Stable Coresets for 1-Median in ℓ_1 via Uniform Sampling

Amir Carmel Robert Krauthgamer

ICLR 2026

The 1-Median Problem

Input: Set of points $P \subset \mathbb{R}^d$

Goal: Find center $c \in \mathbb{R}^d$ minimizing the average cost

$$\mathrm{cost}(c, P) := \frac{1}{|P|} \sum_{p \in P} \|c - p\|_1$$

Optimal value:

$$\mathrm{opt}(P) := \min_{c \in \mathbb{R}^d} \mathrm{cost}(c, P)$$

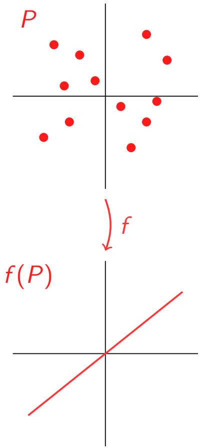
Key point: The median $c^* \in \mathrm{argmin}_{c \in \mathbb{R}^d} \mathrm{cost}(c, P)$ can be **any point in** $\mathbb{R}^d$, not necessarily from P.

Motivation: Coresets as Data Summarization

A coreset is a small subset that approximates the behavior (we wish to preserve) of the full dataset.

Important role in modern algorithmic settings:

- Significant reduction in computational resources.
- Efficient storage and communication.
- Useful in streaming and dynamic algorithms.

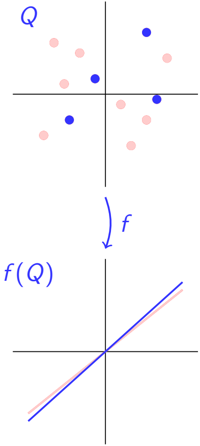


Motivation: Coresets as Data Summarization

A coreset is a small subset that approximates the behavior (we wish to preserve) of the full dataset.

Important role in modern algorithmic settings:

- Significant reduction in computational resources.
- Efficient storage and communication.
- Useful in streaming and dynamic algorithms.



Coreset Definitions

Definition (Strong ϵ-Coreset)

$Q \subseteq P$ is a *strong ϵ-coreset* if

$$\forall c \in \mathbb{R}^d, \quad \mathrm{cost}(c, P) \in (1 \pm \epsilon)\,\mathrm{cost}(c, Q)$$

Definition (Weak ϵ-Coreset)

$Q \subseteq P$ is a *weak ϵ-coreset* if

$$\forall c \in \mathbb{R}^d, \quad \mathrm{cost}(c, Q) \leq (1 + \epsilon)\,\mathrm{opt}(Q) \Rightarrow \mathrm{cost}(c, P) \leq (1 + 2\epsilon)\,\mathrm{opt}(P)$$

Stable Coresets – Our Contribution

Definition (Stable ϵ-Coreset)

$Q \subseteq P$ is a *stable ϵ-coreset* if

$$\forall c_1, c_2 \in \mathbb{R}^d, \quad \mathrm{cost}(c_1, Q) \leq (1+\epsilon)\,\mathrm{cost}(c_2, Q) \Rightarrow \mathrm{cost}(c_1, P) \leq (1+2\epsilon)\,\mathrm{cost}(c_2, P)$$

Intermediate notion: Strong $\Rightarrow$ Stable $\Rightarrow$ Weak

Motivation: Why Study the ℓ_1 Metric?

- ℓ_1 is widely exploited in data analysis in high-dimensional spaces.
- ℓ_1 generalizes ℓ_2 (every ℓ_2 metric can be embedded into ℓ_1).
- There are many intersting metrics that are submetric of ℓ_1.

ℓ_1 is a rich metric!

Known Results: Uniform Sampling for 1-Median

Reference	Coreset Type	Metric	Sample Size
Danos '21	Weak 0	ℓ_1	$\tilde{O}(\epsilon^{-2})$
Huang-Jiang-Lou '23	Weak ϵ	ℓ_2	$\tilde{O}(\epsilon^{-3})$
Our Result	**Stable** ϵ	ℓ_1	$O(\epsilon^{-2} \log d)$

Application: k-Median Algorithm

Corollary (k-Median in ℓ_1)

Let $P \subset \mathbb{R}^d$ and $\epsilon \in (0, 1/5)$.
There exists a $(1 + O(\epsilon))$-approximation algorithm for k-median that runs in time:

$$f(k, \epsilon, s_\epsilon) \cdot n$$

where $s_\epsilon = O(\epsilon^{-2} \log d)$ is the coreset size.

Extends to: Kendall-tau, Hamming, Jaccard, tree metrics, etc.

Open Problems

1. **Removing dimension dependency:**
   - Current bound: $O(\epsilon^{-2} \log d)$
   - **Conjecture:** $O(\epsilon^{-2})$ suffices (dimension-independent)
   - Empirical evidence strongly supports this
2. **Other metric spaces:** Which metrics admit efficient stable coresets?
3. **Other clustering objectives:** Extend stable coresets to other clustering applications (e.g., k-means, robust clustering, facility location).