New Coresets for Projective Clustering and Applications

Murad Tukan
University of Haifa

Xuan Wu
Johns Hopkins University

Samson Zhou
Carnegie Mellon University

Vladimir Braverman
Johns Hopkins University

Dan Feldman
University of Haifa

Abstract

(j, k)-projective clustering is the natural generalization of the family of k-clustering and j-subspace clustering problems. Given a set of points P in $\mathbb{R}^d$, the goal is to find k flats of dimension j, i.e., affine subspaces, that best fit P under a given distance measure. In this paper, we propose the first algorithm that returns an L_∞ coreset of size polynomial in d. Moreover, we give the first strong coreset construction for general M-estimator regression. Specifically, we show that our construction provides efficient coreset constructions for Cauchy, Welsch, Huber, Geman-McClure, Tukey, $L_1 - L_2$, and Fair regression, as well as general concave and power-bounded loss functions. Finally, we provide experimental results based on real-world datasets, showing the efficacy of our approach.

1 INTRODUCTION

Coresets are often used in machine learning, data sciences, and statistics as a pre-processing dimensionality reduction technique to represent a large dataset with a significantly smaller amount of memory, thereby improving the efficiency of downstream algorithms in both running time and working space. Intuitively, a coreset C of a set P of n points in $\mathbb{R}^d$ is a smaller number of weighted representatives of P that can be used to approximate the cost of any query from a set of a given queries. Hence rather than optimizing some predetermined objective on P, it suffices to optimize the objective on C, which has significantly smaller dimension than P. In this paper, we present coresets for projective clustering.

Projective clustering is an important family of clustering problems for applications in unsupervised learning (Procopiuc, 2010), data mining (Aggarwal et al., 1999; Aggarwal and Yu, 2000), computational biology (Procopiuc, 2010), database management (Chakrabarti and Mehrotra, 2000), and computer vision (Procopiuc et al., 2002). Given a set P of n points in $\mathbb{R}^d$, a parameter z for the exponent of the distance, and a parameter k for the number of flats of dimension j, the (j, k)-projective clustering problem is to find a set $\mathcal{F}$ of k j-flats that minimizes the sum of the distances of P from $\mathcal{F}$, i.e., $\min_{\mathcal{F}} \sum_{\mathbf{p}\in P} \text{dist}(\mathbf{p}, \mathcal{F})^z$, where $\text{dist}(\mathbf{p}, \mathcal{F})^z$ denotes the z-th power of the Euclidean distance from $\mathbf{p}$ to the closest point in any flat in $\mathcal{F}$. We abuse notation by defining the projective clustering problem to be $\min_{\mathcal{F}} \max_{\mathbf{p}\in P} \text{dist}(\mathbf{p}, \mathcal{F})$ for $z = \infty$. Projective clustering includes many well-studied problems such as the k-median clustering problem for $z = 1$, $j = 0$, $k \in \mathbb{Z}^+$, the k-means clustering problem for $z = 2$, $j = 0$, $k \in \mathbb{Z}^+$, the k-line clustering problem for $z \geq 0$, $j = 1$, $k \in \mathbb{Z}^+$, the subspace approximation problem for $z \geq 0$, $j \in \mathbb{Z}^+$, $k = 1$, the minimum enclosing ball problem for $z = \infty$, $j = 0$, $k = 1$, the k-center clustering problem for $z = \infty$, $j = 0$, $k \in \mathbb{Z}^+$, the minimum enclosing cylinder problem for $z = \infty$, $j = 1$, $k = 1$, and the k-cylinder problem for $z = \infty$, $j = 1$, $k \in \mathbb{Z}^+$.

1.1 Related Work

Finding the optimal set C for projective clustering is known to be NP-hard Aloise et al. (2009) and even finding a set with objective value that is within a factor of 1.0013 of the optimal value is NP-hard Lee et al. (2017). Procopiuc et al. (2002) implemented a heuristics-based Monte Carlo algorithm for projective clustering while Har-Peled and Varadarajan (2002)

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

introduced a dimensionality reduction technique to decrease the size of each input point, which distorts the cost of the optimal projective clustering. Similarly, Kerber and Raghvendra (2015) used random projections to embed the input points into a lower dimensional space. However, none of these approaches reduces the overall number of input points, whose often causes the main bottleneck for implementing approximation algorithms for projective clustering in big data applications.

Badoiu et al. (2002) first introduced coresets for the k-center and k-median clustering problems in Euclidean space. Their coresets constructions gave $(1+\varepsilon)$-approximations and sampled a number of points with exponential dependency in both $\frac{1}{\varepsilon}$ and k. Their work also inspired a number of coresets for specific projective clustering problems; coresets have subsequently been extensively studied in k-median or k-means clustering (Badoiu et al., 2002; Har-Peled and Mazumdar, 2004; Frahling and Sohler, 2005, 2008; Chen, 2009; Feldman and Schulman, 2012; Braverman et al., 2019; Huang and Vishnoi, 2020), subspace approximation (Deshpande et al., 2006; Deshpande and Varadarajan, 2007; Feldman and Langberg, 2011; Feldman et al., 2010; Clarkson and Woodruff, 2015; Sohler and Woodruff, 2018; Feldman et al., 2020; Tukan et al., 2021b), and a number of other geometric problems and applications (Agarwal et al., 2006; Feldman et al., 2006; Clarkson, 2008; Dasgupta et al., 2008; Ackermann and Blömer, 2009; Phillips and Tai, 2018; Huang et al., 2018; Assadi et al., 2019; Munteanu et al., 2018; Braverman et al., 2020; Mussay et al., 2020; Tukan et al., 2020, 2021a; Jubran et al., 2020; Maalouf et al., 2021). However, these coreset constructions were catered toward specific problems rather than the general (j,k)-projective clustering problem.

Feldman and Langberg (2011) introduced a framework for constructing coresets by sampling each input point with probability proportional to its *sensitivity*, which informally quantifies the importance of the point with respect to the predetermined objective function. Feldman and Langberg (2011) also performed dimensionality reduction for (j,k)-projective clustering by taking the union of two sets $\mathcal{S}$ and $\text{proj}(P,B)$, where P is the input data set of size n. Although the set $\mathcal{S}$ can have size $\text{poly}(j,k,d)$, the set $\text{proj}(P,B)$ still has size n, so their resulting output can actually have *larger* size than the original input. The main point is that $\text{proj}(P,B)$ lies in a low-dimensional space, so their approach should be viewed as a dimensionality reduction technique to decrease the ambient dimension d whereas our coreset construction decreases the input size n. (Clarkson and Woodruff, 2015) suggested approximation algorithms based on matrix sketches for $(1,j)$-projective clustering problems with respect to family of M-estimator functions, and (Clarkson et al., 2019) provided tighter result with respect to the $(1,j)$-projective clustering problems with respect to the Tukey loss function. (Varadarajan and Xiao, 2012c) proved upper bounds for the total sensitivity of the input points for a number of shape fitting problems, including the k-median, k-means, and k-line clustering problems, as well as an L_1 coreset for the integer (j,k)-projective clustering problem. On the other hand, (Har-Peled, 2004) showed that L_∞ coresets for the projective clustering problem does not exist even for $j=k=2$ when the input set consists of points from $\mathbb{R}^d$. When the input is restricted to integer coordinates, (Edwards and Varadarajan, 2005) constructed an L_∞ coreset that gives a $(1+\varepsilon)$-approximation for (j,k)-projective clustering. However, their construction uses a subset of points with size exponential in both k and d, which often prevents practical implementations. Hence, a natural open question is whether there exist L_∞ coreset constructions for integer (j,k)-projective clustering with size polynomial in d.

1.2 Our Contributions

We give the first L_∞ coreset construction for the integer (j,k)-projective clustering problem with size polynomial in d, resolving the natural open question from Edwards and Varadarajan (2005). Specifically, we give an L_∞ ξ-coreset C, so that for any choice $\mathcal{F}$ of k flats with dimension j, the maximum connection cost of C to $\mathcal{F}$ is at most ξ times the maximum connection cost of P. Previously, even in the case of $k=1$ and constant j, the best known L_∞ coreset construction had size $\exp(d)$ (Edwards and Varadarajan, 2005). We first introduce an L_∞ coreset construction for the $(j,1)$-projective clustering problem using Carathéodory's theorem; see Figure 1. We then use our L_∞ coreset for $(j,1)$-projective clustering as a base case to recursively build a coreset D_k for (j,k)-projective clustering from coresets for $(j,k-1)$-projective clustering on the partitions of the input points that have geometrically increasing distances from the affine subspace spanned by the points chosen in the previous steps. We use properties from Edwards and Varadarajan (2005); Feldman et al. (2020) to bound the number of partitions determined by the distances from the input points to each of the affine subspaces, which bounds our coreset size for an input with aspect ratio Δ, i.e., the ratio of the largest and smallest coordinate magnitudes.

Theorem 1.1 (Small L_∞ coreset for (j,k)-projective clustering)**.** *There exists an L_∞ constant-factor approximation coreset for the (j,k)-projective clustering problem with size* $(8j^3\log(d\Delta))^{\mathcal{O}(jk)}$.

Our main technical contribution is the novel L_∞ coreset construction for the $(j,1)$-projective clustering problem that relies on Carathéodory's theorem, which we

crucially use to form the base case in our recursive argument. We then build upon our novel coreset construction by adding a polynomial number of points to the coreset over each step in the inductive argument. By comparison, even the base case for the previous best coreset (Edwards and Varadarajan, 2005) uses exponential space by essentially constructing an epsilon net with $\left(\frac{1}{\epsilon}\right)^{\mathcal{O}(d)}$ points.

We then give the first L_∞ coresets for a number of M-estimator regression problems. Although the framework of Theorem 1.1 immediately gives coreset constructions for Cauchy, Welsch, Huber, Geman-McClure, Tukey, $L_1 - L_2$, and Fair regression, we instead apply sharper versions of the proof of Theorem 1.1 to the respective parameters induced by each of the loss functions to obtain even more efficient coreset constructions. Our constructions give strong coresets so that with high probability, the data structure simultaneously succeeds for all queries. We then apply the framework of Theorem 1.1 to give L_∞ coresets for any non-decreasing concave loss function Ψ with $\Psi(0) = 0$. We generalize this approach to give L_∞ coresets for any non-decreasing concave loss function Ψ with $\Psi(y)/\Psi(x) \leq (y/x)^z$ for a fixed constant $z > 0$, for all $0 \leq x \leq y$. Note that this property essentially states that the loss function $\Psi(x)$ is bounded by some power function x^z. We summarize these results in Table 1.

We also use Theorem 1.1 along with the well-known sensitivity sampling technique to obtain an L_2 coreset for integer (j, k)-projective clustering with approximation $(1 + \varepsilon)$.

Theorem 1.2 (Small L_2 coreset for (j, k)-projective clustering)**.** *There exists an L_2 coreset with approximation guarantee $(1+\varepsilon)$ for the (j, k)-projective clustering problem with size $\mathcal{O}\left((8j^3 \log(d\Delta))^{\mathcal{O}(jk)} \log n\right)$.*

Experiments. Finally, we complement our theoretical results with empirical evaluations on synthetic and real world datasets for regression and clustering problems. We first consider projective clustering on a bike sharing dataset and a 3D spatial network from the UCI machine learning repository (Dua and Graff, 2017). We then generate a synthetic dataset in the two-dimensional Euclidean plane. Since previous coreset constructions with theoretical guarantees are impractical for implementations, we compare our algorithms to a baseline produced by uniform sampling. Our experiments demonstrate that our algorithms have superior performance both across various ranges of j and k for the (j, k)-projective clustering problem as well as across various regression problems, e.g., Cauchy, Huber loss functions.

1.3 Preliminaries

For a positive integer n, we write $[n] := \{1, \ldots, n\}$. We use bold font variables to denote vectors and matrices. For a vector $\mathbf{x} \in \mathbb{R}^d$, we have the Euclidean norm $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$. We use log to denote the base two logarithm. We use the notation $\circ$ to denote vertical concatenation, so that if $\mathbf{u}$ and $\mathbf{v}$ are row vectors with dimension d, then $\mathbf{u} \circ \mathbf{v}$ is the matrix with dimension $2 \times d$ whose first row is $\mathbf{u}$ and second row is $\mathbf{v}$. Recall that for $\mathbf{c} \in \mathbb{R}^d$ and a symmetric positive definite matrix $\mathbf{G} \in \mathbb{R}^{d \times d}$, we define the ellipsoid $E(\mathbf{G}, \mathbf{c})$ to be the set $E(\mathbf{G}, \mathbf{c}) := \left\{\mathbf{x} \in \mathbb{R}^d \,|\, (\mathbf{x} - \mathbf{c})^\top \mathbf{G}(\mathbf{x} - \mathbf{c}) \leq 1\right\}$.

Theorem 1.3 (John-Löwner ellipsoid)**.** *(John, 2014) For a set $L \subseteq \mathbb{R}^d$ of points with nonempty interior, there exists an ellipsoid $E(\mathbf{G}, \mathbf{c})$, where $\mathbf{G} \in \mathbb{R}^{d \times d}$ is a positive definite matrix and $\mathbf{c} \in \mathbb{R}^d$, of minimal volume such that $\frac{1}{d}(E(\mathbf{G}, \mathbf{c}) - \mathbf{c}) + \mathbf{c} \subseteq \text{conv}(L) \subseteq E(\mathbf{G}, \mathbf{c})$.*

The following defines an approximated solution to problem of finding the Löwner ellipsoid.

Definition 1.4 (α-rounding)**.** *(Todd and Yıldırım, 2007) Let $L \subseteq \mathbb{R}^d$ be a finite set such that $\text{span}(L) = \mathbb{R}^d$ and let $\alpha \geq 1$. Then an ellipsoid $E(\mathbf{G}, \mathbf{c})$ is called an α-rounding of $\text{conv}(L)$ if $\frac{1}{\alpha}(E(\mathbf{G}, \mathbf{c}) - \mathbf{c}) + \mathbf{c} \subseteq \text{conv}(L) \subseteq E(\mathbf{G}, \mathbf{c})$.*

Note that if α in the above definition is d (or equiv. $\sqrt{d}$), the corresponding ellipsoid is the Löwner ellipsoid.

In order to define a distance to any affine subspace, we first need the following ingredients.

Definition 1.5 (Orthogonal matrices)**.** *Let $d > j \geq 1$ be integers. We say $\mathbf{X} \in \mathbb{R}^{d \times j}$ is an orthogonal matrix if $\mathbf{X}^\top \mathbf{X} = \mathbb{I}_j$. We use $\mathcal{V}_j \subseteq \mathbb{R}^{d \times j}$ to denote the set of all $d \times j$ orthogonal matrices.*

Definition 1.6 (j-dimensional subspace)**.** *Let $d > j \geq 1$ be integers and let $\mathbf{v} \in \mathbb{R}^d$. Let $\mathbf{X} \in \mathcal{V}_j$ and $\mathbf{Y} \in \mathcal{V}_{d-j}$ such that $\mathbf{Y}^\top \mathbf{X} = 0^{(d-j) \times j}$ and $\mathbf{X}^\top \mathbf{Y} = 0^{j \times (d-j)}$. Let $H(\mathbf{X}, \mathbf{v}) := \{\mathbf{X}\mathbf{X}^\top \mathbf{p} + \mathbf{v} \,|\, \mathbf{p} \in \mathbb{R}^d\}$ denote the j-dimensional affine subspace H that is spanned by the column space of $\mathbf{X}$ and offset by $\mathbf{v}$. Let $\mathcal{H}_j := \{H(\mathbf{X}, \mathbf{v}) \,|\mathbf{X} \in \mathcal{V}_j, \mathbf{v} \in \mathbb{R}^d\}$ denote the set of all j-affine subspaces in $\mathbb{R}^d$.*

We use $\text{dist}(H(\mathbf{X}, \mathbf{v}), \mathbf{p}) := \|(\mathbf{p} - \mathbf{v})^\top \mathbf{Y}\|_2$ to denote the distance between any point $\mathbf{p} \in \mathbb{R}^d$ and the j-dimensional affine subspace $H(\mathbf{X}, \mathbf{v})$, where here $\mathbf{Y} \in \mathbb{R}^{d \times (d-j)}$ such that $\mathbf{Y}^\top \mathbf{X} = 0^{(d-j) \times j}$.

We now define the term *query space* which will aid us in simplifying the proofs as well as the corresponding theorems.

Definition 1.7 (query space)**.** *Let $1 \leq j < d < n$ be positive integers and let $P \subseteq \mathbb{R}^d$ be a set of n points such that $\text{span}(P) = \mathbb{R}^d$. Then for the union of all*

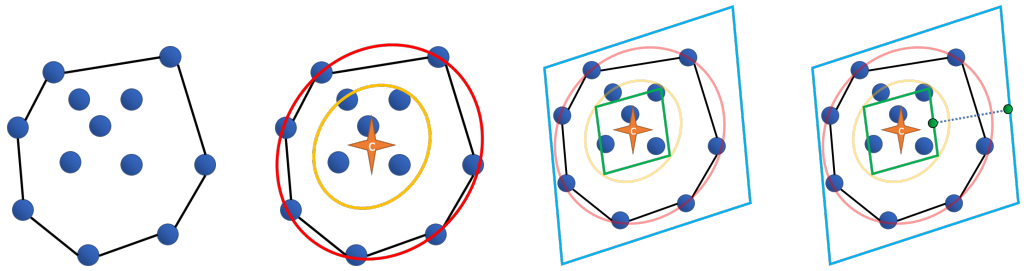


Fig. 1: **Overview of our approach (see Algorithm 1).** Images from left to right: Steps 1 and 2: An approximated ellipsoid E is computed satisfying the conditions of Theorem 1.3 (red ellipsoid), where the ellipsoid E' (in orange) is a dilation of E by a factor of $\alpha = d$ with respect to the center c of E (orange star). Step 3: The vertices of E' (orange ellipsoid) are computed and then dilated such that their convex hull (cyan outline) will contain $\text{conv}(P)$ (black outline). Step 4: A Caratheodory set of $d+1$ points from P is computed for each vertex point of E' (green convex hull). Finally, each point on $\text{conv}(P)$ (black outline) can be represented by a convex combination of a point on the convex hull of the vertices of E' and their dilated points (cyan outline).

Table 1: M-estimator loss functions that can be captured by our coreset construction; d here denotes the dimension of the input data P; all lemmata below can be found at Section B of the supplementary material.

Loss Function Ψ	Formulation	multiplicative error (ℓ_∞-coreset)	Reference
Cauchy	$\left(\lambda^2/2\right)\log\left(1+(x/\lambda)^2\right)$	$8(d+1)^3$	Lemma B.1
Welsch	$\frac{\lambda^2}{2}\left(1-e^{-\left(\frac{x}{\lambda}\right)^2}\right)$	$8(d+1)^3$	Lemma B.3
Huber	$\begin{cases} x^2/2 & \text{If } \lvert x\rvert \le \lambda \\ \lambda\lvert x\rvert - \lambda^2/2 & \text{otherwise} \end{cases}$	$16(d+1)^3$	Lemma B.4
Geman-McClure	$x^2/\left(2+2x^2\right)$	$8(d+1)^3$	Lemma B.5
Concave	$\frac{d^2\Psi}{dx^2} \le 0$	$4(d+1)^{1.5}$	Lemma B.7
Tukey	$\begin{cases} \frac{\lambda^2}{6}\left(1-\left(1-\frac{x^2}{\lambda^2}\right)^3\right) & \text{if } \lvert x\rvert \le \lambda \\ \frac{\lambda^2}{6} & \text{otherwise} \end{cases}$	$8(d+1)^3$	Lemma B.8
$L_1 - L_2$	$2\left(\sqrt{1+x^2/2}-1\right)$	$8(d+1)^3$	Lemma B.9
Fair	$\lambda\lvert x\rvert - \lambda^2\ln\left(1+\lvert x\rvert/\lambda\right)$	$8(d+1)^3$	Lemma B.10
Power Bounded	$\Psi_{Pow}(y)/\Psi_{Pow}(x) \le (y/x)^z$ for all $0 \le x \le y$	$4^z(d+1)^{1.5z}$	Lemma 3.2

j-affine subspaces $\mathcal{H}_j$, the tuple $(P, \mathcal{H}_j, \text{dist})$ is called a query space.

Following the previous definition, we now can define the notion of L_∞ coreset and L_2 coreset.

Definition 1.8 (L_∞ coreset)**.** *Let $j \in [d-1]$, $\varepsilon \in (0,1)$, and $(P, \mathcal{H}_j, \text{dist})$ be a query space. Then a set $C \subseteq P$ is called an L_∞ ε-coreset with respect to the query space $(P, \mathcal{H}_j, \text{dist})$ if for every $\mathbf{X} \in \mathcal{V}_j$ and $\mathbf{v} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \text{dist}(H(\mathbf{X},\mathbf{v}),\mathbf{p}) \le (1+\varepsilon)\max_{\mathbf{p}\in C} \text{dist}(H(\mathbf{X},\mathbf{v}),\mathbf{p})$.*

Definition 1.9 (L_2 coreset)**.** *Let $j \in [d-1]$, $\varepsilon \in (0,1)$, and $(P, \mathcal{H}_j, \text{dist})$ be a query space. Then a set $C \subseteq P$ with a weight function $w : C \to \mathbb{R}$ is called an L_2 ε-coreset with respect to the query space $(P, \mathcal{H}_j, \text{dist})$ if for every $\mathbf{X} \in \mathcal{V}_j$ and $\mathbf{v} \in \mathbb{R}^d$, $\sum_{\mathbf{p}\in P} \text{dist}(H(\mathbf{X},\mathbf{v}),\mathbf{p})^2 \le (1+\varepsilon)\sum_{\mathbf{p}\in C} w(\mathbf{p})\,\text{dist}(H(\mathbf{X},\mathbf{v}),\mathbf{p})^2$.*

Finally, we define a coreset for the k j-cylinders problem, followed by the Carathéodory's theorem which will be used in our proofs and algorithms in computing the L_∞ coreset for the (k,j)-projective clustering problem.

Definition 1.10. *A closed j-cylinder of radius r is a set of points in $\mathbb{R}^d$ whose distance to a certain j-flat is at most r. A set D is an L_∞ C-coreset of $P \subseteq \mathbb{R}^d$ for the (j,k)-projective clustering problem if D is a subset of P such that there exists a union of k j-cylinders of radius Cr that covers P for each union of k j-cylinders of radius r that covers D.*

Theorem 1.11 (Carathéodory's theorem)**.** *(Carathéodory, 1907; Steinitz, 1913) For any $A \subset \mathbb{R}^d$ and $\mathbf{p} \in \text{conv}(A)$, there exists $m \le d+1$ points $\mathbf{p}_1, \ldots, \mathbf{p}_m \in A$ such that $\mathbf{p} \in \text{conv}(\{\mathbf{p}_1, \ldots, \mathbf{p}_m\})$.*

2 L_∞ CORESETS FOR PROJECTIVE CLUSTERING

First, we note that Har-Peled (2004) showed that L_∞ coresets do not exist when the input set is n points from $\mathbb{R}^d$. However in this paper, we consider the integer projective clustering problem, e.g. Edwards and Varadarajan (2005), where the input points lie on a polynomial grid.

We first give an L_∞ coreset for the $(j,1)$-projective clustering problem in Section 2.1. We then use our L_∞ coreset for the $(j,1)$-projective clustering to inductively build an L_∞ coreset for the (j,k)-projective clustering problem. For brevity purposes, proofs of the technical results have been omitted from this manuscript; we refer the reader to the supplementary material for the proofs.

2.1 L_∞ Coreset for $(j,1)$-Projective Clustering

We first give an overview for our algorithm that produces a constant factor approximation coreset for the $(j,1)$-projective clustering problem. We again emphasize that our coreset for the $(j,1)$-projective clustering problem serves as our main technical contribution because we use Carathéodory's theorem to explicitly find a polynomial number of points to add to our coreset. We can then use a natural inductive argument to recursively add a polynomial number of points to create a coreset for the integer (j,k)-projective clustering problem. By contrast, even the base case for the only existing coreset for the integer (j,k)-projective clustering problem already contains an exponential number of points (Edwards and Varadarajan, 2005).

The algorithm takes as input a set $P \subseteq \mathbb{R}^d$ of n points, which are promised to lie on a flat of dimension j, and computes a subset $C \subseteq P$, which satisfies Theorem 2.2. The algorithm appears in full detail in Algorithm 1 and first initializes C to be an empty set. Our algorithm computes $H(\mathbf{W},\mathbf{u})$ to be the j-dimensional flat that contains P and sets Q to be the set of points obtained by projecting P onto the column space of $\mathbf{W}$. The algorithm then defines $E(\mathbf{G},\mathbf{c})$ to be the John-Löwner ellipsoid containing the convex hull of Q and S to be the set of vertices defined the axes of symmetry and the center of the scaled ellipsoid $\frac{1}{j}(E(\mathbf{G},\mathbf{c})-\mathbf{c})+\mathbf{c}$, which can be explicitly and efficiently computed, and note that $|S| \leq 2j$. From Carathéodory's theorem, we can express each point in $S \cup \{\mathbf{c}\}$ as a linear combination of $j+1$ points from Q. We thus define K to be the $\mathcal{O}\left(j^2\right)$ points of Q needed to represent all points in $S \cup \{\mathbf{c}\}$ and set $C = \mu(K)$, where μ is the inverse mapping from Q to P. We first prove the following structural property that follows from Carathéodory's theorem.

Algorithm 1: Coreset for $(j,1)$-Projective Clustering

Input: $P \subseteq \mathbb{R}^d$ of n points that lie on a flat of dimension j
Output: Coreset of size $\mathcal{O}\left(j^2\right)$

$C \leftarrow \emptyset$
Let $H(\mathbf{W},\mathbf{u}) :=$ a j-dimensional flat containing P
$Q := \left\{\mathbf{W}^\top \mathbf{p} \,|\, \mathbf{p} \in P\right\}$
Let μ be function that maps each point $q \in Q$ to its original point in P
Let $E(\mathbf{G},\mathbf{c}) :=$ the John-Löwner ellipsoid of the convex hull of Q
$S :=$ the vertices of the scaled ellipsoid $\frac{1}{j}\left(E(\mathbf{G},\mathbf{c})-\mathbf{c}\right)+\mathbf{c}$
for *each* $\mathbf{s} \in S \cup \{\mathbf{c}\}$ **do**
    $K_{\mathbf{s}} :=$ be at most $j+1$ points from Q whose convex hull contains $\mathbf{s}$
    $C := C \cup \mu\left(K_{\mathbf{s}}\right)$
return C

Lemma 2.1. *Let $d,\ell,m \geq 1$ be integers. Let $\mathbf{p} \in \mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$ be a set of m points with $\mathbf{p} \in \operatorname{conv}(A)$ so that there exists $\alpha : A \to [0,1]$ such that $\sum_{\mathbf{q}\in A} \alpha(\mathbf{q}) = 1$ and $\sum_{\mathbf{q}\in A} \alpha(\mathbf{q}) \cdot \mathbf{q} = \mathbf{p}$. Then for every $\mathbf{Y} \in \mathbb{R}^{d\times\ell}$ and $\mathbf{v} \in \mathbb{R}^\ell$, $\|\mathbf{p}^\top\mathbf{Y} - \mathbf{v}\|_2 \leq \max_{\mathbf{q}\in A} \|\mathbf{q}^\top\mathbf{Y} - \mathbf{v}\|_2$.*

We use Lemma 2.1 to show that Algorithm 1 gives a coreset for the $(j,1)$-projective clustering problem as summarized below. In addition, we show that our ℓ_∞-coreset is also applicable towards the (j,z)-clustering where j denotes the dimensionality of the subspace, and z denotes the power of the distance function. For instance, $z \in [1,2)$ is used for obtaining robust clustering, which is useful against outliers.

Theorem 2.2. *Let $j \in [d-1]$, $z \geq 1$, and let $(P,\mathcal{H}_j,\operatorname{dist})$ be a query space, where P lies in a j-dimensional flat. Let $C \subseteq P$ be the output of Algorithm 1. Then $|C| = \mathcal{O}\left(j^2\right)$ and for every $H(\mathbf{X},\mathbf{v}) \in \mathcal{H}_j$, we have $\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p},H(\mathbf{X},\mathbf{v}))^z \leq 2^{z+1} j^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q},H(\mathbf{X},\mathbf{v}))^z$.*

2.2 L_∞ Coreset for (j,k)-Projective Clustering

Our coreset construction is recursive. Generally speaking, we construct a coreset D_k for (j,k)-projective clustering from a coreset D_{k-1} for $(j,k-1)$-projective clustering. For the base case, we show how to construct a coreset D_1 for $(j,1)$-projective clustering in Theorem 2.2. Now for $k \geq 2$, given a coreset $D_{k-1} \subset P$ for $(j,k-1)$-projective clustering, the construction of D_k has $j+1$ levels and the i-th level will specify

$i+1$ points $\mathbf{v}_0, \ldots, \mathbf{v}_i$ and a corresponding point set $P[\mathbf{v}_0, \ldots, \mathbf{v}_i] \subset P$. We first add D_{k-1} into D_k and separately initialize Level 0 with $\mathbf{v}_0$ being each point of D_{k-1} and define $P[\mathbf{v}_0] = P$. Crucially, each of the $j+1$ levels only adds to the coreset a number of points that is polynomial in $j \leq d-1$ at each level due to the base case using our new coreset for $(j,1)$-projective clustering based on Carathéodory's theorem. Hence, the total number of points is polynomial in d but exponential in j. By contrast, existing coreset constructions of Edwards and Varadarajan (2005) use partitions that must be analyzed over d levels due to their lack of an efficient coreset for their base case; thus their size is exponential in d.

Level 0: Given any choice of $\mathbf{v}_0$ from D_{k-1}, we define $P[\mathbf{v}_0] := P \subset [\Delta]^d$, we have $\operatorname{dist}(\mathbf{p}, \mathbf{v}_0) \in [1, \Delta\sqrt{d}]$ for every $\mathbf{p} \in P[\mathbf{v}_0]$. We can partition $P[\mathbf{v}_0]$ into $\ell = \mathcal{O}(\log(d\Delta))$ sets $K_{0,0}, K_{0,1}, \ldots, K_{0,\ell}$ such that $K_{0,0} = \{\mathbf{v}_0\}$ and $K_{0,i} = \{\mathbf{p} \in P[\mathbf{v}_0] : 2^{i-1} \leq \operatorname{dist}(\mathbf{p}, \mathbf{v}_0) \leq 2^i\}$ for $i \geq 1$. Intuitively, this can be seen as partitioning the points of P into sets with exponentially increasing distance from $\mathbf{v}_0$. For each $K_{0,i}$, we construct an L_∞-coreset $D_{0,i}$ of $K_{0,i}$ for the $(j, k-1)$-projective clustering problem and add $D_{0,i}$ into D_k. We then separately select $\mathbf{v}_1$ to be any point in $D_{0,i}$ across all $i \in [\ell]$ and set $P[v_0, v_1] = \cup_{x=0}^{i} K_{0,x}$.

Level t, for $t \in [1, j]$: Given $\mathbf{v}_0, \ldots, \mathbf{v}_t$ and $P[\mathbf{v}_0, \ldots, \mathbf{v}_t]$, let A_t denote the affine subspace spanned by $\mathbf{v}_0, \ldots, \mathbf{v}_t$. We recall the following structural properties about the convex hull of affine subspaces.

Lemma 2.3 ((Edwards and Varadarajan, 2005), Lemma 45 in (Feldman et al., 2020))**.** *Let $\Delta \geq 2$, k be a positive integer, and $j \leq d-1$ be a positive integer. Let $\mathcal{Q}_{j,k}$ be the family of all sets of k affine subspaces of $\mathbb{R}^d$ with dimension j. Let $\mathbf{A} \in \{-\Delta, \ldots, \Delta\}^{n \times d}$. Then for every $H \in \mathcal{H}_j$, we have either $\operatorname{dist}(H, \mathbf{A}) = 0$ or $\operatorname{dist}(H, \mathbf{A}) \geq \frac{1}{(d\Delta)^{cj}}$, for some universal constant $c > 0$.*

By Lemma 2.3, we have that for every $\mathbf{p} \in P[\mathbf{v}_0, \ldots, \mathbf{v}_t]$, that $\operatorname{dist}(\mathbf{p}, A_t)$ is either 0 or in the range $[1/(d\Delta)^{cj}, 2\Delta\sqrt{d}]$. Thus we can once again partition $P[\mathbf{v}_0, \ldots, \mathbf{v}_t]$ into $\mathcal{O}(j \log(d\Delta))$ subsets $K_{t,0}, \ldots, K_{t,\ell}$ such that $K_{t,0} = P[\mathbf{v}_0, \ldots, \mathbf{v}_t] \cap A_t$ and for each integer $i \in [\ell]$, $K_{t,i} := \{\mathbf{p} \in P[\mathbf{v}_0, \ldots, \mathbf{v}_t] : 2^{i-1} c_j/\Delta^j \leq \operatorname{dist}(\mathbf{p}, A_t) < 2^i c_j/\Delta^j\}$. For each $K_{t,i}$, we construct an L_∞-coreset $D_{t,i}$ of $K_{t,i}$ for $(j, k-1)$-projective clustering and add $D_{t,i}$ to D_k. We then separately select $\mathbf{v}_{t+1}$ to be any point in $D_{t,i}$ across all $i \in [\ell]$ and set $P[v_0, \ldots, \mathbf{v}_{t+1}] = \cup_{x=0}^{i} K_{t,x}$. We remark that we terminate at level $j+1$. Finally, in what follows, we give a bound on the size of our L_∞ coreset.

Lemma 2.4 (Coreset size)**.** *Let $f(k) = |D_k|$ denote the size of the coreset D_k formed at level k for (j,k)-projective clustering. Then $f(k) = \left(8j^3 \log(d\Delta)\right)^{\mathcal{O}(jk)}$.*

To prove that our construction yields an L_∞ constant-factor approximation coreset for the integer (j,k)-projective clustering problem, we use a structural property about the convex hull of affine subspaces. Informally, the property says that if $\mathbf{v}_0, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ are $d+1$ affinely independent vectors that induce a sequence of affine subspaces $\mathbf{A}_0, \ldots, \mathbf{A}_d$, then under certain assumptions, the convex hull formed by $\mathbf{v}_0, \ldots, \mathbf{v}_d$ contains a translation of a scaled hyperrectangle formed by a sequence $\mathbf{u}_0, \ldots, \mathbf{u}_d$ of vectors formed by the orthogonal projection away from $\mathbf{A}_0, \ldots, \mathbf{A}_d$. For more details, see Section A.4 in the supplementary material. Using this structural property, we achieve an L_∞ constant-factor approximation coreset for the integer (j,k)-projective clustering problem with size $(8j^3 \log(d\Delta))^{\mathcal{O}(jk)}$:

Lemma 2.5. *There exists a universal constant $\xi > 0$ such that D_k is a ξ-coreset for the (j,k)-projective clustering problem.*

Theorem 1.1 then follows from Lemma 2.5 and Lemma 2.4 and the observation that $j \leq d-1$. Thus our coresets have size polynomial in d, resolving the natural open question from Edwards and Varadarajan (2005).

Algorithm 2: Coreset for (j,k)-Projective Clustering

Input: $P \subseteq \mathbb{R}^d$ of n points, an integer $j \in [d-1]$, an integer $k \geq 1$, an accuracy parameter $\varepsilon \in (0,1)$ and a failure probability $\delta \in (0,1)$.

Output: A weighted set (C, u)

$P_1 := P$, $i := 1$, $C := \emptyset$

while $|P_i| \geq 1$ **do**

  $S_i :=$ an L_∞-coreset for (j,k)-projective clustering

  for *every* $\mathbf{p} \in S_i$ **do**

    $s(\mathbf{p}) := \frac{1}{i} \cdot |S_i|$

      // $|S_i| = \mathcal{O}\left(j^{1.5}(j \log(d\Delta))^{\mathcal{O}(jk)}\right)$

  $P_{i+1} := P_i \setminus S_i$, $i := i+1$

$t := \sum_{\mathbf{p} \in P} s(\mathbf{p})$

  // $t = \mathcal{O}\left(j^{1.5}(j \log(d\Delta))^{\mathcal{O}(jk)} \log n\right)$

$m := \frac{ct}{\varepsilon^2}\left(djk \log \frac{t}{\delta}\right)$

for m *iterations* **do**

  Sample a point $\mathbf{p} \in P$ with probability $\frac{s(\mathbf{p})}{t}$

  $C := C \cup \{\mathbf{p}\}$, $u(\mathbf{p}) := \frac{t}{m \cdot s(\mathbf{p})}$

return (C, u)

3 APPLICATIONS

In this section, we show that our framework gives an L_∞ coreset for subspace clustering, as well as a large class of M-estimators. To the best of our knowledge, our constructions are the first coresets with size polynomial in d for these M-estimators. Namely, our algorithm achieves approximate regression for the Cauchy, Welsch, Huber, Geman-McClure, Tukey, $L_1 - L_2$, Fair loss functions, as well as general loss functions that are concave or power-bounded; see Table 1.

Beyond traditional projective clustering. First, we present that our L_∞-coreset algorithm is applicable for a family of non-decreasing log-log Lipschitz function.

Theorem 3.1 (L_∞ coreset for log-log Lipschitz loss functions)**.** *Let $j \in [d-1]$, $z \geq 1$, and let $(P, \mathcal{H}_j, \text{dist})$ be a query space, where P lies in a j-dimensional flat. Let $f : [0,\infty) \to [0,\infty)$ such that both (1) f is a monotonically non-decreasing function, i.e., for every $x, y \in [0,\infty)$ with $x \leq y$, it holds that $f(x) \leq f(y)$ and (2) f is log-log Lipschitz, i.e., there exists $\rho \geq 1$ for every $b \geq 1$ such that $f(bx) \leq b^\rho f(x)$. Let C be the output of a call to $L_\infty - \textsc{Coreset}(P)$. Then for every $H \in \mathcal{H}_j$, $\max_{\mathbf{p}\in P} f(\text{dist}(p, H(\mathbf{X}, \mathbf{v}))^z) \leq (2^{z+1} j^{1.5z})^\rho \max_{\mathbf{p}\in C} f(\text{dist}(p, H(\mathbf{X}, \mathbf{v}))^z)$.*

Although the above theorem is applicable to large family of functions, it may not yield tight bounds for each of the loss functions in Table 1. Thus we first prove the following lemma, which guarantees an coreset for power-bounded loss functions $\Psi_{Pow}(x)$.

Lemma 3.2 (L_∞ coreset for regression with power-bounded loss function)**.** *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let $z > 0$ be a fixed constant. Let Ψ_{Pow} denote any non-decreasing loss function with $\Psi_{Pow}(0) = 0$ and $\Psi_{Pow}(y)/\Psi_{Pow}(x) \leq (y/x)^z$ for all $0 \leq x \leq y$. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \,|\, \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \,|\, \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \Psi_{Pow}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 4^z (d+1)^{1.5z} \cdot \max_{\mathbf{q}\in C} \Psi_{Pow}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Since power-bounded loss functions satisfy the conditions of Theorem 3.1, then we can immediately apply Theorem 3.1 to obtain a base case for $z = 1$. Lemma 3.2 then follows by the definition of power-bounded loss functions for general z.

It turns out that many of the loss functions of interest in Table 1 are power-bounded loss functions with specific parameters, so we can apply Theorem 3.1 in the same way as the proof of Lemma 3.2 to obtain the guarantees for Cauchy regression, Huber regression, and Gem-McClure regression. However, in certain cases, we can prove structural properties bounding the growth of these loss functions to obtain guarantees that are sharper than those provided by Theorem 3.1. We prove such structural properties at Section B of the supplementary material to handle Welsch regression, regression with concave loss functions, Tukey regression, $L_1 - L_2$ regression, and Fair regression.

L_∞-coreset to L_2-coreset for integer (j,k)-projective clustering. To construct an ε-coreset, we use our L_∞ coreset along with the framework of sensitivity sampling, in which points are sampled according to their sensitivity, a quantity that roughly captures how important or unique each point is. We give the coreset construction in Algorithm 2 using a standard reduction from an L_2 coreset to an L_∞ coreset based on sensitivity sampling as summarized below.

Theorem 3.3. *With constant probability, Algorithm 2 outputs an L_2 $(1+\epsilon)$-coreset for (j,k)-projective clustering of P.*

Time complexity of our methods. The running time of Algorithm 2, we need to handle two cases – (1) $k = 1$, and (2) $k > 1$. Observe that the time needed for constructing our L_2-coreset for (k,j)-projective clustering where $k = 1$ and any $j \geq 2$ is bounded by $O\left(n\left(n + j^4 \log n\right)\right)$ time. Specifically speaking, the time depends heavily on the time that Algorithm 1. Algorithm 1 depends heavily on the computation of the Löwner ellipsoid and on applying Carathéodory's theorem. The time needed to compute the Löwner ellipsoid of a given set of point $Q \subseteq \mathbb{R}^j$ such that $|Q| = n$ is bounded by $O\left(nj^3 \log n\right)$ Todd and Yıldırım (2007). As for constructing the Caratheodory set, recently provided an algorithm for computing such set in time $O\left(nj + j^4 \log n\right)$. Combining these two methods with the observation that Algorithm 2 has $O\left(\frac{n}{j^2}\right)$ calls to Algorithm 1, results in the upper bound above.

As for $k \geq 2$, following our analyzed steps needed to construct an L_∞-coreset for the (k,j)-projective clustering problem and its variants, the running time is bounded from above by $O\left(nj^4 \left(\log \Delta\right)^{j^2 k}\right)$. Hence, Algorithm 2 requires $O\left(n^2 j^4 \left(\log \Delta\right)^{j^2 k}\right)$ to construct an L_2-coreset for the (k,j)-projective clustering problem.

We note that our algorithm can be boosted theoretically speaking via the use of the merge-and-reduce tree Feldman (2020), resulting in an algorithm that are near-linear in n rather than quadratic in n.

We further note that, our assumption on P being contained in some j-dimensional affine subspace can be dropped as follows.

Remark 3.4. *So far, P was assumed to lie on j-dimensional subspaces, however, one can remove this*

assumption by using Theorem 7 of Varadarajan and Xiao (2012a).

Subspace clustering. We first recall that subspace clustering is a variant of projective clustering where $k = 1$ and $j \in [d-1]$.

M-estimator regression. We present various robust $(1, d-1)$-projective clustering problems for which a strong ε-coreset can be generated using our algorithms. We are given a set P of n points in $\mathbb{R}^d$ and a function $b : P \to \mathbb{R}$, and our goal is to optimize the minimization problem $\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{\mathbf{p}\in P} \Psi\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right)$, where Ψ is any loss function. In particular, the choice of Ψ encompasses many robust regression loss functions that have been designed to reduce the effect of outliers across various optimization problems. We show that Algorithm 1 achieves an L_∞-coreset with accuracy $1 - \frac{1}{\text{poly}(d)}$ for a variety of loss functions; See Section B in the supplementary material.

4 EXPERIMENTS

In this section, we evaluate our coreset against uniform sampling on synthetic and real-world datasets, with respect to the projective clustering problem and its variants.

Software/Hardware. Our algorithms were implemented Tukan et al. (2022) in Python 3.6 (Van Rossum and Drake, 2009) using "Numpy" (Oliphant, 2006), "Scipy" (Virtanen et al., 2020). Tests were performed on 2.59GHz i7-6500U (2 cores total) machine with 16GB RAM.

Datasets. The following datasets used for our experiments were mostly from UCI machine learning repository (Dua and Graff, 2017): (i) **Synthetic** – $20,000$ points in the two dimensional Euclidean space where $19,990$ points lie on the x-axis while the remaining 10 points are generated away from the x-axis. (ii) **Bike Sharing Dataset Data Set** (Dua and Graff, 2017) – consists of 17389 samples, and 17 features of which only 15 were used for the sake of our comparisons. (iii) **Physicochemical Properties of Protein Tertiary Structure Data Set** (Dua and Graff, 2017) – $45,730$ samples, each consisting of 10 features.

Evaluation against uniform sampling. Throughout the experiments, we have chosen 10 sample sizes, starting from 100 till $1,000$ for projective clustering problems and from $1,000$ till $10,000$ for regression problems; see Figure 3. At each sample size, we generate two coresets, where the first is using uniform sampling and the latter is using Algorithm 2. When handling projective clustering problems, for each coreset (S, v), we have computed a suboptimal solution $\tilde{H} \in \mathcal{H}_j$ using an EM-like algorithm (Expectation Maximization) where the number of steps for convergence was 6 while the number of different initializations was set to $1,000$. E.g., in Figure 2c, the goal was to find an suboptimal solution $\tilde{H}$ for the problem $\min_{H\in\mathcal{H}_j} \sum_{p\in S} v(p)\text{dist}\left(p, H\left(X, v\right)\right)^2$. As for regression related problems, we have computed the suboptimal solution using Scipy's (Virtanen et al., 2020) own optimization sub-library which can handle such problem instances, where similarly to the projective clustering settings, we have ran the solver for 100 iterations (at max) while having at max $15,000$ different initializations for the solver. The approximation error ε is set to be the ratio $\sum_{p\in P} f\left(\text{dist}\left(p, \tilde{H}\left(X, v\right)\right)\right)$ to $\left(\min_{H\in\mathcal{H}_j} \sum_{p\in P} f\left(\text{dist}\left(p, H(X, v)\right)\right)\right) - 1$. Finally, the results were averaged across 22 trials, while the shaded regions correspond to the standard deviation.

Choice of baseline. We remark that uniform sampling was selected as the baseline for our algorithm because the only existing coreset construction with theoretical guarantees for the integer (j, k)-projective clustering problem is that of Edwards and Varadarajan (2005). However, their construction is known to be impractical due to the large coreset size. In fact, even the base case requires a number of points that is exponential in d; thus we could not implement the coreset construction of Edwards and Varadarajan (2005). In practice uniform sampling is used due to the observation that real-world data is often not "worst-case" data. Thus it is a natural choice to compare the performance of our algorithm to that of uniform sampling across a number of real-world datasets, even though it is clear that we can generate synthetic data for which uniform sampling can perform arbitrarily badly due to its lack of provable guarantees, while our coreset constructions still maintains its theoretical guarantees.

Discussion. First note that our coresets are generally more accurate than uniform sampling across the experiments, sometimes outperforming uniform sampling by a factor of ≈ 10000, e.g., $(2, 2)$-projective clustering with the Tukey loss function in Figure 2e. Moreover, there exist data distributions in which uniform sampling provably performs *arbitrarily* worse than our coreset construction. For example, consider choosing $k = 2$ centers across n points when $n-1$ points are located at the origin and a single point is located at the position N on the x-axis. Then the optimal clustering has cost zero by choosing a center at the origin and a center at N, but uniform sampling will not find the point at N without $\Omega(N)$ samples and thus incur cost N. Since our coreset finds a multiplicative approximation to the optimal solution, it will also achieve a clustering with cost zero, which is arbitrarily better than N, sampling only polylog(n) points. On the other hand, in

Table 2: **Summary of our results:** Our coreset construction was applied on various application of projective clustering, of which were robust regression as well as robust subspace clustering

Problem type	Loss function	k	j	Dataset	Figure
Regression	Huber	1	$d-1$	(i)	2a
Regression	Cauchy	1	$d-1$	(i)	2b
$(2,2)$-projective clustering	L_2^2	2	2	(ii)	2c
Robust $(2,2)$-projective clustering	Cauchy	2	2	(ii)	2d
Robust $(2,2)$-projective clustering	Tukey	2	2	(iii)	2e
Robust $(2,2)$-projective clustering	Welsch	2	2	(iii)	2f

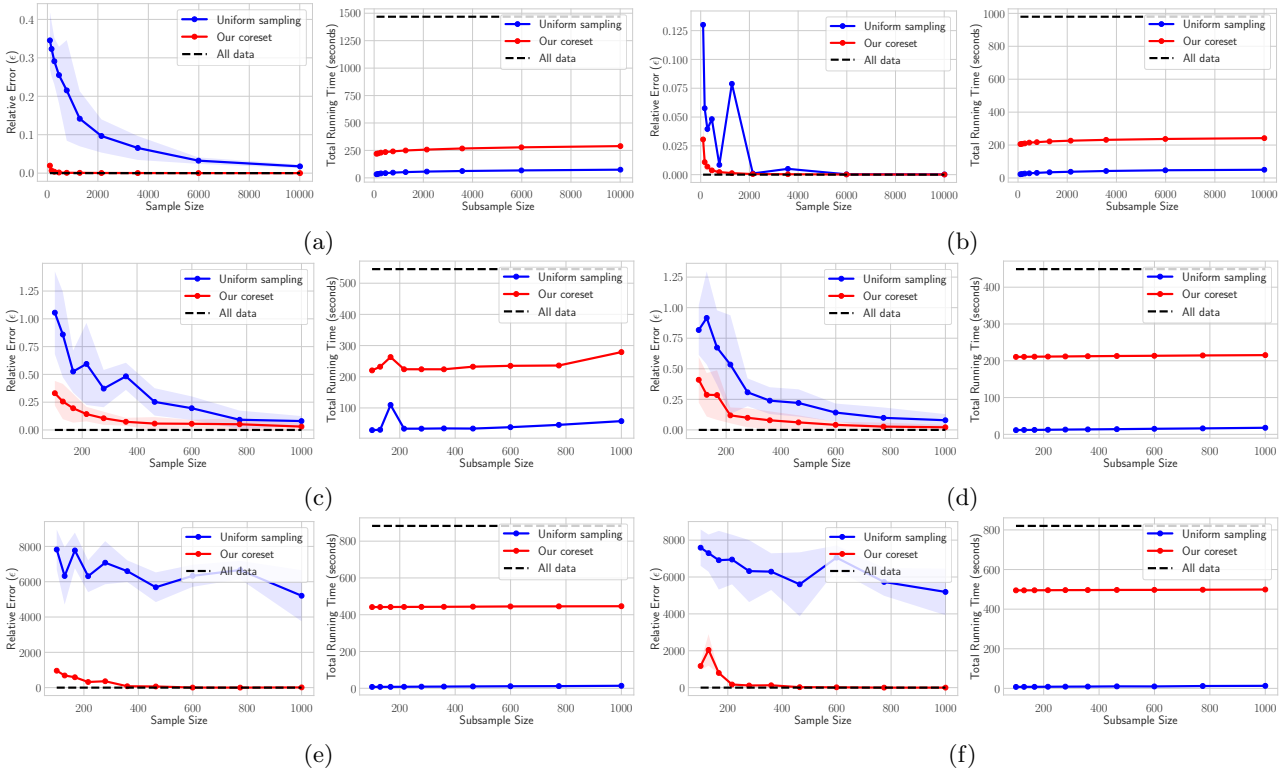


Fig. 2: Our experimental results: evaluating the efficacy of our coreset against uniform sampling.

some of the figures, e.g., Figure 2e, as we increase the sample size, the approximation error that corresponds to our coreset might increase at some sample sizes. This phenomenon is associated with the probabilistic nature of our coreset, as our coreset is a result of a sensitivity sampling technique. This problem can be easily resolved via increasing the number of trials (the number of trials was chosen to be 22). The same holds for uniform sampling.

Although our coreset is generally better in terms of approximation error than uniform sampling, however the running time of our implementation is slow. We strongly believe that our algorithm can achieve faster results using the merge-and-reduce tree on the expense of an increase in the approximation error. For additional results, see Section C at the appendix.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have provided an L_∞ and L_2 coresets for (k,j)-projective clustering problems and its variants, e.g., M-estimators. Our approach leveraged an elegant combination between Löwner ellipsoid and Carathéodory's theorem. This in term sheds light on the use of constant-approximation coresets (our L_∞ coreset) as a stepping stone towards L_2 coresets with ε approximation. We believe that there is room for future work with respect to constructing L_∞-coresets with smaller sizes for constant factor approximation. Finally, the lower bound on the size of constant factor coresets for the (j,k)-projective clustering problem is still unknown. We hope our work presents an important step in resolving the complexity of this problem.

6 ACKNOWLEDGEMENTS

This research was partially supported by the Israel National Cyber Directorate via the BIU Center for Applied Research in Cyber Security, and supported in part by NSF CAREER grant 1652257, NSF grant 1934979, ONR Award N00014-18-1-2364 and the Lifelong Learning Machines program from DARPA/MTO. In addition, Samson Zhou would like to thank National Institute of Health grant 5401 HG 10798-2 and a Simons Investigator Award of David P. Woodruff.

References

Ackermann, M. R. and Blömer, J. (2009). Coresets and approximate clustering for bregman divergences. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1088–1097.

Agarwal, P. K., Har-Peled, S., and Yu, H. (2006). Robust shape fitting via peeling and grating coresets. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 182–191.

Aggarwal, C. C., Procopiuc, C. M., Wolf, J. L., Yu, P. S., and Park, J. S. (1999). Fast algorithms for projected clustering. In *SIGMOD, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 61–72. ACM Press.

Aggarwal, C. C. and Yu, P. S. (2000). Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 70–81.

Aloise, D., Deshpande, A., Hansen, P., and Popat, P. (2009). Np-hardness of euclidean sum-of-squares clustering. *Mach. Learn.*, 75(2):245–248.

Assadi, S., Bateni, M., Bernstein, A., Mirrokni, V. S., and Stein, C. (2019). Coresets meet EDCS: algorithms for matching and vertex cover on massive graphs. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1616–1635.

Badoiu, M., Har-Peled, S., and Indyk, P. (2002). Approximate clustering via core-sets. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing*, pages 250–257. ACM.

Braverman, V., Drineas, P., Musco, C., Musco, C., Upadhyay, J., Woodruff, D. P., and Zhou, S. (2020). Near optimal linear algebra in the online and sliding window models. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 517–528.

Braverman, V., Lang, H., Ullah, E., and Zhou, S. (2019). Improved algorithms for time decay streams. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM*, pages 27:1–27:17.

Carathéodory, C. (1907). Über den variabilitätsbereich der koeffizienten von potenzreihen, die gegebene werte nicht annehmen. *Mathematische Annalen*, 64(1):95–115.

Chakrabarti, K. and Mehrotra, S. (2000). Local dimensionality reduction: A new approach to indexing high dimensional spaces. In *VLDB, Proceedings of 26th International Conference on Very Large Data Bases*, pages 89–100.

Chen, K. (2009). On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947.

Clarkson, K., Wang, R., and Woodruff, D. (2019). Dimensionality reduction for tukey regression. In *International Conference on Machine Learning*, pages 1262–1271. PMLR.

Clarkson, K. L. (2008). Coresets, sparse greedy approximation, and the frank-wolfe algorithm. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 922–931.

Clarkson, K. L. and Woodruff, D. P. (2015). Input sparsity and hardness for robust subspace approximation. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS*, pages 310–329.

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., and Mahoney, M. W. (2008). Sampling algorithms and coresets for ℓ_p regression. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 932–941.

Deshpande, A., Rademacher, L., Vempala, S. S., and Wang, G. (2006). Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(12):225–247.

Deshpande, A. and Varadarajan, K. R. (2007). Sampling-based dimension reduction for subspace approximation. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pages 641–650.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Edwards, M. and Varadarajan, K. R. (2005). No coreset, no cry: II. In *FSTTCS: Foundations of Software Technology and Theoretical Computer Science, 25th International Conference, Proceedings*, pages 107–115.

Feldman, D. (2020). Core-sets: Updated survey. In *Sampling Techniques for Supervised or Unsupervised Tasks*, pages 23–44. Springer.

Feldman, D., Fiat, A., and Sharir, M. (2006). Coresets for weighted facilities and their applications. In *47th Annual IEEE Symposium on Foundations of Computer Science, FOCS Proceedings*, pages 315–324.

Feldman, D. and Langberg, M. (2011). A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC*, pages 569–578.

Feldman, D., Monemizadeh, M., Sohler, C., and Woodruff, D. P. (2010). Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 630–649. SIAM.

Feldman, D., Schmidt, M., and Sohler, C. (2020). Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657.

Feldman, D. and Schulman, L. J. (2012). Data reduction for weighted and outlier-resistant clustering. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 1343–1354.

Frahling, G. and Sohler, C. (2005). Coresets in dynamic geometric data streams. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 209–217.

Frahling, G. and Sohler, C. (2008). A fast k-means implementation using coresets. *Int. J. Comput. Geometry Appl.*, 18(6):605–625.

Har-Peled, S. (2004). No, coreset, no cry. In *FSTTCS: Foundations of Software Technology and Theoretical Computer Science, 24th International Conference, Proceedings*, pages 324–335.

Har-Peled, S. and Mazumdar, S. (2004). On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 291–300. ACM.

Har-Peled, S. and Varadarajan, K. R. (2002). Projective clustering in high dimensions using core-sets. In *Proceedings of the 18th Annual Symposium on Computational Geometry*, pages 312–318.

Huang, L., Jiang, S. H., Li, J., and Wu, X. (2018). Epsilon-coresets for clustering (with outliers) in doubling metrics. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 814–825.

Huang, L. and Vishnoi, N. K. (2020). Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 1416–1429.

John, F. (2014). Extremum problems with inequalities as subsidiary conditions. In *Traces and emergence of nonlinear programming*, pages 197–215. Springer.

Jubran, I., Tukan, M., Maalouf, A., and Feldman, D. (2020). Sets clustering. In *International Conference on Machine Learning*, pages 4994–5005. PMLR.

Kerber, M. and Raghvendra, S. (2015). Approximation and streaming algorithms for projective clustering via random projections. In *Proceedings of the 27th Canadian Conference on Computational Geometry, CCCG*.

Lee, E., Schmidt, M., and Wright, J. (2017). Improved and simplified inapproximability for k-means. *Inf. Process. Lett.*, 120:40–43.

Maalouf, A., Jubran, I., Tukan, M., and Feldman, D. (2021). Coresets for the average case error for finite query sets. *Sensors*, 21(19):6689.

Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. P. (2018). On coresets for logistic regression. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 6562–6571.

Mussay, B., Osadchy, M., Braverman, V., Zhou, S., and Feldman, D. (2020). Data-independent neural pruning via coresets. In *8th International Conference on Learning Representations, ICLR*.

Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

Phillips, J. M. and Tai, W. M. (2018). Improved coresets for kernel density estimates. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 2718–2727.

Procopiuc, C. M. (2010). Projective clustering. In *Encyclopedia of Machine Learning*, pages 806–811. Springer.

Procopiuc, C. M., Jones, M., Agarwal, P. K., and Murali, T. M. (2002). A monte carlo algorithm for fast projective clustering. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 418–427.

Sohler, C. and Woodruff, D. P. (2018). Strong coresets for k-median and subspace approximation: Goodbye dimension. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 802–813.

Steinitz, E. (1913). Bedingt konvergente reihen und konvexe systeme. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1913(143):128–176.

Todd, M. J. and Yıldırım, E. A. (2007). On khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. *Discrete Applied Mathematics*, 155(13):1731–1744.

Tukan, M., Baykal, C., Feldman, D., and Rus, D. (2021a). On coresets for support vector machines. *Theoretical Computer Science*.

Tukan, M., Maalouf, A., and Feldman, D. (2020). Coresets for near-convex functions. *Advances in Neural Information Processing Systems*, 33.

Tukan, M., Maalouf, A., Weksler, M., and Feldman, D. (2021b). No fine-tuning, no cry: Robust svd for compressing deep networks. *Sensors*, 21(16):5599.

Tukan, M., Wu, X., Zhou, S., Braverman, V., and Feldman, D. (2022). Open source code for all the algorithms presented in this paper. Link for open-source code.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Varadarajan, K. and Xiao, X. (2012a). On the sensitivity of shape fitting problems. *arXiv preprint arXiv:1209.4893*.

Varadarajan, K. R. and Xiao, X. (2012b). A near-linear algorithm for projective clustering integer points. In *SODA*.

Varadarajan, K. R. and Xiao, X. (2012c). On the sensitivity of shape fitting problems. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS*, pages 486–497.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.

Supplementary Material: New Coresets for Projective Clustering and Applications

A PROOF OF THE TECHNICAL RESULTS

A.1 Proof of Lemma 2.1

Proof. Since we can write $\mathbf{p}$ as the convex combination of points $\mathbf{q} \in A$ with weight $\alpha(\mathbf{q})$, we have

$$\|\mathbf{p}^\top \mathbf{Y} - \mathbf{v}\|_2 = \| \left(\sum_{\mathbf{q} \in A} \alpha(\mathbf{q}) \mathbf{q}^\top \mathbf{Y} \right) - \mathbf{v} \|_2.$$

Moreover, we have $\sum_{\mathbf{q} \in A} \alpha(\mathbf{q}) = 1$, so we can decompose $\mathbf{v}$ into

$$\|\mathbf{p}^\top \mathbf{Y} - \mathbf{v}\|_2 = \| \sum_{\mathbf{q} \in A} \alpha(\mathbf{q}) \left(\mathbf{q}^\top \mathbf{Y} - \mathbf{v} \right) \|_2.$$

By triangle inequality,

$$\|\mathbf{p}^\top \mathbf{Y} - \mathbf{v}\|_2 \leq \sum_{\mathbf{q} \in A} \alpha(\mathbf{q}) \|\mathbf{q}^\top \mathbf{Y} - \mathbf{v}\|_2 \leq \max_{\mathbf{q} \in A} \|\mathbf{q}^\top \mathbf{Y} - \mathbf{v}\|_2.$$

□

A.2 Proof of Theorem 2.2

Proof. To show the first part of the claim, note that since the ellipsoid $E(\mathbf{G}, \mathbf{c})$ has at most $2j$ vertices and each vertex point of the ellipsoid can be represented a convex combination fo at most $j+1$ points from Q by Carathéodory's theorem, then the number of points in C is at most $2(j+1)^2$, so that $|C| = \mathcal{O}\left(j^2\right)$.

To show the second part of the claim, we first set $H(\mathbf{W}, \mathbf{u})$ to be the j-flat containing P and $\mathbf{Y} \in \mathcal{H}_{d-j}$ so that $\mathbf{Y}^\top \mathbf{X} = 0^{(d-j)\times j}$ and $\mathbf{X}^\top \mathbf{Y} = 0^{j \times (d-j)}$. Notice that each $\mathbf{p} \in P$ satisfies

$$\text{dist}(p, H(\mathbf{X}, \mathbf{v}))^z = \|(\mathbf{p} - \mathbf{v})^\top \mathbf{Y}\|_2^z = \|(\mathbf{p} - \mathbf{u} + \mathbf{u} - \mathbf{v})^\top \mathbf{Y}\|_2^z.$$

Since $\mathbf{p}$ lies in the affine flat $H(\mathbf{W}, \mathbf{u})$, then we have

$$\text{dist}(p, H(\mathbf{X}, \mathbf{v}))^z = \| \left(\mathbf{W}\mathbf{W}^\top (\mathbf{p} - \mathbf{u}) + \mathbf{u} - \mathbf{v} \right)^\top \mathbf{Y} \|_2^z. \tag{1}$$

We now rely on properties of Carathéodory's Theorem and the John-Löwner ellipsoid to bound (1).

First note that

$$\begin{aligned} &\| \left(\mathbf{W}\mathbf{W}^\top (\mathbf{p} - \mathbf{u}) + \mathbf{u} - \mathbf{v} \right)^\top \mathbf{Y} \|_2^z = \\ &\quad \| \left(\mathbf{W}\mathbf{W}^\top \mathbf{p} - \mathbf{W}\mathbf{W}^\top \mathbf{u} + \mathbf{u} - \mathbf{v} \right)^\top \mathbf{Y} \|_2^z. \end{aligned}$$

Recall that for each $\mathbf{p} \in P$, there exists $\mathbf{q} \in Q$ such that $\mathbf{q} = \mathbf{W}^\top \mathbf{p}$ and

$$\begin{aligned} &\| \left(\mathbf{W}\mathbf{W}^\top (\mathbf{p} - \mathbf{u}) + \mathbf{u} - \mathbf{v} \right)^\top \mathbf{Y} \|_2^z = \\ &\quad \| \left(\mathbf{W}\mathbf{q} - \mathbf{W}\mathbf{W}^\top \mathbf{u} + \mathbf{u} - \mathbf{v} \right)^\top \mathbf{Y} \|_2^z. \end{aligned}$$

Since S is the set of vertices of $E(\mathbf{G}, \mathbf{c})$, we have by the definition of the John-Löwner ellipsoid that

$$\frac{1}{j}(E(\mathbf{G}, \mathbf{c}) - \mathbf{c}) + \mathbf{c} \subseteq \text{conv}(Q) \subseteq E(\mathbf{G}, \mathbf{c}).$$

Thus $S \subseteq \text{conv}(S) \subseteq \text{conv}(Q)$ and by Carathéodory's theorem, for each $\mathbf{s} \in S$, there exists a set $K_{\mathbf{s}}$ of at most $j+1$ points such that $\mathbf{s} \in \text{conv}(K_{\mathbf{s}})$. By Lemma 2.1,

$$\|\mathbf{s}\mathbf{W}^\top\mathbf{Y}\|_2^z \leq \max_{\mathbf{q}\in K_{\mathbf{s}}} \|q^\top\mathbf{W}^\top\mathbf{Y}\|_2^z.$$

We also have

$$\frac{1}{\sqrt{j}} \cdot \frac{E(\mathbf{G},\mathbf{c})-\mathbf{c}}{j} + \mathbf{c} \subseteq \text{conv}(S) \subseteq \frac{E(\mathbf{G},\mathbf{c})-\mathbf{c}}{j} + \mathbf{c}.$$

Therefore,

$$\text{conv}(S) \subseteq \text{conv}(Q) \subseteq E(\mathbf{G},\mathbf{c}) \subseteq j^{1.5}(\text{conv}(S)-\mathbf{c})+\mathbf{c}. \tag{2}$$

Thus for every $\mathbf{q} \in Q$, there exists $\mathbf{s} \in \text{conv}(S)$ and $\gamma \in [0,1]$ such that

$$\mathbf{q} = \gamma\mathbf{s} + (1-\gamma)(j^{1.5}(\mathbf{s}-\mathbf{c})+\mathbf{c}).$$

For $\mathbf{a} = \mathbf{u}^\top\mathbf{W}\mathbf{W}^\top\mathbf{Y} - \mathbf{u}^\top\mathbf{Y} - \mathbf{v}^\top\mathbf{Y}$, we then have

$$\|\mathbf{q}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z = \|(\gamma\mathbf{s}+(1-\gamma)(j^{1.5}(\mathbf{s}-\mathbf{c})+\mathbf{c})\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z.$$

Since $z \geq 1$, then $\|\cdot\|_2^z$ is a convex function. Thus by Jensen's inequality,

$$\|\mathbf{q}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z \leq \gamma\|\mathbf{s}\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z +$$
$$(1-\gamma)\|j^{1.5}\mathbf{s}^\top\mathbf{W}^\top\mathbf{Y} + (1-j^{1.5})\mathbf{c}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z.$$

Since $\mathbf{a} = j^{1.5}\mathbf{a} + (1-j^{1.5})\mathbf{a}$, then

$$\|j^{1.5}\mathbf{s}^\top\mathbf{W}^\top\mathbf{Y} + (1-j^{1.5})\mathbf{c}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z \leq$$
$$2^z j^{1.5z}\|\mathbf{s}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z + 2^z(j^{1.5}-1)^z\|\mathbf{c}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z.$$

Since $\mathbf{c} \in \text{conv}(S)$ by (2), then

$$\|\mathbf{c}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z \leq \max_{\mathbf{s}\in\text{conv}(S)} \|\mathbf{s}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z.$$

Since $j^{1.5z} + (j^{1.5}-1)^z \leq 2j^{1.5z}$, then we have that for every $\mathbf{q} \in Q$,

$$\|\mathbf{q}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z \leq 2^{z+1}j^{1.5z}\max_{\mathbf{s}\in S}\|\mathbf{s}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z.$$

Thus we have for every $\mathbf{s} \in S$,

$$\begin{aligned}\|\mathbf{s}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z &\leq \max_{\mathbf{q}\in K}\|\mathbf{q}^\top\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z \\ &\leq \max_{\mathbf{p}\in C}\|\mathbf{p}^\top\mathbf{W}\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z\end{aligned}$$

Because $\mathbf{a} = \mathbf{u}^\top\mathbf{W}\mathbf{W}^\top\mathbf{Y} - \mathbf{u}^\top\mathbf{Y} - \mathbf{v}^\top\mathbf{Y}$, we have

$$\begin{aligned}&\text{dist}(\mathbf{p}, H(\mathbf{X},\mathbf{v}))^z \\ &\leq 2^{z+1}j^{1.5z}\max_{\mathbf{p}\in C}\|\mathbf{p}^\top\mathbf{W}\mathbf{W}^\top\mathbf{Y}+\mathbf{a}\|_2^z \\ &= 2^{z+1}j^{1.5z}\max_{\mathbf{p}\in C}\|(\mathbf{W}\mathbf{W}^\top(\mathbf{p}-\mathbf{u}))^\top\mathbf{Y}+\mathbf{u}^\top\mathbf{Y}-\mathbf{v}^\top\mathbf{Y}\|_2^z.\end{aligned}$$

Since $(\mathbf{p}-\mathbf{u}) \in P$ and P lies within $H(\mathbf{W},\mathbf{u})$, then

$$\begin{aligned}&\text{dist}(\mathbf{p}, H(\mathbf{X},\mathbf{v}))^z \\ &\quad\leq 2^{z+1}j^{1.5z}\max_{\mathbf{p}\in C}\|\mathbf{p}^\top\mathbf{Y}-\mathbf{u}^\top\mathbf{Y}+\mathbf{u}^\top\mathbf{Y}-\mathbf{v}^\top\mathbf{Y}\|_2^z \\ &\quad= \max_{\mathbf{p}\in C}\|\mathbf{p}^\top\mathbf{Y}-\mathbf{v}^\top\mathbf{Y}\|_2^z \\ &\quad= 2^{z+1}j^{1.5z}\max_{\mathbf{p}\in C}\text{dist}(\mathbf{p}, H(\mathbf{X},\mathbf{v}))^z.\end{aligned}$$

□

A.3 Proof of Lemma 2.4

Proof. By Theorem 2.2, we have that $f(1) \leq 2(j+1)^2 \leq 8j^2$. Our construction has $j+1$ levels and each level partitions the data set into $\mathcal{O}(j\log(d\Delta))$ sets. For each of the sets, we construct an L_∞-coreset for $(k-1,j)$-projective clustering and

each of the points in the union of the coresets to be used in the point set $P[\mathbf{v}_0,\ldots,\mathbf{v}_{k+1}]$ for the next level. Thus we have

$$f(k) \leq (\mathcal{O}(j\log(d\Delta)) \cdot f(k-1))^{j+1},$$

so that by induction, $f(k) \leq (8j^3\log(d\Delta))^{\mathcal{O}(jk)}$. □

A.4 Proof of Lemma 2.5

We first require the following structural property about the convex hull of affine subspaces.

Lemma A.1 (Lemma 1 in (Edwards and Varadarajan, 2005))**.** *Let $\mathbf{v}_0,\ldots,\mathbf{v}_d \in \mathbb{R}^d$ be $d+1$ affinely independent vectors and for each $0 \leq i \leq d$, let A_i be the affine subspace spanned by $\mathbf{v}_0,\ldots,\mathbf{v}_i$. Let $\mathbf{w}_i$ be the projection of $\mathbf{v}_i$ onto A_i and let $\mathbf{u}_i = \mathbf{v}_i - \mathbf{w}_i$. Suppose we have $\operatorname{dist}(\mathbf{v}_j, A_i) \leq 2\|\mathbf{u}_i\|_2$ for every $0 \leq i \leq d$ and $j \geq i$. Then there exists an absolute constant c_d that only depends on d, so that the simplex $\operatorname{conv}(\mathbf{v}_0,\ldots,\mathbf{v}_d)$ contains a translation of the hyperrectangle $\{c_d(\alpha_1\mathbf{u}_1 + \ldots + \alpha_d\mathbf{u}_d : \alpha_i \in [0,1]\}$.*

Thus we achieve an L_∞ constant-factor approximation coreset for the integer (j,k)-projective clustering problem with size $(8j^3\log(d\Delta))^{\mathcal{O}(jk)}$:

Proof. Suppose D_k is covered by the k cylinders $S_1,\ldots,S_k$. Then we would like to show that P is covered by a constant-factor C-expansion of $S_1,\ldots,S_k$. Here a x-expansion of a cylinder S is the set $\{xp | p \in S\}$. We first induct on k and then j, noting that the base case $k=1$ is already handled by Theorem 2.2. We then fix $k \geq 2$ and induct on j, first considering stage 0, where we have some $\mathbf{v}_0$ and we define $K_{0,i} = \{\mathbf{p} \in P[\mathbf{v}_0] : 2^{i-1} \leq \operatorname{dist}(\mathbf{p},\mathbf{v}_0) \leq 2^i\}$ for $i \in [\ell]$, where $\ell = \mathcal{O}(\log(d\Delta))$. We then set $D_{0,i}$ to be the corresponding coreset for $K_{0,i}$ for the $(k-1,j)$-projective clustering problem. Let a denote the largest positive integer such that $S_k \cap D_{0,a} \neq \emptyset$, so that by the definition of a, we have that $\cup_{x=a+1}^{\ell} D_{0,x}$ is covered by $S_1,\ldots,S_{k-1}$. Since $D_{0,x}$ is a coreset for the $(k-1,j)$-projective clustering problem, then $\cup_{x=a+1}^{\ell} K_{0,x}$ is covered by a C-expansion of $S_1,\ldots,S_{k-1}$. For any point $\mathbf{v}_1$ in $S_k \cap D_{0,a}$, we enter stage 1 with $\mathbf{v}_0,\mathbf{v}_1$ and so it remains to prove that a C-expansion of $S_1,\ldots,S_k$ covers $P[\mathbf{v}_0,\mathbf{v}_1] = \cup_{x=0}^{a} K_{0,x}$.

For the inductive step, suppose we have fixed $\mathbf{v}_0,\ldots,\mathbf{v}_t$ and for each $i \in [0,t]$, let A_i denote the affine subspace spanned by $\mathbf{v}_0,\ldots,\mathbf{v}_i$, that is $A_i = \left\{\sum_{l=0}^{i}\alpha_l v_l \middle| \forall l \in [i]\, \alpha_i \in \mathbb{R}, \sum_{l=0}^{i}\alpha_l = 1\right\}$. Let $\mathbf{w}_i$ denote the projection of $\mathbf{v}_i$ on A_i and set $\mathbf{u}_i = \mathbf{v}_i - \mathbf{w}_i$. Then for every $\mathbf{p} \in P[\mathbf{v}_0,\ldots,\mathbf{v}_i] \cap A_i$, we have

$$\operatorname{dist}(\mathbf{p}, A_i) \leq 2\operatorname{dist}(\mathbf{v}_i, A_i).$$

Thus for $\mathbf{p} \in P[\mathbf{v}_0,\ldots,\mathbf{v}_t] \cap A_t$, we have that $\mathbf{p}$ is contained in the hyperrectangle

$$\mathcal{M} := \mathbf{v}_0 + \{\alpha_1\mathbf{u}_1 + \ldots + \alpha_t\mathbf{u}_t : \alpha_i \in [-2,2]\}.$$

By Lemma A.1, there exists a constant c_t such that $\operatorname{conv}(\mathbf{v}_0,\ldots,\mathbf{v}_t)$ contains a translation of the hyperrectangle

$$\mathcal{M}_1 := \{c_t(\alpha_1\mathbf{u}_1 + \ldots + \alpha_t\mathbf{u}_t) : \alpha_i \in [0,1]\}.$$

Since S_k covers $\mathbf{v}_0,\ldots,\mathbf{v}_t$, then $\mathcal{M}_1 \subset S_k$. Moreover, we have that for an absolute constant ξ, $\mathcal{M} \subset \xi \cdot \mathcal{M}_1$. Thus, a ξ-expansion of S_k covers $P[\mathbf{v}_0,\ldots,\mathbf{v}_t] \cap A_t$.

Let b denote the largest positive integer such that $S_k \cap D_{t,b} \neq \emptyset$, so that by the definition of b, we have that $\cup_{x=b+1}^{\ell} D_{t,x}$ is covered by $S_1,\ldots,S_{k-1}$. Since $D_{t,x}$ is a coreset for the $(k-1,j)$-projective clustering problem, then $\cup_{x=b+1}^{\ell} K_{t,x}$ is covered by a ξ-expansion of $S_1,\ldots,S_{k-1}$. For any point $\mathbf{v}_{t+1}$ in $S_k \cap D_{t,b}$, we enter stage $t+1$ with $\mathbf{v}_0,\ldots,\mathbf{v}_{t+1}$ and so then by induction, it holds that a ξ-expansion of $S_1,\ldots,S_k$ covers $P[\mathbf{v}_0,\ldots,\mathbf{v}_{t+1}] = \cup_{x=0}^{b} K_{t,x}$. □

A.5 Proof of Theorem 3.1

Proof. Let $H(\mathbf{X},\mathbf{v}) \in \mathcal{H}_j$. Then by Theorem 2.2, we have that

$$\max_{\mathbf{p}\in P} \operatorname{dist}(p, H(\mathbf{X},\mathbf{v}))^z \leq 2^{z+1} j^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(q, H(\mathbf{X},\mathbf{v}))^z.$$

Since f is a monotonically non-decreasing function, then

$$\begin{aligned}
&\max_{\mathbf{p}\in P} f(\operatorname{dist}(p, H(\mathbf{X},\mathbf{v}))^z) \\
&\quad = f\left(\max_{\mathbf{p}\in P} \operatorname{dist}(p, H(\mathbf{X},\mathbf{v}))^z\right) \\
&\quad \leq f\left(2^{z+1} j^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(q, H(\mathbf{X},\mathbf{v}))^z\right).
\end{aligned}$$

Since f is log-log Lipschitz, then

$$\begin{aligned}
&f\left(2^{z+1} j^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(q, H(\mathbf{X},\mathbf{v}))^z\right) \\
&\leq (2^{z+1} j^{1.5z})^{\rho} f\left(\max_{\mathbf{q}\in C} \operatorname{dist}(q, H(\mathbf{X},\mathbf{v}))^z\right) \\
&\leq (2^{z+1} j^{1.5z})^{\rho} \max_{\mathbf{q}\in C} f\left(\operatorname{dist}(q, H(\mathbf{X},\mathbf{v}))^z\right).
\end{aligned}$$

Hence, we have

$$\max_{\mathbf{p}\in P} f(\operatorname{dist}(p, H(\mathbf{X},\mathbf{v}))^z) \leq (2^{z+1} j^{1.5z})^{\rho} \max_{\mathbf{p}\in C} f(\operatorname{dist}(p, H(\mathbf{X},\mathbf{v}))^z)$$

as desired. □

A.6 Proof of Theorem 3.3

Proof. The coreset size follows the bound of Feldman et al. (2020) once the sensitivity and the shattering dimension upper bound are given to us. We actually follow the way of Lemma 3.1 of Varadarajan and Xiao (2012b) to give the sensitivity upper bound $s(p)$. The shattering dimension upper bound $\tilde{O}(djk)$ follows Corollary 34 of Feldman et al. (2020) □

B APPLICATIONS

In what follows, we will show that ℓ_∞-coreset can serve a family of functions including (but not bounded to) M-estimators.

B.1 L_∞ Coreset for Regression with Power-Bounded Loss Function – Proof of Lemma 3.2

Proof. Because the claim is trivially true for $\mathbf{w} = 0^d$, then it suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$ such that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix} \mathbf{p} \\ b(\mathbf{p}) \end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. We also define the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix} \mathbf{w} \\ -1 \end{bmatrix}$. By setting C to be the output of Coreset on $P' = \{\mathbf{p}' \mid \mathbf{p} \in P\}$, then by Theorem 3.1,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \\
&\quad \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z
\end{aligned}$$

Thus for $z = 1$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'| \leq 4(d+1)^{1.5} \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|.$$

Since Ψ_{Pow} is monotonically non-decreasing, then $\Psi_{Pow}(|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|)$ increases as $|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|$ increases. Moreover, we have $\Psi_{Pow}(y)/\Psi_{Pow}(x) \leq (y/x)^z$ for all $0 \leq x \leq y$. Therefore,

$$\max_{\mathbf{p}\in P} \Psi_{Pow}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq \max_{\mathbf{q}\in C} \Psi_{Pow}\left(4(d+1)^{1.5}|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right) \leq 4^z(d+1)^{1.5z} \max_{\mathbf{q}\in C} \Psi_{Pow}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right).$$

□

B.2 L_∞ Coreset for Cauchy Regression

Lemma B.1. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{Cau} denote the Cauchy loss function. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \,|\, \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \,|\, \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \Psi_{Cau}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 8(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{Cau}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Proof. We first observe that the claim is trivially true for $\mathbf{w} = 0^d$. Thus it suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$ such that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix} \mathbf{p} \\ b(\mathbf{p}) \end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. We also define the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix} \mathbf{w} \\ -1 \end{bmatrix}$. By running $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \,|\, \mathbf{p} \in P\}$ to obtain a coreset C, then we have by Theorem 3.1,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \\
&\quad \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z.
\end{aligned}$$

Thus for $z = 2$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'|^2 \leq 8(d+1)^3 \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|^2.$$

The Cauchy loss function is monotonically increasing, so that $\Psi_{Cau}(|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|)$ increases as $|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|$ increases. Thus for every $\mathbf{p} \in P$,

$$\begin{aligned}
&\Psi_{Cau}(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|) \\
&\quad = \Psi_{Cau}\left(\left|(\mathbf{p}')^\top \mathbf{w}'\right|\right) \\
&\quad = \frac{\lambda^2}{2} \log\left(1 + \left(\frac{\left|(\mathbf{p}')^\top \mathbf{w}'\right|}{\lambda}\right)^2\right) \\
&\quad \leq \max_{\mathbf{q}\in C} \frac{\lambda^2}{2} \log\left(1 + 8(d+1)^3 \left(\frac{\left|(\mathbf{q}')^\top \mathbf{w}'\right|}{\lambda}\right)^2\right),
\end{aligned}$$

where the inequality follows from the L_∞-coreset property above and the monotonicity of the Cauchy loss function. Thus by Bernoulli's inequality, we have

$$\begin{aligned}
&\Psi_{Cau}(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|) \\
&\quad \leq \max_{\mathbf{q}\in C} 8(d+1)^3 \cdot \frac{\lambda^2}{2} \log\left(1 + \left(\frac{\left|(\mathbf{q}')^\top \mathbf{w}'\right|}{\lambda}\right)^2\right) \\
&\quad = 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Cau}\left(\left|(\mathbf{q}')^\top \mathbf{w}'\right|\right) \\
&\quad = 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Cau}\left(\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})\right).
\end{aligned}$$

□

B.3 L_∞ Coreset for Welsch Regression

First, we will present the following as a stepping stone towards bounding the approximation error that our L_∞-coreset achieves in the context of Welsch regression problem.

Lemma B.2. *Let $a \geq 1$ be a positive real number. Then for every $x \in \mathbb{R}$,*

$$1 - e^{-a^2x^2} \leq a^2(1 - e^{-x^2}).$$

Proof. Since e^{-x^2} decreases as x^2 increases, then $a^2e^{-x^2} - e^{-a^2x^2}$ is a monotonically non-increasing function that achieves its maximum at $x = 0$. In particular, the value of $a^2e^{-x^2} - e^{-a^2x^2}$ at $x = 0$ is $a^2 - 1$, so that

$$a^2e^{-x^2} - e^{-a^2x^2} \leq a^2 - 1.$$

Thus from rearranging the terms, we have that

$$1 - e^{-a^2x^2} \leq a^2(1 - e^{-x^2}).$$

□

Lemma B.3. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{Wel} denote the Welsch loss function. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \,|\, \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \,|\, \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \Psi_{Wel}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 8(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{Wel}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Proof. We observe that the claim is trivially true for $\mathbf{w} = 0^d$, so that it suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$, so that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$, and for each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix} \mathbf{p} \\ b(\mathbf{p}) \end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. Let $\mathbf{w}'$ denote the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix} \mathbf{w} \\ -1 \end{bmatrix}$. By Theorem 3.1, we have that the output C of $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \,|\, \mathbf{p} \in P\}$ satisfies

$$\begin{aligned} &\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \\ &\quad \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z. \end{aligned}$$

Thus for $z = 2$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'|^2 \leq 8(d+1)^3 \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|^2.$$

The Welsch loss function is monotonically increasing, so that $\Psi_{Wel}(|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|)$ increases as $|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|$ increases. Hence, for every $\mathbf{p} \in P$,

$$\begin{aligned} &\Psi_{Wel}(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|) \\ &\quad = \Psi_{Wel}\left(\left|(\mathbf{p}')^\top \mathbf{w}'\right|\right) \\ &\quad = \frac{\lambda^2}{2}\left(1 - e^{-\left(\frac{\left|(\mathbf{p}')^\top \mathbf{w}'\right|}{\lambda}\right)^2}\right) \\ &\quad \leq \max_{\mathbf{q}\in C} \frac{\lambda^2}{2}\left(1 - e^{-\left(\frac{8(d+1)^3\left|(\mathbf{q}')^\top \mathbf{w}'\right|}{\lambda}\right)^2}\right), \end{aligned}$$

where the inequality results from the L_∞-coreset property above and the monotonicity of the Welsch loss function. By Lemma B.2,

$$\Psi_{Wel}(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|)$$

$$\leq \max_{\mathbf{q}\in C} 8(d+1)^3 \cdot \frac{\lambda^2}{2}\left(1 - e^{-\left(\frac{\left|(\mathbf{q}')^\top \mathbf{w}'\right|}{\lambda}\right)^2}\right)$$
$$= 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Wel}\left(\left|(\mathbf{q}')^\top \mathbf{w}'\right|\right)$$
$$= 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Wel}\left(\left|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})\right|\right).$$

□

B.4 L_∞ coreset for Huber regression

Lemma B.4. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{Hub} denote the Huber loss function. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \,|\, \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \,|\, \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \Psi_{Hub}\left(\left|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})\right|\right) \leq 16(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{Hub}\left(\left|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})\right|\right)$.*

Proof. The claim is trivially true for $\mathbf{w} = 0^d$; it remains to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$, so that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, we use $\mathbf{p}'$ to denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$, $\mathbf{p}' := \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix}\mathbf{p} \\ b(\mathbf{p})\end{bmatrix}$. Similarly, we use $\mathbf{w}'$ to denote the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix}\mathbf{w} \\ -1\end{bmatrix}$. By Theorem 3.1, we have that the output C of $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \,|\, \mathbf{p} \in P\}$ satisfies

$$\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z.$$

Thus for $z = 2$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'|^2 \leq 8(d+1)^3 \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|^2. \tag{3}$$

We now consider casework for whether $|(\mathbf{p}')^\top \mathbf{w}'| \leq \lambda$ or $|(\mathbf{p}')^\top \mathbf{w}'| > \lambda$.

If $|(\mathbf{p}')^\top \mathbf{w}'| \leq \lambda$, then we immediately have from (3) and the fact that $C \subseteq P$ that

$$\Psi_{Hub}\left(\left|(\mathbf{p}')^\top \mathbf{w}'\right|\right) \leq 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Hub}\left(\left|(\mathbf{q}')^\top \mathbf{w}'\right|\right).$$

On the other hand if $|(\mathbf{p}')^\top \mathbf{w}'| > \lambda$, we further consider casework for whether $\max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'| \leq \lambda$ or $\max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'| > \lambda$. If $\max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'| > \lambda$, then we again have from (3) and the fact that $C \subseteq P$ that

$$\Psi_{Hub}\left(\left|(\mathbf{p}')^\top \mathbf{w}'\right|\right) \leq 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Hub}\left(\left|(\mathbf{q}')^\top \mathbf{w}'\right|\right).$$

Finally, if $|(\mathbf{p}')^\top \mathbf{w}'| > \lambda$ but $\max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'| \leq \lambda$, then we observe that from (3) and the assumption that $|(\mathbf{p}')^\top \mathbf{w}'| > \lambda$, we have

$$\frac{\lambda}{\sqrt{8}(d+1)^{1.5}} \leq \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|.$$

Thus if $|(\mathbf{p}')^\top \mathbf{w}'| > \lambda$, then

$$\Psi_{Hub}(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|) = \Psi_{Hub}\left(\left|(\mathbf{p}')^\top \mathbf{w}'\right|\right) = \lambda\left(\left|(\mathbf{p}')^\top \mathbf{w}'\right| - \frac{\lambda}{2}\right) \leq \lambda\left(\left|(\mathbf{p}')^\top \mathbf{w}'\right|\right) \leq \sqrt{8}\lambda(d+1)^{1.5}\left(\max_{\mathbf{q}\in C}\left|(\mathbf{q}')^\top \mathbf{w}'\right|\right),$$

where the last inequality results from the L_∞-coreset property in (3) above. Therefore,

$$\begin{aligned}
&\Psi_{Hub}(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|) \\
&\quad \leq \frac{\lambda}{\sqrt{8}(d+1)^{1.5}} \cdot 8(d+1)^3 \left(\max_{\mathbf{q}\in C} \left|(\mathbf{q}')^\top \mathbf{w}'\right| \right) \\
&\quad \leq 8(d+1)^3 \left(\max_{\mathbf{q}\in C} \left|(\mathbf{q}')^\top \mathbf{w}'\right|^2 \right) \\
&\quad \leq 16(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Hub}(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|).
\end{aligned}$$

Thus in all cases, we have

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \Psi_{Hub}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\
&\quad \leq 16(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{Hub}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right).
\end{aligned}$$

□

B.5 L_∞ coreset for Geman-McClure regression

Lemma B.5. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{GM} denote the Geman-McClure loss function. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \,|\, \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \,|\, \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \Psi_{GM}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 8(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{GM}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Proof. Note that the claim is trivially true for $\mathbf{w} = 0^d$, so it therefore suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$ such that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix} \mathbf{p} \\ b(\mathbf{p}) \end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. We also define the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix} \mathbf{w} \\ -1 \end{bmatrix}$. By setting C to be the output of $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \,|\, \mathbf{p} \in P\}$, then by Theorem 3.1,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \\
&\quad \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z.
\end{aligned}$$

Thus for $z = 2$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'|^2 \leq 8(d+1)^3 \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|^2.$$

The Geman-McClure loss function is monotonically increasing, so that $\Psi_{GM}(|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|)$ increases as $|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|$ increases. Therefore,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \Psi_{GM}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\
&\qquad = \max_{\mathbf{p}\in P} \frac{|(\mathbf{p}')^\top \mathbf{w}|^2}{2 + 2|(\mathbf{p}')^\top \mathbf{w}|^2)} \\
&\qquad \leq \max_{\mathbf{q}\in C} \frac{8(d+1)^3 |(\mathbf{q}')^\top \mathbf{w}|^2}{2 + 2|(\mathbf{q}')^\top \mathbf{w}|^2} \\
&\qquad = 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{GM}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right),
\end{aligned}$$

where the inequality results from the L_∞-coreset property of Theorem 3.1 and the fact that $C \subseteq P$. □

B.6 L_∞ Coreset for Regression with Concave Loss Function

We first recall the following property of concave functions:

Lemma B.6. *Let $f : \mathbb{R} \to \mathbb{R}$ be a concave function with $f(0) = 0$. Then for any $x \leq y$, we have $\frac{f(x)}{x} \geq \frac{f(y)}{y}$.*

Using Lemma B.6, we obtain an L_∞ coreset for regression for any non-decreasing concave loss function Ψ_{Con} satisfying $\Psi_{Con}(0) = 0$.

We obtain an L_∞ coreset for regression for any non-decreasing concave loss function Ψ_{Con} satisfying $\Psi_{Con}(0) = 0$.

Lemma B.7. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{Con} denote any non-decreasing concave loss function with $\Psi_{Con}(0) = 0$. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \mid \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \mid \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p} \in P} \Psi_{Con}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 4(d+1)^{1.5} \cdot \max_{\mathbf{q} \in C} \Psi_{Con}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Proof. The claim is trivially true for $\mathbf{w} = 0^d$, so it suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$ such that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix} \mathbf{p} \\ b(\mathbf{p}) \end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. We also define the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix} \mathbf{w} \\ -1 \end{bmatrix}$. By setting C to be the output of $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \mid \mathbf{p} \in P\}$, then by Theorem 3.1,

$$\begin{aligned} &\max_{\mathbf{p} \in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \\ &\quad \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q} \in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z. \end{aligned}$$

Thus for $z = 1$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'| \leq 4(d+1)^{1.5} \max_{\mathbf{q} \in C} |(\mathbf{q}')^\top \mathbf{w}'|.$$

Since Ψ_{Con} is monotonically non-decreasing, then $\Psi_{Con}(|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p}))$ increases as $|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|$ increases. Thus by Lemma B.6,

$$\begin{aligned} &\max_{\mathbf{p} \in P} \Psi_{Con}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\ &\qquad \leq \max_{\mathbf{q} \in C} \Psi_{Con}\left(4(d+1)^{1.5} |\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right) \\ &\qquad \leq 4(d+1)^{1.5} \max_{\mathbf{q} \in C} \Psi_{Con}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right). \end{aligned}$$

□

B.7 L_∞ Coreset for Tukey Regression

Lemma B.8. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{Tuk} denote the Tukey loss function. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \mid \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \mid \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p} \in P} \Psi_{Tuk}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 8(d+1)^3 \cdot \max_{\mathbf{q} \in C} \Psi_{Tuk}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Proof. We first observe that the claim is trivially true for $\mathbf{w} = 0^d$, so that it suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$ such that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix} \mathbf{p} \\ b(\mathbf{p}) \end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. We define the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix} \mathbf{w} \\ -1 \end{bmatrix}$. By setting C to be the output of $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \mid \mathbf{p} \in P\}$, then by Theorem 3.1,

$$\max_{\mathbf{p} \in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z$$

$$\leq 2^{z+1}(d+1)^{1.5z}\max_{\mathbf{q}\in C}\operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z.$$

Thus for $z=2$, we have for every $\mathbf{p}\in P$,

$$|(\mathbf{p}')^\top\mathbf{w}'|^2 \leq 8(d+1)^3\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'|^2.$$

We first note that if $\frac{|(\mathbf{p}')^\top\mathbf{w}'|}{\sqrt{8}(d+1)^{1.5}} \geq \lambda$, then we trivially have $\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'|^2 \geq \lambda^2$ so that $\max_{\mathbf{q}\in C}\Psi_{Tuk}\left(|\mathbf{q}^\top\mathbf{w}-b(\mathbf{q})|\right) = \frac{\lambda^2}{6} \geq \Psi_{Tuk}(x)$ for all x. Thus, we would have

$$\max_{\mathbf{p}\in P}\Psi_{Tuk}\left(|\mathbf{p}^\top\mathbf{w}-b(\mathbf{p})|\right) \leq \max_{\mathbf{q}\in C}\Psi_{Tuk}\left(|\mathbf{q}^\top\mathbf{w}-b(\mathbf{q})|\right),$$

as desired. Hence, we assume $\frac{|(\mathbf{p}')^\top\mathbf{w}'|}{\sqrt{8}(d+1)^{1.5}} < \lambda$ and consider casework for whether $|(\mathbf{p}')^\top\mathbf{w}'| \leq \lambda$ or $|(\mathbf{p}')^\top\mathbf{w}'| > \lambda$.

If $|(\mathbf{p}')^\top\mathbf{w}'| \leq \lambda$, then since the Tukey loss function is monotonically increasing, we have

$$\begin{aligned}
&\Psi_{Tuk}\left(|(\mathbf{p}')^\top\mathbf{w}'|\right)\\
&\quad= \frac{\lambda^2}{6}\left(1-\left(1-\frac{|(\mathbf{p}')^\top\mathbf{w}'|^2}{\lambda^2}\right)^3\right)\\
&\quad\leq \max_{\mathbf{q}\in C}\frac{\lambda^2}{6}\left(1-\left(1-\frac{8(d+1)^3|(\mathbf{q}')^\top\mathbf{w}'|^2}{\lambda^2}\right)^3\right)
\end{aligned}$$

Unfortunately, the Tukey loss function is not concave, so we cannot directly apply Lemma B.6. However, if we define the function $f(x) := \frac{\lambda^2}{6}\left(1-\left(1-\frac{x}{\lambda^2}\right)^3\right)$, then we have

$$\frac{d^2f}{dx^2} = \frac{x-\lambda^2}{\lambda^4},$$

which is non-positive for all $x \leq \lambda^2$. Thus by Lemma B.6, we have for all $0 \leq x \leq y \leq \lambda^2$ that $\frac{f(x)}{x} \geq \frac{f(y)}{y}$. Since $f(x^2) = \Psi_{Tuk}(x)$, then we have for all $0 \leq x \leq y \leq \lambda$ that $\frac{\Psi_{Tuk}(x)}{x^2} \geq \frac{\Psi_{Tuk}(y)}{y^2}$. Hence by the assumption that $\frac{|(\mathbf{p}')^\top\mathbf{w}'|}{\sqrt{8}(d+1)^{1.5}} < \lambda$,

$$\begin{aligned}
&\Psi_{Tuk}\left(|(\mathbf{p}')^\top\mathbf{w}'|\right)\\
&\quad\leq 8(d+1)^3\max_{\mathbf{q}\in C}\frac{\lambda^2}{6}\left(1-\left(1-\frac{|(\mathbf{q}')^\top\mathbf{w}'|^2}{\lambda^2}\right)^3\right)\\
&\quad\leq 8(d+1)^3\max_{\mathbf{q}\in C}\Psi_{Tuk}\left(|(\mathbf{q}')^\top\mathbf{w}'|\right)\\
&\quad= 8(d+1)^3\max_{\mathbf{q}\in C}\Psi_{Tuk}\left(|\mathbf{q}^\top\mathbf{w}-b(\mathbf{q})|\right).
\end{aligned}$$

On the other hand, if $|(\mathbf{p}')^\top\mathbf{w}'| > \lambda$, then we further consider casework on whether $\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'| > \lambda$ or $\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'| \leq \lambda$. If $\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'| > \lambda$, then we immediately have

$$\begin{aligned}
\Psi_{Tuk}\left(|(\mathbf{p}')^\top\mathbf{w}'|\right) = \frac{\lambda^2}{6} &= \max_{\mathbf{q}\in C}\Psi_{Tuk}\left(|(\mathbf{q}')^\top\mathbf{w}'|\right)\\
&= \max_{\mathbf{q}\in C}\Psi_{Tuk}\left(|\mathbf{q}^\top\mathbf{w}-b(\mathbf{q})|\right).
\end{aligned}$$

Otherwise, suppose $\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'| \leq \lambda$. Note that $|(\mathbf{p}')^\top\mathbf{w}'|^2 \leq 8(d+1)^3\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'|^2$ implies $\max_{\mathbf{q}\in C}|(\mathbf{q}')^\top\mathbf{w}'| > \frac{\lambda}{\sqrt{8}(d+1)^{1.5}}$. Since the Tukey loss function is monotonically increasing, then

$$\max_{\mathbf{q}\in C}\Psi_{Tuk}\left(|(\mathbf{q}')^\top\mathbf{w}'|\right) \geq \Psi_{Tuk}\left(\frac{\lambda}{\sqrt{8}(d+1)^{1.5}}\right).$$

Because $\max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'| \leq \lambda$, then we can again apply the relationship $\frac{\Psi_{Tuk}(x)}{x^2} \geq \frac{\Psi_{Tuk}(y)}{y^2}$ for all $0 \leq x \leq y \leq \lambda$, so that

$$\max_{\mathbf{q}\in C} \Psi_{Tuk}\left(|(\mathbf{q}')^\top \mathbf{w}'|\right) \geq \frac{1}{8(d+1)^3} \Psi_{Tuk}(\lambda).$$

Hence,

$$\begin{aligned}\Psi_{Tuk}\left(|(\mathbf{p}')^\top \mathbf{w}'|\right) &= \Psi_{Tuk}(\lambda) \\ &\leq 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Tuk}\left(|(\mathbf{q}')^\top \mathbf{w}'|\right).\end{aligned}$$

Therefore across all cases, we have

$$\begin{aligned}&\max_{\mathbf{p}\in P} \Psi_{Tuk}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\ &\qquad \leq 8(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{Tuk}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right).\end{aligned}$$

□

B.8 L_∞ Coreset for $L_1 - L_2$ Regression

Lemma B.9. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{LL} denote the $L_1 - L_2$ loss function. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \,|\, \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \,|\, \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \Psi_{LL}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 8(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{LL}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Proof. We first observe that the claim is trivially true for $\mathbf{w} = 0^d$. Therefore, it suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$ such that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix}\mathbf{p} \\ b(\mathbf{p})\end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. We also define the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix}\mathbf{w} \\ -1\end{bmatrix}$. By setting C to be the output of $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \,|\, \mathbf{p} \in P\}$, then by Theorem 3.1,

$$\begin{aligned}&\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \\ &\qquad \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z.\end{aligned}$$

Thus for $z = 2$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'|^2 \leq 8(d+1)^3 \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|^2.$$

The $L_1 - L_2$ loss function is monotonically increasing, so that $\Psi_{LL}(|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p}))$ increases as $|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|$ increases. Therefore,

$$\begin{aligned}&\max_{\mathbf{p}\in P} \Psi_{LL}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\ &\qquad = \max_{\mathbf{p}\in P} 2\left(\sqrt{1 + \frac{|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|^2}{2}} - 1\right) \\ &\qquad \leq \max_{\mathbf{q}\in C} 2\left(\sqrt{1 + \frac{8(d+1)^3 |\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|^2}{2}} - 1\right).\end{aligned}$$

Since the $L_1 - L_2$ loss function is not concave, so we cannot directly apply Lemma B.6. Fortunately, if we define the function $f(x) := 2\left(\sqrt{1 + \frac{x}{2}} - 1\right)$, then we have

$$\frac{d^2 f}{dx^2} = -\frac{2}{16\left(\frac{x}{2} + 1\right)^{3/2}},$$

which is non-positive for all $x \geq 0$. Thus by Lemma B.6, we have for all $0 \leq x \leq y$ that $\frac{f(x)}{x} \geq \frac{f(y)}{y}$. Since $f(x^2) = \Psi_{LL}(x)$, then we have for all $0 \leq x \leq y$ that $\frac{\Psi_{LL}(x)}{x^2} \geq \frac{\Psi_{LL}(y)}{y^2}$. Thus,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \Psi_{LL}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\
&\quad \leq 8(d+1)^3 \max_{\mathbf{q}\in C} 2\left(\sqrt{1 + \frac{|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|^2}{2}} - 1\right) \\
&\quad = 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{LL}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right).
\end{aligned}$$

□

B.9 L_∞ Coreset for Fair Regression

Lemma B.10. *Let $P \subseteq \mathbb{R}^d$ be a set of n points, $b : P \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and let Ψ_{Fair} denote the Fair loss function. Let $P' = \{\mathbf{p} \circ b(\mathbf{p}) \,|\, \mathbf{p} \in P\}$, where $\circ$ denotes vertical concatenation. Let C' be the output of a call to $L_\infty - \textsc{Coreset}(P', d)$ and let $C \subseteq P$ so that $C' = \{\mathbf{q} \circ b(\mathbf{q}) \,|\, \mathbf{q} \in C\}$. Then for every $\mathbf{w} \in \mathbb{R}^d$, $\max_{\mathbf{p}\in P} \Psi_{Fair}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \leq 8(d+1)^3 \cdot \max_{\mathbf{q}\in C} \Psi_{Fair}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right)$.*

Proof. Since the claim is trivially true for $\mathbf{w} = 0^d$, then it suffices to consider nonzero $\mathbf{w} \in \mathbb{R}^d$. Let $Y \in \mathcal{H}_{d-1}$ such that $\mathbf{w}^\top \mathbf{Y} = 0^{d-1}$ and $\mathbf{Y}^\top \mathbf{w} = 0^d$. For each $\mathbf{p} \in P$, let $\mathbf{p}' = \mathbf{p} \circ b(\mathbf{p}) = \begin{bmatrix} \mathbf{p} \\ b(\mathbf{p}) \end{bmatrix}$ denote the vertical concatenation of $\mathbf{p}$ with $b(\mathbf{p})$. We also define the vertical concatenation $\mathbf{w}' = \mathbf{w} \circ (-1) = \begin{bmatrix} \mathbf{w} \\ -1 \end{bmatrix}$. By setting C to be the output of $\textsc{Coreset}$ on $P' = \{\mathbf{p}' \,|\, \mathbf{p} \in P\}$, then by Theorem 3.1,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \operatorname{dist}(\mathbf{p}', H(\mathbf{w}', 0^{d+1}))^z \\
&\qquad \leq 2^{z+1}(d+1)^{1.5z} \max_{\mathbf{q}\in C} \operatorname{dist}(\mathbf{q}', H(\mathbf{w}', 0^{d+1}))^z.
\end{aligned}$$

Thus for $z = 2$, we have for every $\mathbf{p} \in P$,

$$|(\mathbf{p}')^\top \mathbf{w}'|^2 \leq 8(d+1)^3 \max_{\mathbf{q}\in C} |(\mathbf{q}')^\top \mathbf{w}'|^2.$$

The Fair loss function is monotonically increasing, so that $\Psi_{Fair}(|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p}))$ increases as $|\mathbf{p}^\top \mathbf{x} - b(\mathbf{p})|$ increases. Therefore,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \Psi_{Fair}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\
&\qquad \leq \max_{\mathbf{q}\in C} \Psi_{Fair}\left(\sqrt{8}(d+1)^{1.5}|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right).
\end{aligned}$$

The Fair loss function is not concave, so we cannot directly apply Lemma B.6. However, if we define the function $f(x) := \lambda\sqrt{|x|} - \lambda^2 \ln\left(1 + \frac{\sqrt{|x|}}{\lambda}\right)$, then we have

$$\frac{d^2 f}{dx^2} = -\frac{\lambda}{4\sqrt{x}(\lambda + \sqrt{x})^2},$$

which is non-positive for all $x \geq 0$. Thus by Lemma B.6, we have for all $0 \leq x \leq y$ that $\frac{f(x)}{x} \geq \frac{f(y)}{y}$. Since $f(x^2) = \Psi_{Fair}(x)$, then we have for all $0 \leq x \leq y$ that $\frac{\Psi_{Fair}(x)}{x^2} \geq \frac{\Psi_{Fair}(y)}{y^2}$. Thus,

$$\begin{aligned}
&\max_{\mathbf{p}\in P} \Psi_{Fair}\left(|\mathbf{p}^\top \mathbf{w} - b(\mathbf{p})|\right) \\
&\qquad \leq 8(d+1)^3 \max_{\mathbf{q}\in C} \Psi_{Fair}\left(|\mathbf{q}^\top \mathbf{w} - b(\mathbf{q})|\right).
\end{aligned}$$

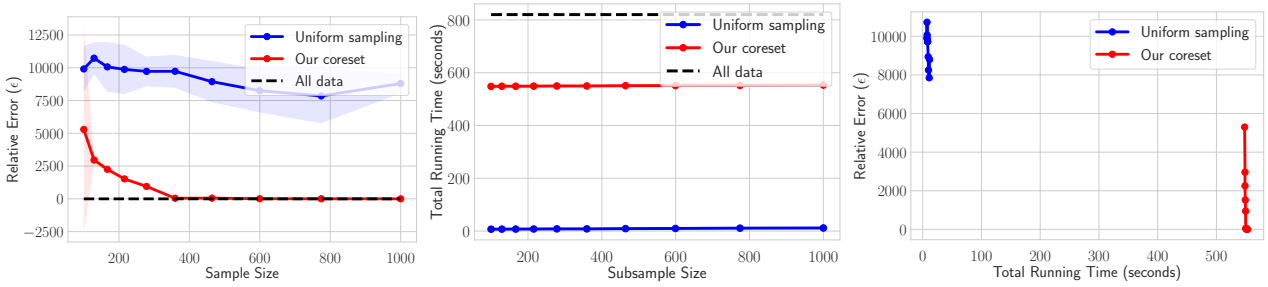
□

C EXPERIMENTS

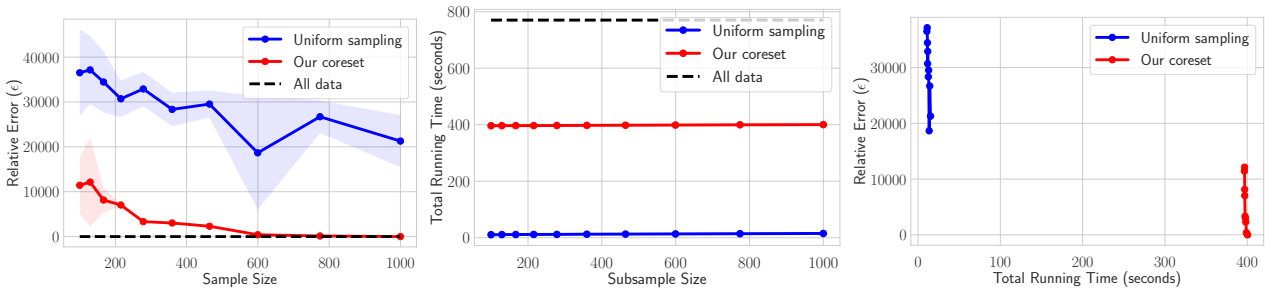
In this section, we carry additional experimental results evaluating our coreset against uniform sampling on real-world datasets, with respect to the projective clustering problem and its variants.

Table 3: **Summary of our results:** Our coreset construction was applied on various application of projective clustering, of which were robust regression as well as robust subspace clustering

Problem type	Loss function	k	j	Dataset	Figure
Robust $(2,2)$-projective clustering	$L_1 - L_2$	2	2	(iii)	3a
Robust $(2,2)$-projective clustering	Huber	2	3	(iii)	3b



(a)



(b)

Fig. 3: Our experimental results: evaluating the efficacy of our coreset against uniform sampling.