Robust Sparsification via Sensitivity

Chansophea Wathanak In[1] Yi Li[1] David Woodruff[2] Xuan Wu[1]

1 NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

2 Carnegie Mellon University

ICML 2025

1. Robust Sparsification Problem

Def: For an optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) \ s.t$$

$$F(x) = F_1(x) + \cdots + F_n(x)$$

Robust Version:

$$\min_{x\in\mathbb{R}^d} F^{(m)}(x) \ s.t$$

$F^{(m)} =$ sum of all but the m largest functions

Motivation: Capture Outliers

Def: $(\varepsilon,\ m)$-coreset for $F^{(m)}(x)$

$$\tilde{F}(x) = \sum_{j=1}^{s} w_j F_{i_j}(x) \quad \textit{Sparse Representation}$$

$$\tilde{F} \textit{ is } (\epsilon, m) - \textit{coreset of } F \textit{ if}$$

$$\forall x \in \mathbb{R}^d, \forall t \le m, \tilde{F}^{(t)}(x) \in (1 \pm \epsilon) \cdot F^{(t)}(x)$$

Sensitivity	Total Sensitivity
$\sigma(F_i) = \sup_{x\in\mathbb{R}^d} \frac{F_i(x)}{F(x)}$	$T = \sum_{i=1}^{n} \sigma(F_i)$

Applications: 1. Robust Clustering

2. Robust Subspace Embedding

3. Trimmed Sqaures Regression

4. Robust PCA

2. Our Contributions

- **Algorithms:**
- For every F that has total sensitivity T and a vanilla coreset of size Q, we construct an $(\varepsilon,\ m)$-coreset of size $\widetilde{O}(mT/\varepsilon + Q)$
- As applications, we give $\exp(O(m/\epsilon + m\log d)) + O(nd)$ time algorithms for robust regression and robust PCA with m outliers
- **Lower Bound:**
- For subspace embedding, we prove mT/ε is necessary $(T = d)$
- **Experiments:**
- Our algorithm outperforms uniform sampling in real-world data

3. Prior Results

Most existing robust coresets are designed for k-clustering:

1. $\exp(k + m)$ [FS12]
2. $m + \text{poly}(mk/\varepsilon)$ [HJLW23,HLLW25]
3. $\min\{mk/\varepsilon, m/\varepsilon^2\} + Q$ [JL25]

Beyond clustering, existing results only satisfy some weaker coreset property:

1. Bi-criteria [FL11, HJLW18]
2. Local coreset [WGD21]

4. Main Novel Techniques

Input: A set A of functions, parameters ε and m
Output: A subset $D \subseteq A$

1: $B \leftarrow \emptyset$
2: for each $f \in A$, with probability $\frac{1}{m}$, add f to B
3: for each $f \in B$, compute the sensitivity $\sigma_B(f)$
4: $D \leftarrow \{f \in B : \sigma_B(f) \ge \frac{\varepsilon}{4}\}$
5: **Return** D

Technical Novelty:

1. We characterize a subset of "contributing functions", whose removal makes a vanilla coreset for remaining data is also robust.
2. A sampling based algorithm to find all contributing functions.
3. An iteraive reduction based anlysis to control the number of contributing functions and the number of iterations.

5. Experiments

- Optimization problem: ℓ_2 subspace embedding

- Result: Outperform uniform sampling

- Dataset:

1. Energy Dataset: 19735x28
2. Emission Dataset: 36733x11

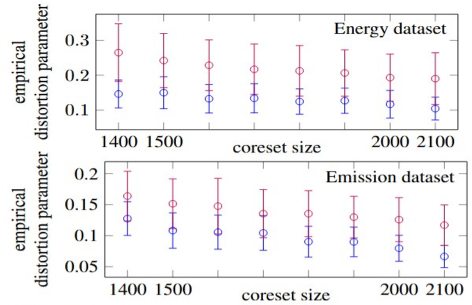

Figure 1: Results of subspace embedding coresets with $m = 10$ and $\varepsilon = 0.25$. Blue plots correspond to our coreset algorithm and the red plots to uniform sampling.