Coresets for Multiple ℓ_p Regression

David P. Woodruff and Taisuke Yasuda

Carnegie Mellon University
Computer Science Department

Coresets for Single ℓ_p Regression

- **ℓ_p linear regression**: given $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$, solve

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p = \sum_{i=1}^{n} |\langle \mathbf{a}_i, \mathbf{x} \rangle - \mathbf{b}_i|^p$$

  - $p = 2$: least squares linear regression
  - $p = 1$: least absolute deviations regression
    - Minimize **average** fitting error, gives **robust** solutions
  - $p = \infty$: Chebyshev regression
    - Minimize **worst-case** fitting error
- **Coresets**: small weighted subset of a dataset
  - Want a subset that is representative of the dataset
  - For ℓ_p linear regression: weighted subset $\mathbf{S}$ s.t.

$$\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p \text{ for every } \mathbf{x} \in \mathbb{R}^d$$

  $\mathbf{S}$: sparse diagonal matrix of weights

  - Goal: choose subset with small size $s = \mathrm{nnz}(\mathbf{S})$
- Prior work on coresets for single response ℓ_p regression
  - We already know how to compute nearly optimal coresets!
  - Technique: ℓ_p Lewis weight sampling [CP15, WY23]

Theorem [CP15, WY23]. Lewis weight sampling gives a coreset $\mathbf{S}$ s.t.

$$\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p \text{ for every } \mathbf{x} \in \mathbb{R}^d$$

with size $s = \mathrm{nnz}(\mathbf{S})$ at most

$$s = \begin{cases} \tilde{O}(\varepsilon^{-2} d^{p/2}) & p > 2 \\ \tilde{O}(\varepsilon^{-2} d) & p \le 2 \end{cases}$$

Furthermore, the bound on s is nearly optimal.

- This settles the picture for single response regression. What about multiple responses?

Question. What if we have m responses rather than just a single response $\mathbf{b}$?

Coresets for Multiple ℓ_p Regression

- **Multiple ℓ_p regression**: given $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$, solve

$$\min_{\mathbf{X} \in \mathbb{R}^{d \times m}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p = \sum_{i=1}^{n} \|\mathbf{a}_i^\top \mathbf{X} - \mathbf{b}_i\|_p^p = \sum_{i=1}^{n} \sum_{j=1}^{m} (\mathbf{A}\mathbf{X} - \mathbf{B})_{i,j}$$

- Prior work:
  - For $p = 2$, Pythagorean theorems allows us to generalize single response result seamlessly: $s = \tilde{O}(\varepsilon^{-2} d)$ [CW13]
  - For $p \neq 2$, straightforward to obtain $s = \mathrm{poly}(\varepsilon^{-1}, d) \cdot m$, that is, linear dependence on m
  - **Challenge**: can we remove the dependence on m?
- Our first main result:

Theorem (Strong coreset). Lewis weight sampling gives a coreset $\mathbf{S}$ s.t.

$$\|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{p,p}^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p \text{ for every } \mathbf{X} \in \mathbb{R}^{d \times m}$$

with size $s = \mathrm{nnz}(\mathbf{S})$ at most

$$s = \begin{cases} \tilde{O}(\varepsilon^{-p} d^{p/2}) & p > 2 \\ \tilde{O}(\varepsilon^{-2} d) & p \le 2 \end{cases}$$

Furthermore, the bound on s is nearly optimal.

- Techniques:
  - Generalization of techniques for active ℓ_p regression [MMWY22] to achieve polylogarithmic dependence on m
  - Averaging argument to completely remove m dependence
- What if we only care about the minimizer?
  - Currently, coreset is required to preserve the entire objective function

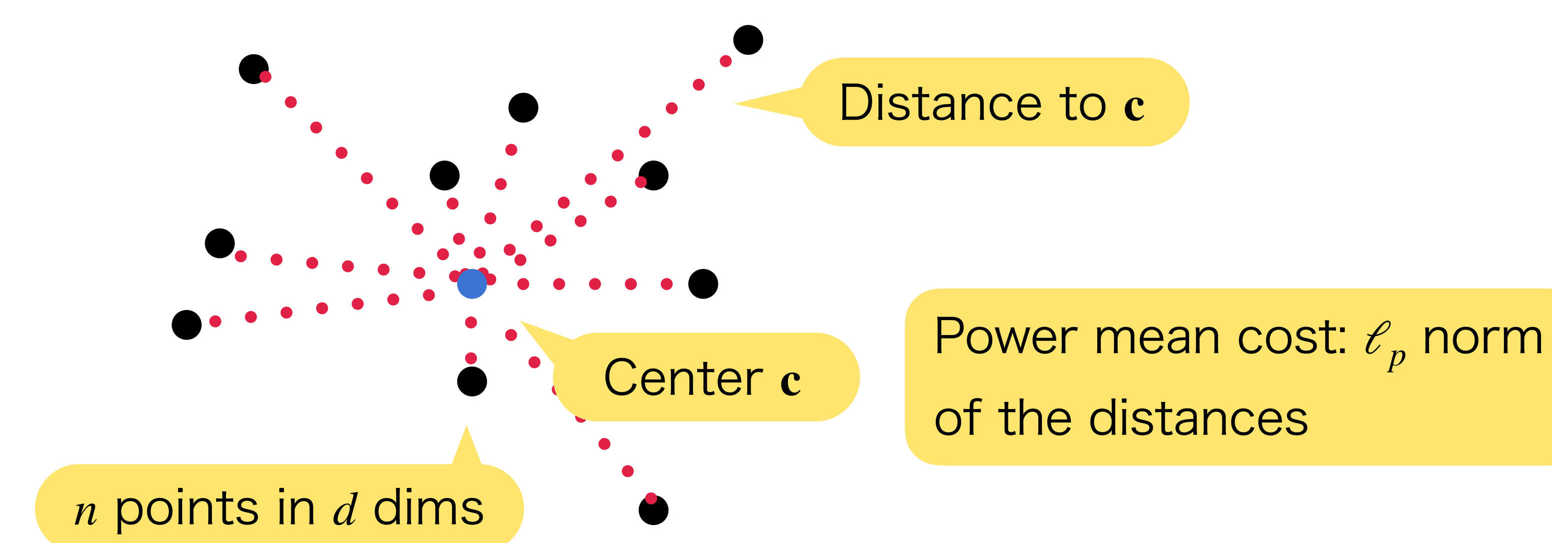
Theorem (Weak coreset). Lewis weight sampling gives a coreset $\mathbf{S}$ s.t. $\tilde{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{p,p}^p$ is a $(1+\varepsilon)$-minimizer with size $s = \mathrm{nnz}(\mathbf{S})$ at most

$$s = \begin{cases} \tilde{O}(\varepsilon^{1-p} d^{p/2}) & p > 2 \\ \tilde{O}(\varepsilon^{-1} d) & 1 < p \le 2 \end{cases}$$

Furthermore, the bound on s is nearly optimal.

Applications

- We give two applications of our techniques
- **Sublinear Euclidean power means**
  - Euclidean power means: generalization of mean



Theorem. The power mean of s uniform samples is a $(1+\varepsilon)$-approximate minimizer of the power mean cost for

$$s = \begin{cases} \tilde{O}(\varepsilon^{1-p}) & p > 2 \\ \tilde{O}(\varepsilon^{-1}) & 1 < p \le 2 \\ \tilde{O}(\varepsilon^{-2}) & p = 1 \end{cases}$$

Furthermore, the bound on s is nearly optimal.

  - Resolves an open question of [CSS21]
  - Think of $\mathbf{B}$ as the input vectors (set $m \leftarrow d$) and $\mathbf{A}$ as the all ones vector (set $d \leftarrow 1$)
  - Dvoretzky's theorem to embed ℓ_2 into ℓ_p
- **Spanning coresets for ℓ_p subspace approximation**
  - ℓ_p subspace approximation: generalization of PCA

Theorem. For $p \in (1,2)$, there is always a subset of s points that spans a $(1+\varepsilon)$ approximation solution, for $s = \tilde{O}(\varepsilon^{-1} k)$. Furthermore, the bound on s is nearly optimal.

  - Improves a $s = \tilde{O}(\varepsilon^{-1} k^2)$ bound of [SV07]

References

[CP15] Cohen, Peng, STOC 2015
[CSS21] Cohen-Addad, Saulpic, Schwiegelshohn, NeurIPS 2021
[CW13] Clarkson, Woodruff, STOC 2013
[MMWY22] Musco, Musco, Woodruff, Yasuda, FOCS 2022
[SV07] Shyamalkumar, Varadarajan, SODA 2007
[WY23] Woodruff, Yasuda, SODA 2023