Data Selection under Low Intrinsic Dimension:
from Interpolatove Decomposition to Ridge Regression

Yijun Dong

Courant Institute of Mathematical Sciences, New York University

Joint Mathematics Meeting, Jan 11, 2025

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU Center for Data Science

INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES ODEN

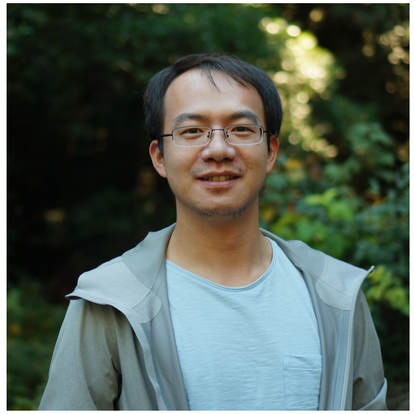
Low Intrinsic Dimension & Data Selection

- **Low intrinsic dimension is ubiquitous in real world**
  - Example: A language model with 341M parameters can be finetuned in a dimension-322 subspace with less than 6K samples [Aghajanyan-Zettlemoyer-Gupta-2020]
- Learning under low intrinsic dimension **with limited data, data selection becomes crucial**

How to **select informative data** for learning under **low intrinsic dimension**?

- Learning without noise: low-rank interpolative decomposition (ID)
- Learning with noise: low-rank approximation (bias) + variance reduction

Robust Blockwise Random Pivoting:
Fast and Accurate Adaptive Interpolative Decomposition



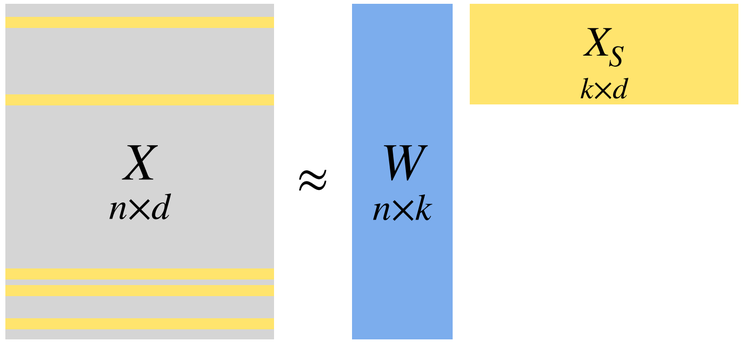
Chao Chen
NCSU

Per-Gunnar Martinsson
UT Austin

Katherine Pearce
UT Austin

Interpolative Decomposition (ID)

- Given a data matrix $X = [x_1, \cdots, x_n]^\top \in \mathbb{R}^{n \times d}$
- A target rank $1 \leq r \leq \text{rank}(X)$
- An error tolerance $\tau > 0$
- Aim to construct an ID of X —— $X \approx (XX_S^\dagger)X_S$ such that

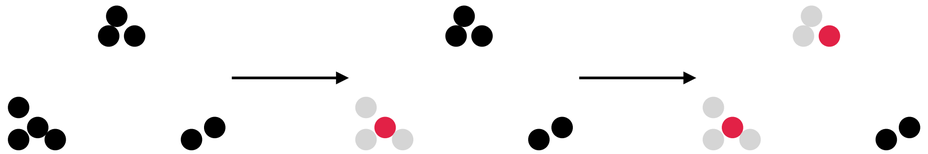


$$\mathscr{E}(S) = \|X - (XX_S^\dagger)X_S\|_F^2 \leq \tau \|X\|_F^2$$

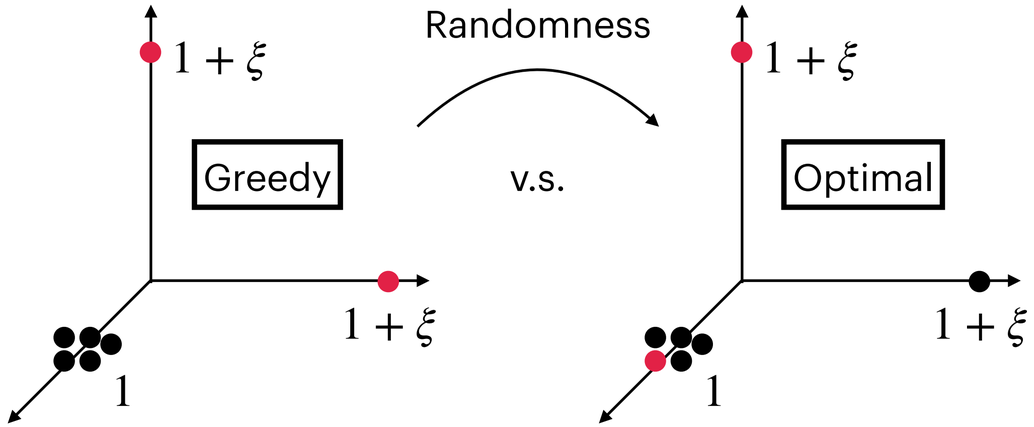
- $S = \{s_1, \cdots, s_k\} \subseteq [n]$ contains indices for a **skeleton subset** of size $|S| = k$ (usually $k \ll n$)
- $X_S = [x_{s_1}, \cdots, x_{s_k}]^\top \in \mathbb{R}^{k \times d}$ is the row skeleton submatrix corresponding to S
- $W = XX_S^\dagger \in \mathbb{R}^{n \times k}$ is an interpolation matrix for the given skeleton subset S

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)



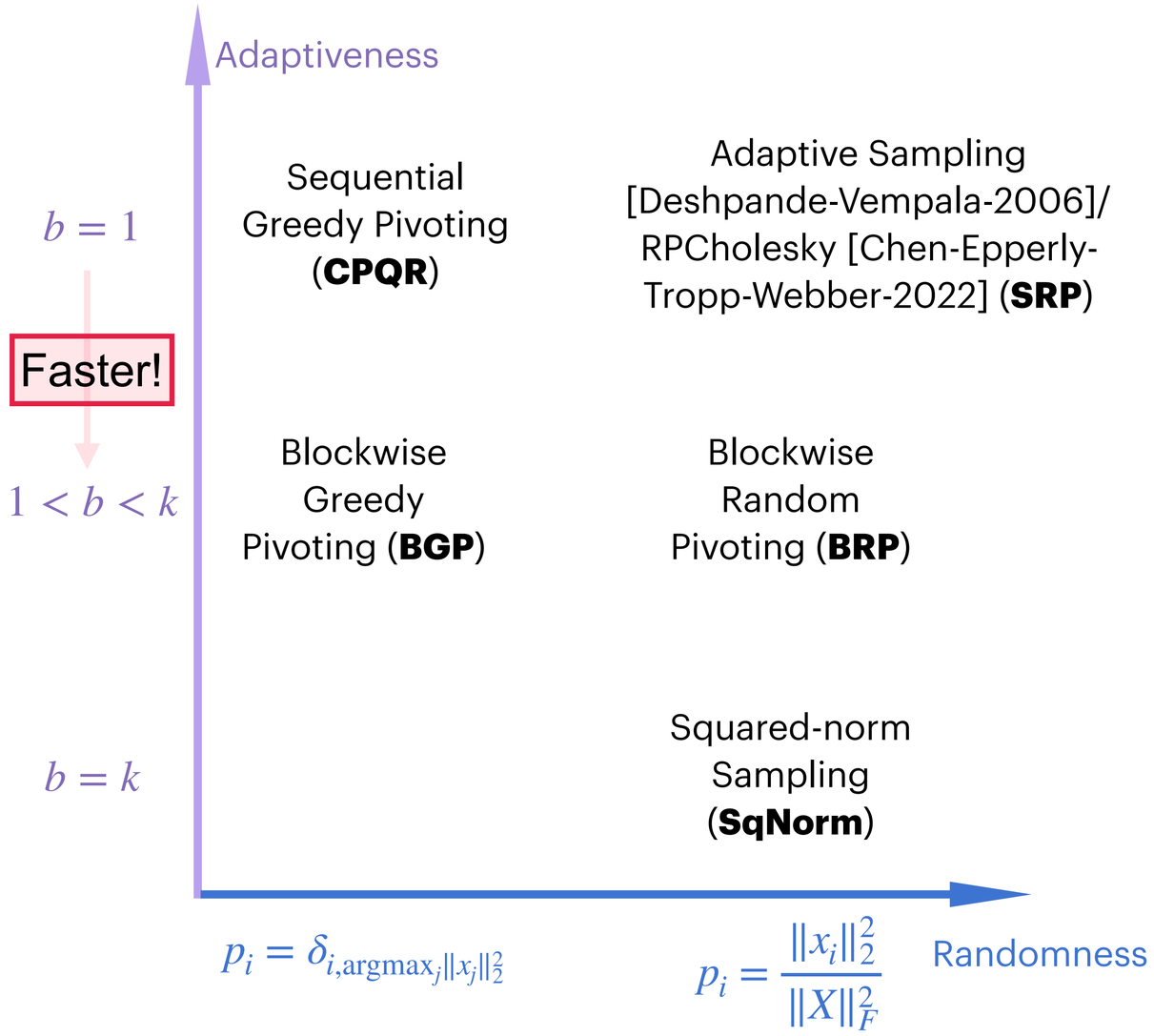
- **Randomness** (in contrast to greedy)
  - Intuition: balance exploitation with exploration
  - Effectively circumvent adversarial inputs for greedy methods
  - Achieve appealing skeleton complexities in expectation
  - Common randomness: sampling, sketching



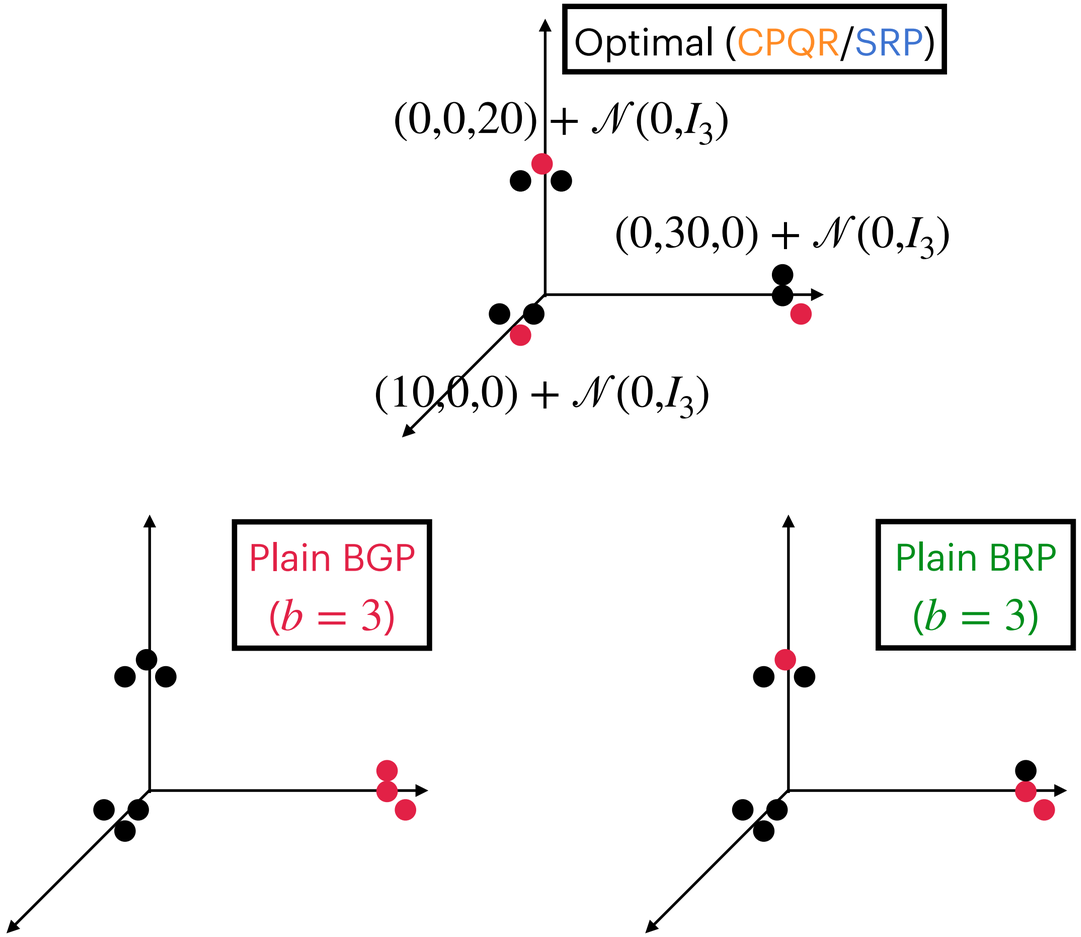
Skeleton Selection: A General Framework

A framework for (blockwise adaptive) skeleton selection

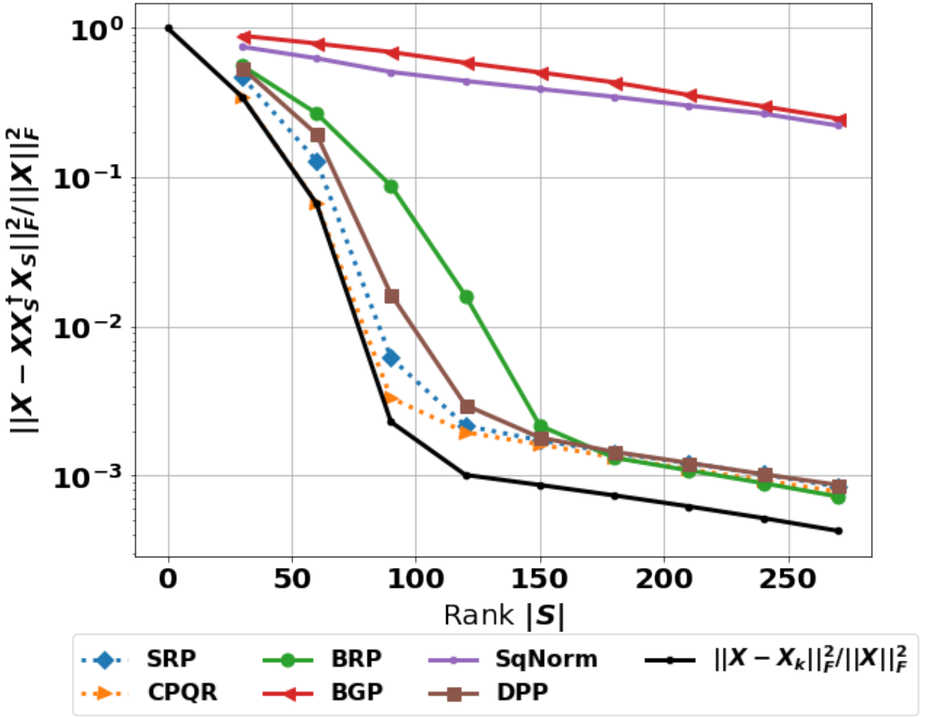
- **Inputs**: $X \in \mathbb{R}^{n \times d}$, $\tau \in (0,1)$
- $X^{(0)} \leftarrow X$, $S^{(0)} \leftarrow \emptyset$, $t \leftarrow 0$
- **while** $\mathscr{E}(S^{(t)}) > \tau \|X\|_F^2$ **do**
  - $t \leftarrow t+1$
  - Select $|S_t| = b$ skeletons S_t based on $\left(p_i\left(X^{(t-1)}\right)\right)_{i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



Pitfall of Plain Blockwise Greedy/Random Pivoting



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

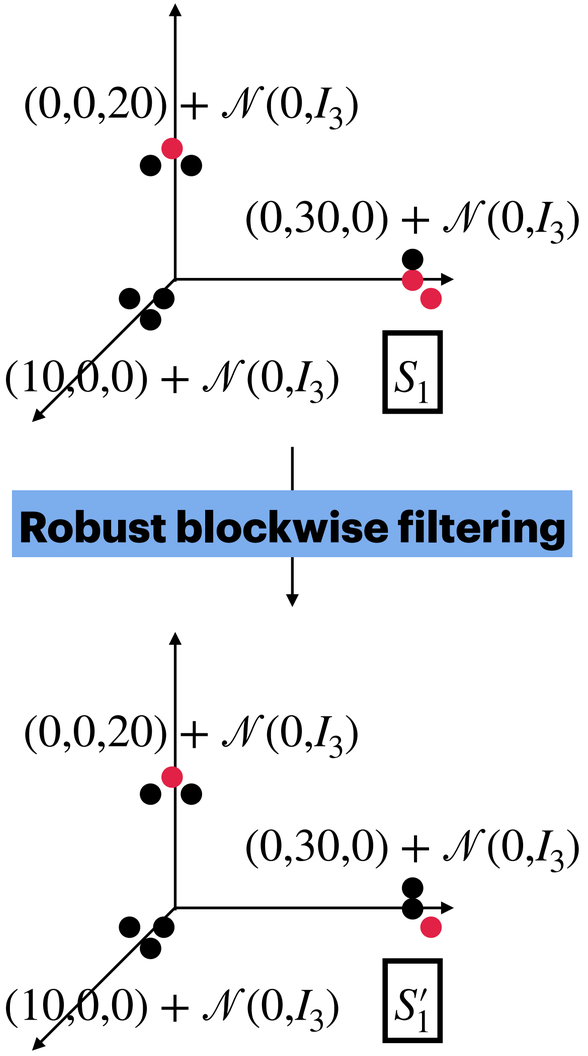


- Sequential pivoting (CPQR & SRP) is nearly optimal
- Plain blockwise pivoting (BRP/BGP, especially BGP) suffers from suboptimal skeleton complexities (up to b times)
- Squared-norm sampling (SqNorm) tends to fail

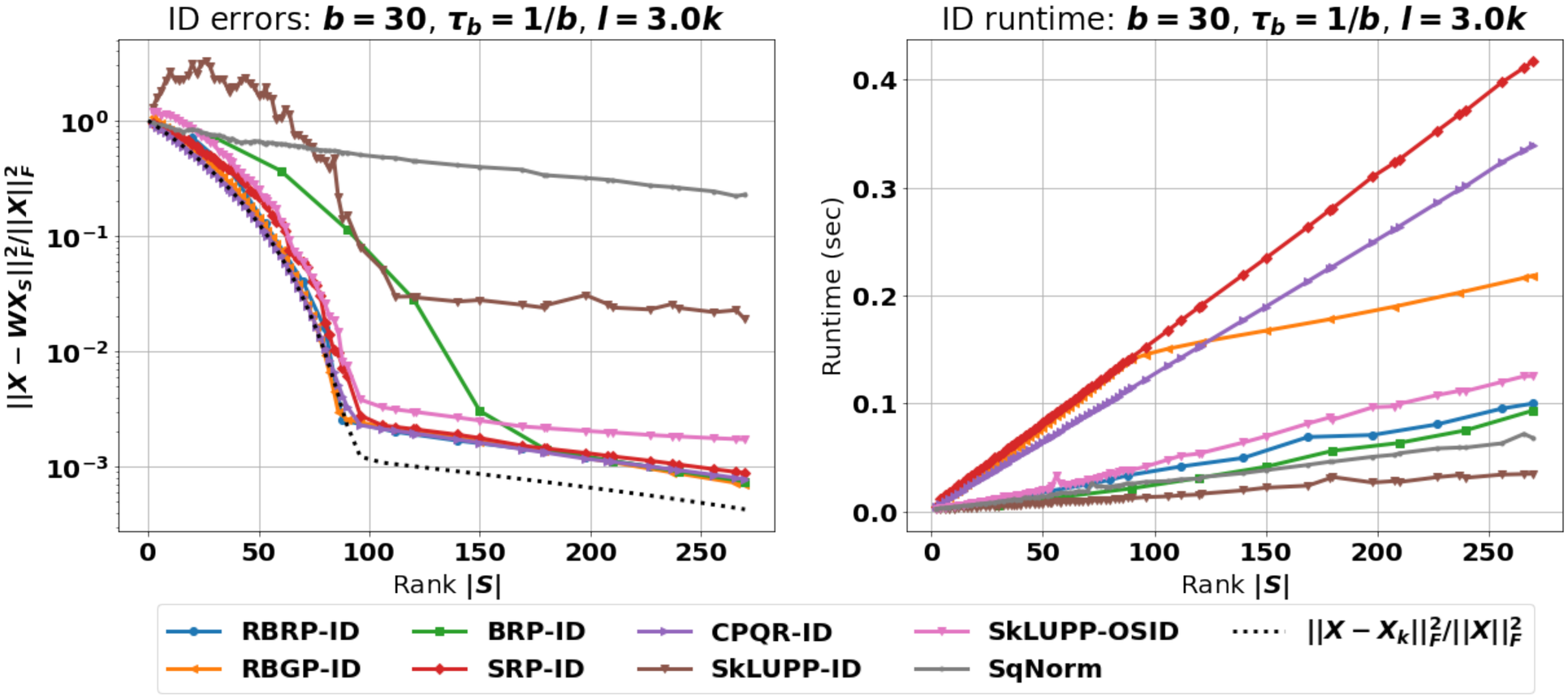
Robust Blockwise Random Pivoting

Robust Blockwise Random Pivoting (RBRP)

- **Inputs**: $X \in \mathbb{R}^{n\times d}$, $\tau \in (0,1)$
- $X^{(0)} \leftarrow X$, $S^{(0)} \leftarrow \emptyset$, $t \leftarrow 0$
- **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)**
    - $\pi \leftarrow \mathrm{CPQR}\left(X_{S_t}^{(t-1)}\right) \in S_b$ (SRP and CPQR both work)
    - $\min_{S_t'=S_t(\pi(1:b'))} b' \text{ s.t. } \|X_{S_t} - X_{S_t'}\|_F^2 < \tau_b\|X_{S_t}\|_F^2$ (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: Efficiency



- 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
- RBGP can be slowed down more significantly than RBRP by robust blockwise filtering

Summary and Questions

- **Blockwise pivoting** exploits the efficiency of Level-3 BLAS, bringing much **better hardware efficiency** than sequential pivoting
- For adversarial inputs, **plain blockwise pivoting can pick up redundant points**
- **Robust Blockwise Random Pivoting (RBRP)** leverages **robust blockwise filtering (RBF)**, a local greedy filtering step with negligible additional cost, as an effective remedy for such vulnerability
- Alternative to RBF, Epperly-Tropp-Webber-2024 showed that **rejective sampling** can also serve as a remedy for a closely related problem of Cholesky decomposition

With the shared virtue of **low intrinsic dimension**, are there connections between ID and finetuning?

Beyond low-rank approximation, are “redundant” points necessarily bad?

Data Selection for Finetuning

- Large full dataset $X = [x_1, \cdots, x_N]^\top \subset \mathcal{X}^N$, $y = [y_1, \cdots, y_N] \in \mathbb{R}^N$ drawn i.i.d. from unknown distribution P
- Finetuning function class $\mathcal{F} = \{f(\cdot\,; \theta) : \mathcal{X} \to \mathbb{R} \mid \theta \in \Theta\}$ with parameters $\Theta \subset \mathbb{R}^r$
- Pre-trained initialization $0_r \in \mathbb{R}^r$ (without loss of generality)
- Ground truth $\theta_* \in \Theta$ such that $\mathbb{E}[y \mid x] = f(x; \theta_*)$ and $\mathbb{V}[y \mid x] \leq \sigma^2$

Select a small coreset $(X_S, y_S) \subset \mathcal{X}^n \times \mathbb{R}^n$ of size n indexed by $S \subset [N]$ such that:

$$(1) \quad \theta_S = \arg\min_{\theta \in \Theta} \frac{1}{n} \|f(X_S; \theta) - y_S\|_2^2 + \alpha \|\theta\|_2^2$$

- Low-dimensional data selection: $r \leq n$, (1) = linear regression ($\alpha = 0$)
- **High-dimensional data selection**: $r > n$, (1) = ridge regression ($\alpha > 0$)

Finetuning falls in the Kernel Regime

- Finetuning dynamics fall in the **kernel regime**:

$$f(x;\theta) \approx f(x;0_r) + \nabla_\theta f(x;0_r)^\top \theta$$

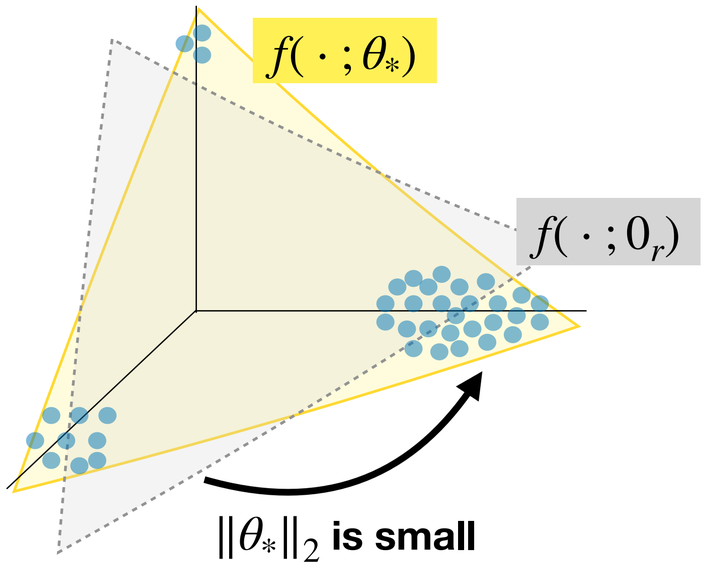
- With a **suitable pre-trained initialization** (i.e. $f(\cdot,0_r)$ is close to $f(\,\cdot\,,\theta_*)$), $\|\theta_*\|_2$ is small

- Let $G = \nabla_\theta f(X;0_r) \in \mathbb{R}^{N\times r}$ and $G_S = \nabla_\theta f(X_S;0_r) \in \mathbb{R}^{n\times r}$, (1) is well approximated by:

$$(2) \quad \theta_S = \arg\min_{\theta\in\Theta} \frac{1}{n}\|G_S\theta - (y_S - f(X_S;0_r))\|_2^2 + \alpha\|\theta\|_2^2$$

- Aim to control the excess risk $\mathrm{ER}(\theta_S) = \|\theta_S - \theta_*\|_\Sigma^2$ where $\Sigma = \mathbb{E}_{x\sim P}[\,\nabla_\theta f(x;0_r)\,\nabla_\theta f(x;0_r)^\top] \in \mathbb{R}^{r\times r}$

- Let $\Sigma_S = G_S^\top G_S/n \succeq 0$



r = number of finetunable parameters

($n < r$ in overparametrized regime)

Qs: Are there connections between ID and finetuning?

- In the **noiseless setting** $\sigma = 0$, the generalization error is controlled by the bias:

$$\mathbb{E}[\mathrm{ER}(\theta_S)] \leq \mathrm{tr}(\Sigma - \Sigma G_S^\dagger G_S)\|\theta_*\|_2^2$$

Low-rank approximation error of ID!

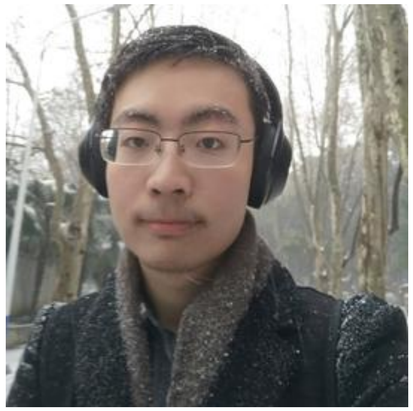
Theorem (**Variance-bias tradeoff**): Given a coreset S of size n, let $P_{\mathcal{S}} \in \mathbb{R}^{r\times r}$ be the orthogonal projector onto any subspace $\mathcal{S} \subset \mathrm{Range}(\Sigma_S)$, and $P_{\mathcal{S}}^\perp = I_r - P_{\mathcal{S}}$. There exists $\alpha > 0$ such that (2) satisfies

$$\mathbb{E}[\mathrm{ER}(\theta_S)] \leq \min_{\mathcal{S}\subset \mathrm{Range}(\Sigma_S)} \underbrace{\frac{2\sigma^2}{n}\mathrm{tr}(\Sigma(P_{\mathcal{S}}\Sigma_S P_{\mathcal{S}})^\dagger)}_{\text{variance}} + \underbrace{2\mathrm{tr}(\Sigma P_{\mathcal{S}}^\perp)\|\theta_*\|_2^2}_{\text{bias}}$$

- For a noiseless finetuning problem, accurate ID brings good data selection
- In high-dimensional data selection, **bias** is controlled by the **low-rank approximation error**
- Will see: learning with noise $\sigma > 0$, "redundant" points are critical for **variance reduction**!

Sketchy Moment Matching: Toward Fast and Provable Data Selection for Finetuning

Hoang Phan
NYU



Xiang Pan
NYU

Qi Lei
NYU

In Low Dimension: Variance Reduction

- Consider **fixed design** for simplicity: $\Sigma = \mathbb{E}_{x\sim P}[\nabla_\theta f(x; 0_r) \nabla_\theta f(x; 0_r)^\top] = G^\top G/N$
- **Low-dimensional** data selection: $\mathrm{rank}(G_S) = r \le n$ such that $\Sigma_S = G_S^\top G_S/n \succ 0$
- **V(ariance)-optimality** characterizes generalization: $\mathbb{E}[\mathrm{ER}(\theta_S)] \le \frac{\sigma^2}{n}\mathrm{tr}(\Sigma\Sigma_S^{-1})$

Uniform sampling achieves nearly optimal sample complexity in low dimension: Assuming $\|\nabla_\theta f(\cdot\,; 0_r)\|_2 \le B$ and $\Sigma \succeq \gamma I_r$. With probability $\ge 1-\delta$, X_S sampled uniformly from X satisfies

$$\Sigma \preceq c_S\Sigma_S \text{ for any } c_S > 1 \text{ when } n \gtrsim \frac{B^4}{\gamma^2(1-c_S^{-1})^2}(r + \log(1/\delta))$$

Optimal rank-t approximation (truncated SVD)

Assumption (**Low intrinsic dimension**): For $\Sigma = G^\top G/N$, let $\bar{r} = \min\{t \in [r] \mid \mathrm{tr}(\Sigma - \langle\Sigma\rangle_t) \le \mathrm{tr}(\Sigma)/N\}$ be the intrinsic dimension of the learning problem. Assume $\bar{r} \ll \min\{N, r\}$

Can the **low intrinsic dimension** of fine-tuning be leveraged when $r > n$ (Σ_S is low-rank)?

Explore Low Intrinsic Dimension: Gradient Sketching

- **Gradient sketching**: Randomly projecting the high-dimensional gradients $G = \nabla_\theta f(X; 0_r) \in \mathbb{R}^{N\times r}$ with $r > n$ to a lower-dimension $m = O(\bar{r}) \ll r$ via a Johnson-Lindenstrauss transform (JLT) $\Gamma \in \mathbb{R}^{r\times m}$
  - Common JLT: a Gaussian random matrix with i.i.d entries $\Gamma_{ij} \sim \mathcal{N}(0, 1/m)$

Theorem (**Gradient sketching**): For Gaussian embedding $\Gamma \in \mathbb{R}^{r\times m}$ with $m \geq 11\bar{r}$, let $\widetilde{\Sigma} = \Gamma^\top \Sigma \Gamma$ and $\widetilde{\Sigma}_S = \Gamma^\top \Sigma_S \Gamma$. If the coreset $S \subset [N]$ satisfies $\mathrm{rank}(\Sigma_S) = n > m$ and the $\lceil 1.1\bar{r} \rceil$-th largest eigenvalue $s_{\lceil 1.1\bar{r} \rceil}(\Sigma_S) \geq \gamma_S > 0$, then with probability at least 0.9 over Γ, there exists $\alpha > 0$ such that

$$\mathbb{E}[\mathrm{ER}(\theta_S)] \lesssim \underbrace{\frac{\sigma^2}{n}\mathrm{tr}(\widetilde{\Sigma}(\widetilde{\Sigma}_S)^\dagger)}_{\text{variance}} + \underbrace{\frac{\sigma^2}{n}\frac{1}{m\gamma_S}\|\widetilde{\Sigma}(\widetilde{\Sigma}_S)^\dagger\|_2 \mathrm{tr}(\Sigma)}_{\text{sketching error}} + \underbrace{\frac{1}{n}\|\widetilde{\Sigma}(\widetilde{\Sigma}_S)^\dagger\|_2 \mathrm{tr}(\Sigma)\|\theta_*\|_2^2}_{\text{bias}}$$

- If S further satisfies $\widetilde{\Sigma} \preceq c_S \widetilde{\Sigma}_S$ for some $c_S \geq n/N$, with $m = \max\{\sqrt{\mathrm{tr}(\Sigma)/\gamma_S}, 11\bar{r}\}$,

$$\mathbb{E}[\mathrm{ER}(\theta_S)] \lesssim \frac{c_S}{n}(\sigma^2 m + \mathrm{tr}(\Sigma)\|\theta_*\|_2^2)$$

Control Variance: Sketchy Moment Matching (SkMM)

Gradient sketching

- Draw a (fast) JLT (e.g. Gaussian random matrix) $\Gamma \in \mathbb{R}^{r\times m}$
- Sketch the gradients $\widetilde{G} = \nabla_\theta f(X; 0_r)\Gamma \in \mathbb{R}^{N\times m}$

Moment matching

- Spectral decomposition $\widetilde{\Sigma} = \widetilde{G}^\top \widetilde{G}/N = V\Lambda V^\top$ with $V = [v_1, \cdots, v_m]$, $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$
- Initialize $s = [s_1, \cdots, s_N]$ with $i \in [N]$ and $s_i = 0$ otherwis
- Sample a size-n coreset $S \subset$ that solves the optimization p

$$\min_{s\in[0,1/n]^N} \min_{\gamma=[\gamma_1,\cdots,\gamma_m]\in\mathbb{R}^m} \sum_{j=1}^m (v_j^\top \widetilde{G}^\top \mathrm{diag}(s)\, \widetilde{G} v_j - \gamma_j\lambda_j)^2$$

$$\text{s.t.} \quad \|s\|_1 = 1, \quad \gamma_j \geq 1/c_S \;\forall\, j \in [m]$$

Select $S \subset [N]$ of size $|S| = n$ that reduces the sketched V-optimality:

$$\mathrm{tr}(\widetilde{\Sigma}(\widetilde{\Sigma}_S)^\dagger)$$

Efficiency of SkMM: (recall $m \ll \min\{N, r\}$)

- **Gradient sketching** is parallelizable with input-sparsity time: for $\mathrm{nnz}(G)$ = #nonzeros in G
  - Gaussian embedding: $O(\mathrm{nnz}(G)m)$
  - Fast JLT (sparse sign): $O(\mathrm{nnz}(G)\log m)$
- **ing** takes $O(m^3)$ for spectral The optimization takes $O(Nm)$

Relaxation of $\Sigma \preceq c_S\widetilde{\Sigma}_S$:

- $\widetilde{\Sigma} \preceq c_S\widetilde{\Sigma}_S \iff V^\top(\widetilde{G})_S^\top\widetilde{G})_S/n)V \succeq \Lambda/c_S$
- Assume Σ, Σ_S commute such that imposing m diagonal constraints is sufficient

SkMM on Synthetic Data: Regression

Synthetic high-dimensional linear probing

- Gaussian mixture model (GMM) $G \in \mathbb{R}^{N \times r}$
- $N = 2000$, $r = 2400 > N$
- $\bar{r} = 8$ well separated clusters of random sizes
- Grid search for the nearly optimal $\alpha > 0$

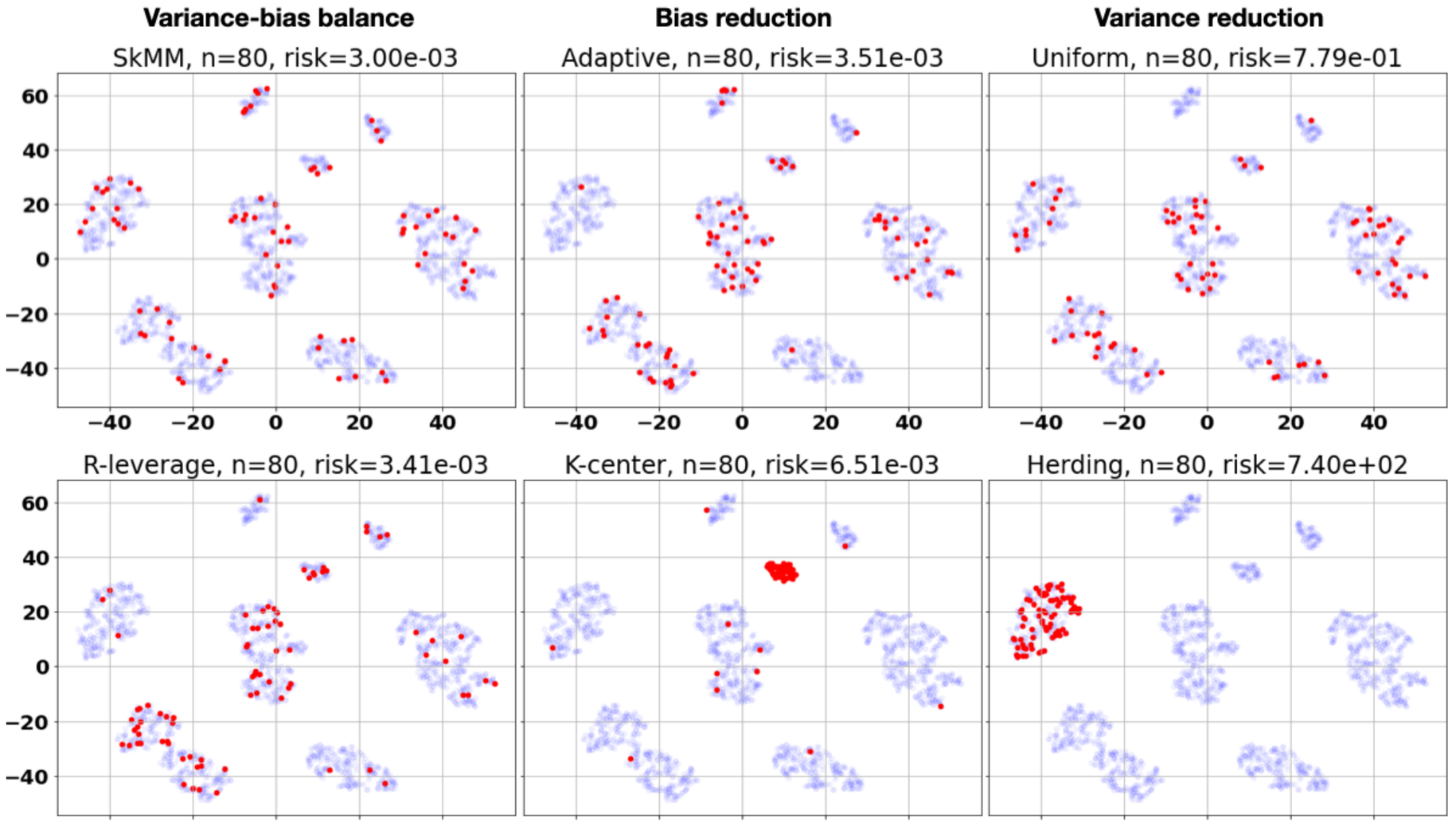
Baselines

- Herding
- Uniform sampling
- K-center greedy
- Adaptive sampling/random pivoting
- T(runcated)/R(idge) leverage score sampling

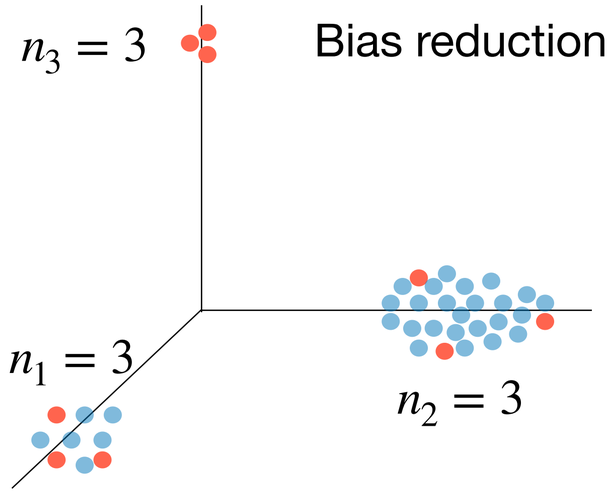
Table 1: Empirical risk $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_S)$ on the GMM dataset at various n, under the same hyperparameter tuning where ridge regression over the full dataset $\mathcal{D}$ with $N = 2000$ samples achieves $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{[N]}) =$ **2.95e-3**. For methods involving sampling, results are reported over 8 random seeds.

n	48	64	80	120	400	800	1600
Herding	7.40e+2	7.40e+2	7.40e+2	7.40e+2	7.38e+2	1.17e+2	**2.95e-3**
Uniform	(1.14 ± 2.71)e-1	(1.01 ± 2.75)e-1	(3.44 ± 0.29)e-3	(3.13 ± 0.14)e-3	(2.99 ± 0.03)e-3	**(2.96 ± 0.01)e-3**	**(2.95 ± 0.00)e-3**
K-center	(1.23 ± 0.40)e-2	(9.53 ± 0.60)e-2	(1.12 ± 0.45)e-2	(2.73 ± 1.81)e-2	(5.93 ± 4.80)e-2	(1.18 ± 0.64)e-1	(1.13 ± 0.70)e+0
Adaptive	(3.81 ± 0.65)e-3	(3.79 ± 1.37)e-3	(4.83 ± 1.90)e-3	(4.03 ± 1.35)e-3	(3.40 ± 0.67)e-3	(7.34 ± 3.97)e-3	(3.19 ± 0.16)e-3
T-leverage	(0.99 ± 1.65)e-2	(3.63 ± 0.49)e-3	(3.30 ± 0.30)e-3	(3.24 ± 0.14)e-3	**(2.98 ± 0.01)e-3**	**(2.96 ± 0.01)e-3**	**(2.95 ± 0.00)e-3**
R-leverage	(4.08 ± 1.58)e-3	(3.48 ± 0.43)e-3	(3.25 ± 0.31)e-3	(3.09 ± 0.06)e-3	(3.00 ± 0.02)e-3	(2.97 ± 0.01)e-3	**(2.95 ± 0.00)e-3**
SkMM	**(3.54 ± 0.51)e-3**	**(3.31 ± 0.15)e-3**	**(3.12 ± 0.07)e-3**	**(3.07 ± 0.08)e-3**	**(2.98 ± 0.02)e-3**	**(2.96 ± 0.01)e-3**	**(2.95 ± 0.00)e-3**

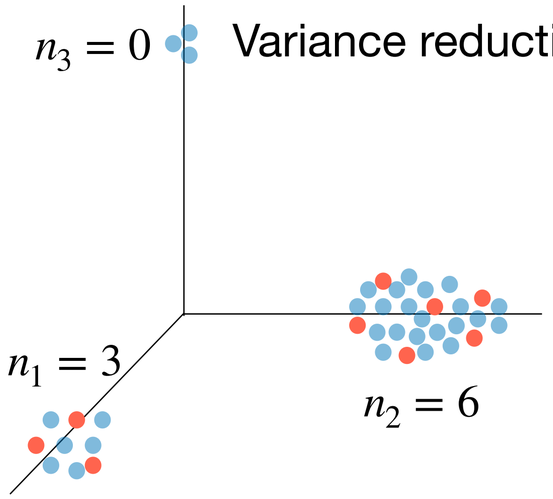
SkMM on Synthetic Data: Regression



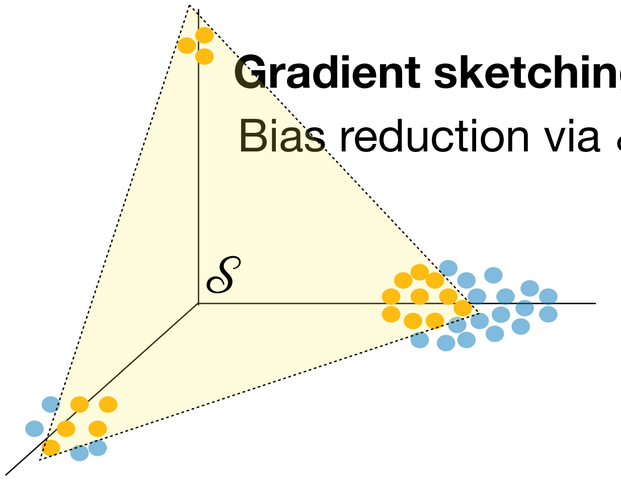
SkMM simultaneously controls variance and bias



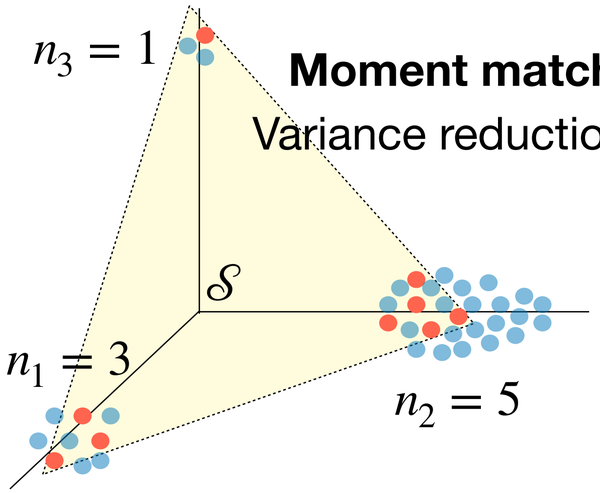
Bias reduction



Variance reduction



Gradient sketching:
Bias reduction via $\mathcal{S}$



Moment matching:
Variance reduction in $\mathcal{S}$

Thank You! Happy to take any questions

Dong, Y., Phan, H., Pan, X., & Lei, Q. Sketchy Moment Matching: Toward Fast and Provable Data Selection for Finetuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dong, Y., Chen, C., Martinsson, P. G., & Pearce, K. (2023). Robust blockwise random pivoting: Fast and accurate adaptive interpolative decomposition. *arXiv preprint arXiv:2309.16002*.