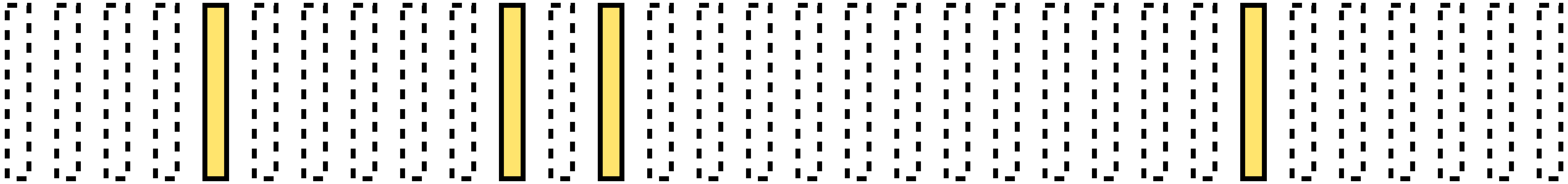


New Subset Selection Algorithms for Low Rank Approximation: Offline and Online

Taisuke Yasuda and David P. Woodruff

Carnegie Mellon University
Computer Science Department

STOC 2023

Low rank approximation

- **Low rank approximation**: given an input matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, find a rank k matrix $\hat{\mathbf{A}} \in \mathbb{R}^{n \times d}$ that is "close" to $\mathbf{A}$
  - $\min_{\mathrm{rank}(\hat{\mathbf{A}}) \le k} \mathscr{L}(\mathbf{A}, \hat{\mathbf{A}})$ for some loss function $\mathscr{L}$

- Frobenius norm low rank approximation:

$$\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \sum_{j=1}^{d} (\mathbf{A} - \hat{\mathbf{A}})_{i,j}^2 = \|\mathbf{A} - \hat{\mathbf{A}}\|_{2,2}^2$$

- Entrywise ℓ_p low rank approximation:

$$\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \sum_{j=1}^{d} |(\mathbf{A} - \hat{\mathbf{A}})_{i,j}|^p = \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,p}^p$$

Entrywise g-norm low rank approximation:

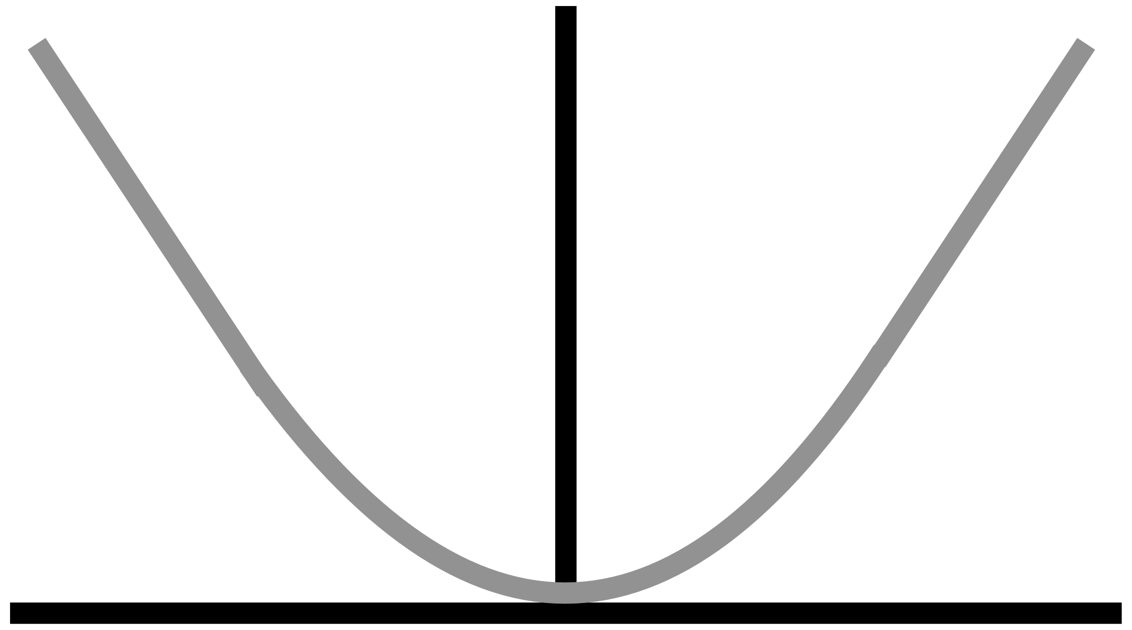
$$\boxed{\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \sum_{j=1}^{d} g((\mathbf{A} - \hat{\mathbf{A}})_{i,j}) = \|\mathbf{A} - \hat{\mathbf{A}}\|_g}$$

- ℓ_p subspace approximation:

$$\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \|\mathbf{e}_i^\top \mathbf{A} - \mathbf{e}_i^\top \hat{\mathbf{A}}\|_2^p = \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,2}^p$$

Entrywise g-norm low rank approximation

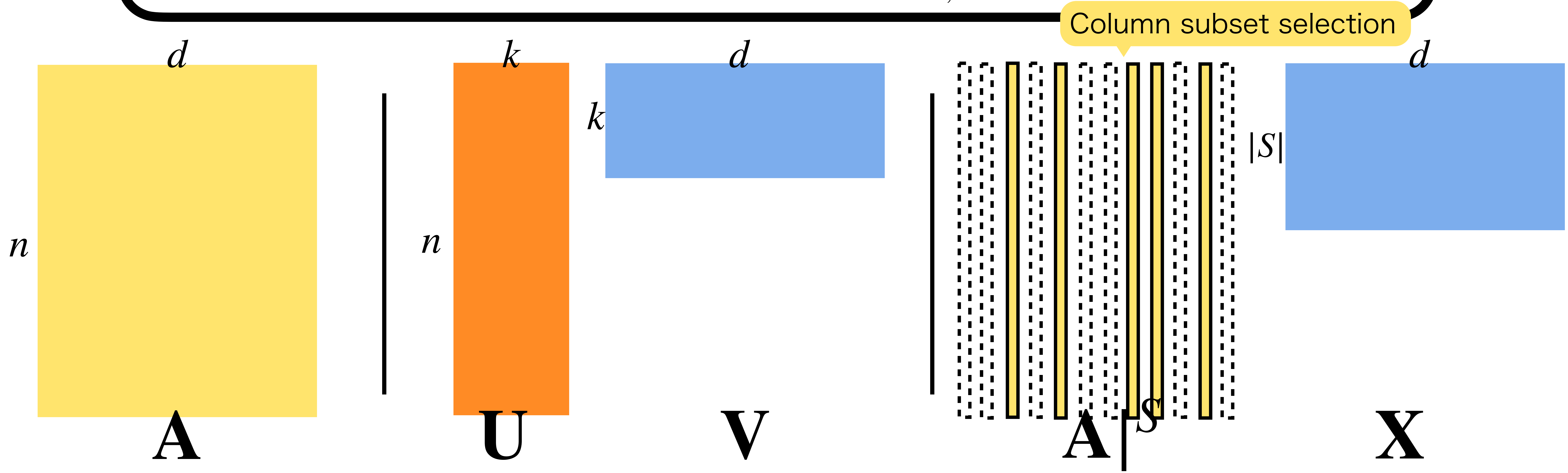
- **Assumptions on** g:
  - Symmetry: $g(x) = g(|x|)$
  - Approximate triangle inequality: $g\left(\sum_{i=1}^{t} x_i\right) \leq \mathsf{ati}_{g,t} \sum_{i=1}^{t} g(x_i)$
  - Linear growth: $|x| \leq |y| \implies \frac{g(y)}{g(x)} \geq \frac{|y|}{|x|}$



Entrywise g-norm low rank approximation

Theorem (Song-Woodruff-Zhong 2019). There is an efficient algorithm for computing a set S of $O(k\log d)$ columns of $\mathbf{A}$ such that

$$\min_{\mathbf{X}\in\mathbb{R}^{s\times d}} \|\mathbf{A}-\mathbf{A}|^S\mathbf{X}\|_g \le \tilde{O}(k)\cdot \mathsf{ati}_{g,O(k)} \min_{\mathbf{U}\in\mathbb{R}^{n\times k},\mathbf{V}\in\mathbb{R}^{k\times d}} \|\mathbf{A}-\mathbf{U}\mathbf{V}\|_g$$



Entrywise g-norm low rank approximation

Our results

Theorem (Song-Woodruff-Zhong 2019). There is an efficient algorithm for computing a set S of $O(k \log d)$ columns of $\mathbf{A}$ such that

$$\min_{\mathbf{X} \in \mathbb{R}^{s \times d}} \|\mathbf{A} - \mathbf{A}|^S \mathbf{X}\|_g \le \tilde{O}(k) \cdot \mathrm{ati}_{g,O(k)} \min_{\mathbf{U} \in \mathbb{R}^{n \times k}, \mathbf{V} \in \mathbb{R}^{k \times d}} \|\mathbf{A} - \mathbf{U}\mathbf{V}\|_g$$

Theorem (Woodruff-Y 2023). There is an efficient algorithm for computing a set S of $O(k(\log \log k)\log d)$ columns of $\mathbf{A}$ such that

$$\min_{\mathbf{X} \in \mathbb{R}^{s \times d}} \|\mathbf{A} - \mathbf{A}|^S \mathbf{X}\|_g \le \tilde{O}(\sqrt{k}) \cdot \mathrm{ati}_{g,\tilde{O}(k)} \min_{\mathbf{U} \in \mathbb{R}^{n \times k}, \mathbf{V} \in \mathbb{R}^{k \times d}} \|\mathbf{A} - \mathbf{U}\mathbf{V}\|_g$$

- Main technique: the power of relaxing **linear bases** to **spanning sets**
- Second application: nearly optimal oblivious ℓ_p subspace embeddings

Entrywise g-norm low rank approximation

Song-Woodruff-Zhong 2019: algorithm

Goal: show that "typically", this is $O(k/d) \cdot \mathrm{ati}_{g,O(k)} \cdot \|\mathbf{\Delta}\|_g$

1. Randomly sample a set H of $2k$ columns of $\mathbf{A}$
2. For each remaining column $\mathbf{a}^i$ for $i \in [d] \backslash H$, compute $\mathrm{cost}(i) := \min_{\mathbf{x}} \|\mathbf{A}|^H \mathbf{x} - \mathbf{a}^i\|_g$
3. Remove the top 0.1% of columns $i \in [d] \backslash H$ with the lowest $\mathrm{cost}(i)$
4. Repeat

$\Rightarrow$ terminate after $O(\log d)$ rounds

Notation.

Let $\mathbf{A}_*$ satisfy $\|\mathbf{A} - \mathbf{A}_*\|_g = \min_{\mathrm{rank}(\hat{\mathbf{A}}) \le k} \|\mathbf{A} - \hat{\mathbf{A}}\|_g$.

Let $\mathbf{\Delta} := \mathbf{A} - \mathbf{A}_*$.

Entrywise g-norm low rank approximation

Song-Woodruff-Zhong 2019: well-conditioned basis

- Consider a set of n vectors $\{\mathbf{a}^i\}_{i=1}^n$ in k dimensions
- **Well-conditioned basis**: subset of $|S| = k$ vectors maximizing $\det(\{\mathbf{a}^i\}_{i\in S})$
  - For any $i \in [n]$, write $\mathbf{a}^i = \sum_{j\in S} x_j \cdot \mathbf{a}^j$
  - By Cramer's rule, $|x_j| = \dfrac{|\det(\{\mathbf{a}^l\}_{l\in S-j+i})|}{\det(\{\mathbf{a}^l\}_{l\in S})} \le 1$
  - That is, any $\mathbf{a}^i$ can be written in this basis with **small coefficients in** ℓ_∞

Entrywise g-norm low rank approximation

Song-Woodruff-Zhong 2019: correctness

Goal: show that "typically", this is $O(k/d) \cdot \mathrm{ati}_{g,O(k)} \cdot \|\mathbf{\Delta}\|_g$

1. Randomly sample a set H of $2k$ columns of $\mathbf{A}$
2. For each remaining column $\mathbf{a}^i$ for $i \in [d] \backslash H$, compute $\mathrm{cost}(i) := \min_{\mathbf{x}} \|\mathbf{A}|^H \mathbf{x} - \mathbf{a}^i\|_g$
3. Remove the top 0.1% of columns $i \in [d] \backslash H$ with the lowest $\mathrm{cost}(i)$
4. Repeat

Notation.

Let $\mathbf{A}_*$ satisfy $\|\mathbf{A} - \mathbf{A}_*\|_g = \min_{\mathrm{rank}(\hat{\mathbf{A}}) \le k} \|\mathbf{A} - \hat{\mathbf{A}}\|_g$.

Let $\mathbf{\Delta} := \mathbf{A} - \mathbf{A}_*$.

Approximate triangle inequality + linear growth

$$\min_{\mathbf{x}} \|\mathbf{A}|^H \mathbf{x} - \mathbf{a}^i\|_g \le \|\mathbf{A}|^H \mathbf{x}_* - \mathbf{a}^i\|_g = \|\mathbf{\Delta}|^H \mathbf{x}_* - \mathbf{\Delta}^i\|_g \le \mathrm{ati}_{g,2k+1} \sum_{j \in H+i} |(\mathbf{x}_*)_j| \|\mathbf{\Delta}^j\|_g \lesssim \frac{2k+1}{d} \cdot \mathrm{ati}_{g,2k+1} \sum_{j=1}^{d} \|\mathbf{\Delta}^j\|_g$$

Well-conditioned basis $\Rightarrow$ $\mathbf{a}^i_* = \mathbf{A}_*|^H \mathbf{x}_*$ for $\|\mathbf{x}_*\|_\infty \le 1$

If i is random, $H + i$ is a set of $2k+1$ random columns

Entrywise g-norm low rank approximation

Woodruff-Y 2023: well-conditioned ~~basis~~ spanning set

- Consider a set of n vectors $\{\mathbf{a}^i\}_{i=1}^n$ in k dimensions
- **Well-conditioned basis**: $|S| = k$ vectors s.t. $\mathbf{a}^i = \mathbf{A}|^S\mathbf{x}$ with $\|\mathbf{x}\|_\infty \leq 1$
- **Well-conditioned spanning set**: $|S| = \tilde{O}(k)$ vectors s.t. $\mathbf{a}^i = \mathbf{A}|^S\mathbf{x}$ with $\|\mathbf{x}\|_2 \leq 1.1$
  - Construction: coresets for John ellipsoids [Kumar-Yildirim 2005, Todd 2016]

Entrywise g-norm low rank approximation

Woodruff-Y 2023: correctness

1. Randomly sample a set H of $2k$ columns of $\mathbf{A}$
2. For each remaining column $\mathbf{a}^i$ for $i \in [d]\backslash H$, compute $\mathrm{cost}(i) := \min_{\mathbf{x}} \|\mathbf{A}|^H\mathbf{x} - \mathbf{a}^i\|_g$
3. Remove the top 0.1% of columns $i \in [d]\backslash H$ with the lowest $\mathrm{cost}(i)$
4. Repeat

Goal: show that "typically", this is $O(\sqrt{k}/d)$ ~~$O(k/d)$~~ $\cdot\ \mathsf{ati}_{g,O(k)} \cdot \|\mathbf{\Delta}\|_g$

Notation.

Let $\mathbf{A}_*$ satisfy $\|\mathbf{A} - \mathbf{A}_*\|_g = \min_{\mathrm{rank}(\hat{\mathbf{A}}) \le k} \|\mathbf{A} - \hat{\mathbf{A}}\|_g$.

Let $\mathbf{\Delta} := \mathbf{A} - \mathbf{A}_*$.

$$\min_{\mathbf{x}} \|\mathbf{A}|^H\mathbf{x} - \mathbf{a}^i\|_g \le \|\mathbf{A}|^H\mathbf{x}_* - \mathbf{a}^i\|_g = \|\mathbf{\Delta}|^H\mathbf{x}_* - \mathbf{\Delta}^i\|_g \le \mathsf{ati}_{g,2k+1} \sum_{j \in H+i} |(\mathbf{x}_*)_j| \|\mathbf{\Delta}^j\|_g \le \mathsf{ati}_{g,2k+1} \|\mathbf{x}_*\|_2 \left(\sum_{j \in H+i} \|\mathbf{\Delta}^j\|_g^2\right)^{1/2}$$

Cauchy-Schwarz

If $\|\mathbf{\Delta}^j\|_g$ are ~same

$$\lesssim \mathsf{ati}_{g,2k+1} \frac{1}{\sqrt{k}} \sum_{j \in H+i} \|\mathbf{\Delta}^j\|_g \lesssim \mathsf{ati}_{g,2k+1} \frac{\sqrt{k}}{d} \sum_{j=1}^{d} \|\mathbf{\Delta}^j\|_g$$

Entrywise g-norm low rank approximation

Our results

Theorem (Song-Woodruff-Zhong 2019). There is an efficient algorithm for computing a set S of $O(k \log d)$ columns of $\mathbf{A}$ such that

$$\min_{\mathbf{X}\in\mathbb{R}^{s\times d}} \|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\|_g \leq \tilde{O}(k) \cdot \mathrm{ati}_{g,O(k)} \min_{\mathbf{U}\in\mathbb{R}^{n\times k},\mathbf{V}\in\mathbb{R}^{k\times d}} \|\mathbf{A} - \mathbf{U}\mathbf{V}\|_g$$

Theorem (Woodruff-Y 2023). There is an efficient algorithm for computing a set S of $O(k(\log\log k)\log d)$ columns of $\mathbf{A}$ such that

$$\min_{\mathbf{X}\in\mathbb{R}^{s\times d}} \|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\|_g \leq \tilde{O}(\sqrt{k}) \cdot \mathrm{ati}_{g,\tilde{O}(k)} \min_{\mathbf{U}\in\mathbb{R}^{n\times k},\mathbf{V}\in\mathbb{R}^{k\times d}} \|\mathbf{A} - \mathbf{U}\mathbf{V}\|_g$$

- Main technique: the power of relaxing **linear bases** to **spanning sets**
- Second application: nearly optimal oblivious ℓ_p subspace embeddings

Nearly optimal ℓ_p oblivious subspace embeddings

Applications of well-conditioned spanning sets

Theorem (Woodruff-Wang 2019). Let $1 < p < 2$. There is a distribution over matrices $\mathbf{S} \in \mathbb{R}^{\tilde{O}(d)\times n}$ such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$, w.p. ≥ 0.99,

$$\forall \mathbf{x} \in \mathbb{R}^d, \|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p \leq \tilde{O}(d)\|\mathbf{A}\mathbf{x}\|_p$$

Theorem (Woodruff-Y 2023). Let $1 < p < 2$. There is a distribution over matrices $\mathbf{S} \in \mathbb{R}^{\tilde{O}(d)\times n}$ such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$, w.p. ≥ 0.99,

$$\forall \mathbf{x} \in \mathbb{R}^d, \|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p \leq \tilde{O}(d^{1/p})\|\mathbf{A}\mathbf{x}\|_p$$

- Apply well-conditioned spanning sets to the set $\{\mathbf{A}\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_p \leq 1\}$
- Closes a long line of work initiated by Sohler-Woodruff 2011

Low rank approximation

- **Low rank approximation**: given an input matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, find a rank k matrix $\hat{\mathbf{A}} \in \mathbb{R}^{n \times d}$ that is "close" to $\mathbf{A}$
  - $\min_{\mathrm{rank}(\hat{\mathbf{A}}) \le k} \mathscr{L}(\mathbf{A}, \hat{\mathbf{A}})$ for some loss function $\mathscr{L}$

- Frobenius norm low rank approximation:

$$\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \sum_{j=1}^{d} (\mathbf{A} - \hat{\mathbf{A}})_{i,j}^2 = \|\mathbf{A} - \hat{\mathbf{A}}\|_{2,2}^2$$

- Entrywise ℓ_p low rank approximation:

$$\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \sum_{j=1}^{d} |(\mathbf{A} - \hat{\mathbf{A}})_{i,j}|^p = \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,p}^p$$

- Entrywise g-norm low rank approximation:

$$\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \sum_{j=1}^{d} g((\mathbf{A} - \hat{\mathbf{A}})_{i,j}) = \|\mathbf{A} - \hat{\mathbf{A}}\|_g$$

ℓ_p subspace approximation:

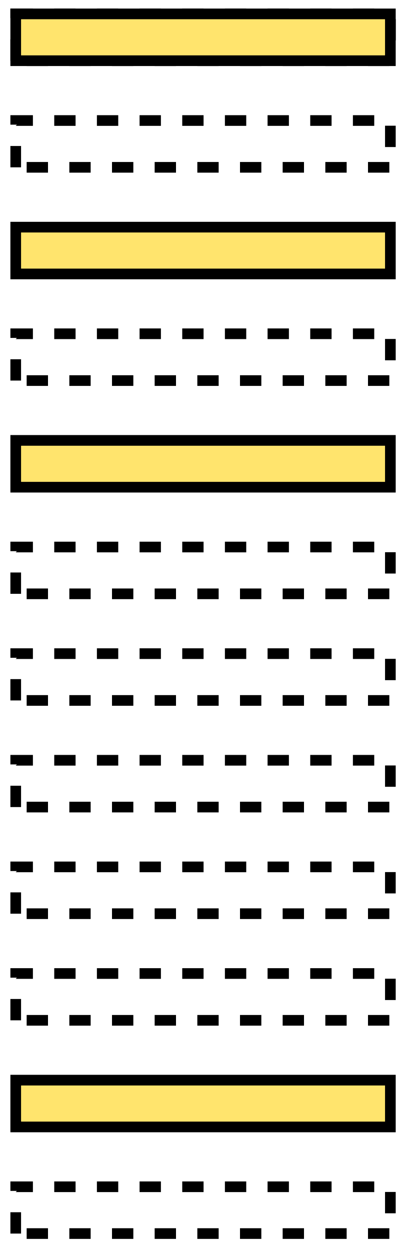
$$\mathscr{L}(\mathbf{A}, \hat{\mathbf{A}}) = \sum_{i=1}^{n} \|\mathbf{e}_i^\top \mathbf{A} - \mathbf{e}_i^\top \hat{\mathbf{A}}\|_2^p = \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,2}^p$$

Online coresets for ℓ_p subspace approximation

- ℓ_p **subspace approximation:**
  - For a rank k subspace $F \subseteq \mathbb{R}^d$, let $\mathbf{P}_F$ be the orthogonal projection matrix
  - $\min_{\operatorname{rank}(\hat{\mathbf{A}}) \le k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,2}^p = \min_{\operatorname{rank}(F) \le k} \|\mathbf{A} - \mathbf{A}\mathbf{P}_F\|_{p,2}^p$
- **Coresets:** weighted subset of $\{\mathbf{a}_i\}_{i=1}^n$ s.t. for all rank k subspaces $F \subseteq \mathbb{R}^d$,

$$\sum_{i=1}^n w_i \|\mathbf{e}_i^\top(\mathbf{A} - \mathbf{A}\mathbf{P}_F)\|_2^p = (1 \pm \varepsilon) \sum_{i=1}^n \|\mathbf{e}_i^\top(\mathbf{A} - \mathbf{A}\mathbf{P}_F)\|_2^p = (1 \pm \varepsilon)\|\mathbf{A} - \mathbf{A}\mathbf{P}_F\|_{p,2}^p$$

- **Online coresets:** rows $\{\mathbf{a}_i\}_{i=1}^n$ arrive one by one, select subset online



Question. Do small online coresets for ℓ_p subspace approximation exist?

Online coresets for ℓ_p subspace approximation

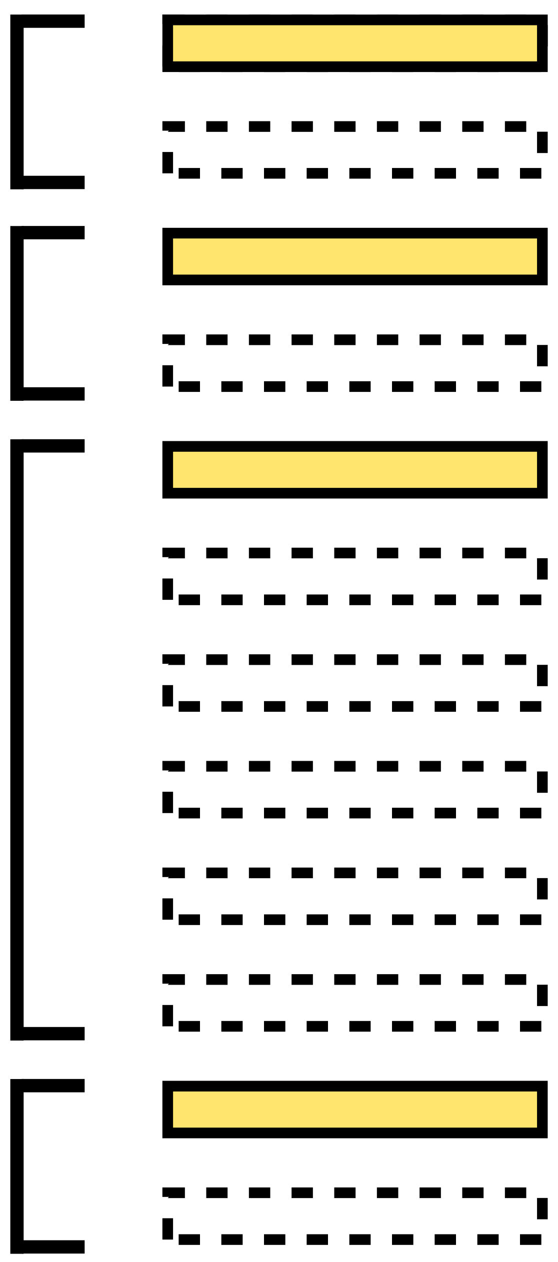
- **Coresets ℓ_p subspace approximation:**
  - Typical coreset algorithm:
    1. Compute a constant factor solution to ℓ_p subspace approximation
    2. Sample $\mathrm{poly}(k/\varepsilon)$ rows proportionally to the residual cost

This is a sequential procedure — how can this be implemented online?

Preserves only an approximately optimal solution (rather than all candidates)

Idea. Reduction to constant factor "weak" online coresets
- A constant factor online coreset of size s partitions the stream into s "phases"
- In each "phase", the online coreset defines a fixed constant factor solution
- Can sample proportionally to the residual cost in each "phase"

Lemma. Weak online coresets = Johnson—Lindenstrauss lemma + online coreset for ℓ_p subspace embeddings (Woodruff-Y 2023)



Online coresets for ℓ_p subspace approximation

Theorem (Woodruff-Y 2023). Let $1 \leq p < \infty$. There is an online coreset algorithm that computes weights w_i with at most $s = \mathrm{poly}(k, \varepsilon^{-1}, \log(n\kappa^{\mathsf{OL}}))$ nonzero weights s.t.

$$\sum_{i=1}^{n} w_i \|\mathbf{e}_i^\top(\mathbf{A} - \mathbf{A}\mathbf{P}_F)\|_2^p = (1 \pm \varepsilon) \sum_{i=1}^{n} \|\mathbf{e}_i^\top(\mathbf{A} - \mathbf{A}\mathbf{P}_F)\|_2^p = (1 \pm \varepsilon)\|\mathbf{A} - \mathbf{A}\mathbf{P}_F\|_{p,2}^p$$

Online condition number κ^{OL}

Subset Selection for Low Rank Approximation

Summary

- We study new subset selection algorithms for low rank approximation
  - Entrywise g-norm low rank approximation
    - We give a new structural result on a **relaxation of well-conditioned basis** to a **well-conditioned spanning set** with better conditioning properties
    - We improve the distortion of prior subset selection algorithms by a $\sqrt{k}$ factor
    - We improve the distortion of oblivious ℓ_p subspace embeddings from $\tilde{O}(d)$ to a nearly optimal $\tilde{O}(d^{1/p})$
  - Online coresets for ℓ_p subspace approximation
    - We give the first $\mathrm{poly}(k, \varepsilon^{-1}, \log(n\kappa^{\mathsf{OL}}))$-sized strong coreset for ℓ_p subspace approximation
    - We show a reduction to constant factor weak online coresets, which we obtain via Johnson-Lindenstrauss together with online coresets for ℓ_p subspace embeddings