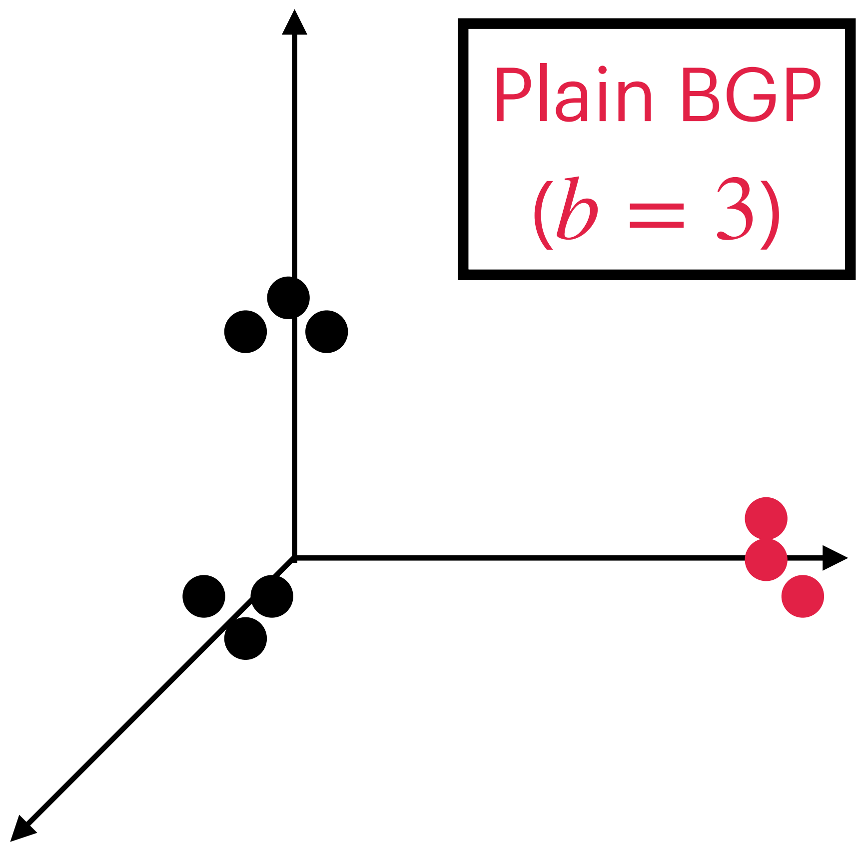
Algorithm	Skeleton Complexity	Asymp. Cost + Parallelizability	Error-reveal	ID-reveal
Greedy Pivoting	$k \geq (1+(1+\epsilon)\eta_r)n$	$O(ndk)$ sequential	✓	**Exact**
Squared-norm Sampling	$k \geq \frac{r-1}{\epsilon\eta_r} + \frac{1}{\epsilon}$	$O(nd)$ parallel	✗	Non
Random Pivoting	$k \geq k_{RP} := \frac{r}{\epsilon} + r\log\left(\min\left\{\frac{1}{\epsilon\eta_r}, \frac{2^{r+1}}{\epsilon}\right\}\right)$	$\boxed{O(ndk) \text{ sequential}}$	✓	**Exact**
Sketchy Pivoting	Conjecture: $k \gtrsim k_{RP}$	$O(ndk)$ parallel	✗	Inexact
RBRP	Conjecture: $k \gtrsim k_{RP}$	$\boxed{O(ndk) \text{ parallel}}$	✓	**Exact**

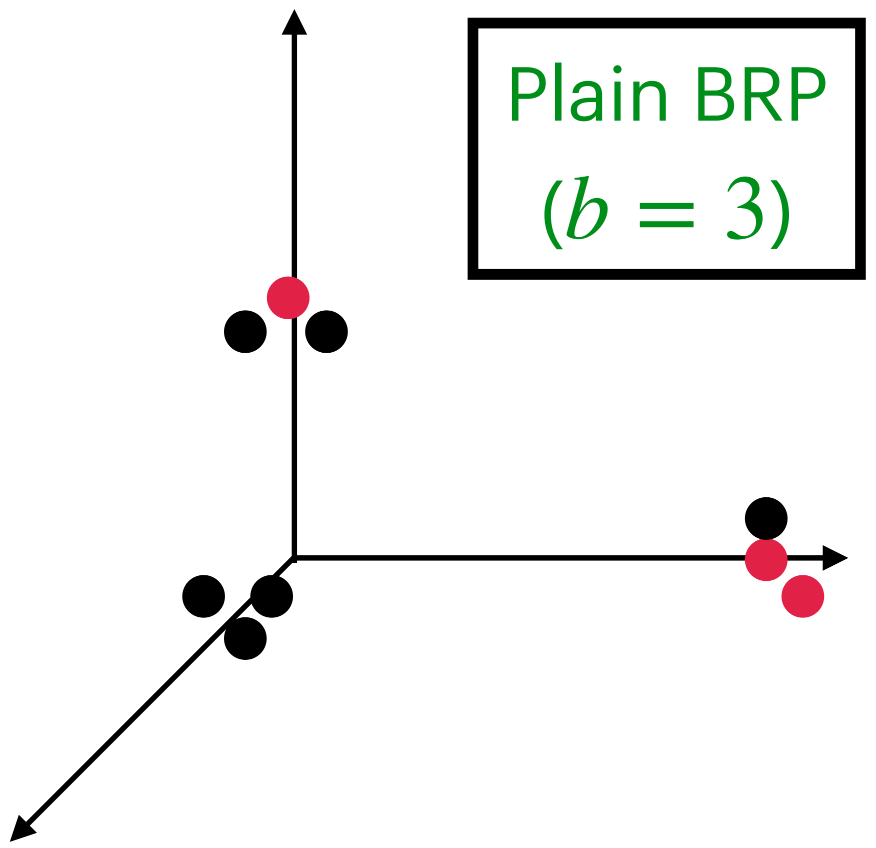
* $\eta_r = \|X - X_{<r>}\|_F^2 / \|X\|_F^2$ quantifies the relative optimal rank-r approximation error of X

Question: How to parallelize random pivoting?
Answer: Blockwise random pivoting

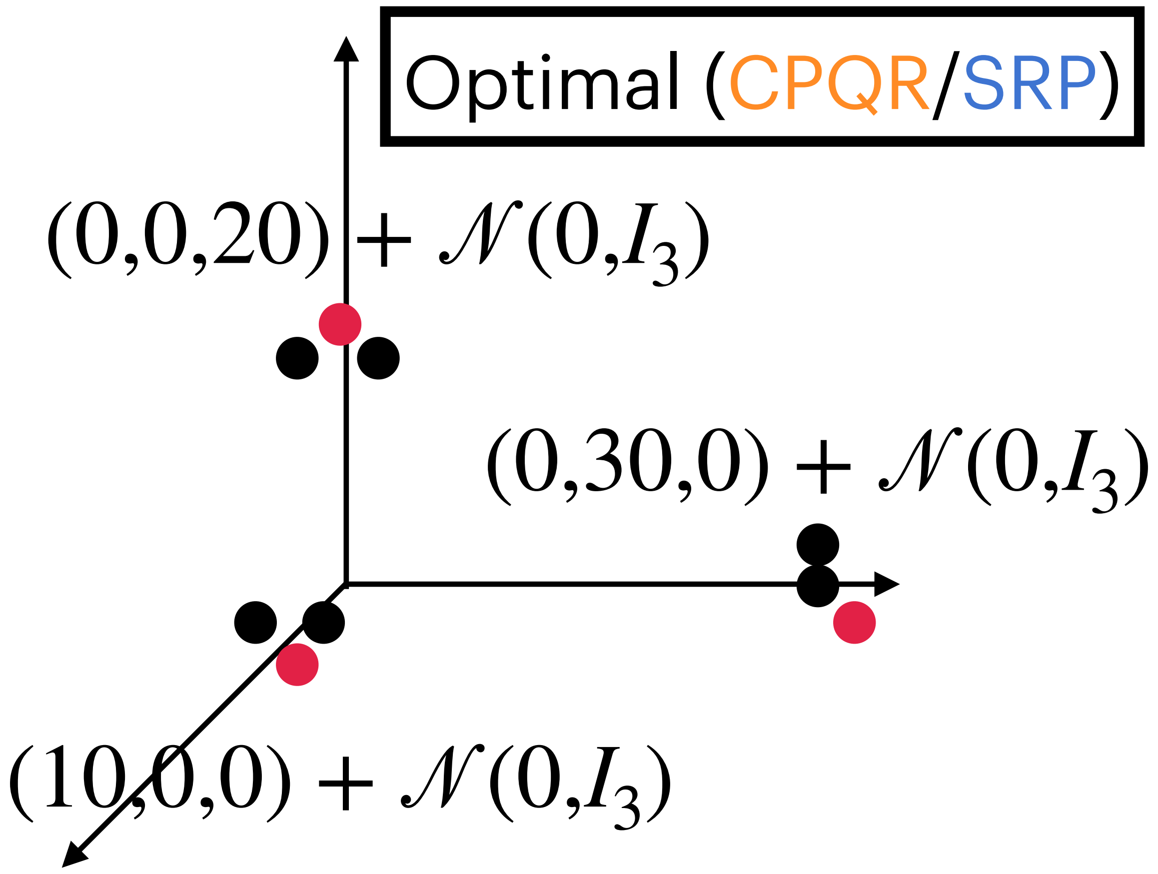
Pitfall of Plain Blockwise Greedy/Random Pivoting

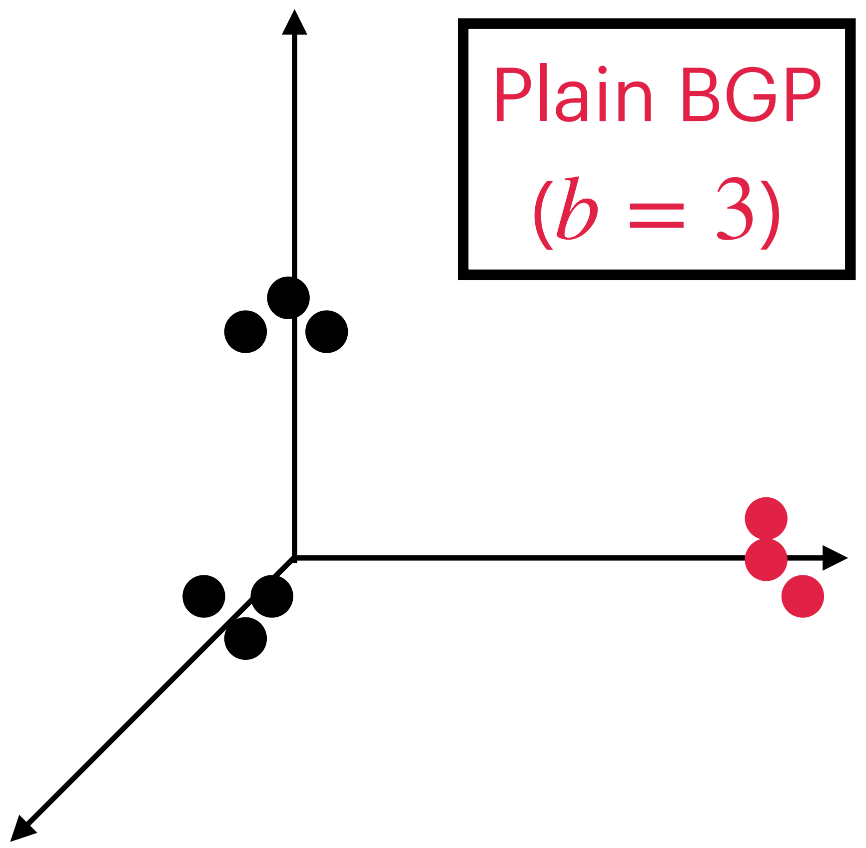


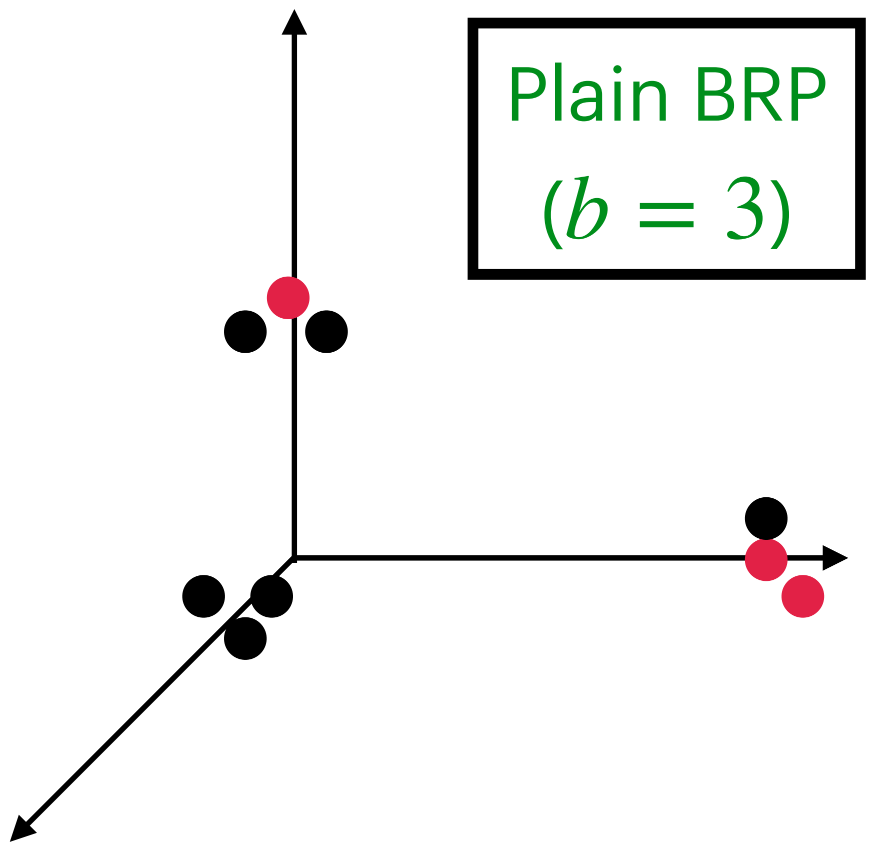




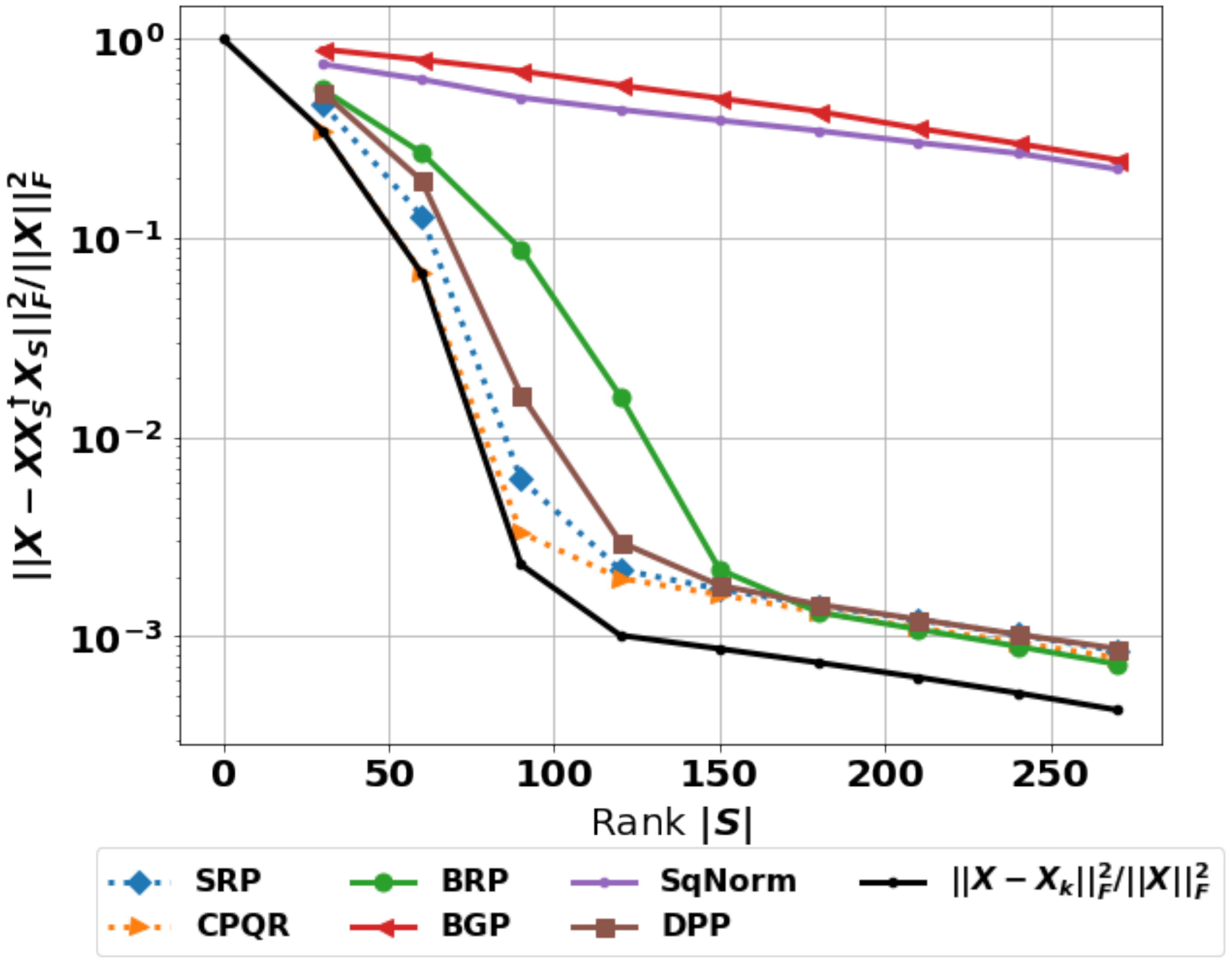
Pitfall of Plain Blockwise Greedy/Random Pivoting







$k = 100$ clusters centered at $\{10j \cdot e_j\}_{j \in [k]}$, $n = 20k$, $d = 500$, $b = 30$

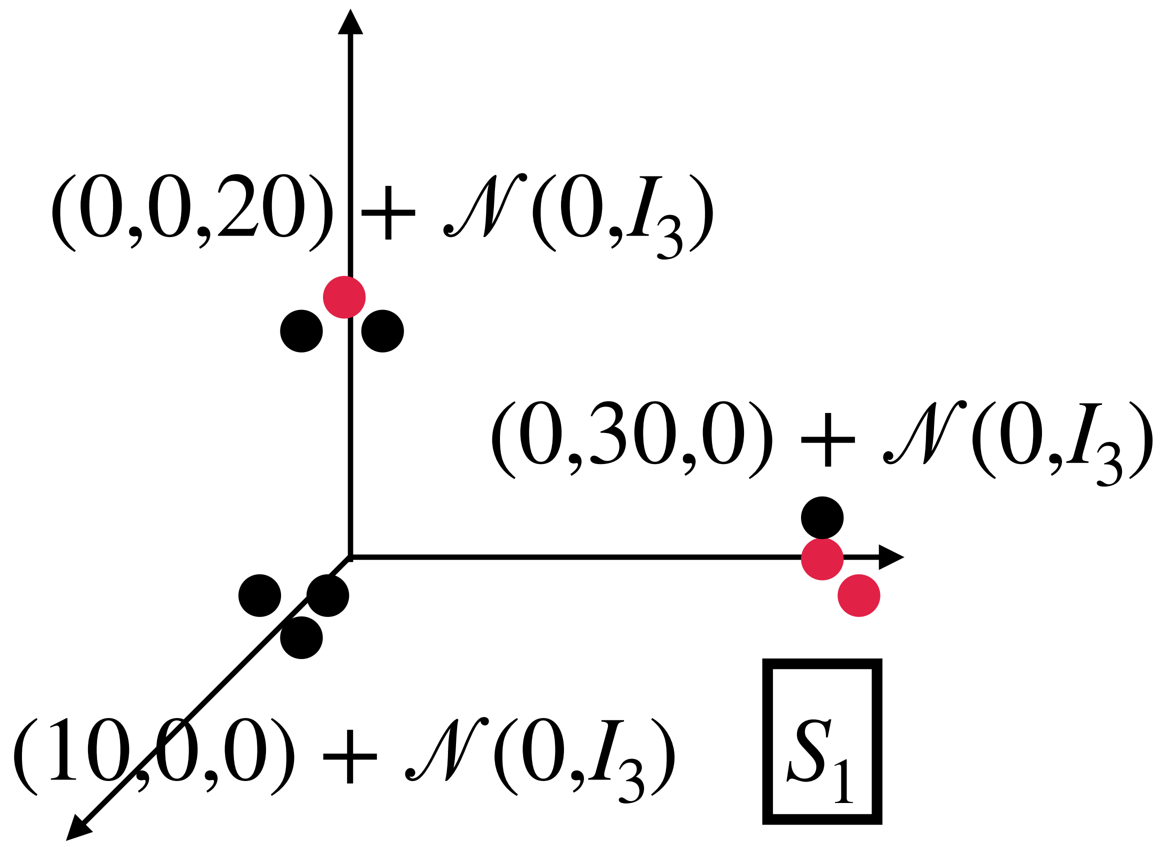


- Sequential pivoting (CPQR & SRP) is nearly optimal
- Plain blockwise pivoting (BRP/BGP, especially BGP) suffers from suboptimal skeleton complexities (up to b times)
- Squared-norm sampling (SqNorm) tends to fail

Robust Blockwise Random Pivoting

Robust **B**lockwise **R**andom **P**ivoting (**RBRP**)

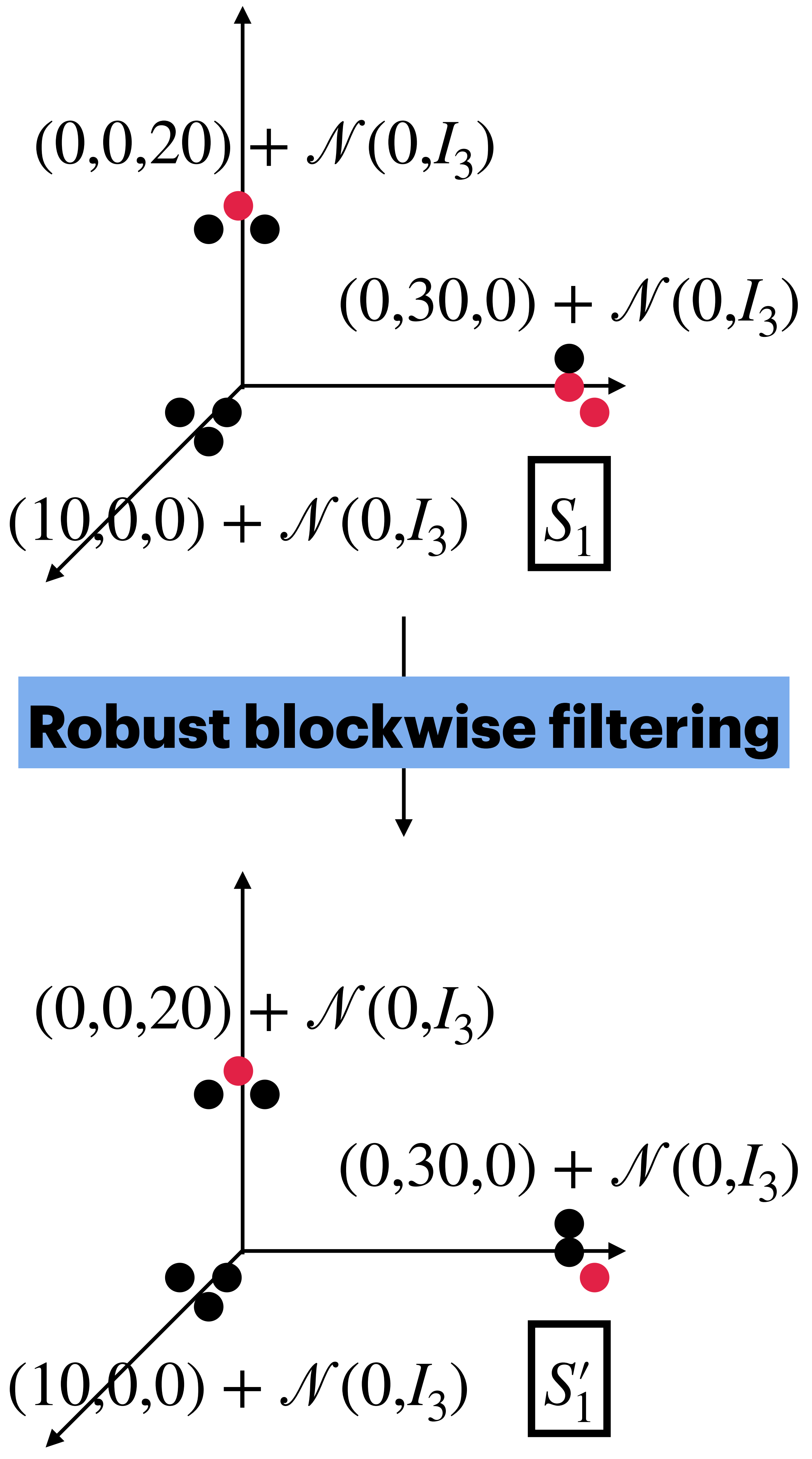
- **Inputs**: $X \in \mathbb{R}^{n \times d}$, $\tau = (1+\epsilon)\eta_r \in (0,1)$
- $X^{(0)} \leftarrow X$, $S^{(0)} \leftarrow \emptyset$, $t \leftarrow 0$
- **while** $\mathcal{E}(S^{(t)}) > \tau \|X\|_F^2$ $(t \leftarrow t+1)$ **do**
  - Select $|S_t| = b$ skeletons S_t based on $\left(p_i\left(X^{(t-1)}\right)\right)_{i \in [n]}$
  - **Robust blockwise filtering (RBF)**
    - $\pi \leftarrow \mathrm{CPQR}\left(X_{S_t}^{(t-1)}\right) \in S_b$ (SRP and CPQR both work)
    - $\min_{S_t' = S_t(\pi(1:b'))} b' \ \mathrm{s.t.}\ \|X_{S_t} - X_{S_t'}\|_F^2 < \tau_b \|X_{S_t}\|_F^2$ (e.g., $\tau_b = \frac{1}{b}$)
  - $S^{(t)} \leftarrow S^{(t-1)} \cup S_t'$ and $X^{(t)} \leftarrow X^{(t-1)}\left(I_d - X_{S_t'}^{\dagger} X_{S_t'}\right)$
- $S \leftarrow S^{(t)}$, $k = |S|$



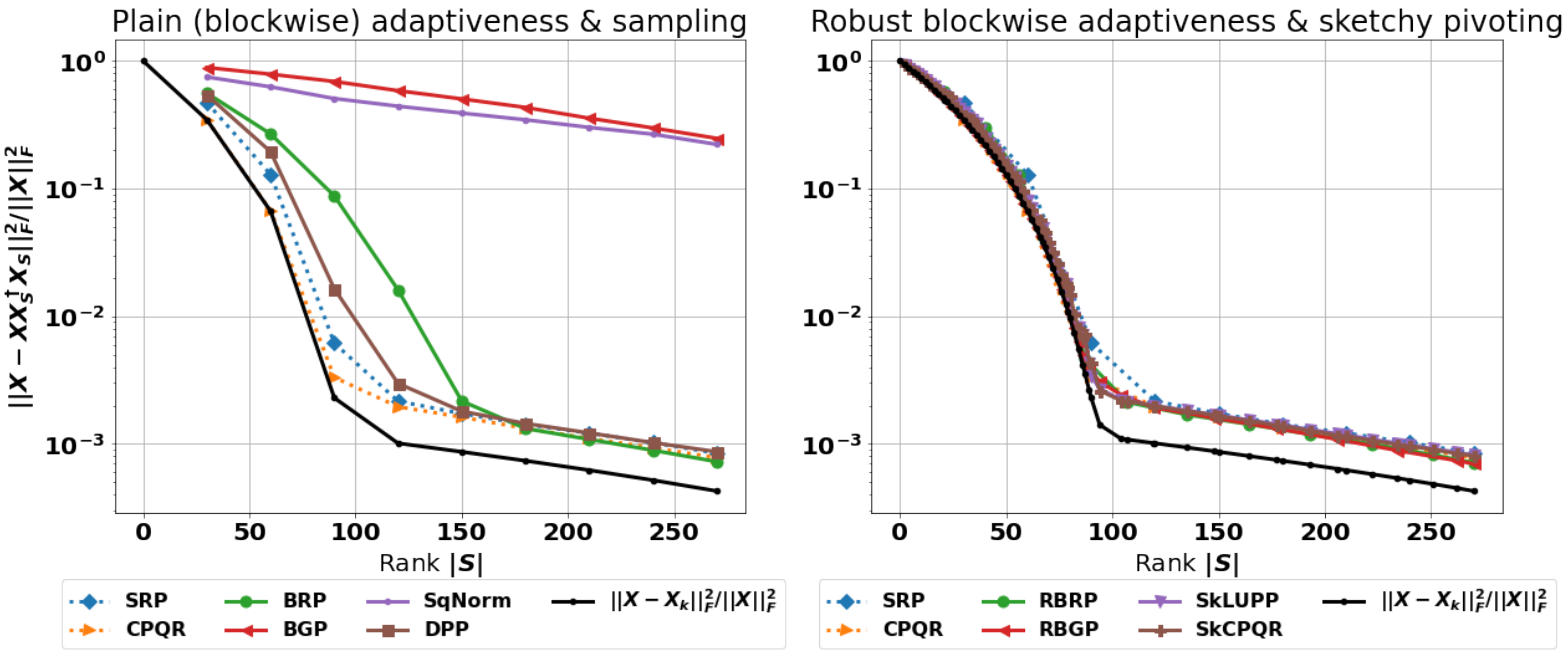
Robust Blockwise Random Pivoting

Robust Blockwise Random Pivoting (RBRP)

- **Inputs**: $X \in \mathbb{R}^{n\times d}$, $\tau = (1+\epsilon)\eta_r \in (0,1)$
- $X^{(0)} \leftarrow X$, $S^{(0)} \leftarrow \varnothing$, $t \leftarrow 0$
- **while** $\mathscr{E}(S^{(t)}) > \tau \|X\|_F^2$ $(t \leftarrow t+1)$ **do**
  - Select $|S_t| = b$ skeletons S_t based on $\left(p_i\left(X^{(t-1)}\right)\right)_{i\in[n]}$
  - **Robust blockwise filtering (RBF)**
    - $\pi \leftarrow \mathrm{CPQR}\left(X^{(t-1)}_{S_t}\right) \in S_b$ (SRP and CPQR both work)
    - $\min_{S_t'=S_t(\pi(1:b'))} b' \ \mathrm{s.t.}\ \|X_{S_t} - X_{S_t'}\|_F^2 < \tau_b \|X_{S_t}\|_F^2$ (e.g., $\tau_b = \frac{1}{b}$)
  - $S^{(t)} \leftarrow S^{(t-1)} \cup S_t'$ and $X^{(t)} \leftarrow X^{(t-1)}\left(I_d - X_{S_t'}^{\dagger} X_{S_t'}\right)$
- $S \leftarrow S^{(t)}$, $k = |S|$

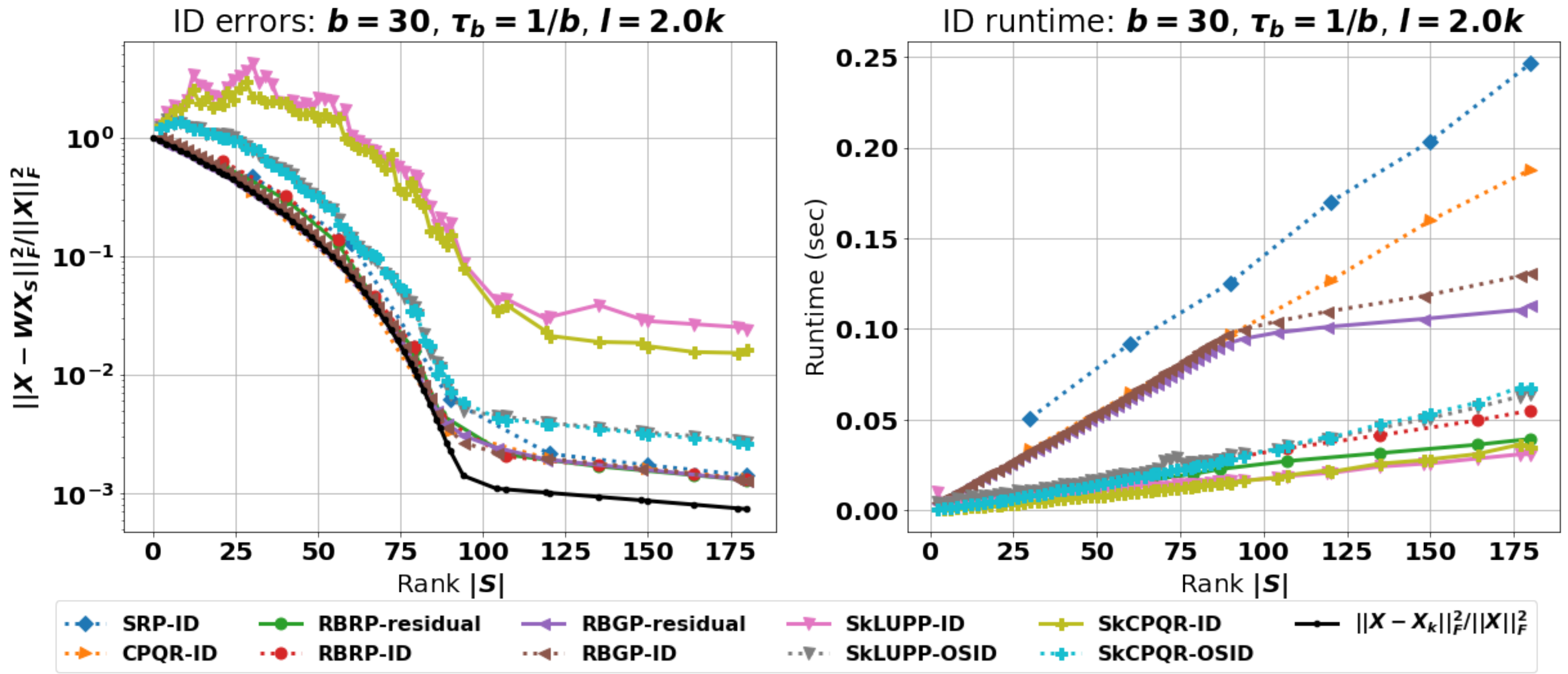


Robust Blockwise Random Pivoting: Robustness



- GMM with $k = 100$ clusters centered at $\{10j \cdot e_j\}_{j \in [k]}$, $\Sigma = I_d$, $n = 20k$, $d = 500$, $b = 30$
- Robust blockwise filtering (RBRP and RBGP) brings nearly optimal skeleton complexities

Robust Blockwise Random Pivoting: Efficiency

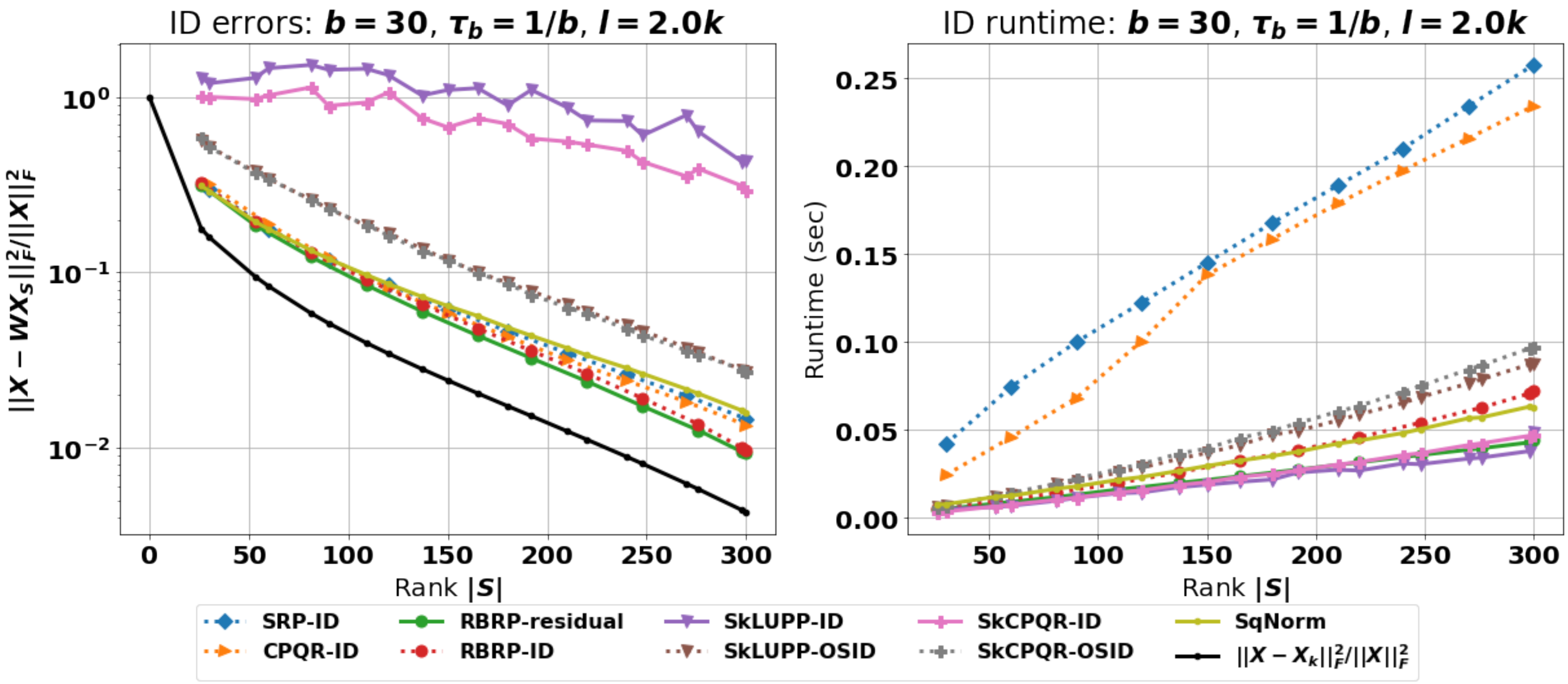


- Robust blockwise filtering (RBRP and RBGP) brings nearly optimal skeleton complexities
- RBGP tends to be slowed down much more significantly than RBRP by robust blockwise filtering
- For ID: RBRP-ID is almost as fast as sketchy pivoting (SkLUPP-ID/SkCPQR-ID), while enjoying much better interpolation error $\mathscr{E}_X(W \mid S) = \mathscr{E}_X(S)$ thanks to its exact-ID-revealing property.

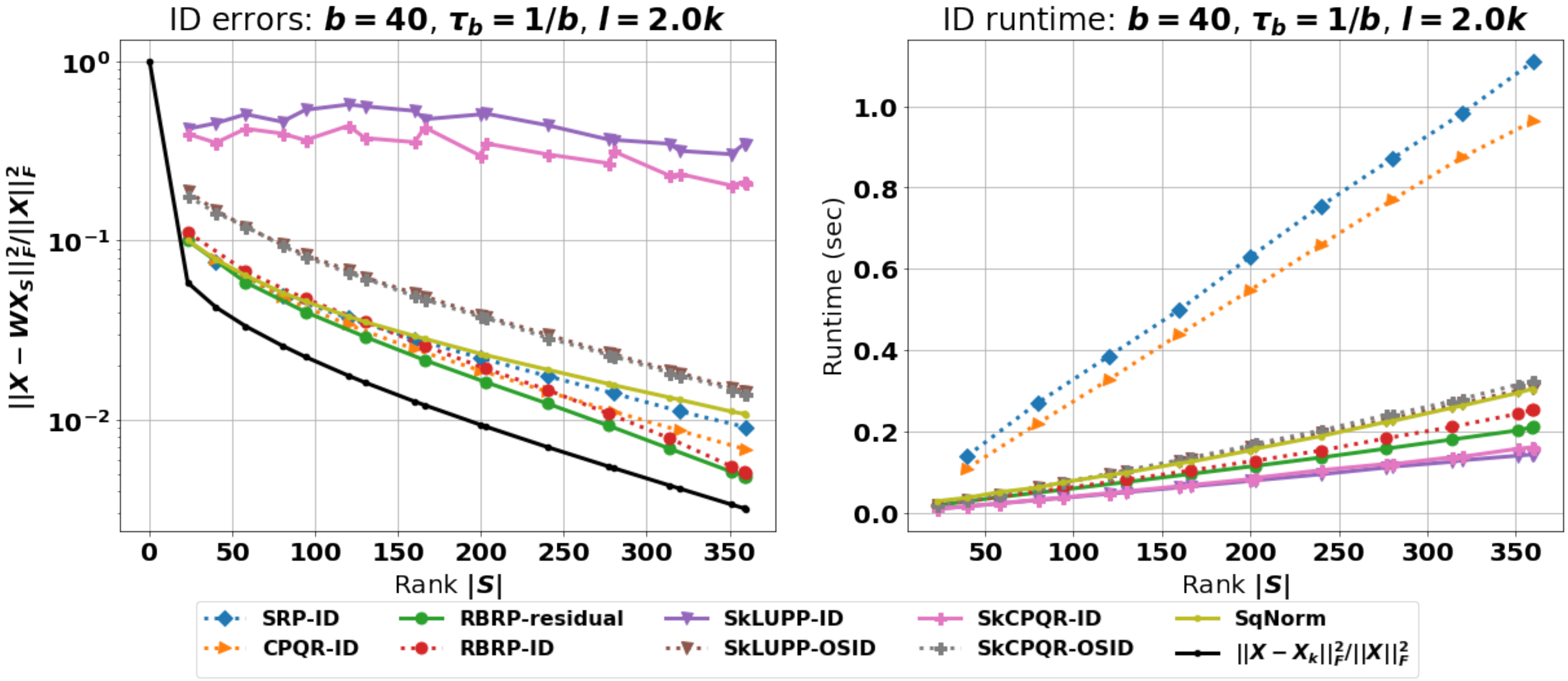
Exact- v.s. Inexact- ID-revealing Algorithms

- **Exact-ID-revealing algorithms**
  - Sequential/blockwise random/greedy pivoting algorithms (SRP, CPQR, BRP, BGP, RBRP, RBGP)
  - The skeleton selection process generates sufficient information for solving the least square problem $\min_{W \in \mathbb{R}^{n \times k}} \|X - WX_S\|_F^2$ in $O(nk^2)$ time
- **Inexact-ID-revealing algorithms**
  - Sketchy pivoting algorithms (SkLUPP, SkCPQR)
  - The skeleton selection process generates sufficient information for solving the **sketched** least square problem $\min_{W \in \mathbb{R}^{n \times k}} \|X\Omega - WX_S\Omega\|_F^2$ in $O(nk^2)$ time
  - Oversampled sketchy ID (**OSID**): for $|S| = k$
    - Sketching with oversampling $Y = X\Omega \in \mathbb{R}^{n \times l}$ such that $l = O(k)$
    - $W = YY_S^\dagger$ can be computed in $O(nlk) = O(nk^2)$ time
  - Suboptimal interpolation error: $\mathscr{E}_X(W \mid S) - \mathscr{E}_X(S) = O(k/l)$

More Numerical Comparisons: MNIST



More Numerical Comparisons: CIFAR-10



Summary

- A fast & accurate ID algorithm that finds $\|X - WX_S\|_F^2 \leq (1+\epsilon)\|X - X_{\langle r \rangle}\|_F^2$
  - With **nearly optimal skeleton complexity** in practice
  - Computationally efficient in terms of both **asymptotic complexity** and **parallelizability**
  - **Error-revealing** without requiring prior knowledge of the target skeleton subset size
  - **Exact-ID-revealing** where the optimal interpolation matrix can be computed efficiently
- **Combining adaptiveness and randomness** is a key for designing robust skeleton selection algorithms with competitive skeleton complexity
- A critical challenge is to **relax the sequential natural of adaptive selection**
- We introduced **Robust Blockwise Random Pivoting (RBRP)**, a parallelizable blockwise adaptive selection scheme that achieves comparable skeleton complexity as its sequential counterpart

Thank You!

arXiv: https://arxiv.org/abs/2309.16002

GitHub: https://github.com/dyjdongyijun/Robust_Blockwise_Random_Pivoting