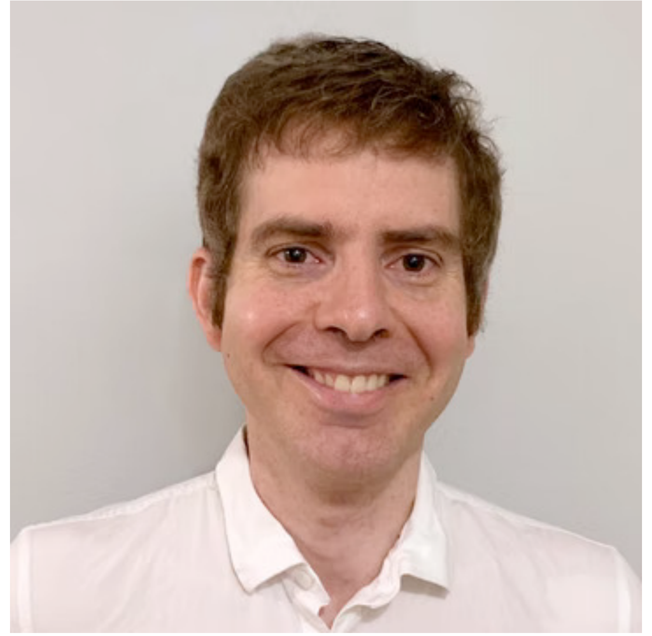
Online Lewis Weight Sampling

Taisuke Yasuda
Carnegie Mellon University

based on work with

David P. Woodruff
Carnegie Mellon University



Subset Selection

ℓ_p Approximation

How small can T be?

- Input: n vectors $\{a_i\}_{i=1}^n \subseteq \mathbb{R}^d$ in d dimensions, $n \gg d$
  - Goal: approximate $\{a_i\}_{i=1}^n$ by a smaller weighted subset
- Spectral approximation: subset $T \subseteq [n]$ and weights s_i for $i \in T$ such that…

$$\text{for every } x \in \mathbb{R}^d, \qquad \|Ax\|_2^2 = \sum_{i=1}^n |\langle a_i, x\rangle|^2 = (1 \pm \varepsilon) \sum_{i \in T} s_i |\langle a_i, x\rangle|^2$$

- ℓ_p subspace embedding: subset $T \subseteq [n]$ and weights s_i for $i \in T$ such that…

$$\text{for every } x \in \mathbb{R}^d, \qquad \|Ax\|_p^p = \sum_{i=1}^n |\langle a_i, x\rangle|^p = (1 \pm \varepsilon) \sum_{i \in T} s_i |\langle a_i, x\rangle|^p$$

- Applications: linear regression, low rank approximation, …

Subset Selection

Algorithms for ℓ_p Approximation: Sampling

- Natural approach for estimating a sum: sampling
  1. For each $i \in [n]$, compute sampling probabilities p_i
  2. Set $s_i = \begin{cases} \frac{1}{p_i} & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}$
  3. Set $T = \{i \in [n] : s_i > 0\}$
- T is a random variable with $\mathbf{E}|T| = \sum_{i=1}^{n} p_i$

Subset Selection

ℓ_p Lewis Weight Sampling

Theorem (ℓ_p Lewis Weight Sampling, Cohen—Peng 2015). There exist sampling probabilities $\{p_i\}_{i=1}^n$ that samples an ε-approximate ℓ_p subspace embedding with probability at least $2/3$, and

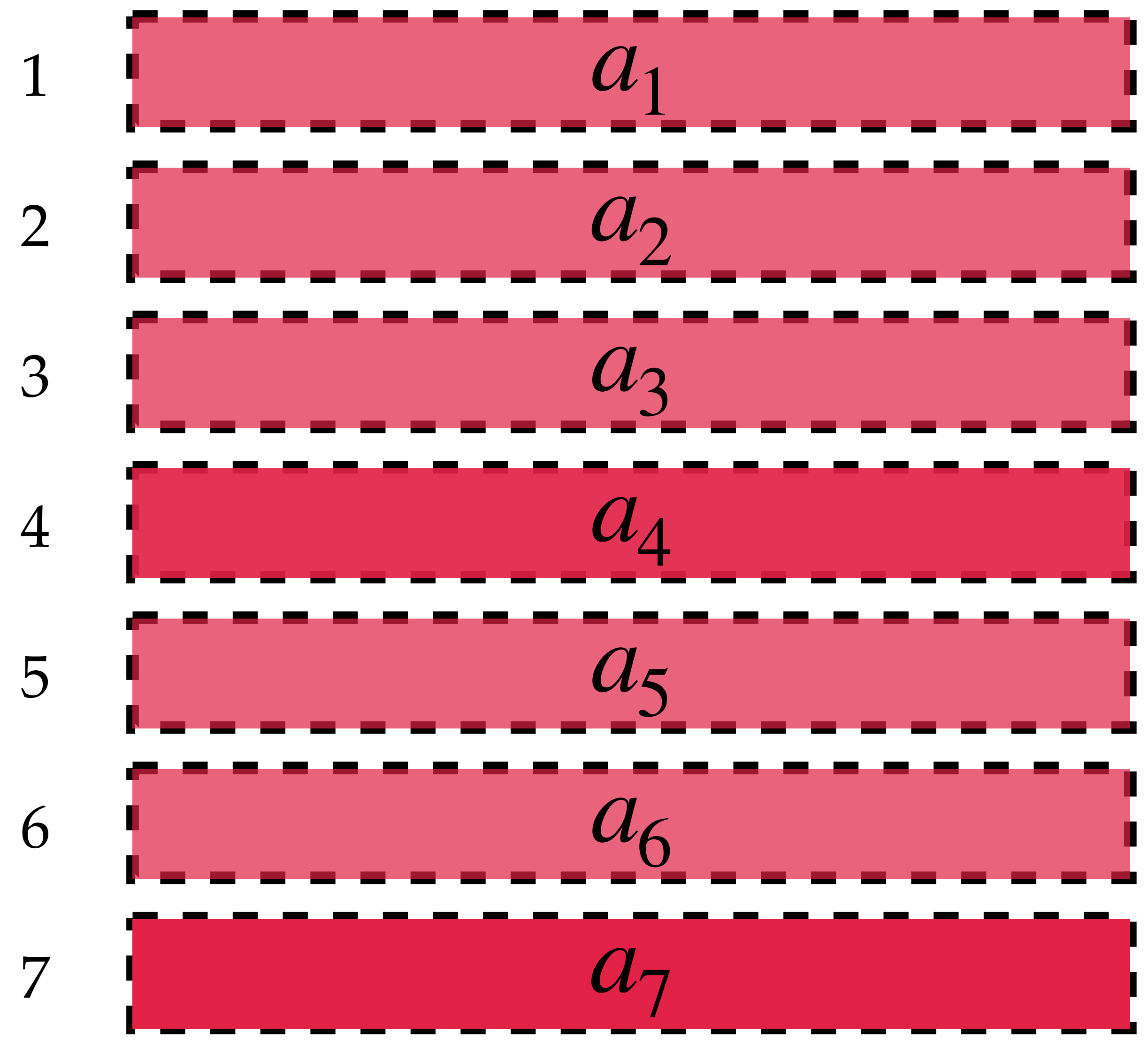
$$\mathbf{E}|T| = \sum_{i=1}^n p_i = \begin{cases} \tilde{O}(\varepsilon^{-2}d) & p \leq 2 \\ \tilde{O}(\varepsilon^{-5}d^{p/2}) & p > 2 \end{cases}$$

- Based on work by Lewis (1978)
- Developed by Bourgain—Lindenstrauss—Milman (1989), Ledoux—Talagrand (1991), and others

Online Subset Selection

ℓ_p Approximation

- Input: n vectors $\{a_i\}_{i=1}^n \subseteq \mathbb{R}^d$ in d dimensions, $n \gg d$, that arrive one by one
- Output: subset $T \subseteq [n]$ and weights s_i for $i \in T$ chosen online
  - At time step $i \in [n]$, either:
    - Irrevocably assign a weight s_i and keep row i, OR
    - Irrevocably discard row i
- Can we output an ε-approximate spectral approximation?
- Can we output an ε-approximate ℓ_p subspace embedding?



Online Subset Selection

Online Spectral Approximation

Theorem (Online Spectral Approximation, Cohen—Musco—Pachocki 2016). There exist sampling probabilities $\{p_i\}_{i=1}^n$ that can be computed online that samples an ε-approximate spectral approximation with probability at least 2/3, and

$$\mathbf{E}|T| = \sum_{i=1}^{n} p_i = \tilde{O}(\varepsilon^{-2} d \log \kappa^{\mathsf{OL}}) \qquad \kappa^{\mathsf{OL}} = \text{“online condition number”}$$

- Main question we study: does a similar result hold for ℓ_p subspace embeddings?
- Several challenges encountered by previous attempts

Online Subset Selection

Main Result 1: Online ℓ_p Lewis Weight Sampling

Theorem (Online ℓ_p Lewis Weight Sampling, Woodruff—Y 2023). There exist sampling probabilities $\{p_i\}_{i=1}^n$ that can be computed online that samples an ε-approximate ℓ_p subspace embedding with probability at least 2/3, and

$$\mathbf{E}|T| = \sum_{i=1}^n p_i = \begin{cases} \tilde O(\varepsilon^{-2} d \log(n\kappa^{\mathsf{OL}})) & p \le 2 \\ \tilde O(\varepsilon^{-2} d^{p/2} \log(n\kappa^{\mathsf{OL}})) & p > 2 \end{cases}$$

κ^{OL} = "online condition number"

- We match offline ℓ_p Lewis weight sampling up to a necessary $\log \kappa^{\mathsf{OL}}$ factor
- Answers an open question of Braverman—Drineas—Musco—Musco—Upadhyay—Woodruff (2020)
- We improve the exponent on ε from ε^{-5} to ε^{-2}, also using ideas from online ℓ_p Lewis weight sampling!

Subset Selection

Main Result 2: Offline ℓ_p Lewis Weight Sampling

Theorem (ℓ_p Lewis Weight Sampling, Cohen—Peng 2015 / Woodruff—Y 2023). There exist sampling probabilities $\{p_i\}_{i=1}^n$ that samples an ε-approximate ℓ_p subspace embedding with probability at least 2/3, and

$$\mathbf{E}|T| = \sum_{i=1}^n p_i = \begin{cases} \tilde{O}(\varepsilon^{-2}d) & p \leq 2 \\ \cancel{\tilde{O}(\varepsilon^{-5}d^{p/2})} & p > 2 \end{cases}$$

$\tilde{O}(\varepsilon^{-2}d^{p/2})$

Technical Discussion

Subset Selection

Leverage Score Sampling for Spectral Approximation

- Sampling algorithm for offline spectral approximation: *leverage score* sampling
  - For each $i \in [n]$, define the ith leverage score: $\tau_i(A) = \sup_{x \in \mathbb{R}^d} \frac{|\langle a_i, x\rangle|^2}{\|Ax\|_2^2} = a_i^\top (A^\top A)^{-1} a_i$
  - Set sampling probability $p_i \leftarrow \varepsilon^{-2}\tau_i(A)$
  - $\mathbf{E}|T| = \varepsilon^{-2}\sum_{i=1}^{n}\tau_i(A) = \varepsilon^{-2}d$
- Intuition: sample "heavier" rows with higher probability

The largest fraction of ℓ_2 mass occupied by the ith coordinate

Online Subset Selection

Online Leverage Score Sampling (Cohen—Musco—Pachocki 2016)

- At time $i \in [n]$, let $A_i \in \mathbb{R}^{i \times d}$ denote the first i rows of A
  - Submatrix of A formed by the rows we have seen so far
- *Online leverage scores* $\tau_i^{\mathsf{OL}}(A) := \tau_i(A_i)$
  - $\tau_i(A_i) \geq \tau_i(A)$, so this also gives an ε-approximate spectral approximation
  - Can also show $\sum_{i=1}^{n} \tau_i^{\mathsf{OL}}(A) \leq O(d \log \kappa^{\mathsf{OL}})$, so only $\tilde{O}(\varepsilon^{-2} d \log \kappa^{\mathsf{OL}})$ rows are sampled

How can we compute $\tau_i^{\mathsf{OL}}(A)$ without storing A?

Online Subset Selection

Online Leverage Score Sampling (Cohen—Musco—Pachocki 2016)

- Computing $\tau_i^{\mathsf{OL}}(A) = \tau_i(A_i)$:
  - Cohen—Musco—Pachocki 2016: approximate A_i using previously sampled rows
    - Analysis using martingale argument, difficult to extend to $p \neq 2$
  - Woodruff—Y 2023: deterministically maintain $A_i^\top A_i$ using d^2 space
    - $\tau_i^{\mathsf{OL}}(A) = a_i^\top (A_i^\top A_i)^- a_i$
    - Decouple the *computation* of sampling probabilities p_i from *sampling*
    - This generalizes well to $p \neq 2$!

Online Subset Selection

Online ℓ_p Lewis Weight Sampling

- Offline ℓ_p Lewis weights (Cohen—Peng 2015): compute weights w_i that satisfy

$$w_i = \left(a_i^\top (A^\top \mathrm{diag}(w)^{1-2/p} A)^- a_i\right)^{p/2}$$

- Online ℓ_p Lewis weights (Woodruff—Y 2023): compute weights w_i that satisfy

$$w_i = \left(a_i^\top (A_{i-1}^\top \mathrm{diag}(w)_{i-1}^{1-2/p} A_{i-1})^- a_i\right)^{p/2}$$

  - Maintain $A_i^\top \mathrm{diag}(w)_i^{1-2/p} A_i$ deterministically
  - We show this works!

Subset Selection

How do online ℓ_p Lewis weights help with offline ℓ Lewis weight sampling?

- Key difficulty of ℓ_p Lewis weights for $p > 2$: *non-monotonicity*
  - When rows are added to A, leverage scores can only decrease: monotonicity
  - When rows are added to A, ℓ_p Lewis weights can *increase*
  - Leads to $\tilde{O}(\varepsilon^{-5}d^{p/2})$ bound rather than $\tilde{O}(\varepsilon^{-2}d^{p/2})$ bound
- Observation: online ℓ_p Lewis weights are monotonic, even for $p > 2$!
  - We show that this observation indeed can be used to obtain $\tilde{O}(\varepsilon^{-2}d^{p/2})$ bound

Conclusion

- We study sampling algorithms for ℓ_p subspace embeddings in the online setting
    - We introduce *online ℓ_p Lewis weight sampling*, which obtains nearly optimal sample complexity bounds
    - This answers an open question of Braverman—Drineas—Musco—Musco—Upadhyay—Woodruff (2020)
- In the offline setting, we use online ℓ_p Lewis weights to obtain the first $\tilde{O}(\varepsilon^{-2}d^{p/2})$ bound for ℓ_p Lewis weight sampling for $p > 2$
    - This improves a previous $\tilde{O}(\varepsilon^{-5}d^{p/2})$ bound by Cohen—Peng (2015)

arXiv