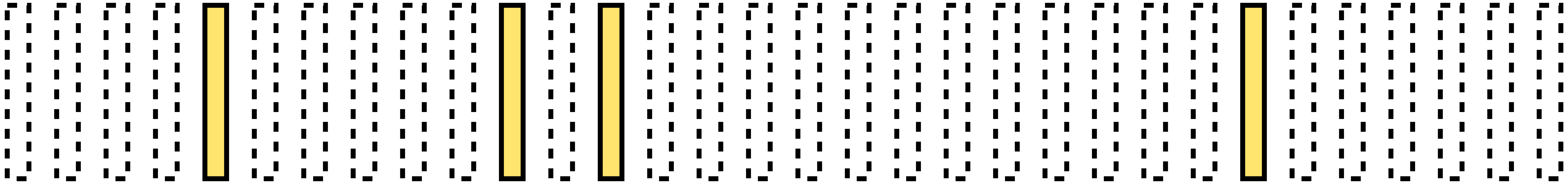


Thesis Defense

Algorithms for Matrix Approximation:

Sketching, Sampling, and Sparse Optimization

Taisuke (Tai) Yasuda

Carnegie Mellon University
Computer Science Department

Thesis Committee:

- David P. Woodruff (Chair)
- Anupam Gupta
- Richard Peng
- Cameron Musco (University of Massachusetts Amherst)

Matrix Approximation

"Turning big data to tiny data"

- Matrix approximation algorithms
  - **Input**: a large matrix
  - **Output**: a small/structured matrix which "resembles" the input matrix
- Why is matrix approximation interesting?
  - Widespread use in real world **engineering applications**
    - Efficient algorithms for processing large datasets
  - Numerous **connections to other fields** of theoretical computer science
    - Optimization, computational geometry, sublinear algorithms, etc
  - Information and communication theoretic **lower bounds**
    - Understand how good our algorithms are

Matrix Approximation

Randomized Numerical Linear Algebra

- **Numerical linear algebra**: deterministic algorithms for solving linear algebra to machine precision
- **Randomized numerical linear algebra**: randomized approximation algorithms for numerical linear algebra
    - Randomized: succeed every time → succeed with 99% probability
    - Approximation: solve exactly → solve up to (small) error
    - **This flexibility leads to extraordinary improvements in efficiency!**
    - Key techniques: **sketching** (dimensionality reduction) and **sampling**

Matrix Approximation

Sparse Optimization

- Approximation: (often convex) **optimization**
- Small/structured matrix: **sparsity** in an appropriate sense
- Matrix approximation problems are often captured by sparse optimization

Minimize $f(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^d$

s.t. $\mathbf{x}$ has at most k nonzero entries

- Key techniques: **greedy algorithms**, **convex relaxations**

Matrix Approximation

Overview of the Talk

- Part I. **Sketching**
  - Oblivious ℓ_p subspace embeddings
- Part II. **Sampling**
  - Active ℓ_p linear regression
  - Coresets for ℓ_p subspace approximation
  - Streaming computational geometry
- Part III. **Sparse Optimization**
  - Column subset selection

Part I. Sketching

Linear Regression

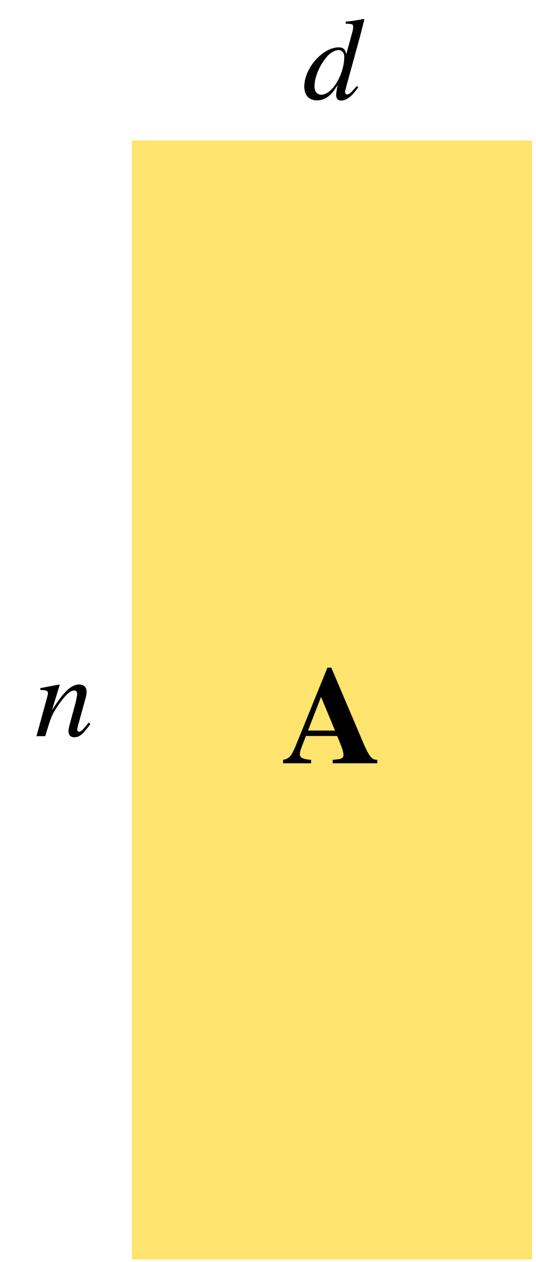
- **Linear regression**: given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and a vector $\mathbf{b} \in \mathbb{R}^n$, solve

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

  - Widely used model for supervised learning
  - Building block for complex models and algorithms
  - Can we design efficient approximation algorithms for linear regression?

Part I. Sketching

Linear Regression



$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$$

- One approach: approximate the original instance by a smaller instance

Goal. Replace $\mathbf{A}$ and $\mathbf{b}$ by a smaller $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$ s.t.

$\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 \approx \|\tilde{\mathbf{A}}\mathbf{x}-\tilde{\mathbf{b}}\|_2$ for every $\mathbf{x}\in\mathbb{R}^d$

- There are many possible ways to choose $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$!

Idea. Choose $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$ and $\tilde{\mathbf{b}} = \mathbf{S}\mathbf{b}$ for some $\mathbf{S}\in\mathbb{R}^{r\times n}$, $r \ll n$

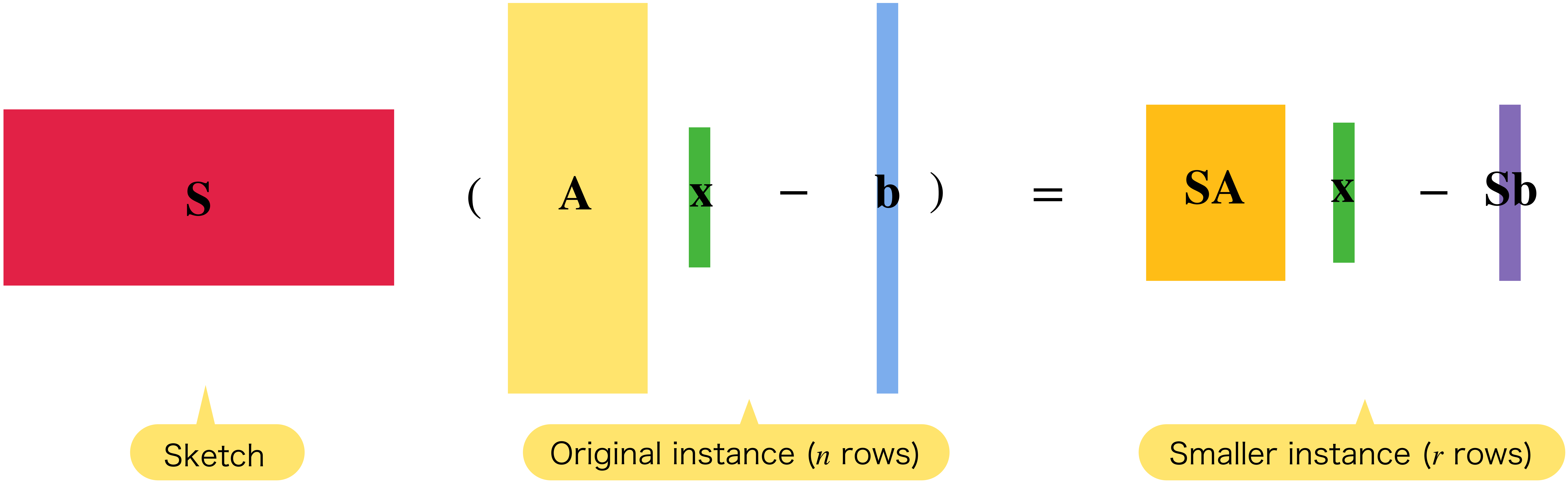
"Sketch"

Part I. Sketching

Linear Regression

Goal. Replace $\mathbf{A}$ and $\mathbf{b}$ by a smaller $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$ s.t. $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \approx \|\tilde{\mathbf{A}}\mathbf{x} - \tilde{\mathbf{b}}\|_2$ for every $\mathbf{x} \in \mathbb{R}^d$

Idea. Choose $\tilde{\mathbf{A}} = \mathbf{S}\mathbf{A}$ and $\tilde{\mathbf{b}} = \mathbf{S}\mathbf{b}$ for some $\mathbf{S} \in \mathbb{R}^{r \times n}$, $r \ll n$



Part I. Sketching

Linear Regression

Definition (Sarlos 2006). $\mathbf{S} \in \mathbb{R}^{r \times n}$ is a subspace embedding of $\mathbf{A} \in \mathbb{R}^{n \times d}$ if

$$\|\mathbf{Ax}\|_2 \le \|\mathbf{SAx}\|_2 \le \kappa \|\mathbf{Ax}\|_2$$

for every $\mathbf{x} \in \mathbb{R}^d$.

Distortion/error

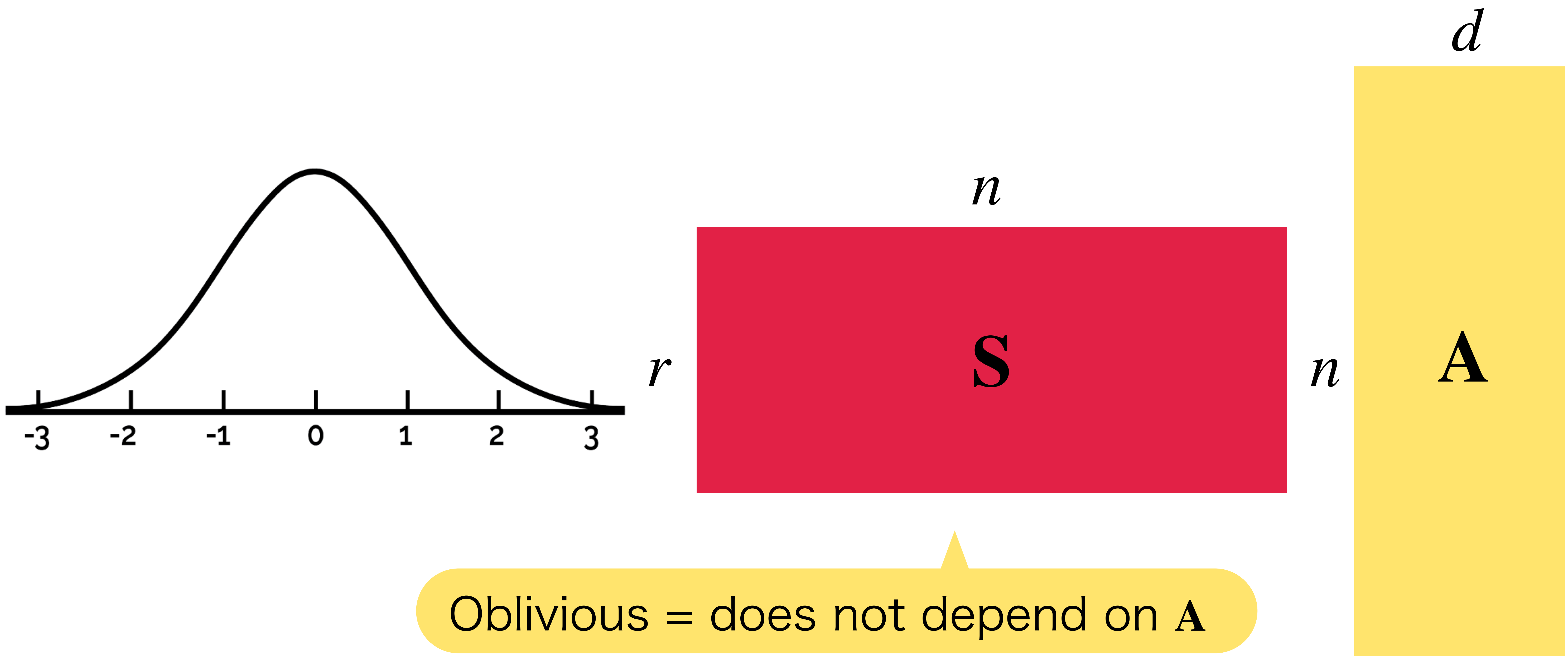
- Why is this useful?
  - Let $\mathbf{A}' = [\mathbf{A}\ \mathbf{b}] \in \mathbb{R}^{n \times (d+1)}$ and let $\mathbf{S} \in \mathbb{R}^{r \times n}$ be a subspace embedding of $\mathbf{A}'$
  - $\mathbf{Ax} - \mathbf{b} = \mathbf{A}'\mathbf{x}'$ for $\mathbf{x}' = [\mathbf{x}; -1] \in \mathbb{R}^{d+1}$
  - $\|\mathbf{Ax} - \mathbf{b}\|_2 \le \|\mathbf{SAx} - \mathbf{Sb}\|_2 \le \kappa \|\mathbf{Ax} - \mathbf{b}\|_2$

    Cost of $\mathbf{SA}$ and $\mathbf{Sb}$ approximates the cost of $\mathbf{A}$ and $\mathbf{b}$
  - Solve linear regression on $\mathbf{SA}$ and $\mathbf{Sb}$ instead of $\mathbf{A}$ and $\mathbf{b}$

Part I. Sketching

Linear Regression

Theorem (Sarlos 2006). Let $\kappa = (1+\varepsilon)$. Let $r = \tilde{O}(\varepsilon^{-2} d)$. If $\mathbf{S}$ is an $r \times n$ Gaussian matrix, then for every $\mathbf{A}$, $\mathbf{S}$ is a subspace embedding for $\mathbf{A}$ with distortion κ, with probability 99%.



Why oblivious embeddings?

- Useful when $\mathbf{A}$ is unknown
  - Turnstile streaming
  - Distributed computation

So ℓ_2 linear regression is resolved.

What's next?

Part I. Sketching

ℓ_p Linear Regression

ℓ_1 linear regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_1$$

- Minimize the **average** error
- **Robust** loss function

ℓ_2 linear regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$$

- Minimize the **sum of squares** of errors

ℓ_∞ linear regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_\infty$$

- Minimize the **worst-case** error
- **Sensitive** loss function

ℓ_p linear regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p$$

Question. What trade-offs are possible for oblivious ℓ_p subspace embeddings?

Part I. Sketching

Oblivious ℓ_p Subspace Embeddings

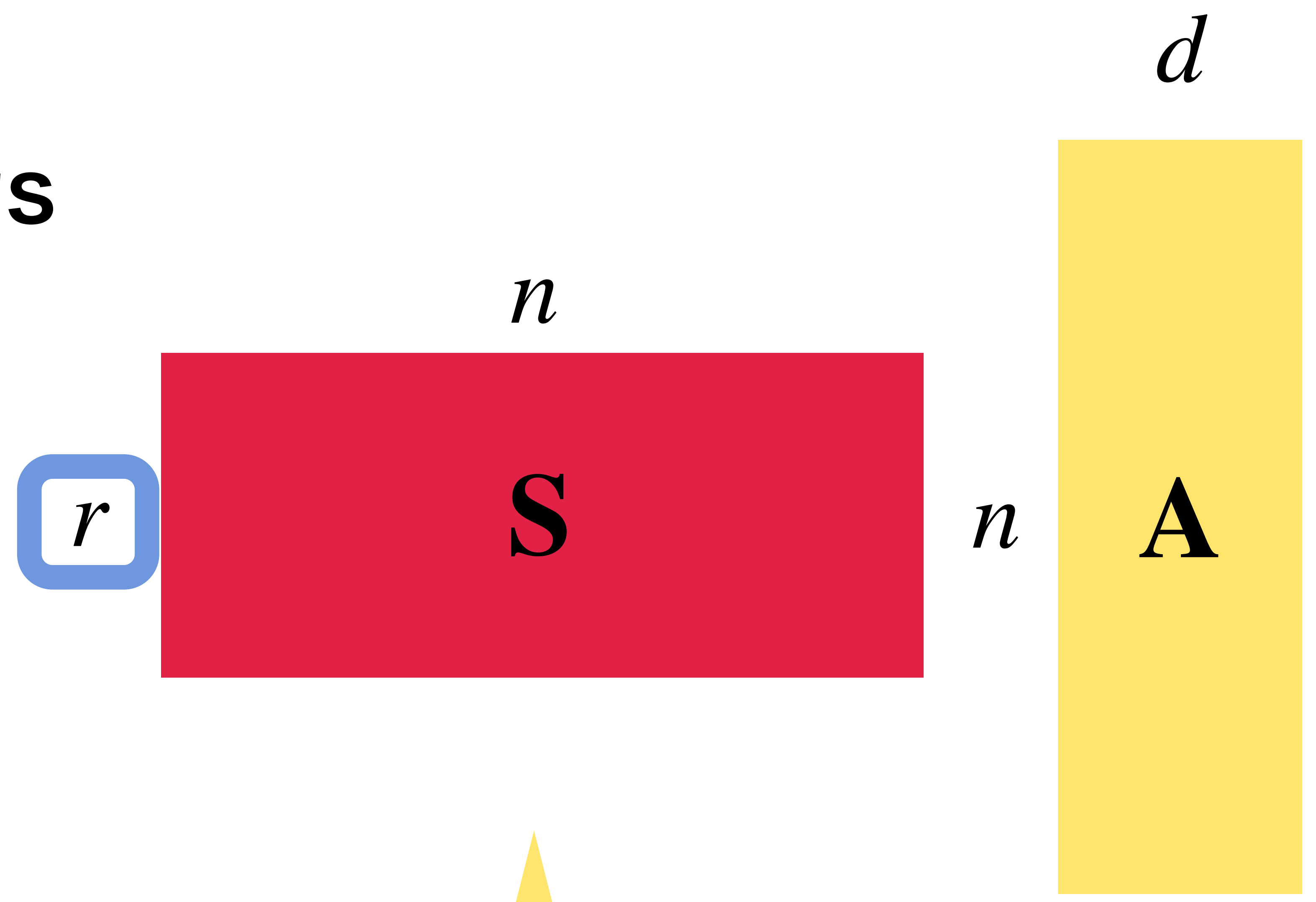
Some bad news…

Theorem (Wang—Woodruff 2019). If $\mathbf{S}$ is an oblivious ℓ_p subspace embedding for $p \in [1,2)$, then…

- $r \leq \mathrm{poly}(d) \Rightarrow \kappa \gtrsim d^{1/p}$
- $\kappa \leq O(1) \Rightarrow r \gtrsim \exp(\sqrt{d})$

Could be refined to $(1+\varepsilon)$ via iterative optimization, sampling

Theorem (Li—Lin—Woodruff—Zhang 2022). If $\mathbf{S}$ is an oblivious ℓ_p subspace embedding for $p \in (2,\infty)$, then $r \cdot \kappa^2 \gtrsim n^{1-2/p}$



Oblivious = does not depend on $\mathbf{A}$

s.t. $\|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p \leq \boxed{\kappa}\|\mathbf{A}\mathbf{x}\|_p$

for every $\mathbf{x} \in \mathbb{R}^d$

Part I. Sketching

Oblivious ℓ_p Subspace Embeddings

Some bad news...

Theorem (Wang—Woodruff 2019). If $\mathbf{S}$ is an oblivious ℓ_p subspace embedding for $p \in [1,2)$, then...

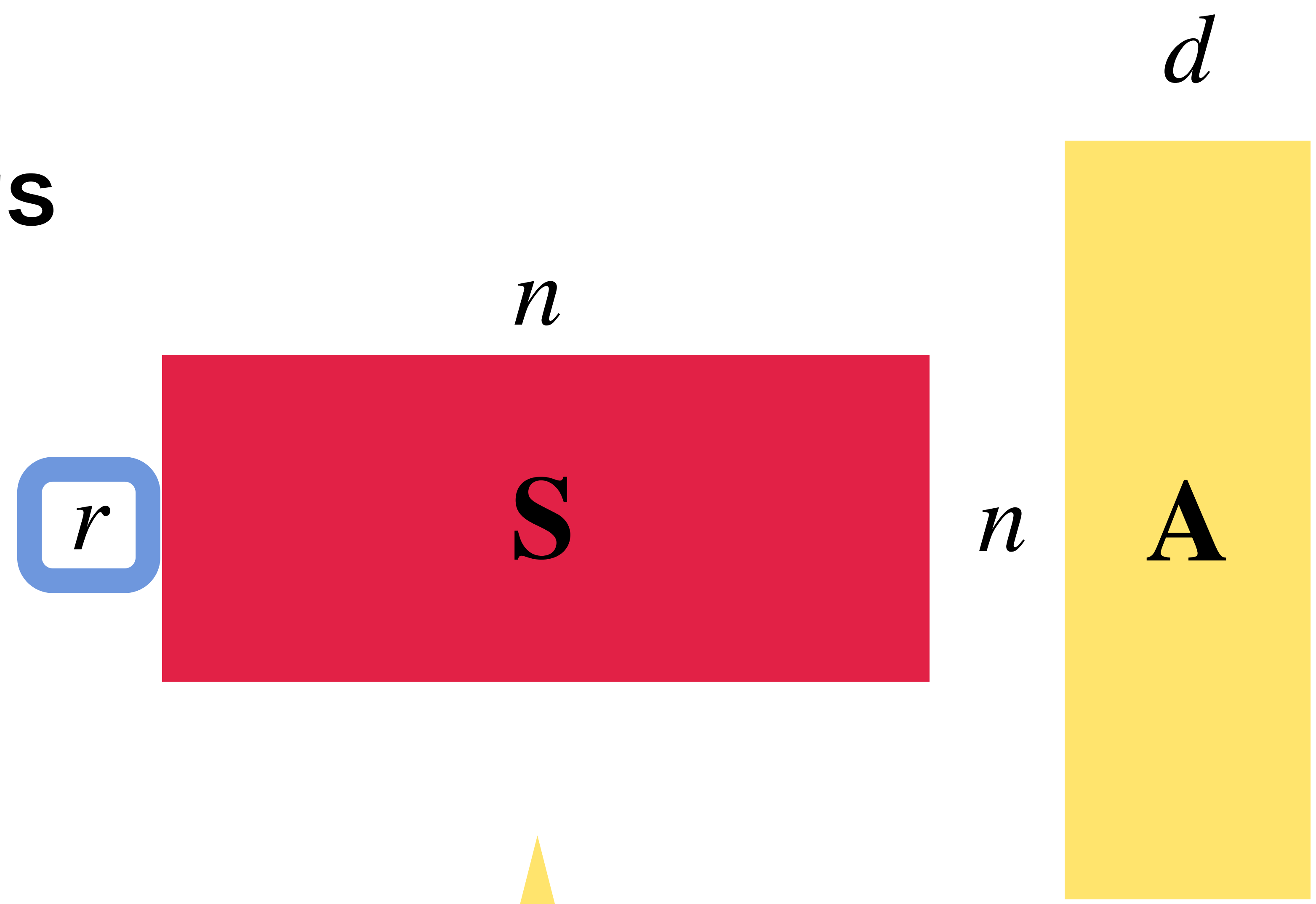
- $r \leq \mathrm{poly}(d) \Rightarrow \kappa \gtrsim d^{1/p}$
- $\kappa \leq O(1) \Rightarrow r \gtrsim \exp(\sqrt{d})$

Could be refined to $(1+\varepsilon)$ via iterative optimization, sampling

How close are these bounds from the truth? Pretty close!

Theorem. There exist oblivious ℓ_p subspace embeddings for $p \in [1,2)$, s.t. $\cdots$

- $r = \tilde{O}(d)$, $\kappa \lesssim d^{1/p}$ **[Woodruff—Y 2023]**
- $\kappa = (1+\varepsilon)$, $r \lesssim \exp(\varepsilon^{-1} d)$ **[Li—Woodruff—Y 2021]**



Oblivious = does not depend on $\mathbf{A}$

s.t. $\|\mathbf{Ax}\|_p \leq \|\mathbf{SAx}\|_p \leq \kappa \|\mathbf{Ax}\|_p$

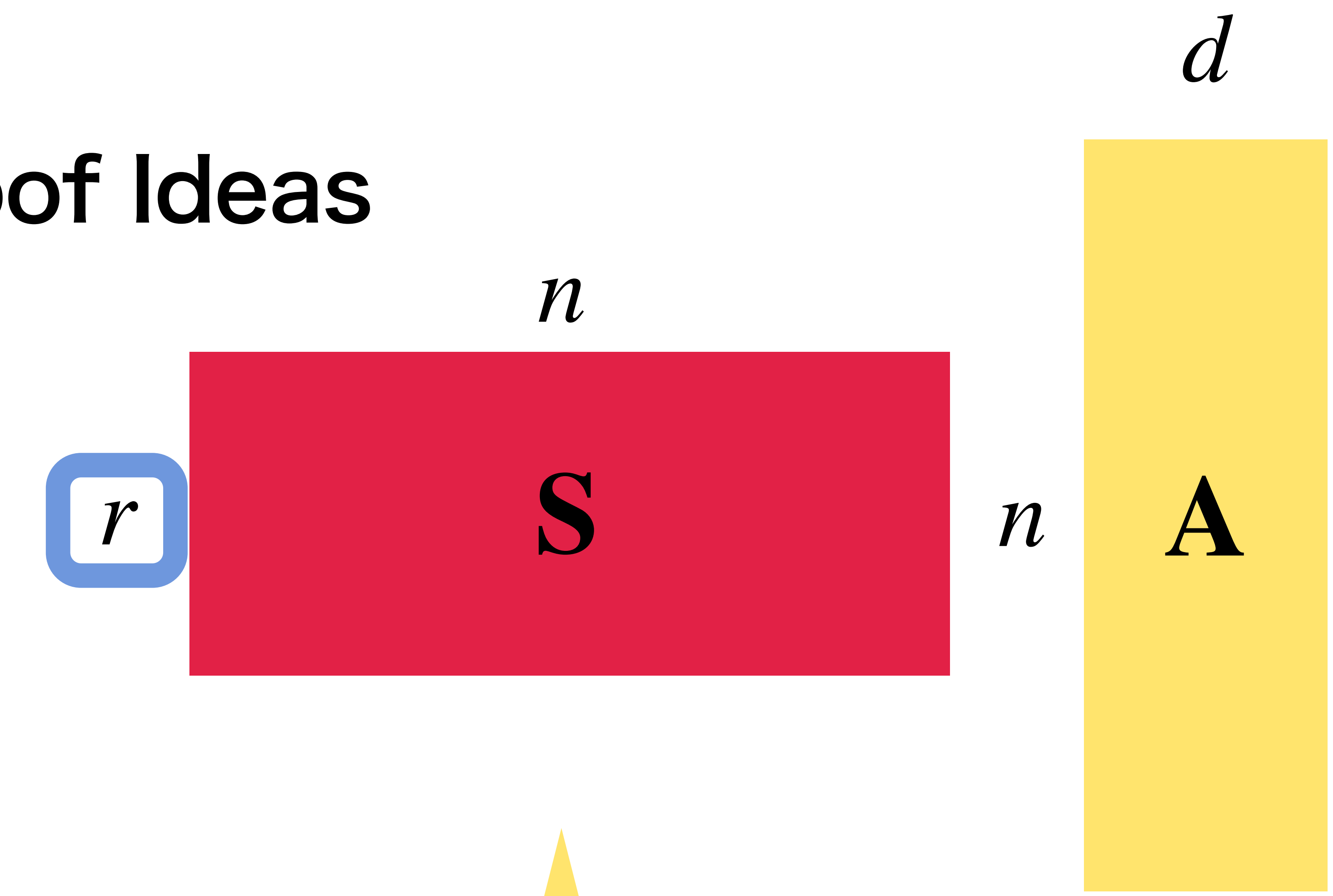
for every $\mathbf{x} \in \mathbb{R}^d$

Part I. Sketching

Oblivious ℓ_p Subspace Embeddings: Proof Ideas

Fact. Oblivious ℓ_p subspace embeddings reduce to constructing **well-conditioned bases** for subspaces

≈ orthonormal bases for ℓ_p norms



Oblivious = does not depend on $\mathbf{A}$

- Let $\mathbf{U}$ be an ~~orthonormal basis~~ well-conditioned basis for $\mathbf{A}$
  - ~~$\|\mathbf{U}\|_F \le d^{1/2}$ (with equality)~~
    $\|\mathbf{U}\|_{p,p} \le \alpha$ entrywise ℓ_p norm
  - ~~$\|\mathbf{U}\mathbf{x}\|_2 \ge \|\mathbf{x}\|_2$ for every $\mathbf{x} \in \mathbb{R}^d$ (with equality)~~
    $\|\mathbf{U}\mathbf{x}\|_p \ge \|\mathbf{x}\|_q$ for every $\mathbf{x} \in \mathbb{R}^d$

Hölder conjugate, $\dfrac{1}{p} + \dfrac{1}{q} = 1$

s.t. $\|\mathbf{A}\mathbf{x}\|_p \le \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p \le \boxed{\kappa}\|\mathbf{A}\mathbf{x}\|_p$

for every $\mathbf{x} \in \mathbb{R}^d$

$\boxed{r = \tilde{O}(d),\ \kappa \lesssim d^{1/p} \text{ [WY 2023]}}$

$\mathbf{S}$ is a random p-stable matrix

Part I. Sketching

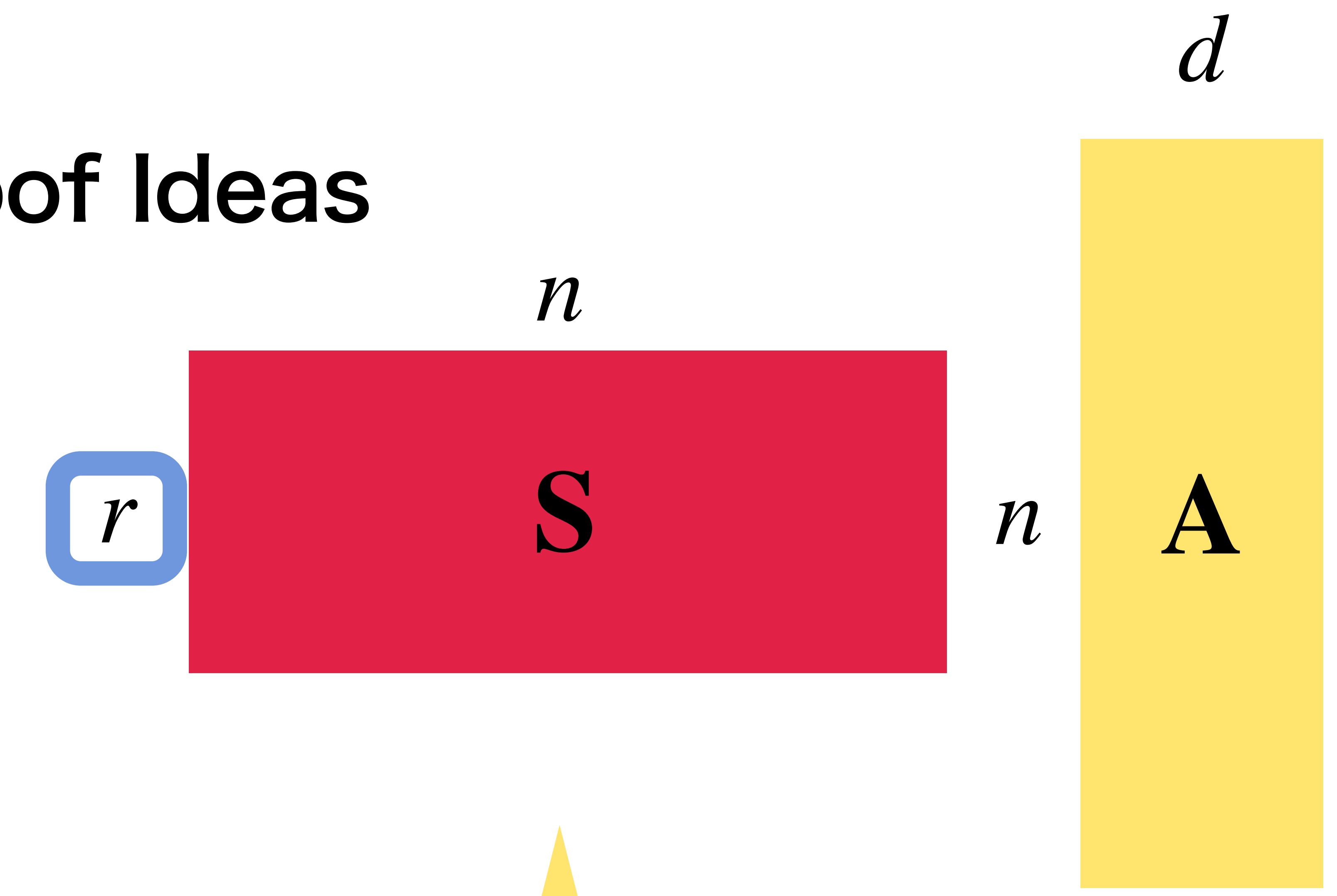
Oblivious ℓ_p Subspace Embeddings: Proof Ideas

Fact. Oblivious ℓ_p subspace embeddings reduce to constructing **well-conditioned bases** for subspaces

- $\|\mathbf{U}\|_{p,p} \leq \alpha$, $\|\mathbf{U}\mathbf{x}\|_p \geq \|\mathbf{x}\|_q$ for every $\mathbf{x} \in \mathbb{R}^d$

Theorem (Auerbach 1930).
For any $\mathbf{A}$, there is $\mathbf{U}$ with $\alpha = d$.

Conjecture. ???
For any $\mathbf{A}$, there is $\mathbf{U}$ with $\alpha = d^{1/p}$.



Oblivious = does not depend on $\mathbf{A}$

s.t. $\|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p \leq \boxed{\kappa}\|\mathbf{A}\mathbf{x}\|_p$

for every $\mathbf{x} \in \mathbb{R}^d$

$r = \tilde{O}(d)$, $\kappa \lesssim d^{1/p}$ [WY 2023]

$\mathbf{S}$ is a random p-stable matrix

Part I. Sketching

Oblivious ℓ_p Subspace Embeddings: Proof Ideas

Fact. Oblivious ℓ_p subspace embeddings reduce to constructing **well-conditioned bases** for subspaces

- $\|\mathbf{U}\|_{p,p} \leq \alpha$, $\|\mathbf{U}\mathbf{x}\|_p \geq \|\mathbf{x}\|_q$ for every $\mathbf{x} \in \mathbb{R}^d$

Theorem (Auerbach 1930). For any $\mathbf{A}$, there is $\mathbf{U}$ with $\alpha = d$.

Conjecture. ???
For any $\mathbf{A}$, there is $\mathbf{U}$ with $\alpha = d^{1/p}$.

Idea. Relax well-conditioned **bases** to well-conditioned **spanning sets**

Theorem (Woodruff—Y 2023). For any $\mathbf{A}$, there is $\mathbf{U} \in \mathbb{R}^{n\times s}$ for $s = O(d)$ such that

- $\|\mathbf{U}\|_{p,p} \leq \alpha$ for $\alpha = s^{1/p}$
- For every $\mathbf{z} \in \mathbb{R}^s$, there is $\mathbf{x} \in \mathbb{R}^s$ s.t. $\mathbf{A}\mathbf{z} = \mathbf{U}\mathbf{x}$ and $\|\mathbf{U}\mathbf{x}\|_p \geq \|\mathbf{x}\|_2 \geq \|\mathbf{x}\|_q$

$\Rightarrow$ $r = \tilde{O}(d)$, $\kappa \lesssim d^{1/p}$ [WY 2023]

Matrix Approximation

Overview of the Talk

- **Sketching**
  - Oblivious ℓ_p subspace embeddings

Sampling

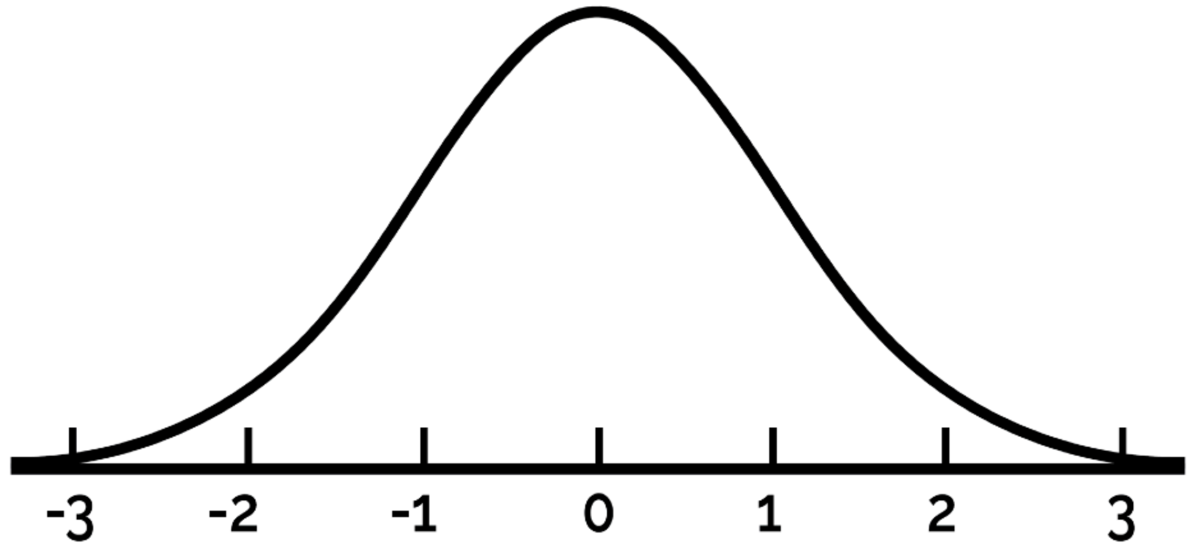
  - Active ℓ_p linear regression
  - Coresets for ℓ_p subspace approximation
  - Streaming computational geometry

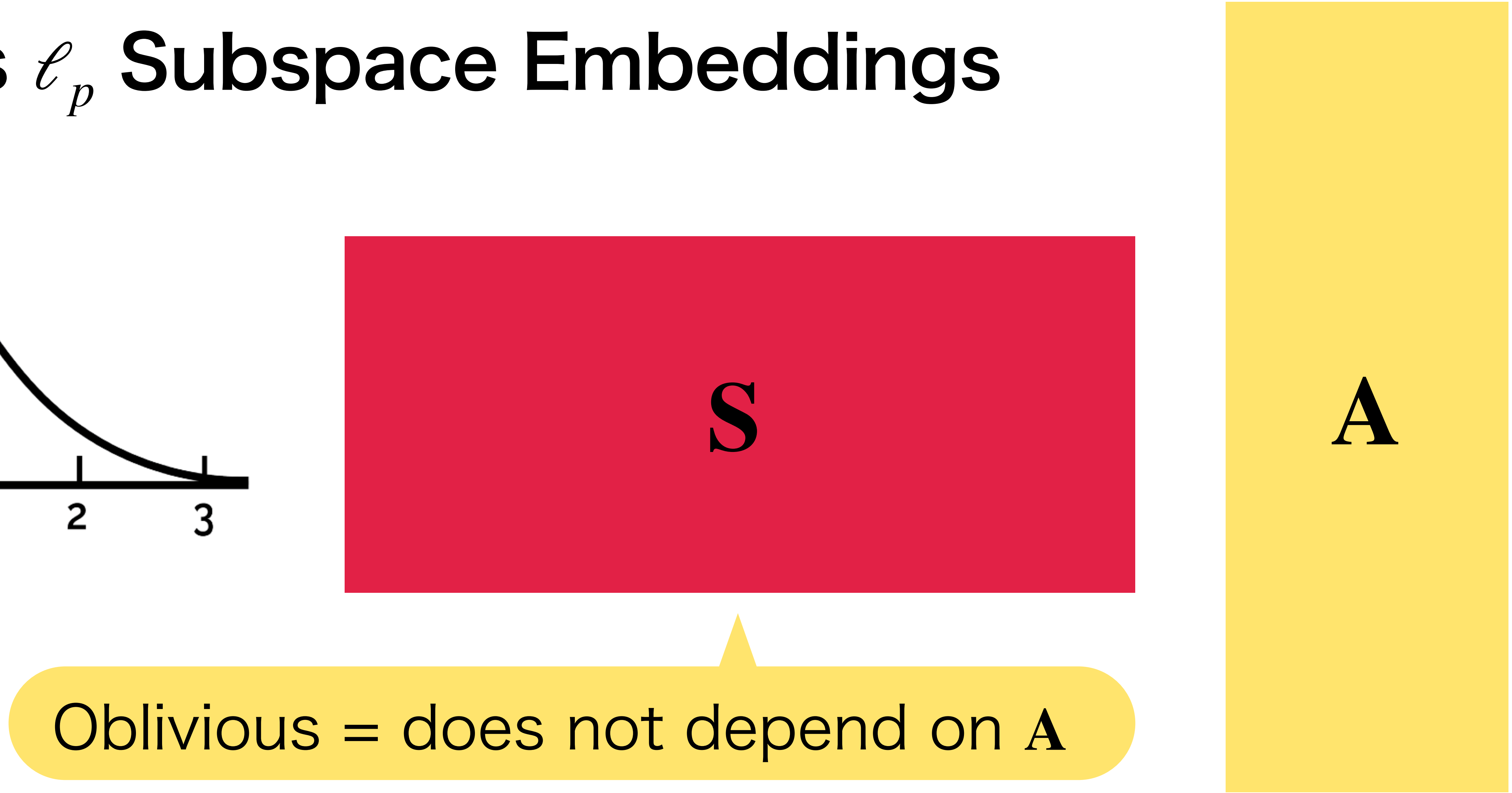
- **Sparse Optimization**
  - Column subset selection

Part II. Sampling

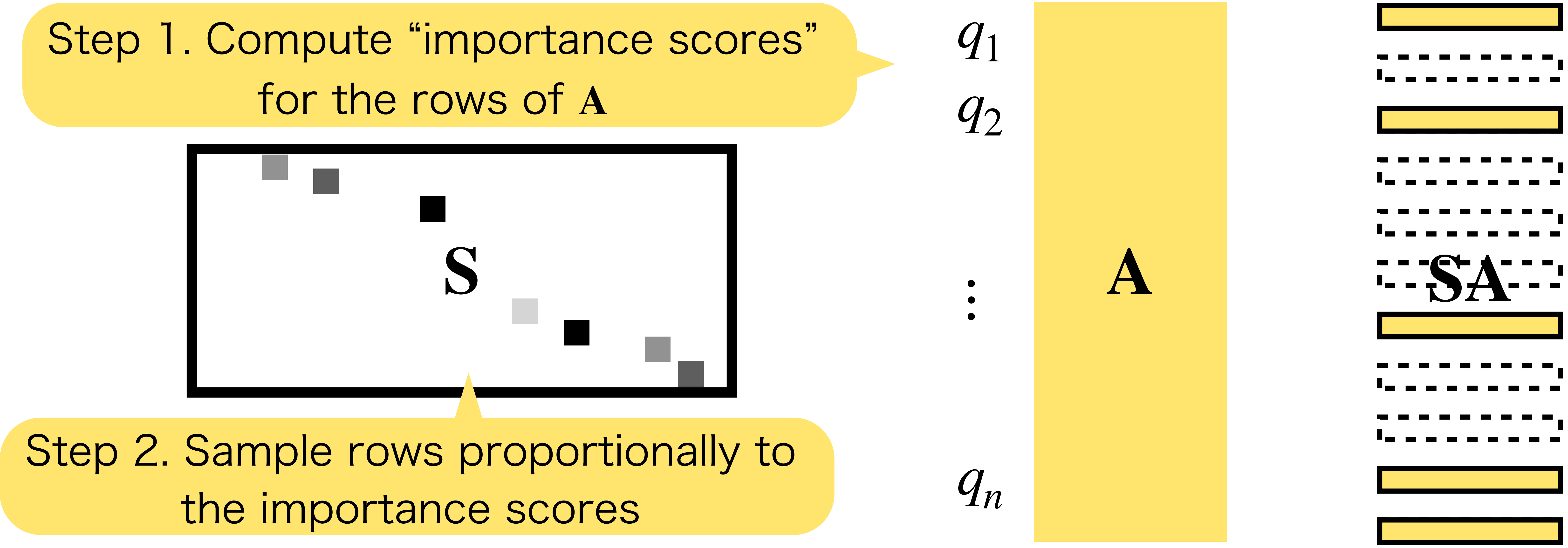
Non-oblivious ℓ_p Subspace Embeddings

Oblivious





Non-oblivious/ Sampling



Part II. Sampling

Non-oblivious ℓ_p Subspace Embeddings

- Why sampling?
  - **Better trade-offs** between compression size r and the distortion κ for ℓ_p subspace embeddings
  - Applications to **query-efficient algorithms**
  - Applications to **coresets**
  - Applications to **streaming algorithms**

Theorem (Lewis weight sampling). For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, there are probabilities $q_1, q_2, \ldots, q_n$ that sample r weighted rows of $\mathbf{A}$ that form an ℓ_p subspace embedding with distortion $\kappa = (1 + \varepsilon)$ with probability 99%, for

$$r = \begin{cases} \tilde{O}(\varepsilon^{-2} d) & p \in (0,2] \quad \textbf{[Cohen—Peng 2015]} \\ \tilde{O}(\varepsilon^{-2} d^{p/2}) & p \in [2,\infty) \quad \textbf{[Woodruff—Y 2023]} \end{cases}$$

Part II. Sampling

Leverage scores: how to choose sampling probabilities?

Definition (Leverage scores). For $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $i \in [n]$, the i-th leverage score is

$$\tau_i(\mathbf{A}) = \sup_{\|\mathbf{A}\mathbf{x}\|_2 = 1} \langle \mathbf{a}_i, \mathbf{x} \rangle^2$$

How "big" the i-th row can be, after normalization

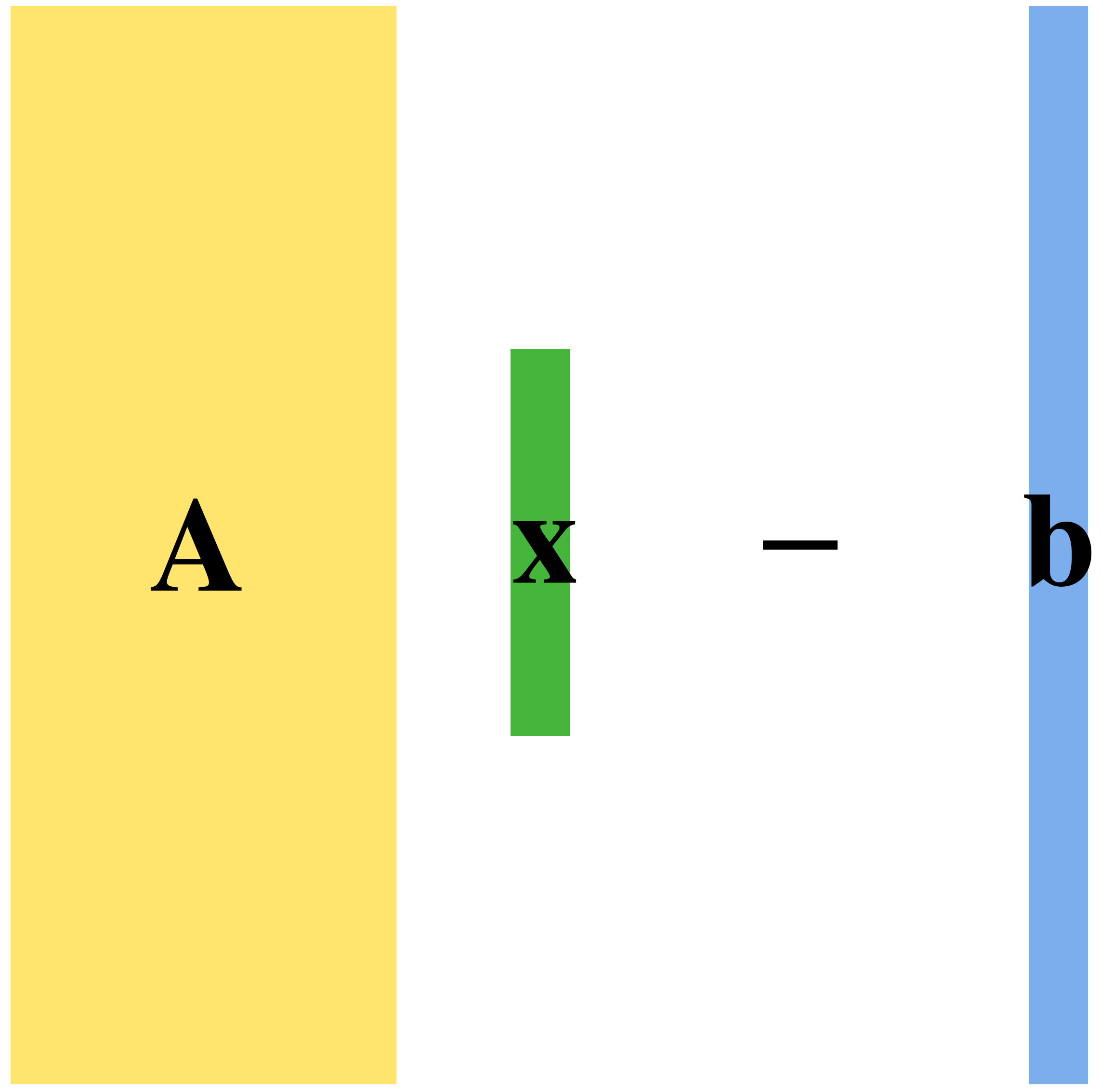
- Many generalizations
  - ℓ_p sensitivity scores [Langberg—Schulman 2010]
  - ℓ_p Lewis weights [Lewis 1978]
  - Ridge leverage scores [El Alaoui—Mahoney 2015, Cohen—Musco—Musco 2017]
  - Online leverage scores [Cohen—Musco—Pachocki 2016]

Part II. Sampling

Active ℓ_p Linear Regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p$$

- **Active learning:** machine learning when **label acquisition** is the most expensive resource
  - Labeling could require···
    - Manual labor
    - Purchasing information
    - Running expensive experiments
  - Goal: **minimize the # of labels read**



Full access to features $\mathbf{A}$

Query access to labels $\mathbf{b}$

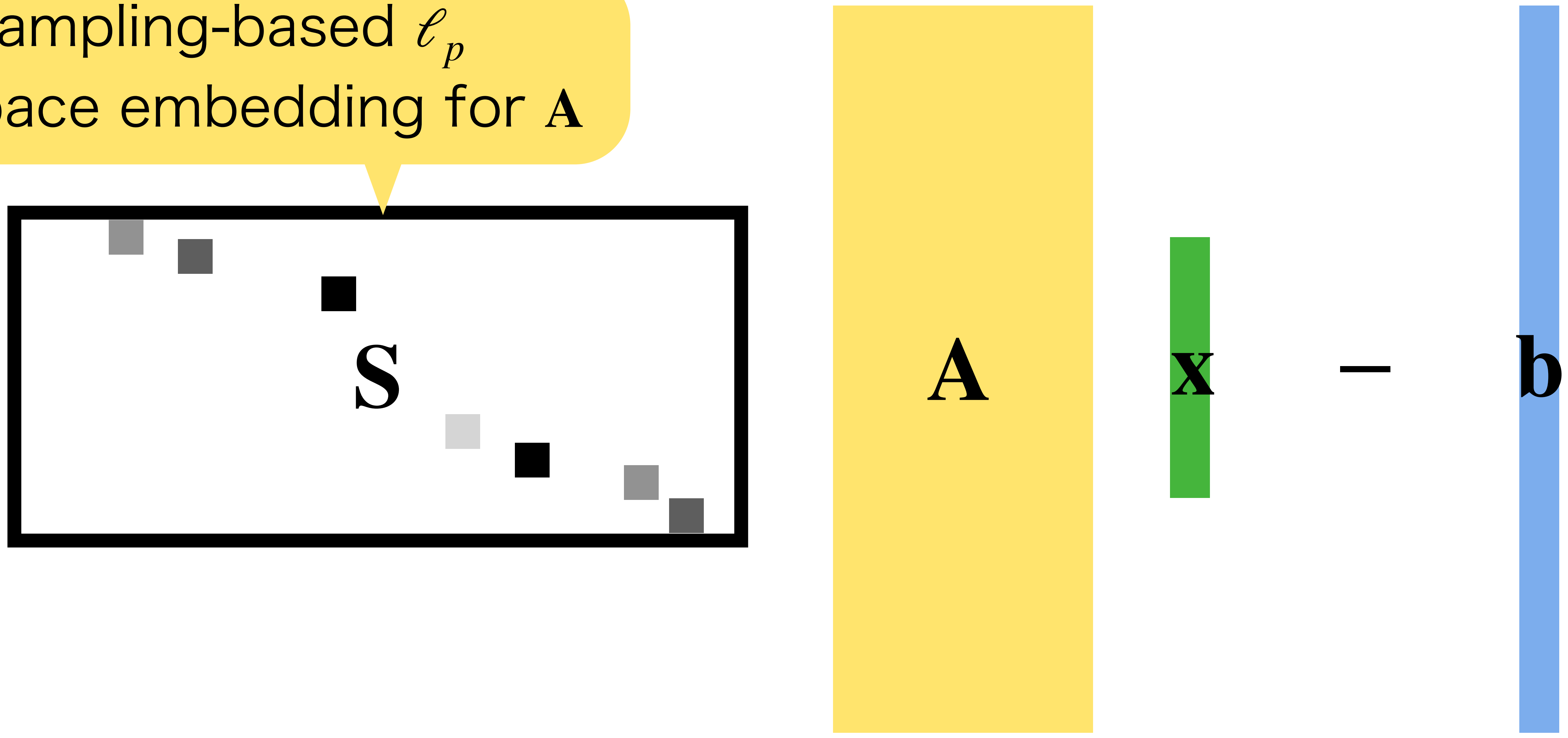
How many entries of $\mathbf{b}$ need to be read?

Part II. Sampling

Active ℓ_p Linear Regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p$$

Sampling-based ℓ_p subspace embedding for $\mathbf{A}$



Full access to features $\mathbf{A}$

Query access to labels $\mathbf{b}$

How many entries of $\mathbf{b}$ need to be read?

Part II. Sampling

Active ℓ_p Linear Regression

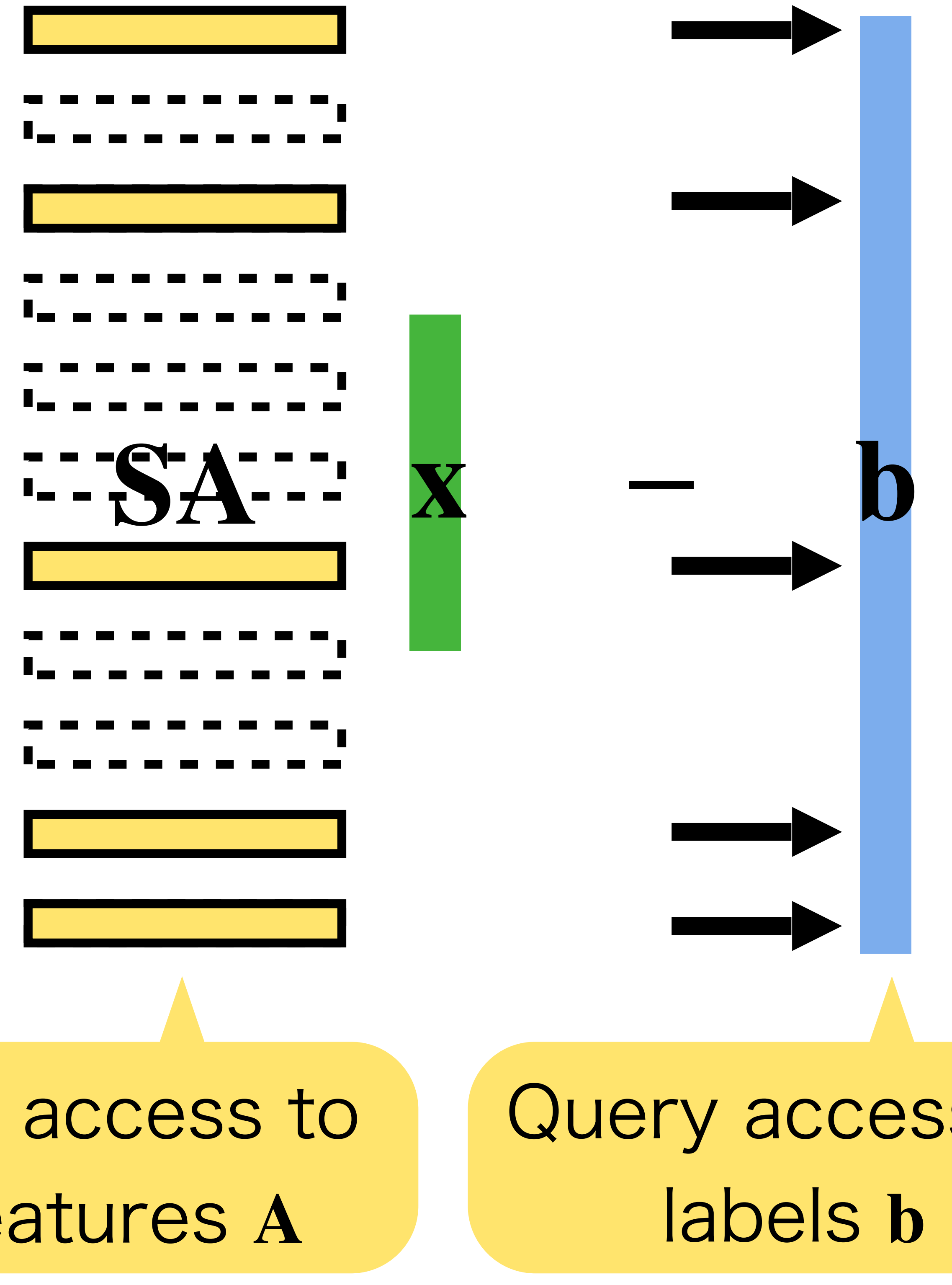
Theorem. There is an active ℓ_p regression algorithm that outputs a $(1+\varepsilon)$-approximate solution with probability 99% and reads at most r entries of $\mathbf{b}$, for

$$r = \begin{cases} \tilde{O}(\varepsilon^{-2} d) & p = 1 \\ \tilde{O}(\varepsilon^{-1} d) & p \in (1,2) \\ O(\varepsilon^{-1} d) & p = 2 \\ \tilde{O}(\varepsilon^{1-p} d^{p/2}) & p \in (2,\infty) \end{cases}$$

- $p = 1$: **[CD 2021, PPP 2021]**
- $p \in (1,2)$: **[MMWY 2023]**
- $p = 2$: **[CP 2019]**
- $p \in (2,\infty)$: **[WY 2023]**

Furthermore, these bounds are nearly optimal.

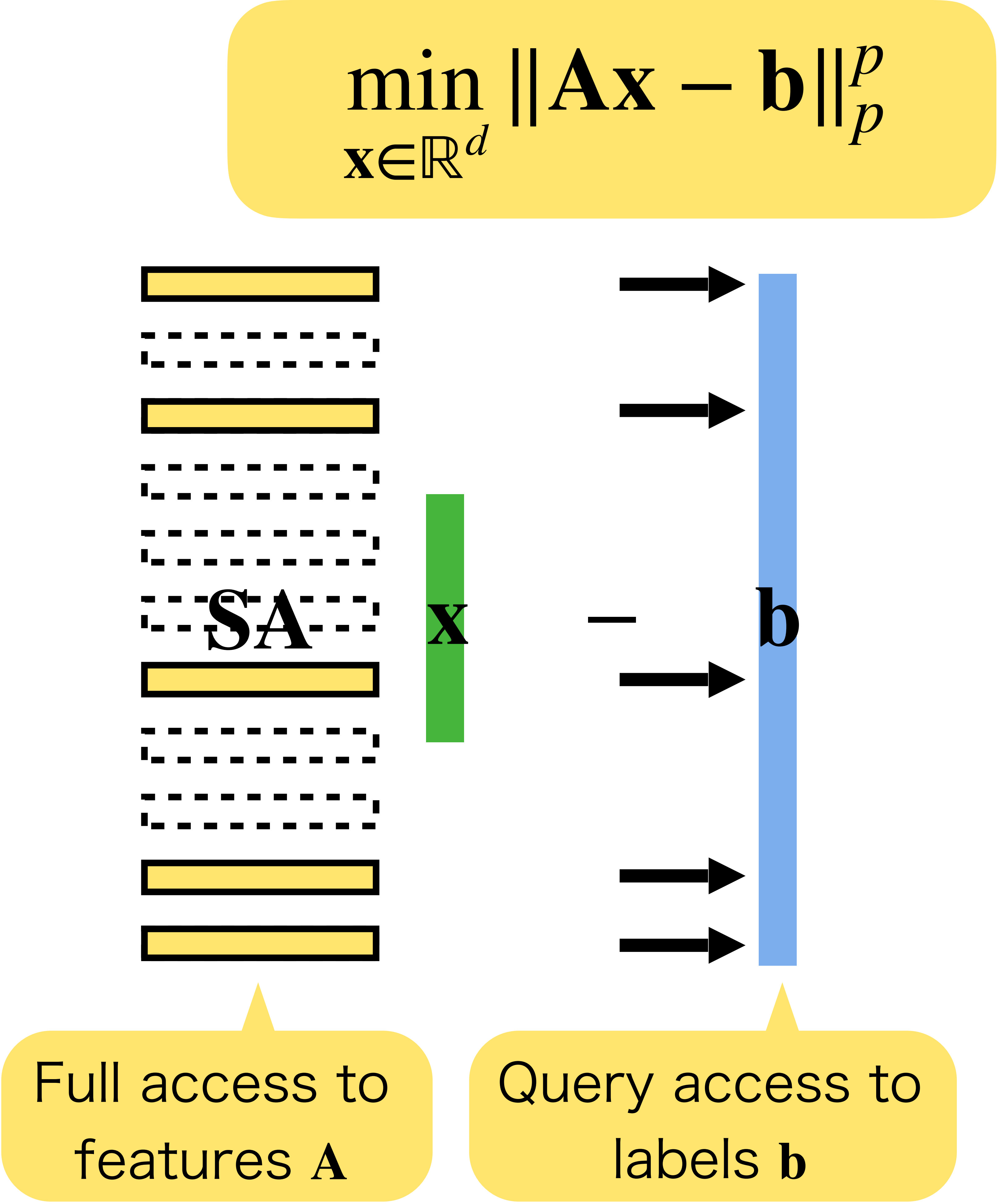
$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p$$



How many entries of $\mathbf{b}$ need to be read?

Part II. Sampling

Active ℓ_p Linear Regression: Proof Ideas



- **Problem:** $\mathbf{S}$ samples row i when $|[\mathbf{A}\mathbf{x}](i)|^p$ is big, but not necessarily when $|[\mathbf{A}\mathbf{x} - \mathbf{b}](i)|^p$ is big
- **Idea:**
  - WLOG restrict to $\|\mathbf{A}\mathbf{x}\|_p^p = O(1)$ and $\|\mathbf{b}\|_p^p = O(1)$
  - If $|\mathbf{b}(i)|^p \lesssim |[\mathbf{A}\mathbf{x}](i)|^p$, then analysis is ok
  - If $|\mathbf{b}(i)|^p \gg |[\mathbf{A}\mathbf{x}](i)|^p$, then $|[\mathbf{A}\mathbf{x} - \mathbf{b}](i)|^p \approx |\mathbf{b}(i)|^p$
    - These coordinates are just adding constants to the objective

How many entries of $\mathbf{b}$ need to be read?

Part II. Sampling

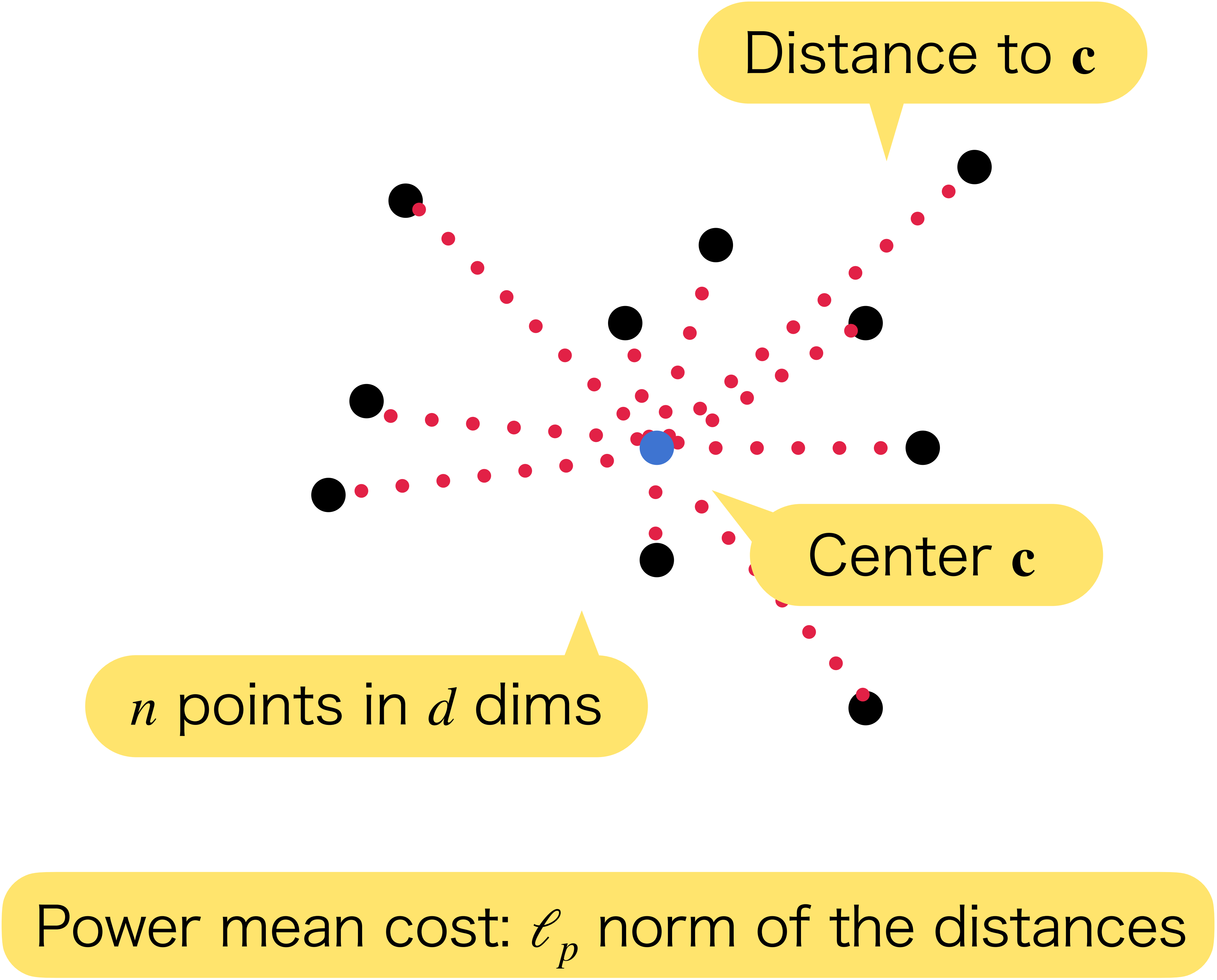
Active ℓ_p Linear Regression: Applications to Power Means

Theorem (Woodruff—Y 2024). In a set of n vectors, a uniform sample of r points is sufficient for a $(1+\varepsilon)$-approximate of the Euclidean p-power mean, for

$$r = \begin{cases} \tilde{O}(\varepsilon^{-2}) & p = 1 \\ \tilde{O}(\varepsilon^{-1}) & p \in (1,2) \\ \tilde{O}(\varepsilon^{1-p}) & p \in (2,\infty) \end{cases}$$

Furthermore, these bounds are nearly optimal.

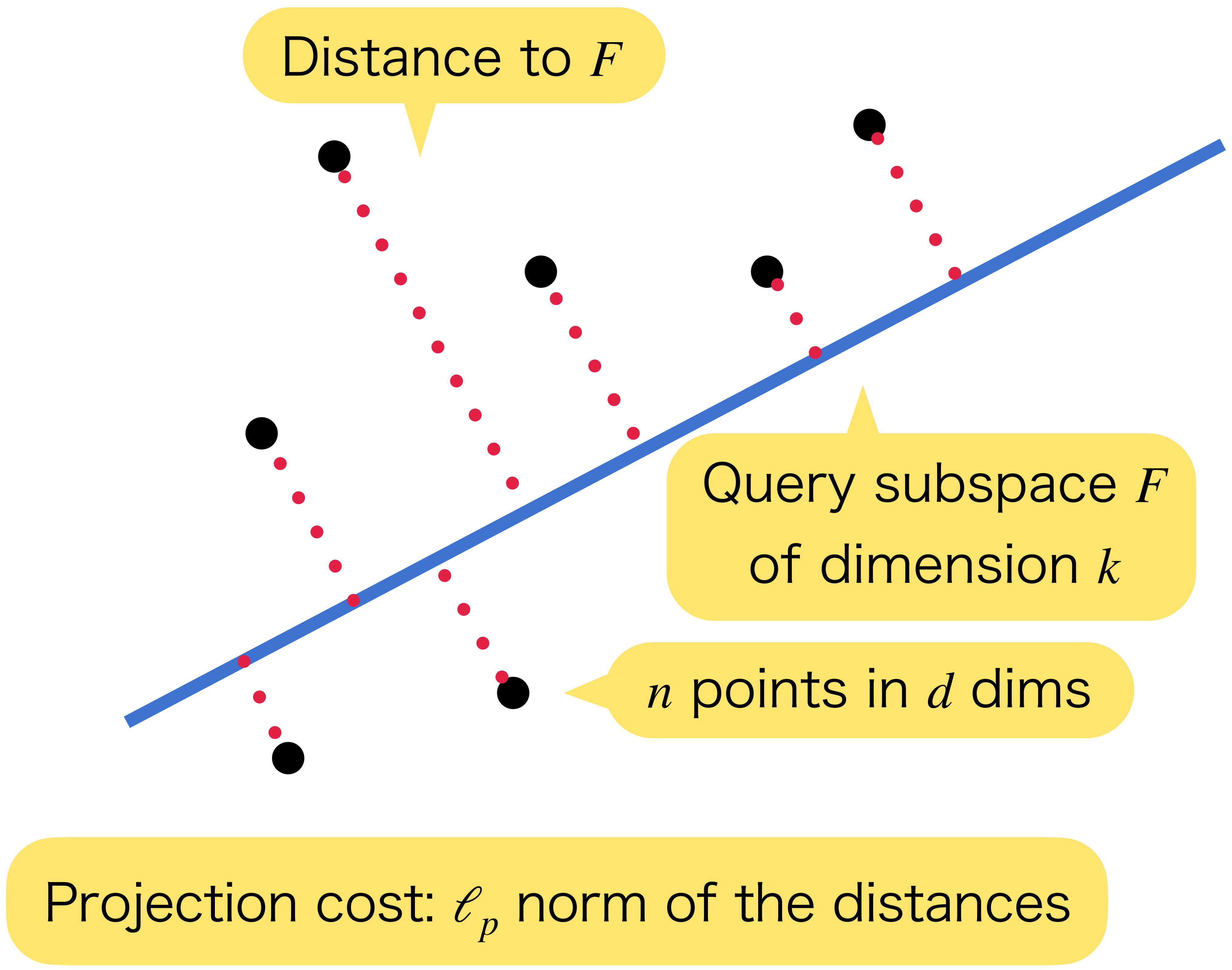
- **Idea:** think of $\mathbf{b}$ as the n vectors, $\mathbf{A}$ as all ones



How many uniform samples do we need?

Part II. Sampling

Coresets for ℓ_p Subspace Approximation



Is there a weighted subset of points that approximates the cost of every subspace F?

Theorem (Woodruff—Y 2024). There is always a weighted subset of r points that approximates the cost of every k-dimensional subspace F, where

$$r = \begin{cases} \tilde{O}(k)\mathrm{poly}(\varepsilon^{-1}) & p \in [1,2) \\ \tilde{O}(k^{p/2})\mathrm{poly}(\varepsilon^{-1}) & p \in (2,\infty) \end{cases}$$

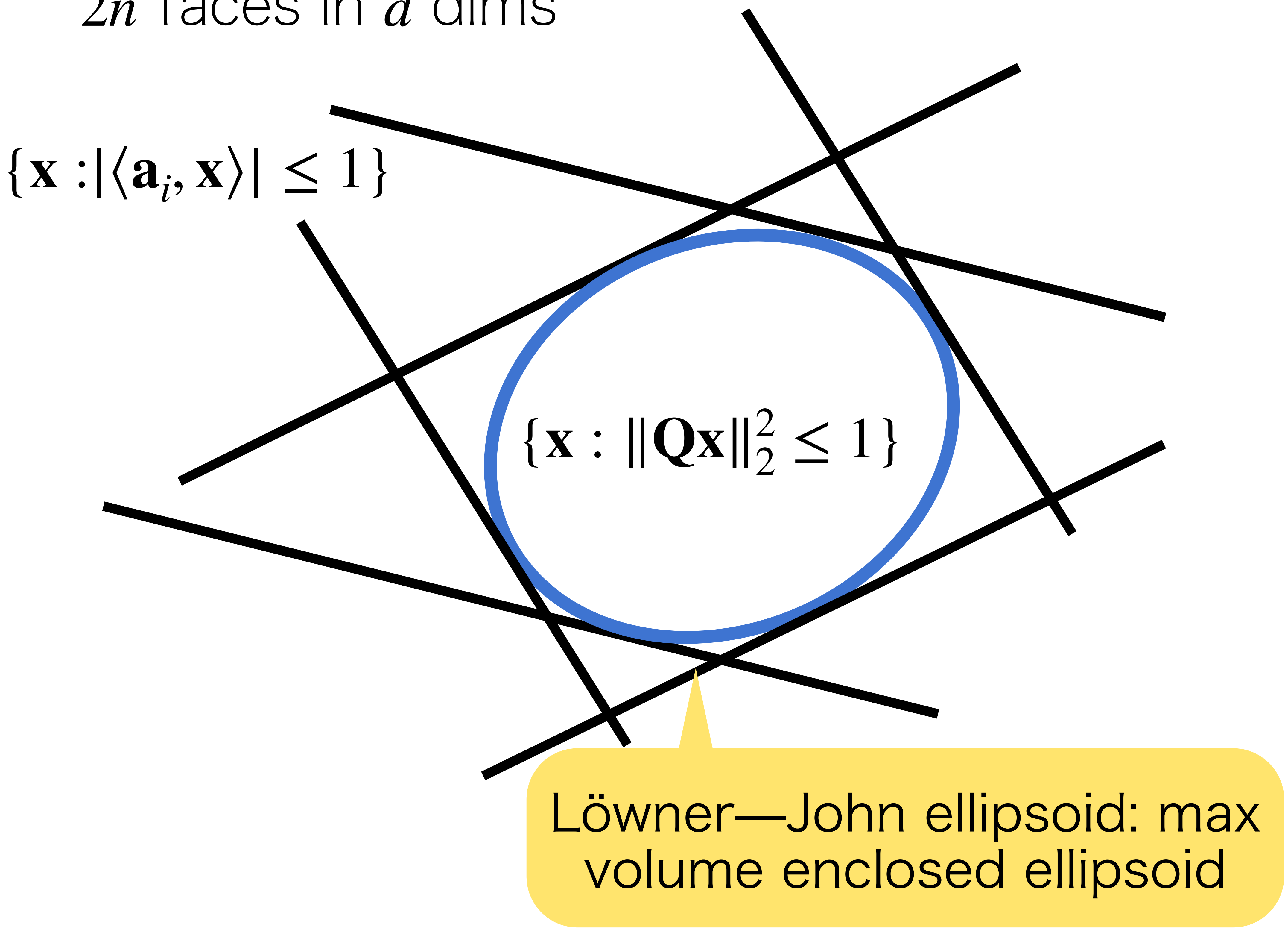
Furthermore, the dependence on k is nearly optimal.

- **Idea:** ridge leverage scores [CMM 2017]

Part II. Sampling

Streaming Computational Geometry: Löwner—John Ellipsoids

Input: symmetric polytope with $2n$ faces in d dims



Can Löwner—John ellipsoids be maintained in $\mathrm{poly}(d, \log n)$ bits of space?

Theorem (Woodruff—Y 2022). There is an algorithm that maintains Löwner—John ellipsoids in $O(d^2 \log^2 n)$ bits of space.

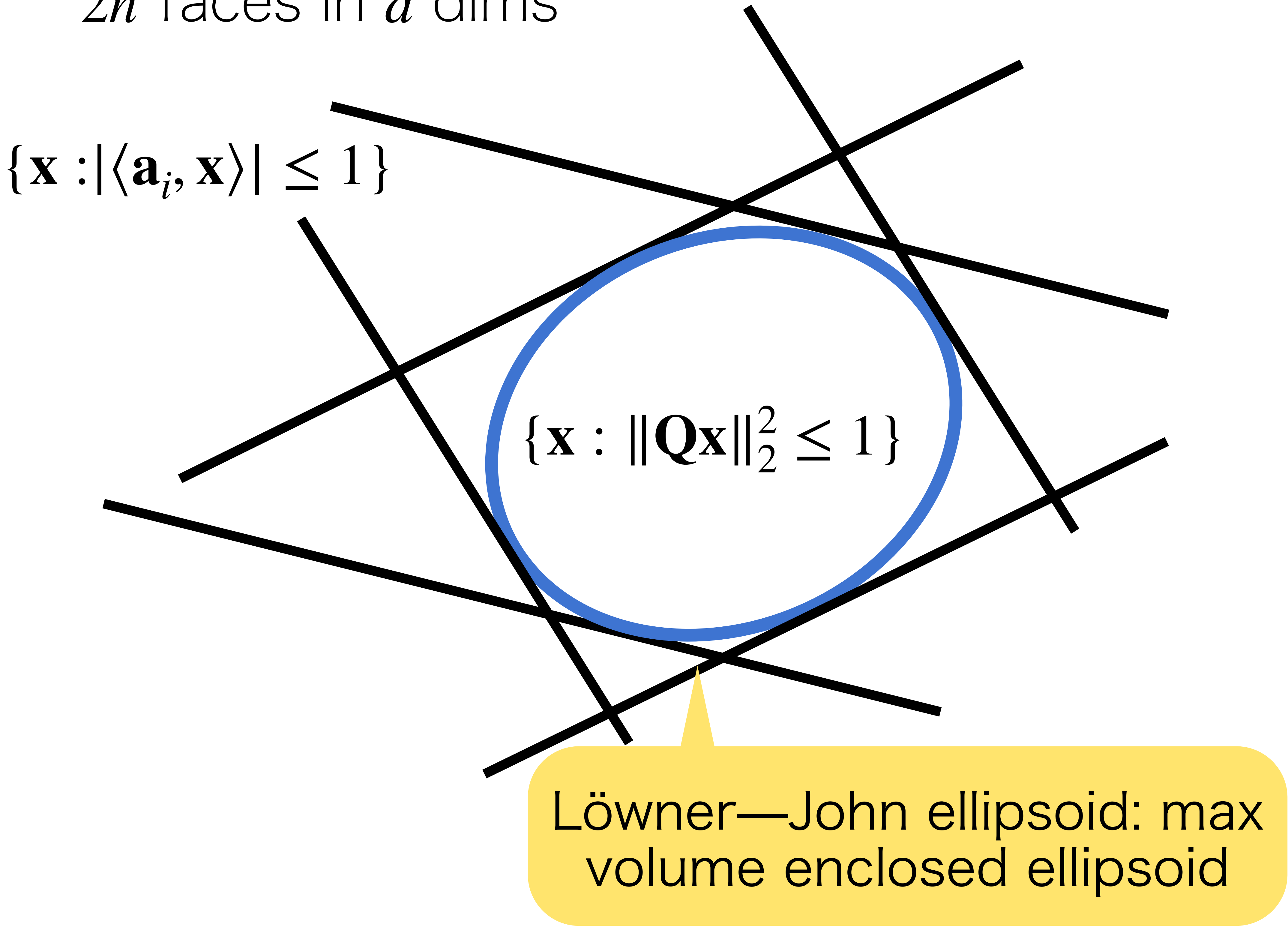
- **Ideas:**
  - Deterministically select a subset of constraints
  - Use **online** leverage scores to⋯
    - Test if an ellipsoid respects a new constraint $\mathbf{a}_i$
    - Bound # of times we keep a new constraint $\mathbf{a}_i$

Ellipsoid respects $\mathbf{a}_i$ iff $\sup_{\|\mathbf{Q}\mathbf{x}\|_2 = 1} \langle \mathbf{a}_i, \mathbf{x} \rangle^2 \leq 1$

Part II. Sampling

Streaming Computational Geometry: Löwner—John Ellipsoids

Input: symmetric polytope with $2n$ faces in d dims



Can Löwner—John ellipsoids be maintained in $\mathrm{poly}(d, \log n)$ bits of space?

Theorem (Woodruff—Y 2022). There is an algorithm that maintains Löwner—John ellipsoids in $O(d^2 \log^2 n)$ bits of space.

- **Corollary:** first polynomial space algorithms for⋯
  - Robust directional width
  - Convex hull approximation
  - Volume maximization
  - Min-width spherical shells
  - Linear programming
  - ⋯

Matrix Approximation

Overview of the Talk

- **Sketching**
  - Oblivious ℓ_p subspace embeddings
- **Sampling**
  - Active ℓ_p linear regression
  - Coresets for ℓ_p subspace approximation
  - Streaming computational geometry
- **Sparse Optimization**
  - Column subset selection

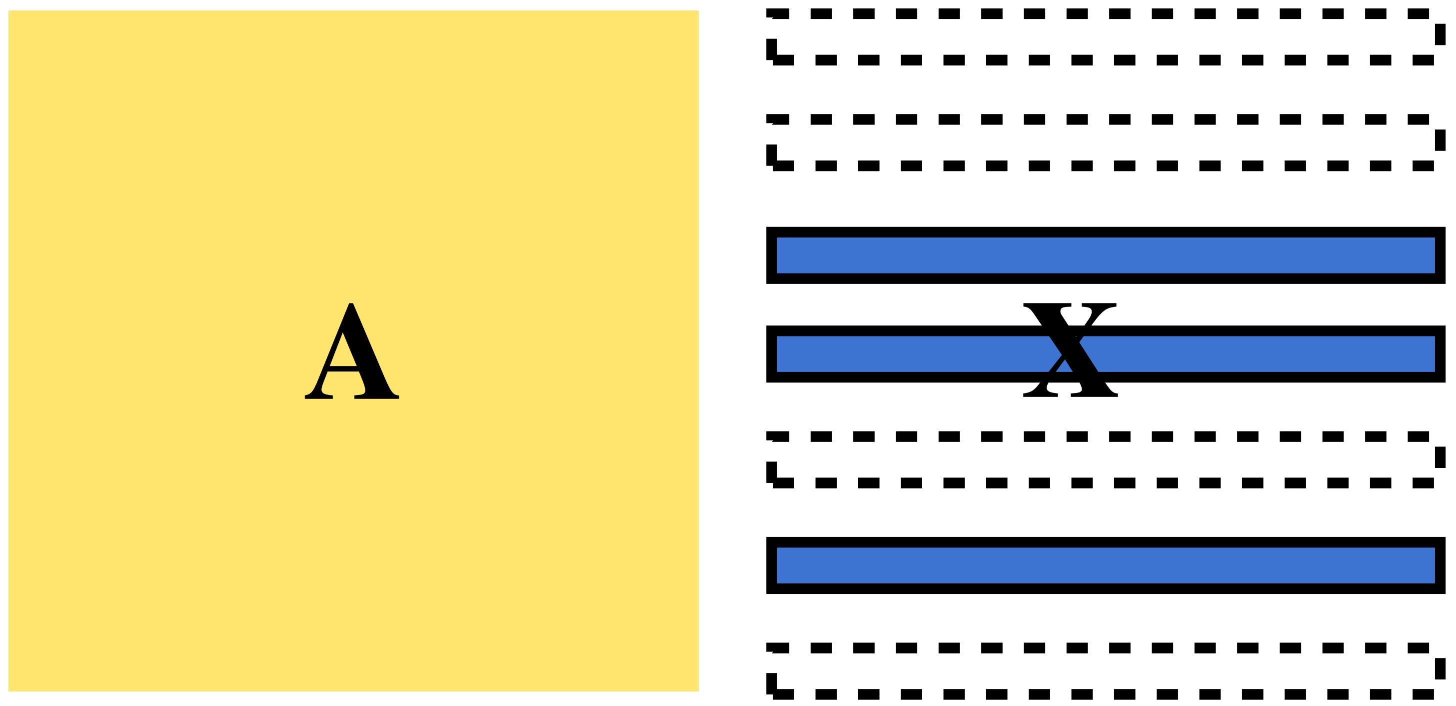
Part III. Sparse Optimization

Column Subset Selection

- **Column subset selection:** select a subset of the columns of a matrix which minimizes the reconstruction error

Minimize $f(\mathbf{A} - \mathbf{A}\mathbf{X})$ over $\mathbf{X} \in \mathbb{R}^{d \times d}$

s.t. $\mathbf{X}$ has at most k nonzero rows



- Entrywise losses f: ℓ_p, non-norm losses
  - New guarantees for greedily fitting columns [Woodruff—Y 2023]
  - Improved analysis via **well-conditioned spanning sets**
- General convex functions f with restricted smoothness and strong convexity
  - New guarantees for group LASSO [Axiotis—Y 2023]

Matrix Approximation

Conclusion

- In this thesis, we studied **matrix approximation** problems from a variety of perspectives, in particular **sketching**, **sampling**, and **sparse optimization** techniques
- We develop and improve foundational tools in matrix approximation, including⋯
  - Subspace embeddings and linear regression
  - Low rank approximation
- Our results also resolve important questions in related areas, including⋯
  - Sublinear algorithms
  - Computational geometry
  - Streaming and online algorithms

Matrix Approximation

Featured Works

- Sketching
  - New Subset Selection Algorithms for Low Rank Approximation: Offline and Online [STOC'23]
  - Exponentially Improved Dimensionality Reduction for ℓ_1: Subspace Embeddings and Independence Testing [COLT'21]
- Sampling
  - Coresets for Multiple ℓ_p Regression [ICML'24]
  - Nearly Linear Sparsification of ℓ_p Subspace Approximation [preprint]
  - Online Lewis Weight Sampling [SODA'23]
  - High-Dimensional Geometric Streaming in Polynomial Space [FOCS'22]
  - Active Linear Regression for ℓ_p Norms and Beyond [FOCS'22]
- Sparse optimization
  - Performance of ℓ_1 Regularization for Sparse Convex Optimization [preprint]

Matrix Approximation

Other Works

- Sketching
  - Sketching Algorithms for Sparse Dictionary Learning: PTAS and Turnstile Streaming [NeurIPS'23]
- Sampling
  - Sharper Bounds for ℓ_p Sensitivity Sampling [ICML'23]
- Sparse optimization
  - SequentialAttention++ for Block Sparsification: Differential Pruning Meets Combinatorial Optimization [preprint]
  - Sequential Attention for Feature Selection [ICLR'23]
  - Improved Algorithms for Low Rank Approximation from Sparsity [SODA'22]
- Other
  - Reweighted Solutions for Weighted Low Rank Approximation [ICML'24]