Thesis for the degree
Master of Science

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Matan Danos

מאת
מתן דנוס

קבוצות ליבה עבור אשכול באמצעות דגימה אחידה ובעיית איגוד דירוגים מוכללת

Coresets for Clustering by Uniform Sampling and Generalized Rank Aggregation

Advisors:
Prof. Robert Krauthgamer

Prof. Shaofeng Jiang
(Peking University)

מנחים:
פרופ' רוברט קראוטגמר

פרופ' שאופנג ג'יאנג
(אוניברסיטת פקינג)

November 2021

כסלו ה'תשפ"ב

Abstract

The *k-Median* clustering problem of a point set A asks to find a set of k center points that minimizes the sum of distances to the points in A. A *coreset* for a k-median instance is a compressed set of data points that succinctly represents the input and is often used to improve the efficiency of algorithms in terms of space and running time. We use two definitions: a *strong coreset* preserves the cost of every center set compared to the input instance, and a *weak coreset* preserves the cost only for certain center sets (optimal for the coreset).

The first part of the thesis investigates the construction of a weak coreset by the simple and natural method of *uniform sampling*. In Euclidean space $\mathbb{R}^d$, it is known [MS18] that a large enough uniform sample from the input can serve as a weak coreset for the 1-median problem under the ℓ_2 distance. We improve this bound, which depends on the dimension d, and obtain a weak coreset for 1-median by a uniform sample whose size is dimension-*independent*, under both ℓ_1 and ℓ_2 distances. We then construct a weak coreset for 2-median, again by uniform sampling, for inputs that are "well-separated" [ABS12, MOP04]. In these inputs, the optimal clusters are far away and approximately of the same size, in which case we can apply our analysis for 1-median to each optimal cluster separately.

The rank-aggregation problem can be viewed as 1-median over the metric space of permutations equipped with Kendall-tau distance [Ken38, Kem59]. We study its generalization to a k-median problem and design for it a strong coreset by applying results from [JKS16, BJKW21]. We then use our coreset to obtain a PTAS for two special cases, when k is small and when the input is "well-separated". In the former, our PTAS computes a coreset and then chooses the best performing medians among all possible partitions of the coreset. This method has running time exponential in k. In the latter case, when the input satisfies a separation property proposed in [ABS10], we use a coreset and a modification of an algorithm from [ABS10] (for a slightly different setting) to find a $(1+\varepsilon)$-factor approximate solution in polynomial time.

Acknowledgements

First and foremost, I wish to express my deepest appreciation to Prof. Robert Krauthgamer. His expertise, insights, and guidance were fundamental for the success of this thesis. Second, great gratitude is to Prof. Shaofeng Jiang for fruitful collaboration and knowledge contribution that assisted in every step of the way.

Many thanks also to friends from Weizmann. Even during a time of a pandemic, everyone kept the atmosphere fun and welcoming. It was wonderful learning side by side with you, from brainstorming to funny lunchtime on the lawn. Last but not least, I am incredibly thankful for my family for always being there when I needed them. Especially to my partner Shay, that made it all possible with her truthful and honest advice and never-ending support.

Contents

Chapter 1

Introduction

Clustering is a ubiquitous and essential tool in the field of data analysis and unsupervised learning that is well-known to be computationally hard to solve optimally. In the problem of k-clustering, the goal is to partition the dataset into k subsets (called clusters), where the items in each cluster have similar characteristics, often captured by a distance measure. This thesis focuses on a special case of k-clustering, called *metric k-Median*, that is defined by specifying a metric space $M = (V, D)$ and a point set $A \subseteq V$ of size n, with the goal to optimize

$$\min_{\substack{\mathcal{C} \subset V \\ |\mathcal{C}|=k}} \sum_{x \in A} \min_{c \in \mathcal{C}} D(x, c).$$

An optimal solution, i.e., an optimal set of k centers for A, which need not be unique, is denoted by $\mathcal{C}^*(A)$. The objective value of a set A with respect to a set of centers $\mathcal{C}$ is denoted by $\text{cost}(A, \mathcal{C})$, and the optimal objective for A by $\mathsf{OPT}(A, k) := \text{cost}(A, \mathcal{C}^*(A))$. It is emphasized, that in center-based k-clustering, a solution with k center points naturally induces a k-partition of A, formed by grouping points that have the same nearest center in $\mathcal{C}$.

In the last few decades, huge amounts of data are regularly collected and stored ("big data"), and thus efficient methods to find approximate solutions and store the input are increasingly in the spotlight. Approximation algorithms for the k-median problem are abundant and it is evident that the complexity of the problem exacerbates when the size of the data set A gets extremely large, in both number of data points and possibly their dimension. See the related work (Section 1.2) for more details. Naturally, when storing a small portion of the dataset for later use, one hopes that it can be used to provide a similar result as if one used the input set. This property is usually compromised with the simplicity of the reduction method, and we will strive to prove that uniform sampling, which is a very simple method, is effective in this regard.

A *coreset construction* is a data-reduction (compression) method for clustering problems, that has emerged in the last two decades, and since became prevalent in theoretical computer science and machine learning [AHPV04, HPM04, Che09, FL11, SW18, BJKW21, Fel20, CSS21] (see Section 1.2 for a full picture). Intuitively, one can view a coreset S as a weighted set of points, usually a subset of the input set A, which correctly captures the structure of the input. The usual definition of a coreset for clustering problems, also known as a strong coreset, requires that *every* solution will have a similar cost when evaluated on both the coreset and the input.

Definition 1.0.1 (Strong Coreset)**.** Let A be a set of n points in some metric space (V, D). A weighted set $S \subseteq A$ is called a *strong* (ε, k)*-coreset* of A if

$$\forall \mathcal{C} \subseteq V, |\mathcal{C}| = k, \quad \text{cost}_w(S, \mathcal{C}) \in (1 \pm \varepsilon)\,\text{cost}(A, \mathcal{C})$$

where $\text{cost}_w(S, \mathcal{C}) := \sum_{x \in S} w(x) \min_{c \in \mathcal{C}} D(x, c)$.

In contrast, one can have a weaker definition, known as a weak coreset, that only preserves the cost of optimal k centers of S.

Definition 1.0.2 (Weak Coreset)**.** Let A be a set of n points in some metric space (V, D). A set $S \subseteq A$ is called a *weak* (ε, k)*-coreset* of A if for every optimal k-median solution for S, denoted by $\mathcal{C}^*(S)$, it holds that

$$\text{cost}(A, \mathcal{C}^*(S)) \in (1 \pm \varepsilon)\mathsf{OPT}(A, k).$$

Remark. This definition of a weak coreset is taken from [MS18]. However, there are multiple definitions in the literature, generally capturing similar ideas [FMS07, FL11, Phi16, HV20].

For the reasoning described above, the most fundamental performance measure for coreset design is its size, seeking to minimize it. The primary framework for constructing small-size strong coresets, was introduced by [LS10, FL11, FSS20], and is based on the importance sampling technique. Intuitively, the algorithm performs independent sampling of the input according to a distribution that gives higher probability to points that might contribute more to the cost.

In the main part of the thesis we strive to understand the power of uniform sampling for coreset construction and ask, *whether uniform sampling produces weak coresets for clustering.* Concretely, we aim to find tight size bounds for a weak $(\varepsilon, 1)$-coreset that is constructed by uniform sampling in both ℓ_1 and ℓ_2 settings and improve the dependency in the dimension for both cases [MS18]. We emphasize that the motivation for using uniform sampling lies mainly in its simplicity, both theoretically and in practice. Theoretically, uniform sampling is data-oblivious, simple to describe, and its power for coreset constructions is not so well understood. In practice, it is very easy to implement uniform sampling.

The above question is phrased for weak coresets, since constructing a (small) strong coreset using only uniform sampling is impossible even for the one-dimensional real line and $k = 1$. For example, consider an input with n points, where two points are located at opposite sides, that are very far from the rest of the points in A. To preserve the cost of *every* candidate center, one must sample these two points, and if one uses uniform sampling, the number of samples must be $\Omega(n)$ [MS18]. Uniform sampling for clustering was examined in [MOP01, CS04]. They generate a uniform sample and then apply a known metric k-median approximation algorithm on the sample. This method achieves a constant factor approximation in sublinear time. However, they assumes a finite metric space with bounded diameter, and the sample size depends on this value. Additionally, there are approximation algorithms that use uniform sampling in certain regions of space. For instance, Chen's algorithm [Che09] uses uniform sampling in rings around approximate centers that are computed beforehand. Nevertheless, this can be seen as a precursor to the importance sampling technique, and in this work we avoid the step of pre-computing approximate centers, and just sample uniformly from the entire input set.

It is still unknown whether uniform sampling on its own can yield a weak (ε, k)-coreset for $k > 1$, hence, to ease the analysis, we assume henceforth that the input satisfies an additional property of "well-separated" optimal clusters. This assumption is a common relaxation approach since it allows for simpler and better clustering algorithms, see surveys by [Ben15, BR14]. Intuitively, "well-separated" instance is structured such that each optimal cluster has small radius compared to the inter-cluster distances. Therefore we ask, *can one construct a smaller coreset for "well-separated" instance, by using a uniform sample?* This notion will be formalized using definitions proposed by Awasthi, Blum, and Sheffet [ABS12, ABS10] (to be formally presented in the next section). Note that there are several other competing definitions that capture similar separability ideas. An overview and references are found in the related work section.

A secondary, somewhat independent, goal of this thesis is to present an application of strong coresets to the metric space of permutations over N items, equipped with the Kendall-tau distance, denoted by $(\mathcal{P}_N, K_\tau)$. A permutation $\sigma \in \mathcal{P}_N$ can be viewed as a ranking, which is bijection $\sigma : [N] \to [N]$, where $\sigma[i]$ denotes the location (or rank) of alternative $i \in [N]$. The Kendall-tau distance measure was introduced by Kendall [Ken38] as a measure of rank correlation, and was later generalized into a distance by Kemeny [Kem59]. It is defined by counting the number of pairwise disagreements between the permutations of every two alternatives (also known as the bubble-sort distance).

Definition 1.0.3 (Kendall-tau Distance)**.** The *Kendall-tau distance* between two rankings $\sigma_1, \sigma_2 \in \mathcal{P}_N$, is defined by

$$K_\tau(\sigma_1, \sigma_2) = \sum_{i<j\in[N]} \mathbb{1}\left[\operatorname{sign}\left(\sigma_1[i] - \sigma_1[j]\right) \neq \operatorname{sign}\left(\sigma_2[i] - \sigma_2[j]\right)\right].$$

Rank-aggregation is a generic name for the problem of finding, given a set $A \subseteq \mathcal{P}_N$, a single ranking over N candidates that correctly represents the combined preferences given by n voters. Here, we focus on the specific case (also known as Kemeny-aggregation [Kem59, DKNS01]) where the objective is to find a ranking $\pi^* \in \mathcal{P}_N$ that minimizes the sum of Kendall-tau distances to all input rankings in A, i.e.,

$$\min_{\pi\in\mathcal{P}_N} \sum_{\sigma\in A} K_\tau(\sigma, \pi). \tag{1.1}$$

This definition is consistent with the definition of 1-median used in the first part, and is computationally prohibitive to solve exactly. Research regarding the computational complexity of the rank-aggregation problem was initiated by Bartholdi et al. [BTT89a, BTT89b] who proved that Kemeny-aggregation (1.1) is NP-hard. A synopsis of improvements and approximation algorithms can be found in the related work Section 1.2.

We consider a generalization of rank-aggregation, that we call the *k-clustering rank-aggregation*, or *k-rank-aggregation* in short. In this problem, the goal is to partition a set of n voters that each ranks N possible candidates, into k disjoint sets (clusters), with a representative ranking (the median) for each set. As in the k-median definition, we would like to find a set $\Pi \subseteq \mathcal{P}_N$ of size k that minimizes the objective

$$\sum_{\sigma\in A} \min_{\pi\in\Pi} K_\tau(\sigma, \pi).$$

As one could expect, 1-rank-aggregation corresponds to the already NP-hard rank-aggregation problem [Kem59], hence, we do not strive to solve its generalization exactly, and we resort to approximation algorithms. There is a known $(1+\varepsilon)$-approximation algorithm (PTAS) for $k = 1$, by Schudy and Mathieu [KS07, Sch12], and there is no approximation algorithm for $k > 1$. Hence we ask, *can we use coresets to approximate the k-rank-aggregation problem for general k?*

1.1 Our Results

This section proceeds to partially answer the questions outlined above. The results focus on *coresets* for the *k-Median* problem under various metric spaces, and comprise of two parts.

Weak Coresets by Uniform Sampling

This part contains three results that deal with constructing weak coresets by sampling uniformly and independently with replacement from the input. The first two results, present a weak coreset for the 1-median objective. They guarantee a weak $(\varepsilon, 1)$-coreset in $\mathbb{R}^d$ of dimension-*independent* size under either ℓ_2 or ℓ_1 distances.

Theorem 1.1.1 (Informal version of Theorems 2.1.5 and 2.2.9)**.** *Let $A \subseteq \mathbb{R}^d$ be a set of n points, equipped with the ℓ_p-distance for $p \in \{1, 2\}$. Let $\varepsilon \in (0, \frac{1}{10})$ and let $S \subseteq A$ be a multiset obtained by $s = \tilde{O}(\varepsilon^{-2p})$ independent and uniform draws (with replacement) from A. Then, with high probability, S is a weak $(\varepsilon, 1)$-coreset of A.*

This theorem extends and improves some previously known dimension-dependent results from [Tho05, MS18, McG18]. Our improved analysis for ℓ_2 uses a Johnson-Lindenstraus (JL)

[JL84] dimensionality reduction, with target dimension independent of n and d, and relies on a result of [MMR19] that the JL-transform preserves the cost of every partition. Intuitively, when the target dimension is independent of d, we can perform dimension-dependent uniform sampling in the target space to obtain a dimension-independent weak coreset in the input space. For ℓ_1, we perform a fine-grained analysis, based on considering each dimension separately and dividing it into geometrically decreasing-size intervals. For each interval, we estimate the probability that the sample median falls in that interval, and the approximation factor in this case. Formal statements and their analyses are written in Sections 2.1 and 2.2.

The third result considers "well-separated" inputs, and construct a weak (ε, k)-coreset for $k = 2$. Intuitively, the optimal clusters for these inputs are separated in space and have roughly the same size. This allows us to employ the previous results of $k = 1$ for each optimal cluster separately. The separation notion is captured by the α-center proximity property from [ABS12] which states that in every optimal clustering, every input point is a factor-α closer to its own optimal center than to any other center.

Definition 1.1.2 (α-center proximity [ABS12])**.** A k-median instance $A \subseteq V$ satisfies the *α-center proximity* property for $\alpha > 1$, if for every optimal center set $\mathcal{C}^*$, every $p \in A$ with its closest center $c_i^* \in \mathcal{C}^*$, and every other optimal center $c_j^* \neq c_i^*$, it holds that

$$\alpha D(p, c_i^*) < D(p, c_j^*).$$

The second property the input needs to satisfy, is that the "average" size of a cluster in an optimal solution is approximately $\frac{n}{k}$.

Definition 1.1.3 (η-Balanced Optimal Clusters)**.** Let $A \subseteq V$ be a k-median instance. The instance is said to have *η-balanced* optimal clusters for $\eta \leq 1$ if for every optimal partition $(A_1^*, \ldots, A_k^*)$ and every $i \in [k]$, the size of A_i^* is at least $\eta \frac{n}{k}$.

In [MOP04] there is a discussion about the necessity of the above assumption and some reasonable justifications for uniform sampling. In a nutshell, if we have a small-size optimal cluster, say $O(\sqrt{n})$, then there is a very low probability of sampling a point in this cluster (unless the sample size is depends on n). However, if that optimal cluster is also very far away from the rest of the points, then the points that belong to this cluster contribute a lot to the cost, and thus it is un-affordable to sample none of them. Since the goal of clustering is to divide the input into distinct classes, this assumption can also be viewed as a justification to "neglecting" a small number of outliers that might not fit into any of the classes.

Overall, for the special case of instances that satisfy the two assumptions specified above, a weak $(\varepsilon, 2)$-coreset is constructed by uniform sampling.

Theorem 1.1.4 (Informal version of Theorem 2.3.1)**.** *Let $\varepsilon > 0$ and $p \in \{1, 2\}$ and $k = 2$. Let A be a set of n points in $(\mathbb{R}^d, \ell_p)$, that satisfy the α-center proximity property for a large enough fixed $\alpha > 1$ and has η-balanced optimal clusters. Then with constant probability, a uniform sample S of size $s = \tilde{O}(\varepsilon^{-2p}\eta^{-1}k \log k)$ is a weak (ε, k)-coreset.*

The proof is based on the separability property and the 1-median result above for each optimal cluster separately. Roughly speaking, the separability property implies that an optimal "local" center for each cluster can serve as a "global" center in an optimal 2-median solution. We emphasize that this method results in dimension-independent coreset size. The formal theorem and its analysis can be found in Chapter 2.3, and it immediately raises the obvious question: how can we generalize the result to $k \geq 2$? Finally, we remark that if datasets in practice satisfy the above conditions, then this theorem provides theoretical justification for why uniformly sampling is effective in reducing the dataset size, while preserving the cost of the optimal solution. To the best of our knowledge, this is the first result that relies on a separability assumption to get an improved coreset construction.

PTAS for k-Clustering Rank-Aggregation

Now, we present two applications of strong (ε, k)-coresets, in order to $(1+\varepsilon)$-approximate the k-rank-aggregation problem. This part of the thesis is independent of the first part, and the coresets are not based on uniform sampling, but rather on recent results that use the importance sampling framework. In particular, a strong coreset for $(\mathcal{P}_N, K_\tau)$ is obtained using a known dimension-independent coreset construction by [BJKW21] and an isometric embedding of $\mathcal{P}_N$ into ℓ_2^2 [JKS16]. Another key component in our result is a PTAS for 1-rank-aggregation (Corollary 7.3 by [Sch12]). Our algorithms invoke it (in a black-box manner) to find a 1-median of each set in certain partitions.

The first result is for small k (e.g. constant).

Theorem 1.1.5 (Informal version of Theorem 3.2.1)**.** *Let $\varepsilon, \delta \in (0,1)$. Let $A \subseteq \mathcal{P}_N$ of size n be a k-rank-aggregation instance. Then, there is a $(1+\varepsilon)$-approximation randomized algorithm that succeeds with probability at least $1-\delta$ and in time*

$$O\left(\mathrm{poly}(n, N, \frac{1}{\varepsilon}) \cdot k^{\mathrm{poly}\left(k, \frac{1}{\varepsilon}, \frac{1}{\delta}\right)}\right).$$

If ε is treated as a constant, the above running time is actually polynomial in n and N, and thus this algorithm is actually an FPT-PTAS in k. The algorithm operates by computing a strong (ε, k)-coreset of $A \subseteq \mathcal{P}_N$, and then trying all possible partitions of the coreset, where for each partition it uses the 1-median PTAS from [Sch12] to find an approximate center for each set (cluster). See section 3.2 for full details.

The second result does not restrict the value of k, but rather considers a special subset of inputs that satisfy a separability condition (a different one than in the results above) known as the β-distributed property.

Definition 1.1.6. (β-distributed [ABS10]) A k-median instance A is called *β-distributed* for $\beta > 0$, if for every optimal center c_i^* of an optimal cluster A_i^*, and every data point $p \notin A_i^*$, it holds that

$$D(p, c_i^*) \geq \beta \frac{\mathsf{OPT}(A,k)}{|A_i^*|}.$$

This definition implies that in a β-distributed instance, for every optimal cluster, the optimal center is far away from every outside point. Using this definition, we have the following theorem.

Theorem 1.1.7. *(Informal version of Theorem 3.3.1) Let $\varepsilon \in (0,1)$. Let $A \subseteq \mathcal{P}_N$ of size n be a k-rank-aggregation instance that satisfies the β-distributed property. Then there is a $(1+\varepsilon)$-approximation algorithm that returns a solution in time*

$$\tilde{O}\left(\left(\frac{n}{\varepsilon}\right)^{O(\varepsilon^{-5}\beta^{-1})} k^{O(\beta^{-1})} \mathrm{poly}\left(n, N, \frac{1}{\varepsilon}\right)\right).$$

Our algorithm is an adaptation of an algorithm from [ABS10]. The setting in [ABS10] is different since their algorithm only deals with center points that are part of the input and thus it is affordable to enumerate all possible centers and take the best one. In our setting, the centers are chosen from a prohibitively large ambient space ($N!$ points to be exact) and the algorithm cannot efficiently try all possible options. Our modification to the algorithm from [ABS10] is an added step that guesses, by exhaustive search, a coreset (that is known to exist) for every cluster with high cost and then applies on this coreset the 1-rank-aggregation PTAS [KS07, Sch12] to find an approximate center. We remark that since the instance is separated, it is not possible to have many clusters with high cost, and thus this exhaustive search still results in a reasonable running time. Additionally, the algorithm guesses the coresets since it does not know in advance what are the sets from which it needs to extract a coreset. We note that if ε and β are treated as constants, then this algorithm's running time avoids the exponential dependency in k. See Section 3.3 for the specifics.

1.2 Related Work

k-Clustering. The notoriously difficult k-median problem generated a huge amount of research that considered many variants. It is a special case of the general k-clustering problem or the facility location problem, and two of its most important versions are the "continuous" and "discrete" k-median. In the former, which is also the focus of this thesis, the centers $\mathcal{C}$ can be any arbitrary set of k points from the ambient metric space V, while in the latter, the k centers come from a specific set of points that is given as input, e.g., the input set A. Along with the many settings for k-median problem, there are also plenty of algorithmic approaches for approximating a solution, to name a few, linear programming rounding [LV92, CGTS02], primal-dual methods [JV01, ANFSW19], dual fitting [JMS02], local search [CG99, KPR00, AGK+04], sampling [MOP01, CS04, MP04, MOP04] and combinations of these methods [Ind99, LS13]. For the discrete case, the state of the art is $(2.675+\varepsilon)$-approximation [BPR+15], which relies on two improved steps from previous works of [AGK+04, LS13]. It is possible to derive a $(5.35+\varepsilon)$-approximation for the continuous case by only considering input points as centers and losing a factor of 2. Additionally, we mention approximation algorithms in noteworthy metric spaces. In Euclidean space, the problem is APX-hard when both k and d are part of the input [GI03], however it admits a PTAS if we fix one of these parameters [ARR98, HPM04, Che09, CKM19, FRS19]. A PTAS also exists for bounded doubling metrics, as shown in recent results [FRS19, CFS19].

Regarding the hardness of the problem, [JMS02] showed that the discrete metric k-median is NP-hard to approximate better than $1+\frac{2}{e}$ factor, which is an improvement to a classic result by [GK99]. For the continuous case, it was recently shown by [CKL21] to be NP-hard to approximate within a $2-o(1)$-factor.

Coreset Constructions. Strong coreset constructions for k-median (and k-clustering in general), saw many advances since the first definitions in [AHPV04, HPM04]. They are notably useful, when there are constraints on the available amount of storage, or in streaming and distributed settings, using the merge-and-reduce framework [BS80, AHPV04]. The first construction was geometric-based for Euclidean spaces of dimension d by [HPM04], of size $O\left(k\varepsilon^{-d}\log n\right)$ which is exponential in d, and it was later shown that the factor of $\log n$ can be avoided in Euclidean spaces by [HPK07]. Subsequently, Chen [Che09] designed strong (ε,k)-coresets for general metric spaces, which employ a few uniform sampling steps from rings around approximate centers and achieve size $O\left(k^2\varepsilon^{-2}\log^2 n\right)$. We remark that in general metric spaces it is shown in [BBH+20] that the $\log n$ factor in the size of strong (ε,k)-coresets is inevitable, even for $k=1$. This fact rules out the possibility of removing the $\log n$ factor in general metric spaces even for separable inputs (using definitions that implicitly assume $k>1$). Succeeding Chen, a framework based on importance sampling and VC-dimension arguments was introduced by [LS10, FL11]. The importance sampling technique is different compared to uniform sampling, since the algorithm constructs a sampling distribution, where "important" points are sampled with higher probability. Compared to uniform sampling, this distribution is harder to estimate in practice, hence the advantage of using uniform sampling in this thesis. Until recently, this framework produced the best coreset sizes by improvements and refinements in multiple papers for different settings and metrics spaces, [SW18, BBH+20, HV20, FSS20, BJKW21]. For Euclidean spaces, the works of [BFL16, SW18, HV20, BJKW21] removed the dependency in the dimension d for the coreset size using different proof techniques and running times. Coresets were found in additional metric spaces. For example, in metric spaces of bounded doubling dimension, denoted $ddim$, [HJLW18] showed a construction of coreset of size $O\left(k^3\cdot ddim\cdot\varepsilon^{-2}\right)$. In graph metrics, [BBH+20] constructs a coreset for graphs with bounded treewidth of size linear in the treewidth, and in [BJKW21] a framework for excluded-minor or bounded highway dimension graphs was designed.

The current state of the art is by [CSS21], which propose a completely new construction framework. In this framework, one need only to show the existence of a centroid set in order to obtain a small (logarithmic in the size of the centroid set) coreset. Their framework applies to Euclidean spaces, achieving coreset size of $O(k \log k\varepsilon^{-4})$ (independent of both n and dimension d), to general metric spaces achieving coreset size of $O(k\varepsilon^{-2} \log n)$, and to the special graph families mentioned above.

For a comprehensive surveys of coresets and their construction techniques, see [MS18, Fel20]. Finally, we remark that some of the papers mentioned above also leverage the coresets to design approximation algorithms for the corresponding clustering problem.

Separability Conditions. There are about five central, up to small variants, competing separability conditions for the k-clustering problem, that capture approximately the same notion of separated optimal clusters. In the papers listed below, the authors used these definitions to restrict the k-clustering problem, and suggested improved (exact or approximation) algorithms for these cases. Here, we briefly discuss the conditions themselves and emphasize that some of them imply the conditions used in this thesis, and thus motivate our choice of these definitions. The first condition is known as γ-perturbation resilient [BL12]. The intuition around this property was the fact that most distance measures used on real-life datasets are based on some heuristic measure, and thus do not exactly capture the dissimilarity between the objects. Hence, if we assume the input is γ-perturbation stable, then the optimal clusters should stay invariant if we perturb the distances between the objects a little (by up to a factor of γ). This definition is a successfull one, since it implies the α-center proximity condition (for $\gamma = \alpha$) that is more useful for proving algorithms correctness [ABS12, ACM+20]. An important property of the α-center proximity condition, is that for large enough α, specifically for $\alpha \geq 1+\sqrt{2}$, the k-median problem can be optimally solved in polynomial time using an algorithm that is based the single-linkage algorithm and dynamic programming to find the best partition [ABS12, BL16]. Two additional conditions, are known as the ε-separated [ORSS13] and (c, ε)-approximation stability [BBG09]. The former assumes that the optimal cost of clustering using k-centers is factor ε smaller then the optimal clustering using $k-1$ centers. The latter definition provides a guarantee that an approximation to the cost is actually an approximation to the true target clustering. Both of [ORSS13] and [BBG09] definitions imply the β-distributed condition (for slightly different parameters), which allows for designing a PTAS for the discrete k-medians [ABS10].

For surveys on these separability notions, see [ABD09, BR14, Ben15] and references therein. Additionally, the separability definitions were surveyed from a more practical perspective in [Ben15]. They discuss if these definitions can get us closer to resolving the "clustering is difficult only when it doesn't matter" interpretation [DLS12].

Rank Aggregation. The rank aggregation problems have deep roots in the fields of social choice theory, voting and statistics [Ken38, Kem59, YL78, You88], and we only present here an overview from a computational point of view. As mentioned above, it is known to be NP-hard to optimize exactly [BTT89b, BTT89a]. The case of $n = 2$ voters is easy to solve by simply choosing one of the voters (or any point on one of their shortest paths) as an optimal solution. It was later shown to be NP-hard for even integer $n \geq 4$ number of voters [DKNS01, BBD05]. The complexity of constant odd number of votes was open until recently, when [BBG+19] showed that it is NP-hard for every $n \geq 7$. Thus, exact polynomial algorithms for $n = 3$ and $n = 5$ number of voters is still unresolved. See [MHP18] for a recent investigation for the case of $n = 3$.

There is a diverse landscape of approximation algorithms for the 1-rank-aggregation problem and we briefly survey the main results. First, we consider randomized constant factor approximation algorithms. Two 2-approximation algorithms were suggested by [ACN08]. The simple one is a 2-approximation achieved by picking any input permutation at random. A more intricate approach leverages a reduction from the set of rankings A to a directed tournament

graph, which allows for divide-and-conquer algorithm (performing "quick-sort" type iteration on the vertices) to create an output ranking (the graph topological sort). It is also shown by [ACN08] that choosing the best output among the two above algorithms guarantees an even better 11/7-approximation. Additionally, an analysis by [CFR10] guarantees that a simple positional method, attributed as early as to Borda [Bor84], provides a 5-approximation. Another approach uses a distance measure called the Spearman-footrule distance [DG77]. It was shown in [DKNS01] that it is possible to optimally solve the corresponding rank-aggregation using this distance measure, and since for every two permutations, their Spearman-footrule distance is at most a factor 2 from their Kendall-tau distance [DG77], it immediately yields yet another 2-approximation. The best constant factor approximation of 4/3 was given by [ACN08], using randomized LP rounding. The randomized pivot-based and linear programming algorithms were derandomized by [ZW07].
Somewhat in parallel to the above results, the method of local search (swap adjacent candidates until local minimum is reached) from [DKNS01] was used to formulate a PTAS (that guarantees $(1+\varepsilon)$-approximation) [KS07, Sch12]. Their idea was to reduce the problem into a directed tournament graph, and begin from a known constant factor approximation. Then, they improve by divide and conquer split and local search on each part. We use their result in a black-box manner every time our algorithm will search for a 1-median of a certain subset of the input.

Finally, we also mention that there are notions of partial rankings (such as q-ratings and top-r lists) [FKS03, FKM$^+$04, FKM$^+$06, Ail10]. A PTAS was also designed for top-r lists by combining two of the above methodologies [MM20]. It starts with an initial constant factor approximation using positional or comparison-sort algorithm, and then performs incremental improvements using local search.

1.3 Open Questions

In this thesis, we mention multiple open questions along the way, and we repeat here the main questions that are left open. For the first part, the most obvious future work is to extend the weak $(\varepsilon, 2)$-coreset result to general $k > 2$. In general, the fundamental power of the uniform sampling for weak coresets constructions is interesting. Can this technique work for other metric spaces? Regarding separability notions, it will be useful to understand whether large enough uniform sample preserves the separability property of the input set for $k \geq 2$, even at a cost of deteriorating the separability parameter. Finally, do we really need the separability condition to construct a weak coreset for $k \geq 2$? If so, what is the weakest assumption that can make it work?

From the second part of the thesis, the most glaring open question, is to design a PTAS for k-rank-aggregation without the β-distributed separability assumption. For example, is there a different algorithm that leverages the structure of the permutation space $(\mathcal{P}_N, K_\tau)$ rather than just using a 1-median approximation as a black-box?

Chapter 2

Weak Coresets by Uniform Sampling

In this chapter, we study weak coresets, denoted by S, that are constructed using uniform sampling from an input set A. The sampling is performed by s independent draws, with replacement, from the input set A, where every point $x \in A$ is sampled with uniform probability of $\frac{1}{n}$. In Sections 2.1 and 2.2 we obtain a weak $(\varepsilon, 1)$-coresets of dimension-*independent* size in both the ℓ_2 and ℓ_1 metrics spaces respectively (with different size bounds). Section 2.3 considers inputs that satisfy separability assumptions (Definitions 1.1.2 and 1.1.3), to obtain a weak $(\varepsilon, 2)$-coreset. These assumptions allow us to argue about each optimal cluster independently and to use the above results for $k = 1$.

All through this chapter, we denote by $m(X)$ an optimal median of a set X. If the median is not unique, then $m(X)$ refers to an arbitrary one. Additionally, $\hat{m}(X)$ denotes a point that is a $(1 \pm \varepsilon)$-approximate median. We emphasize that both $m(\cdot)$ and $\hat{m}(\cdot)$ are not easily computed in general.

2.1 Dimension-Independent Weak $(\varepsilon, 1)$-Coreset in ℓ_2

In this section we consider a set $A \subset \mathbb{R}^d$ equipped with the Euclidean distance ℓ_2, and wish to construct a weak $(\varepsilon, 1)$-coreset of size that is dimension-*independent* by using only uniform sampling from A (Theorem 2.1.5). This is the first bound that is dimension independent, and improves over [MS18]. Before proving it, we need to review the dimension-dependent bound.

Dimension-Dependent Bound

The following is a known result for dimension-dependent size weak $(\varepsilon, 1)$-coreset obtained by using uniform sampling.

Theorem 2.1.1 (Theorem 4 of [MS18]). *Let $\varepsilon \in (0, 1)$. Let $A \subseteq \mathbb{R}^d$ and let S be a uniform sample of A of size $|S| \geq \Omega(\varepsilon^{-2} d \log \frac{d}{\varepsilon} \log(\frac{1}{\delta}))$. Then with probability at least $1 - \delta$, S is a weak $(\varepsilon, 1)$-coreset of A i.e.,*

$$\sum_{x \in A} \|x - m(S)\|_2 \leq (1 + \varepsilon) \sum_{x \in A} \|x - m(A)\|_2 .$$

We now show that this theorem extends also to a $(1 + \varepsilon)$-approximation of $m(S)$.

Theorem 2.1.2 (Weak Coreset Approximate Center). *Let $\varepsilon \in (0, \frac{1}{45}), \delta \in (0, \frac{1}{4})$. Let $A \subseteq \mathbb{R}^d$ and let S be a uniform sample of A of size $|S| \geq \max\{\Omega(\varepsilon^{-2} d \log \frac{d}{\varepsilon} \log \frac{1}{\delta}), \frac{4}{\delta}\}$. Then with probability at least $1 - \delta$, every $(1 + \varepsilon)$-approximate median of S, denoted $\hat{m}(S)$, is a $(1 + 37\varepsilon)$-approximate median of A, i.e.,*

$$\sum_{x \in A} \|x - \hat{m}(S)\|_2 \leq (1 + 37\varepsilon) \sum_{x \in A} \|x - m(A)\|_2 .$$

The proof of this theorem is similar to the proof of Theorem 2.1.1, but requires adjusting the constants in their technical lemma (see Lemma 6 in [MS18]) and using an additional observation.

Lemma 2.1.3 (Adaptation of Lemma 6 in [MS18]). *Let $\varepsilon \in (0, \frac{1}{45})$. Let $A \subset \mathbb{R}^d$ and $S \subseteq A$ be a uniform sample of A. Then for every point b with $\sum_{x\in A} \|x-b\|_2 \geq (1+36\varepsilon)\sum_{x\in A}\|x-m(A)\|_2$,*

$$\mathbb{P}\left[\sum_{x\in S}\|x-b\|_2 < \sum_{x\in S}\|x-m(A)\|_2 + \frac{9\varepsilon|S|}{|A|}\sum_{x\in A}\|x-m(A)\|_2\right] \leq \exp\left(-\frac{\varepsilon^2|S|}{4}\right).$$

Proof. The proof is similar to the proof of Lemma 6 in [MS18]. We only need to change the constants of their defined random variable to the following

$$X = \sum_{x\in S}\frac{\|x-b\| - \|x-m(A)\| + \|m(A)-b\|}{2\big(\,\|m(A)-b\| + \frac{9\varepsilon}{|A|}\sum_{x\in A}\|x-m(A)\|\,\big)}.$$

Rearranging the condition for a point b as follows

$$\sum_{x\in A}(\|x-b\| - \|x-m(A)\|) > 9\varepsilon\sum_{x\in A}\|x-m(A)\| + 9\varepsilon\big(|A|\,\|b-m(A)\| + 9\varepsilon\sum_{x\in A}\|x-m(A)\|\,\big)$$

and using Chernoff bound completes the proof. ∎

Lemma 2.1.4. *For every $\delta \in (0, \frac{1}{2})$,*

$$\mathbb{P}\big[\frac{1}{|S|}\sum_{x\in S}\|x-m(A)\|_2 \leq \frac{1}{\delta|A|}\sum_{x\in A}\|x-m(A)\|_2\,\big] \geq 1-\delta.$$

Proof. Follows by using Markov's inequality since

$$\mathbb{E}\left[\frac{1}{|S|}\sum_{x\in S}\|x-m(A)\|_2\right] = \frac{1}{|A|}\sum_{x\in A}\|x-m(A)\|_2\,.$$

∎

We are now ready to prove the extended theorem (its proof structure is based on [MS18]).

Proof of Theorem 2.1.2. We denote $\mathsf{OPT} := \sum_{x\in A}\|x-m(A)\|_2$. By Markov's inequality and a union bound over S, all points in S will be contained in the ball B of radius $r = \frac{|S|}{\delta'}\frac{\mathsf{OPT}}{n}$ around $m(A)$ w.p at least $1-\delta'$. Let $\hat{m}(S)$ a $(1+\varepsilon)$-approximate median of S. To show that $\hat{m}(S)$ is also inside the ball B, we use the fact that it is $(1+\varepsilon)$-approximation of $m(S)$, triangle inequality and Lemma 2.1.4, and obtain with probability at least $1-\delta'$

$$\sum_{x\in S}[\|m(S)-\hat{m}(S)\|_2 - \|x-m(S)\|_2] \leq \sum_{x\in S}\|x-\hat{m}(S)\|_2 \leq (1+\varepsilon)\sum_{x\in S}\|x-m(S)\|_2$$
$$\Rightarrow \|m(S)-\hat{m}(S)\|_2 \leq \frac{(2+\varepsilon)}{|S|}\sum_{x\in S}\|x-m(S)\|_2 \leq \frac{(2+\varepsilon)}{\delta'|A|}\mathsf{OPT}.$$

In addition, by Theorem 2.1.1 we have with probability at least $1-\delta'$, and by using the triangle inequality,

$$\sum_{x\in A}\|x-m(S)\|_2 \leq (1+\varepsilon)\sum_{x\in A}\|x-m(A)\|_2 \Rightarrow \|m(S)-m(A)\|_2 \leq \frac{2+\varepsilon}{|A|}\mathsf{OPT}.$$

Using triangle inequality and assuming $|S| > \frac{1}{2\delta'}$, conclude that $\hat{m}(S)$ is contained in B:

$$\|\hat{m}(S) - m(A)\|_2 \leq \|\hat{m}(S) - m(S)\|_2 + \|m(S) - m(A)\|_2 \leq \frac{4\mathsf{OPT}}{\delta'|A|} < \frac{8|S|\mathsf{OPT}}{|A|}.$$

Let C be an $\varepsilon' = \varepsilon\frac{\mathsf{OPT}}{|A|}$ ball-cover of B (i.e. for every $x \in B$, there exists $p \in C$ such that $\|x - p\|_2 \leq \varepsilon'$. It is known that $|C| = O(3/\varepsilon)^d$). Using union bound and setting $|S| = c \cdot d\varepsilon^{-2} \log\frac{d}{\varepsilon} \log\frac{1}{\delta'}$ for sufficiently large constant c, Lemma 2.1.3 holds for every point $b \in C$ with probability at least $1 - \delta'$. Assume this event happens. Let $p \in C$ be the closest point in the ball-cover to $\hat{m}(S)$. We have by Lemma 2.1.4,

$$\begin{aligned}
\sum_{x\in S} \|x - p\|_2 &\leq \sum_{x\in S} \left[\, \|x - \hat{m}(S)\|_2 + \|\hat{m}(S) - p\|_2 \,\right] \\
&\leq (1+\varepsilon) \sum_{x\in S} \|x - m(S)\|_2 + |S|\varepsilon\frac{\mathsf{OPT}}{|A|} \\
&\leq \sum_{x\in S} \|x - m(A)\|_2 + \varepsilon\frac{8|S|\mathsf{OPT}}{|A|} + |S|\varepsilon\frac{\mathsf{OPT}}{|A|} \\
&= \sum_{x\in S} \|x - m(A)\|_2 + \varepsilon\frac{9|S|}{|A|} \sum_{x\in A} \|x - m(A)\|_2 .
\end{aligned}$$

Now, by Lemma 2.1.3 we have that $\sum_{x\in A} \|x - c\|_2 \leq (1 + 36\varepsilon) \sum_{x\in A} \|x - m(A)\|_2$. Finally, using triangle inequality and union bound over all events mentioned above (where $\delta' = \delta/4$), concludes the proof:

$$\sum_{x\in A} \|x - \hat{m}(S)\|_2 \leq \sum_{x\in A} \|x - c\|_2 + |A|\, \|c - \hat{m}(S)\|_2 \leq (1 + 37\varepsilon) \sum_{x\in A} \|x - m(A)\|_2 .$$

■

Dimension-Independent Bound

The following theorem is the main result of this section, a weak $(\varepsilon, 1)$-coreset in ℓ_2 distance using only uniform sampling, and of size that is dimension-independent.

Theorem 2.1.5. *Let $\varepsilon \in (0, \frac{1}{45})$ and $\delta \in (0, \frac{1}{4})$. Consider the 1-median clustering problem with input set $A \subset \mathbb{R}^d$ and a uniform sample $S \subseteq A$ of size*

$$|S| = \max\left\{\Omega\left(\varepsilon^{-4} \log^2\left(\frac{1}{\varepsilon\delta}\right) \log\left(\varepsilon^{-3} \log\frac{1}{\varepsilon\delta}\right)\right), \frac{8}{\delta}\right\}.$$

Then with probability at least $1 - \delta$, S is a weak $(\varepsilon, 1)$-coreset of A, i.e.,

$$\sum_{x\in A} \|x - m(S)\|_2 \in (1 \pm \varepsilon) \sum_{x\in A} \|x - m(A)\|_2 .$$

Preserving Cost using Johnson-Lindenstrauss Transform. The following is a general definition for sub-gaussian dimension reduction, typically referred to as Johnson-Lindenstrauss (JL) transform.

Definition 2.1.6 (Sub-Gaussian Dimension Reduction [MMR19])**.** Let $\varepsilon, \delta \in (0, 1/2)$. A family of random linear maps $\mathcal{L}_{d,t} : \mathbb{R}^d \to \mathbb{R}^t$ is called *sub-gaussian dimension reduction* if for every $d \geq 1$, and all $x \in \mathbb{R}^d$, we have

$$\mathbb{P}_{L\sim\mathcal{L}_{d,t}} \left[\|Lx\|_2 \in (1 \pm \varepsilon) \cdot \|x\|_2\right] \geq 1 - \delta$$

where $t = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$. Also, for every unit vector $x \in \mathbb{R}^d$ and every $\Delta \geq 0$, it holds,

$$\mathbb{P}_{L\sim\mathcal{L}_{d,t}}\left[\|Lx\|_2 \geq 1+\Delta\right] \leq \exp\left(-\Omega(\Delta^2 t)\right).$$

A key step in our proof is a recent result by [MMR19] that shows the above transforms with target dimension $t = \tilde{O}(\varepsilon^{-2}\log k)$ preserves the cost of every partition to k subsets within $(1+\varepsilon)$ factor.

Theorem 2.1.7 (Theorem 1.3 of [MMR19])**.** *Let $\varepsilon, \delta \in (0, 1/4)$ and consider any family of random maps $\mathcal{L}_{d,t} : \mathbb{R}^d \to \mathbb{R}^t$ that satisfy Definition 2.1.6. Then for every $d \geq 1$, and integer $k \geq 1$ the following holds. For every finite $A \subset \mathbb{R}^d$,*

$$\mathbb{P}_{L\sim\mathcal{L}_{d,t}}\left[\text{cost}(L(\mathcal{A})) \in (1\pm\varepsilon)\,\text{cost}(\mathcal{A}) \text{ for all partitions } \mathcal{A} = (A_1, A_2, \ldots, A_k) \text{ of } A\right] \geq 1-\delta$$

where $t = O\left(\frac{\log(\frac{k}{\varepsilon\delta})}{\varepsilon^2}\right)$ and $\text{cost}(\mathcal{A}) = \sum_{i=1}^{k} \min_{c_i\in\mathbb{R}^d} \sum_{x\in A_i} \|x - c_i\|_2$.

In [MMR19] it is asserted that the constructions in [JL84, DG03, Ach03] satisfy Definition 2.1.6, and thus apply to the above theorem, making it constructive. An important remark is that the cost of all possible *partitions* is preserved, and *not* the cost for all possible center sets. Using this theorem, we will mainly need the fact that the optimal clustering cost is preserved. In particular, for the case of $k = 1$, we have that the cost of the optimal median is preserved under these types of JL dimension reduction.

Corollary 2.1.8. *For a set $A \subset \mathbb{R}^d$, and using a JL transform L satisfying Definition 2.1.6, with target dimension $t = O\left(\varepsilon^{-2}\log(\frac{1}{\varepsilon\delta})\right)$, then with probability at least $1-\delta$,*

$$\sum_{x\in A} \|L(x) - m(L(A))\|_2 \in (1\pm\varepsilon)\sum_{x\in A}\|x - m(A)\|_2.$$

A valuable property of Theorem 2.1.7, is that the target dimension t is independent of the input dimension d and the number of input points n, especially when considering the original JL lemma [JL84] that require target dimension $O(\varepsilon^{-2}\log n)$. Hence, another consequence of Theorem 2.1.7 is that the sum of distances is preserved after a dimension reduction, while also saving the factor of $\log n$ in the target dimension. Intuitively, we only care to preserve the value of the entire sum, and not all pairwise distances, which allows preserving the extra $\log n$ factor.

Lemma 2.1.9. *Let L be a JL transform according to the conditions in Theorem 2.1.7 with target dimension $t = O\left(\varepsilon^{-2}\log(\frac{1}{\varepsilon\delta})\right)$. Then, for every set $A \subset \mathbb{R}^d$, with probability at least $1-\delta$,*

$$\sum_{x\in A}\|L(x)\|_2 \in (1\pm\varepsilon)\sum_{x\in A}\|x\|_2.$$

Proof. This proof was suggested to us by Lingxiao Huang and is based on a reduction of A to a symmetric set so we can apply Theorem 2.1.7 with $k = 1$. Let $A \subset \mathbb{R}^d$. Define $A' = A \cup \{-x \mid x \in A\}$. Then:

$$\sum_{x\in A'}\|x\|_2 = 2\sum_{x\in A}\|x\|_2 \quad , \quad \sum_{x\in A'}\|L(x)\|_2 = 2\sum_{x\in A}\|L(x)\|_2$$

Therefore, it is equivalent to prove

$$\sum_{x\in A'}\|L(x)\|_2 \in (1\pm\varepsilon)\sum_{x\in A'}\|x\|_2. \tag{2.1}$$

For a symmetric multi-set A', where $x \in A'$ implies $-x \in A'$, the optimal solution is the origin (the 0 vector) since it is the point that minimizes the distance to all pairs of points $\{-x, x\}$. In our case, $m(A') = \mathbf{0} \in \mathbb{R}^d$ and $m(L(A')) = \mathbf{0} \in \mathbb{R}^t$, because both A' and $L(A')$ are symmetric. Thus, by Corollary 2.1.8, with probability at least $1-\delta$, Equation (2.1) holds. ■

Remark. A different proof, that might be of independent interest, using sub-gaussian properties of the JL-transform, is given in Theorem A.3.1. See Appendix A for details and comparison.

Concluding the Proof of Theorem 2.1.5. The proof uses a dimension reduction L satisfying Definition 2.1.6. The reduction is used only for the sake of analysis in order to use the dimension-dependent weak $(\varepsilon, 1)$-coreset guarantee by Theorem 2.1.2 in lower dimension. The following two additional lemmas that use the results previously mentioned are needed. The first lemma implies that $L(m(A))$ is a $(1 \pm \varepsilon)$-approximate for the median of $L(A)$.

Lemma 2.1.10. *Consider a JL transform L that satisfies the conditions and result of Theorem 2.1.7. Then*

$$\sum_{y \in L(A)} \|y - L(m(A))\|_2 \in (1 \pm 3\varepsilon) \sum_{y \in L(A)} \|y - m(L(A))\|_2 .$$

Proof. First, by Corollary 2.1.8:

$$\sum_{y \in L(A)} \|y - m(L(A))\|_2 \in (1 \pm \varepsilon) \sum_{x \in A} \|x - m(A)\|_2 .$$

On the other hand, by applying Lemma 2.1.9 on the set A with respect to $m(A)$,

$$\sum_{x \in A} \|L(x) - L(m(A))\|_2 \in (1 \pm \varepsilon) \sum_{x \in A} \|x - m(A)\|_2 .$$

Conclude the lemma by combining both of the above inequalities. ■

The second lemma, shows that $L(m(S))$ is an $(1 \pm \varepsilon)$-approximate for the median of $L(A)$.

Lemma 2.1.11. *Consider a JL transform L that satisfies the conditions and result of Theorem 2.1.7. Then:*

$$\sum_{x \in A} \|L(x) - L(m(S))\|_2 \in (1 \pm O(\varepsilon)) \sum_{x \in A} \|L(x) - m(L(A))\|_2$$

Proof. By Lemma 2.1.10 we have that $L(m(S))$ can serve as $\hat{m}(L(S))$. Then, by applying Theorem 2.1.2 with respect to the set $L(S) \subseteq L(A)$, we obtain:

$$\sum_{x \in A} \|L(x) - L(m(S))\|_2 \leq (1 + O(\varepsilon)) \sum_{x \in A} \|L(x) - m(L(A))\|_2$$

■

Proof of Theorem 2.1.5. Let $\delta' = \frac{\delta}{2}$, pick $t = c\varepsilon^{-2} \log(\frac{1}{\varepsilon\delta'})$ for sufficiently large constant c and draw $L \sim \mathcal{L}_{d,t}$ satisfying the conditions of Definition 2.1.6. Then, combining all of the above results and using union bound, we have that Theorem 2.1.2, Theorem 2.1.5 and their corollaries hold. Thus, taking S of size,

$$|S| = \max\left\{\Omega\left(\varepsilon^{-2} t \log\frac{t}{\varepsilon}\right), \frac{8}{\delta}\right\} = \max\left\{\Omega\left(\varepsilon^{-4} \log^2(\frac{1}{\varepsilon\delta}) \log(\varepsilon^{-3} \log\frac{1}{\varepsilon\delta})\right), \frac{8}{\delta}\right\}$$

that was computed by plugging t into the sample size specified by Theorem 2.1.2, suffices to obtain a probability of at least $1 - \delta$ for the following to hold.

$$\begin{aligned}
\sum_{x \in A} \|x - m(S)\|_2 &\leq (1 + \varepsilon) \sum_{x \in A} \|L(x) - L(m(S))\|_2 && \text{Lemma 2.1.9} \\
&\leq (1 + O(\varepsilon)) \sum_{x \in A} \|L(x) - m(L(A))\|_2 && \text{Lemma 2.1.11} \\
&\leq (1 + O(\varepsilon)) \sum_{x \in A} \|x - m(A)\|_2 . && \text{Corollary 2.1.8}
\end{aligned}$$

■

Remark. We do not know whether the above theorem and proof technique can be extended to a $(1 + \varepsilon)$-approximation of $m(S)$. Currently, the difficulty lies in using Theorem 2.1.7, which require an optimal center for every partition.

2.2 Dimension-Independent Weak $(\varepsilon, 1)$-Coreset in ℓ_1

In this section, we obtain a weak $(\varepsilon, 1)$-coreset under the ℓ_1 distance using only uniform sampling of size $O(\varepsilon^{-2} \log \varepsilon^{-1})$ (Theorem 2.2.9). We emphasize that the size is independent of the number of points n and the dimension d. To prove it, we begin by showing a dimension-dependent weak $(\varepsilon, 1)$-coreset and then proceed to perform a finer-grained analysis to obtain a dimension-independent weak $(\varepsilon, 1)$-coreset. The analysis is based on considering each dimension independently, and dividing each dimension into geometrically decreasing size intervals. For each interval, we estimate the probability and approximation factor for the case that the median of the sample lands in that interval. To have a well-defined unique median, we assume for simplicity that n and s are odd numbers. This leads us to first show the following well known fact, that one can compute the 1-median under ℓ_1 distance efficiently, by considering each coordinate separately.

Fact 2.2.1. *The 1-median of a set of points $A \subset \mathbb{R}^d$ in ℓ_1 is given by the median of the points in each coordinate separately.*

Specifically, throughout this section, given $A \subset \mathbb{R}^d$ equipped with the ℓ_1 distance, when we need to compute the 1-median of A, denoted by $m(A)$, we assume that in each coordinate $i \in [d]$, denoted by $m(A)_i$, we have median of the multiset $A_i = \{x_i \mid x \in A\} \subset \mathbb{R}$ where x_i is the i'th entry of the point x, i.e.,

$$m(A) = (\text{median}(A_1), \text{median}(A_2), \ldots, \text{median}(A_d)).$$

The proof works by changing the order of summation, and arguing that the median minimize the sum of absolute errors on each dimension separately. Additionally, we use the following definitions for numbers in $\mathbb{R}$.

Definition 2.2.2. The *rank* of an element $y \in A \subset \mathbb{R}$ is $\text{rank}(y) = |\{x \in A \ : \ x \leq y\}|$.

Definition 2.2.3. An *ε-rank-approximate median* of a set $A \subset \mathbb{R}$ is a point $y \in A$ s.t.

$$\text{rank}(y) \in \left(\frac{n+1}{2} \pm \varepsilon n \right).$$

Dimension-Dependent Bound

As a prerequisite to the main result of this section, we begin by showing a weak $(\varepsilon, 1)$-coreset in $(\mathbb{R}^d, \ell_1)$ metric space, that will be constructed using uniform sampling, and has size which is *dependent* of the dimension d.

Theorem 2.2.4. *Let $A \subset \mathbb{R}^d$ and $\varepsilon \in \left(0, \frac{1}{10}\right)$. Let S be a sample multi-set of size $s = 2\varepsilon^{-2} \log(8d)$ drawn independently and uniformly (with replacement) from A. Then with probability at least $\frac{3}{4}$, S is a weak $(\varepsilon, 1)$-coreset of A, i.e., for $m(S) \in \mathbb{R}^d$ the 1-median of S,*

$$\sum_{x \in A} \|x - m(S)\|_1 \leq (1 + O(\varepsilon)) \sum_{x \in A} \|x - m(A)\|_1 .$$

The analysis is based on the following three lemmas, in which we consider each dimension separately. Thus, we switch our focus to the 1-dimensional case of the real line, and temporarily use the notation of $A \subset \mathbb{R}$ and $S \subseteq A$ a uniform sample. The first lemma states the median of a uniform samples is an ε-rank-approximate.

Lemma 2.2.5. *Let $A \subset \mathbb{R}$ a set of n real distinct numbers and $\varepsilon \in \left(0, \frac{1}{10}\right)$. Let $S \subseteq A$ be a uniform sample multiset of size $s = 2\varepsilon^{-2} \log \frac{2}{\delta}$. Then with probability at least $1 - \delta$, $m(S)$ is an ε-rank-approximate median of A.*

Proof. The proof is based on [McG18] and is repeated here for completeness. We begin by partitioning the set A into three disjoint sets using the rank:

$$A_L = \left\{ x \in A \ : \ \text{rank}(x) \leq \frac{n+1}{2} - \varepsilon n \right\}$$
$$A_M = \left\{ x \in A \ : \ \text{rank}(x) \in \left[\frac{n+1}{2} \pm \varepsilon n \right] \right\}$$
$$A_U = \left\{ x \in A \ : \ \text{rank}(x) \geq \frac{n+1}{2} + \varepsilon n \right\}.$$

Observe that $m(S)$ is an ε-rank-approximate median of A if less than $\frac{s}{2}$ points are sampled from *both* A_L and A_U, i.e., we require

$$|A_L \cap S| < \frac{s}{2} \text{ and } |A_U \cap S| < \frac{s}{2}.$$

We show this holds with high probability for A_L, and the case for A_U follows symmetrically. Define the random variable $X = \sum_{x \in S} \mathbb{1}[x \in A_L]$ as a sum of independent indicators. Notice that $|A_L \cap S| \geq \frac{s}{2}$ if and only if $X \geq \frac{s}{2}$. Also, by linearity of expectation, $\mathbb{E}[X] = s\frac{(\frac{n+1}{2} - \varepsilon n)}{n} = s\left(\frac{1}{2} + \frac{1}{2n} - \varepsilon\right)$.
Using Hoeffding's inequality, and assuming $\frac{1}{n} \leq \varepsilon$ and $s = 2\varepsilon^{-2} \log \frac{2}{\delta}$ we have:

$$\mathbb{P}\left[X \geq \frac{s}{2}\right] = \mathbb{P}\left[X - \mathbb{E}X \geq s \cdot \varepsilon - \frac{s}{2n}\right] \leq \mathbb{P}\left[X - \mathbb{E}X \geq \frac{s \cdot \varepsilon}{2}\right] \leq \exp\left(-\frac{1}{2}\varepsilon^2 s\right) \leq \frac{\delta}{2}. \quad (2.2)$$

Overall, by union bound, we have that

$$\mathbb{P}\Big[|A_L \cap S| < \frac{s}{2} \wedge |A_U \cap S| < \frac{s}{2}\Big] \geq 1 - \delta.$$

■

The second lemma shows that for every point in A, it is possible to upper bound the difference between the cost of using that point to the cost of using the median, by associating the difference in ranks and distance between the point and the median.

Lemma 2.2.6. *For every point $y \in A \subset \mathbb{R}$ it holds that,*

$$\sum_{x \in A} \left[\, |x - y| - |x - m(A)| \, \right] \leq 2 \, |\text{rank}(y) - \text{rank}(m(A))| \cdot |y - m(A)| \, .$$

Proof. The proof is inspired by the proof of Lemma 2.2.5, defining three different sets.
Let $y \in A$. Assume that $\text{rank}(y) \neq \text{rank}(m(A))$, (otherwise both sides are equal to zero). Assume $y < m(A)$, and the other direction follows symmetrically. Define:

$$A_L = \{x \in A \ : \ \text{rank}(x) \leq \text{rank}(y)\}$$
$$A_M = \{x \in A \ : \ \text{rank}(y) < \text{rank}(x) \leq \text{rank}(m(A))\}$$
$$A_U = \{x \in A \ : \ \text{rank}(m(A)) < \text{rank}(x)\}$$

Clearly, by definition,

$$|A_L| = \text{rank}(y) \ , \ |A_U| = n - \text{rank}(m(A)) \ , \ |A_M| = \text{rank}(m(A)) - \text{rank}(y).$$

Now, depending on which set $x \in A$ is in, we have the following cases:

- If $x \in A_L$, then $|x - y| - |x - m(A)| = -\,|y - m(A)|$.

- If $x \in A_U$, then $|x-y| - |x-m(A)| = |y-m(A)|$.
- If $x \in A_M$, then by triangle inequality, $|x-y| - |x-m(A)| \leq |y-m(A)|$.

Using the first two items and the fact that

$$|A_U| - |A_L| = n - \text{rank}(m(A)) - \text{rank}(y) \leq \text{rank}(m(A)) - \text{rank}(y),$$

we obtain,

$$\sum_{x \in A_L \cup A_U} [|x-y| - |x-m(A)|] = (\text{rank}(m(A)) - \text{rank}(y)) \, |m(A)-y| \, .$$

The last item implies that

$$\sum_{x \in A_M} [|x-y| - |x-m(A)|] \leq (\text{rank}(m(A)) - \text{rank}(y)) \cdot |y-m(A)| \, .$$

Overall, using the symmetric case of $y > m(A)$, we have

$$\begin{aligned}\sum_{x \in A} [|x-y| - |x-m(A)|] = &\sum_{x \in A_L \cup A_U} [|x-y| - |x-m(A)|] \\ &+ \sum_{x \in A_M} [|x-y| - |x-m(A)|] \\ &\leq 2 \, |\text{rank}(y) - \text{rank}(m(A))| \cdot |y-m(A)| \, .\end{aligned}$$

■

Continue to follow notation on the 1-dimensional real line, and denote by $\mathsf{OPT}_1 \coloneqq \sum_{x \in A} |x - m(A)|$, and $\mathsf{ALG}_1 \coloneqq \sum_{x \in A} |x - m(S)|$. Now, we relate a difference in the rank to an approximation factor of the objective cost.

Lemma 2.2.7. *Let $y \in A \subset \mathbb{R}$ and $\frac{1}{n} \leq a < \frac{1}{2}$ s.t. $|\text{rank}(m(A)) - \text{rank}(y)| \leq (\frac{n+1}{2} - an)$. Then,*

$$\sum_{x \in A} |x-y| \leq \frac{1 - a + \frac{1}{2n}}{a - \frac{1}{2n}} \mathsf{OPT}_1.$$

Proof. Applying Lemma 2.2.6 and triangle inequality, we have,

$$\begin{aligned}\sum_{x \in A} |x-y| &\leq 2 \, |\text{rank}(y) - \text{rank}(m(A))| \cdot |y-m(A)| + \mathsf{OPT}_1 \\ &\leq 2(\frac{1}{2} + \frac{1}{2n} - a)n \, |y-m(A)| + \mathsf{OPT}_1 \\ &\leq (1 + \frac{1}{n} - 2a)(\sum_{x \in A} |x-y| + \mathsf{OPT}_1) + \mathsf{OPT}_1.\end{aligned}$$

Rearranging to obtain,

$$\sum_{x \in A} |x-y| \leq \frac{1 - a + \frac{1}{2n}}{a - \frac{1}{2n}} \mathsf{OPT}_1.$$

■

Corollary 2.2.8. *If y is an ε-rank-approximate median, for $\varepsilon \in \left(0, \frac{1}{10}\right)$, then,*

$$\sum_{x \in A} |x-y| \leq (1 + 5\varepsilon)\mathsf{OPT}_1.$$

Proof. By definition, since y is an ε-rank-approximate median then $|\text{rank}(y) - \text{rank}(m(A))| \leq \varepsilon n$. To use Lemma 2.2.7, we notice that $an = (\frac{n+1}{2} - \varepsilon n)$ and thus $a = (\frac{1}{2} + \frac{1}{n} - \varepsilon)$. Plugging in the formula provide the required result. ■

To complete the proof, we return to $\mathbb{R}^d$, again writing $S \subseteq A \subset \mathbb{R}^d$.

Proof of Theorem 2.2.4. Using Lemma 2.2.1 we know we can find the median of each coordinate separately. Using Lemma 2.2.5 with $\delta = 1/4d$, we obtain a sample size of $s = 2\varepsilon^{-2} \log 8d$. Then, using union bound over all d coordinates, we have that with probability at least $3/4$, for every coordinate $i \in [d]$, $m(S)_i$ is an ε-rank-approximate median of A_i. Then, if we assume this event happens, we have by Corollary 2.2.8 that all coordinates are $(1 + 5\varepsilon)$-approximation and thus in total also $m(S)$,

$$\begin{aligned}\sum_{x \in A} \|x - m(S)\|_1 &= \sum_{x \in A} \sum_{i=1}^{d} |x_i - m(S)_i| = \sum_{i=1}^{d} \sum_{x \in A} |x_i - m(S)_i| \\ &\leq \sum_{i=1}^{d} (1 + 5\varepsilon) \sum_{x \in A} |x_i - m(A)_i| = (1 + 5\varepsilon) \sum_{x \in A} \|x - m(A)\|_1 .\end{aligned}$$

■

We remark that the above lemmas and approach will also be useful now when we extend to the dimension-independent case.

Dimension-Independent Bound

Now, the main theorem of this section is a weak $(\varepsilon, 1)$-coreset under the ℓ_1 distance, with dimension-*independent* size which is obtained using uniform sampling.

Theorem 2.2.9. *Let $A \subset \mathbb{R}^d$ and $\varepsilon \in \left(0, \frac{1}{10}\right)$. Let S be a sample multi-set of size $s = 18\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon}$ points drawn independently and uniformly (with replacements) from A. Then with probability at least $\frac{3}{4}$, S is a weak $(\varepsilon, 1)$-coreset of A, i.e., for $m(S) \in \mathbb{R}^d$ the 1-median of S,*

$$\sum_{x \in A} \|x - m(S)\|_1 \leq (1 + O(\varepsilon)) \sum_{x \in A} \|x - m(A)\|_1 .$$

The analysis is again based on proving general results in one dimension, and in the end we sum over all of the coordinates. Given a uniform sample $S \subseteq A \subset \mathbb{R}$, and its median $m(S)$, we divide the set of possible ranks of $m(S)$ (with respect to A) into geometrically decreasing size intervals depending on their distance from $rank(m(A))$. For each interval, we compute the probability that $m(S)$ lands in the interval, and the approximation factor in such case. These are quantified in the following lemmas.

Lemma 2.2.10. *Let a, b be two real numbers s.t. $\frac{1}{n} \leq a < b < \frac{1}{2} + \frac{1}{2n}$. Let $I = [an, bn]$ be their corresponding interval (to the left of the $\text{rank}(m(A))$). If a multiset $S \subseteq A \subset \mathbb{R}$ is drawn uniformly at random, then,*

$$\mathbb{P}\left[\text{rank}(m(S)) \in I\right] \leq b^{|S|/2} 2^{|S|} = (4b)^{\frac{|S|}{2}}.$$

Proof. Denote by $B \sim \text{Binom}(|S|, b)$ and use direct computation using the binomial cumulative distribution function,

$$\begin{aligned}\mathbb{P}\left[\text{rank}(m(S)) \in I\right] &\leq \mathbb{P}\left[\text{rank}(m(S)) \leq bn\right] \leq \mathbb{P}\left[B \geq \frac{|S|}{2}\right] \\ &\leq \sum_{r=\frac{|S|+1}{2}}^{|S|} \binom{|S|}{r} b^r (1-b)^{|S|-r} \leq b^{|S|/2} \cdot 2^{|S|}.\end{aligned}$$

■

Now, assuming $\varepsilon \leq 0.1 < \frac{3}{8} + \frac{1}{2n}$, we define the following intervals to cover $[1, n]$.

Definition 2.2.11 (Covering Intervals)**.** Let $\varepsilon \leq 0.1$. The following intervals are called *covering intervals* of $[1, n]$.

1. $I^M = [\frac{n+1}{2} - \varepsilon n, \frac{n+1}{2} + \varepsilon n]$.
2. $I^L = [\frac{n}{2^3}, \frac{n+1}{2} - \varepsilon n]$ and $I^R = [\frac{n+1}{2} + \varepsilon n, n - \frac{n}{2^3} + 1]$.
3. For every integer $\lambda \in [3, \log n - 1]$, write

$$I_\lambda = I^L_\lambda \cup I^R_\lambda = [\frac{n}{2^{\lambda+1}}, \frac{n}{2^\lambda}] \cup [n - \frac{n}{2^\lambda} + 1, n - \frac{n}{2^{\lambda+1}} + 1].$$

Remark. The endpoints of the intervals might not be integral, but it should not matter for the analysis as long as the union of all intervals cover $[1, n]$.

Again, we denote by $\mathsf{OPT}_1 \coloneqq \sum_{x \in A} |x - m(A)|$, and $\mathsf{ALG}_1 \coloneqq \sum_{x \in A} |x - m(S)|$. For the following results, we consider points with rank smaller than $\frac{n+1}{2}$. Analogues results can be proven symmetrically for the rest of the points with rank larger than $\frac{n+1}{2}$.

Corollary 2.2.12. *For the covering intervals, if* $\mathrm{rank}(m(S)) \in I^L_\lambda$*, then* $\mathsf{ALG}_1 \leq 2^{\lambda+3}\mathsf{OPT}_1$.

Proof. Let λ be an integer in $[3, \log n - 1]$. Notice that if $\mathrm{rank}(m(S)) \in I^L_\lambda$, then

$$|\mathrm{rank}(m(S)) - \mathrm{rank}(m(A))| \leq (\frac{n+1}{2} - \frac{n}{2^{\lambda+1}}).$$

Thus, taking $a = \frac{1}{2^{\lambda+1}}$ for Lemma 2.2.7 concludes the proof by bounding the approximation factor as follows,

$$\frac{1 - \frac{1}{2^{\lambda+1}} + \frac{1}{2n}}{\frac{1}{2^{\lambda+1}} - \frac{1}{2n}} = \frac{2^{\lambda+1} - 1 + \frac{2^\lambda}{n}}{1 - \frac{2^\lambda}{n}} \leq \frac{2^\lambda(2 + \frac{1}{n})}{0.5} \leq 2^{\lambda+3}.$$

■

Additionally, we will need the following technical lemma.

Lemma 2.2.13. *For* $|S| \geq 2\log\frac{1}{\varepsilon} + 18$*, and* $\lambda \geq 3$*, it holds,* $2^{\lambda+3} \cdot \left(\frac{1}{2^{\lambda-2}}\right)^{|S|/2} \leq \frac{1}{2^\lambda}\varepsilon$

Proof. One can write using $|S| = 2\log\frac{1}{\varepsilon} + 18$,

$$\left(\frac{1}{2^{\lambda-2}}\right)^{|S|/2} = \left(\frac{1}{2^{\lambda-2}}\right)^5 \left(\frac{1}{2^{\lambda-2}}\right)^3 \left(\frac{1}{2^{\lambda-2}}\right)^{\log\varepsilon^{-1}} \leq \frac{1}{2^{\lambda+3}} \cdot \frac{1}{2^\lambda} \left(\frac{1}{2}\right)^{\log\varepsilon^{-1}} \leq \frac{1}{2^{\lambda+3}} \cdot \frac{1}{2^\lambda} \cdot \varepsilon$$

Multiplying by $2^{\lambda+3}$ concludes the proof. ■

Lemma 2.2.14. *For every multiset* $S \subset A$ *of real numbers, drawn uniformly at random of size* $|S| = 18\varepsilon^{-2}\log\varepsilon^{-1}$*, it holds,*

$$\mathbb{E}\left[\mathsf{ALG}_1\right] \leq (1 + O(\varepsilon))\mathsf{OPT}_1.$$

Proof. Dividing the possible options for $rank(m(S))$ according to the above covering intervals and using the law of total expectation with symmetry w.r.t to the median rank $\frac{n+1}{2}$, we have,

$$\begin{aligned}\mathbb{E}\left[\mathsf{ALG}_1\right] \leq\ & \mathbb{E}\left[\mathsf{ALG}_1 \mid \operatorname{rank}(m(S)) \in I^M\right] \cdot \mathbb{P}\left[\operatorname{rank}(m(S)) \in I^M\right] + \\ & 2\mathbb{E}\left[\mathsf{ALG}_1 \mid \operatorname{rank}(m(S)) \in I^L\right] \cdot \mathbb{P}\left[\operatorname{rank}(m(S)) \in I^L\right] + \\ & 2\sum_{\lambda=3}^{\log n-1} \mathbb{E}\left[\mathsf{ALG}_1 \mid \operatorname{rank}(m(S)) \in I^L_\lambda\right] \cdot \mathbb{P}\left[\operatorname{rank}(m(S)) \in I^L_\lambda\right].\end{aligned}$$

The first term can be bounded by using Lemma 2.2.7,

$$\mathbb{E}\left[\mathsf{ALG}_1 | \operatorname{rank}(m(S)) \in I_M\right] \cdot \mathbb{P}\left[\operatorname{rank}(m(S)) \in I_M\right] \leq (1+5\varepsilon)\mathsf{OPT}_1.$$

Bounding the second term follows by using Lemma 2.2.7 with $a = \frac{1}{2^3}$ for the expectation term, and using Equation (2.2) for the probability,

$$2\mathbb{E}\left[\mathsf{ALG}_1 | \operatorname{rank}(m(S)) \in I_L\right] \cdot \mathbb{P}\left[\operatorname{rank}(m(S)) \in I_L\right] \leq 32\mathsf{OPT}_1 \cdot \varepsilon.$$

To bound the third term, we use Corollary 2.2.12 and Lemma 2.2.10 with $b = \frac{1}{2^\lambda}$,

$$2\sum_{\lambda=3}^{\log n-1} \mathbb{E}\left[\mathsf{ALG}_1 | \operatorname{rank}(m(S)) \in I^L_\lambda\right] \cdot \mathbb{P}\left[\operatorname{rank}(m(S)) \in I^L_\lambda\right] \leq 2\sum_{\lambda=3}^{\log n-1} 2^{\lambda+3}\mathsf{OPT}_1 \cdot \left(\frac{1}{2^{\lambda-2}}\right)^{|S|/2}.$$

Since $|S| = 18\varepsilon^{-2}\log\varepsilon^{-1} > 2\log\varepsilon^{-1} + 18$ for small enough ε, then by Lemma 2.2.13 each summand is at most $\frac{1}{2^\lambda}\varepsilon\mathsf{OPT}_1$, and their sum is bounded by $4\varepsilon\mathsf{OPT}_1$.
Overall,

$$\mathbb{E}\left[\mathsf{ALG}_1\right] \leq (1+5\varepsilon)\mathsf{OPT}_1 + 32\varepsilon\mathsf{OPT}_1 + 4\varepsilon\mathsf{OPT}_1 \leq (1+41\varepsilon)\mathsf{OPT}_1.$$

■

Finally, to conclude the proof of the main theorem, we switch back to $\mathbb{R}^d$. The 1-dimensional values of the i'th coordinate are denoted by $\mathsf{ALG}^{(i)} := \sum_{x\in A_i} |x - m(S)|_1$ and $\mathsf{OPT}^{(i)} := \sum_{x\in A_i} |x - m(A)|_1$. Hence, we can write the values of the d-dimensional case with $\mathsf{OPT} := \sum_{i=1}^d \mathsf{OPT}^{(i)}$, and $\mathsf{ALG} := \sum_{i=1}^d \mathsf{ALG}^{(i)}$.

Proof of Theorem 2.2.9. Let $A \subset \mathbb{R}^d$. Using linearity of expectation with Lemma 2.2.14, we obtain,

$$\mathbb{E}\left[\mathsf{ALG}\right] = \sum_{i=1}^d \mathbb{E}\left[\mathsf{ALG}^{(i)}\right] \leq \sum_{i=1}^d (1 + \underbrace{41\varepsilon}_{=:\varepsilon'})\mathsf{OPT}^{(i)} = (1+\varepsilon')\mathsf{OPT}.$$

Now, given $t \geq 1$, we have by Markov inequality,

$$\begin{aligned}\mathbb{P}[\mathsf{ALG} \geq (1+t\varepsilon')\mathsf{OPT}] &= \mathbb{P}\left[\mathsf{ALG} - \mathsf{OPT} \geq t\varepsilon'\mathsf{OPT}\right] \\ &\leq \frac{\mathbb{E}\left[\mathsf{ALG} - \mathsf{OPT}\right]}{t\varepsilon'\mathsf{OPT}} \leq \frac{\varepsilon'\mathsf{OPT}}{t\varepsilon'\mathsf{OPT}} = \frac{1}{t}.\end{aligned}$$

Conclude, by choosing $t = 4$, that with probability at least $\frac{3}{4}$,

$$\sum_{x\in A} \|x - m(S)\|_1 \leq (1+4\varepsilon')\mathsf{OPT}.$$

■

Remark (Amplification). It is possible to perform a "median-trick" to amplify the constant success probability to at least $1-\delta$ probability. It is performed by sampling $q = \log 1/\delta$ independent uniform samples $S_1, \ldots, S_q \subseteq A$, and then choosing the sample S_j, s.t. $\operatorname{cost}(A, m(S_j))$ is the median objective score with respect to all other samples.

2.3 Weak $(\varepsilon, 2)$-Coreset for Separable Instance

This section considers a set $A \subset \mathbb{R}^d$ of size n under either ℓ_2 or ℓ_1 distance. The main result of this section (Theorem 2.3.1) is a weak (ε, k)-coreset for $k = 2$ with dimension-*independent* size. Construction is again performed by using only uniform sampling, and assuming the instance satisfies the α-center proximity (Definition 1.1.2) and η-balanced (Definition 1.1.3) conditions. The analysis will use the results from previous sections (2.1, 2.2). Denote by s_1, s_2 the sizes of a weak $(\varepsilon, 1)$-coreset obtained by uniform sampling in either ℓ_1 or ℓ_2 respectively (specified by Theorems 2.1.5 ,2.2.9), when using a failure probability of $\delta = \frac{1}{8k}$. Specifically, for $p \in \{1, 2\}$,

$$s_p = \begin{cases} O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon} \log k\right) & p = 1 \\ O\left(\max\left\{\varepsilon^{-4} \log^2\left(\varepsilon^{-1}k\right) \log\left(\varepsilon^{-3} \log\left(\varepsilon^{-1}k\right)\right), k\right\}\right) & p = 2 \end{cases}.$$

The most relevant weak (ε, k)-coreset construction that we are aware of, with a slightly different definition, is due to [FMS07]. There, the algorithm performs uniform sampling in a small ring around approximate centers, and importance sampling of points farther away from the approximate centers. In our approach, we consider inputs that satisfies two additional intertwined conditions, and each sample is chosen *uniformly at random*, without the need for computing approximate centers beforehand.

1. *Balanced* optimal clusters. The input instance A satisfy the η-balanced optimal property (Definition 1.1.3). That is, for $k = 2$, and every optimal clustering A_1^*, A_2^*, each optimal cluster has size at least $\eta\frac{n}{2}$, for some $\eta \in (0, 1]$.

2. *Separable* optimal clusters: For $k = 2$, the input instance satisfy the α-center proximity property (Definition 1.1.2), for $\alpha \geq \frac{9k}{\eta} + 5$.

We emphasize that the definitions require that *every* optimal solution is balanced and separable. Additionally, the balanced clusters condition is similar to the assumption used [MOP04]. In short, they showed [MOP04, Section 7] that if we do not have a lower bound on the smallest optimal cluster, then the probability of uniformly sampling points from the small optimal cluster is also very small. Thus, the structure of that clusters is not captured, and it might result in large contribution to the cost if the inter-cluster distance is large.

We are now ready to state the following main result of this section.

Theorem 2.3.1. *Let $A \subseteq \mathbb{R}^d$ in $(\mathbb{R}^d, \ell_p)$ for $p \in \{1, 2\}$, let $k = 2$, and let A_1^*, A_2^* be any two optimal clusters. Assume A satisfies the η-balanced optimal clusters for $\eta \in (0, 1]$ and α-center proximity for $\alpha \geq \frac{9k}{\eta} + 5$. Then, a uniform sample multiset $S \subseteq A$ of size $s = 27k \log(16k)\frac{1}{\eta}s_p$ is a weak $(\varepsilon, 2)$-coreset with probability at least $\frac{3}{4}$, i.e., for every optimal 2-median solution $\{m_1^*(S), m_2^*(S)\}$ of S, it holds,*

$$\text{cost}(A, \{m_1^*(S), m_2^*(S)\}) \leq (1 + \varepsilon)\mathsf{OPT}(A, 2).$$

The analysis is split into two stages. At first, we show that with good probability, every sampled cluster $S_i := A_i^* \cap S$ is large enough and is a weak $(\varepsilon, 1)$-coreset of the corresponding optimal cluster A_i^*. The second stage shows that the set $\{m(S_1), m(S_2)\}$ is an optimal 2-median solution of S. Then by the separability condition, each S_i is actually an optimal cluster for 2-median on S. Combining the above two steps, will conclude the proof.

Probability Analysis. Let $k \geq 2$ be an integer. The following lemma show that for every $i, j \in [k]$, the two sampled clusters S_i, S_j are large enough, i.e., $|S_i| \geq s_p$, and that approximately the same size, up to some constant, i.e., $|S_i| \approx |S_j|$.

Lemma 2.3.2. *Let $A \subset \mathbb{R}^d$, k a positive integer and $p \in \{1,2\}$. Assume that for every $i \in [k]$, the optimal cluster A_i^* is of size at least $\eta\frac{n}{k}$ for $\eta \in (0,1]$. Then with probability at least $\frac{7}{8}$, a uniform sample $S \subseteq A$ of size $s = 27k\log(16k)\frac{1}{\eta}s_p$ samples at least $\frac{2}{3}\eta\frac{s}{k}$ points from every optimal cluster A_i^*.*

Proof. Consider a single optimal cluster A_i^*. Define the number of points sampled from the optimal cluster A_i^* by $N_i = \sum_{j=1}^{s} \mathbb{1}[x_j \in A_i^*]$ where x_j is the j'th point sampled into S. Every A_i^* is of size at least $\eta\frac{n}{k}$, thus we sample a point from A_i^* with probability at least $\frac{\eta}{k}$ and we have that $\mathbb{E}N_i \geq s\frac{\eta}{k}$. Hence, by Chernoff bound with precision parameter of $\frac{1}{3}$,

$$\mathbb{P}\left[N_i \notin \left[\frac{2}{3},\frac{4}{3}\right]\mathbb{E}N_i\right] \leq 2\exp\left(-\frac{\mathbb{E}N_i}{27}\right) \leq \frac{1}{8k}.$$

Now, using union bound, we claim that with probability at least $\frac{7}{8}$, the number of samples in every optimal cluster fall in the above range, i.e.,

$$\mathbb{P}\left[\forall i \in [k]\ ,\ |S_i| \in \left[\frac{2}{3},\frac{4}{3}\right]EN_i\right] \geq \frac{7}{8}$$

■

Corollary 2.3.3. *With probability at least $\frac{7}{8}$, for every $i \in [k]$, it holds that $|S_i| \geq \frac{2}{3}|S|\frac{\eta}{k} \geq s_p$.*

We can now show, that with constant probability, for every $i \in [k]$ the sample set S_i is a weak $(\varepsilon,1)$-coreset of its respective optimal cluster A_i^*.

Lemma 2.3.4. *Let $A \subset \mathbb{R}^d$, k a positive integer and $p \in \{1,2\}$. Assume that for every $i \in [k]$, the optimal cluster A_i^* is of size at least $\eta\frac{n}{k}$ for $\eta \in (0,1]$. Then with probability at least $\frac{3}{4}$, a uniform sample $S \subseteq A$ of size $s = 27k\log(16k)\frac{1}{\eta}s_p$ can be partitioned into $S = S_1 \cup S_2 \cup \cdots \cup S_k$, where for every $i \in [k]$, it holds that S_i is a weak $(\varepsilon,1)$-coreset A_i^*.*

Proof. By Corollary 2.3.3, with probability at least $\frac{7}{8}$, there are at least s_p samples from every optimal cluster A_i^*, and thus the weak $(\varepsilon,1)$-coreset condition is satisfied. Since the failure probability of $\delta = \frac{1}{8k}$ was chosen for the size of s_p, then, using union bound, with probability at least $\frac{7}{8}$, for every $i \in [k]$, the sample set S_i is a weak $(\varepsilon,1)$-coreset of A_i^*. Conclude in total, again by union bound, that with probability at least $\frac{3}{4}$, for every $i \in [k]$, the sample set S_i is a 1-median weak coreset for the optimal cluster A_i^*. ■

Remark. The probability analysis was performed for general k, and might be useful assuming the approximation analysis can be generalized for $k > 2$.

Approximation Analysis. Let $k = 2$ and $p \in \{1,2\}$. We use the following notation. Let $i \in \{1,2\}$. Then $R_i \coloneqq \max_{x \in A_i^*} \|x - m(A_i^*)\|_p$ is the distance of $m(A_i^*)$ to the farthest point in its cluster, and write $R \coloneqq \max_{j\in\{1,2\}} R_j$. For $\gamma = \frac{3k}{\eta} + 1$ we define $B_i \coloneqq Ball(m(A_i^*), \gamma R) = \{x \in \mathbb{R}^d \ :\ \|x - m(A_i^*)\|_p \leq \gamma R\}$. The balls of radius γR around the centers $m(A_i^*)$, will allow us to argue about ambient space centers (it will soon be clear how to find the constraints on γ with respect to α and their value). Clearly, by definition, $S_i \subseteq A_i \subseteq B_i$. We assume that $R > 0$. Otherwise, for $R = 0$, the dataset is comprised of two distinct points with multiplicities, and thus the 2-median is just these two points.

We begin by stating useful properties of the α-center proximity assumption (Definition 1.1.2). The first one is a lower bound of the distance between any two optimal centers.

Claim 2.3.5. *Let $A \subset \mathbb{R}^d$ be a 2-median instance satisfying the α-center proximity property. Then, $(\alpha - 1)R < \|(m(A_1^*) - m(A_2^*)\|_p$.*

Proof. Let $i \neq j \in [2]$. By α-center proximity and triangle inequality, for every $x \in A_i^*$,

$$\alpha \left\|x - m(A_i^*)\right\|_p < \left\|x - m(A_j^*)\right\|_p \leq \left\|x - m(A_i^*)\right\|_p + \left\|m(A_i^*) - m(A_j^*)\right\|_p .$$

Hence, for every $x \in A_i^*$ and every center $j \neq i$, we have,

$$(\alpha - 1) \left\|x - m(A_i^*)\right\|_p < \left\|m(A_i^*) - m(A_j^*)\right\|_p .$$

Conclude the proof by taking the point x realizing R. ■

Using this lower bound, we have that any two points from distinct balls are far away. Then, conclude that B_1, B_2 are disjoint for large enough α, i.e., if $\gamma < \frac{1}{2}(\alpha - 1)$.

Claim 2.3.6. *Let $A \subset \mathbb{R}^d$ be a 2-median instance satisfying the α-center proximity property. For every $y \in B_1$, $z \in B_2$, $\|y - z\|_p > (\alpha - 1 - 2\gamma)R$*

Proof. Immediately by triangle inequality and Claim 2.3.5,

$$\|y - z\|_p \geq \|m(A_1^*) - m(A_2^*)\|_p - \|y - m(A_1^*)\|_p - \|z - m(A_2^*)\|_p > (\alpha - 1)R - 2\gamma R.$$

■

An additional required property, is that using the nearest neighbour rule, with arbitrary tie breaking, $m(S_1), m(S_2)$ separates A into A_1^*, A_2^*, and S into S_1, S_2.

Lemma 2.3.7. *Let $A \subset \mathbb{R}^d$ be a 2-median instance satisfying the α-center proximity property. Assume $\gamma \leq \frac{\alpha-2}{3}$. Then every $y \in B_1, z \in B_2$ partition A into A_1^*, A_2^* and S into S_1, S_2.*

Proof. Without loss of generality, let $x \in A_1^*$ (the case of $x \in A_2^*$ is symmetric).
First, by triangle inequality and the definition of R and the balls,

$$\|x - y\|_p \leq \|x - m(A_1)\|_p + \|m(A_1) - y\|_p \leq R + \gamma R \leq (\gamma + 1)R.$$

On the other hand, since $x \in B_1$, $z \in B_2$, and using Claim 2.3.6, we have

$$\|x - z\|_p > (\alpha - 1 - 2\gamma)R.$$

Since $\gamma \leq \frac{\alpha-2}{3}$, it holds that $\|x - z\|_p \geq \|x - y\|_p$. Overall, A is partitioned into A_1^* and A_2^*, and similarly S is partitioned into S_1 and S_2 since for every $i \in [2]$, $S_i \subseteq A_i$. ■

The following corollary requires an additional bound of the distance between the optimal and approximate centers. Since $S_i \subseteq A_i \subseteq Ball(m(A_i^*), R_i)$, then for $\gamma \geq 1$, and for every $i \in [2]$, and using convexity, $m(S_i) \in Ball(m(A_i^*), R_i) \subseteq B_i$. Overall, $m(A_i), m(S_i) \in B_i$, and using Lemma 2.3.7, we obtain the following corollary.

Corollary 2.3.8. *Let $A \subset \mathbb{R}^d$ be a 2-median instance satisfying the α-center proximity property. Assume $\gamma \in \left[1, \frac{\alpha-2}{3}\right]$. Then both sets of 2-centers $\{m(A_1^*), m(A_2^*)\}$ and $\{m(S_1), m(S_2)\}$ partition A into A_1^*, A_2^* and S into S_1, S_2.*

Additional corollary is the following bound for the cost of S using the centers $\{m(S_1), m(S_2)\}$.

Corollary 2.3.9. *Let $A \subset \mathbb{R}^d$ be a 2-median instance satisfying the α-center proximity property. Assume $\gamma \in \left[1, \frac{\alpha-2}{3}\right]$. Then cost of S with centers $\{m(S_1), m(S_2)\}$ satisfies*

$$\text{cost}(S, \{m(S_1), m(S_2)\}) = \mathsf{OPT}(S_1, 1) + \mathsf{OPT}(S_2, 1) \leq 2|S_1|R_1 + 2|S_2|R_2 \leq 2|S|R.$$

Now, we turn to the two main lemmas that show that only 2-medians of the form $\{m(S_1), m(S_2)\}$ realize the optimal cost. Hence, the final conclusion is that any $\{m(S_1), m(S_2)\}$ induce an optimal partition S_1^*, S_2^*. By Corollary 2.3.8, this partition is S_1, S_2, and we have that $S_1^* = S_1$ and $S_2^* = S_2$ (up to permutation).

Lemma 2.3.10. *Let $A \subset \mathbb{R}^d$ be a 2-median instance satisfying the α-center proximity and the η-balanced conditions. Let $S \subseteq A$ be a uniform sample, where $m(S_i)$ is the median of $S_i = S \cap A_i^*$. Assume that the event described in Lemma 2.3.2 occur and take $\gamma \in \left[\frac{6}{\eta} + 1, \frac{\alpha-2}{3}\right]$. Then $m(S_1), m(S_2)$ realize the optimal 2-median cost of S.*

Proof. The goal of the proof is to show that for every $c_1, c_2 \in \mathbb{R}^d$, it holds that $\{m(S_1), m(S_2)\}$ achieves at least the same cost, i.e., need to show:

$$\forall c_1, c_2 \in \mathbb{R}^d \ , \ \text{cost}(S, \{c_1, c_2\}) \geq \text{cost}(S, \{m(S_1), m(S_2)\}.$$

Let $c_1, c_2 \in \mathbb{R}^d$. We divide into two cases, according to the way we choose c_1, c_2, depending if they belong to B_1, B_2.

- Case 1: Each ball has a representative center, i.e. $c_1 \in B_1, c_2 \in B_2$.
  By Lemma 2.3.7, c_1, c_2 partition S into S_1, S_2. Thus, we can lower bound the cost using c_1, c_2, since $m(S_1), m(S_2)$ minimize the sum of distances over S_1, S_2 respectively.

$$\begin{aligned}
\text{cost}(S, \{c_1, c_2\}) &= \sum_{x \in S_1} \|x - c_1\| + \sum_{x \in S_2} \|x - c_2\|_p && \text{Lemma 2.3.7} \\
&\geq \sum_{x \in S_1} \|x - m(S_1)\|_p + \sum_{x \in S_2} \|x - m(S_2)\|_p && \text{Optimality} \\
&= \text{cost}(S, \{m(S_1), m(S_2)\}). && \text{Corollary 2.3.9}
\end{aligned}$$

- Case 2: There is at least one ball without a representative center. Assume without loss of generality that $c_1, c_2 \notin B_2$.
  In this case, all points of S_2 need to be represented by either c_1 or c_2. In either case, since both c_1, c_2 do not belong in B_2, by triangle inequality it holds that for every $x \in S_2$ and every $i \in \{1, 2\}$,

$$\|x - c_i\|_p \geq \|c_i - m(A_2)\|_p - \|x - m(A_2)\|_p > (\gamma - 1)R.$$

  Summing up to obtain,

$$\text{cost}(S, \{c_1, c_2\}) > |S_2|(\gamma - 1)R.$$

  On the other hand, by Corollary 2.3.9,

$$\text{cost}(S, \{m(S_1), m(S_2)\}) \leq 2|S|R.$$

  Finally, for $k = 2$, and using Corollary 2.3.3 that $|S|$ is at most $\frac{3k}{2\eta}|S_2|$, it holds that for $\gamma \geq \frac{3k}{\eta} + 1$, the required inequality $2|S|R \leq |S_2|(\gamma - 1)R$ is satisfied.

Overall, the goal is achieved for both cases, hence for every $c_1, c_2 \in \mathbb{R}^d$. ■

Now, we need to show that under the separability assumption, every optimal 2-centers of S for the 2-median objective are actually two 1-medians for S_1 and S_2 separately.

Lemma 2.3.11. *Let $A \subset \mathbb{R}^d$ be a 2-median instance satisfying the α-center proximity and the η-balanced conditions. Let $S \subseteq A$ be a uniform sample, where $m_1^*(S), m_2^*(S)$ are optimal 2-median for S. Assume that the event described in Lemma 2.3.2 occur and take $\gamma \in \left[\frac{6}{\eta} + 1, \frac{\alpha-2}{3}\right]$. Then for every $i \in [2]$, one of the optimal 2-centers is a 1-median of S_i.*

Proof. By the analysis of Lemma 2.3.10, we claim that $m_1^*(S), m_2^*(S)$ belong to B_1 and B_2 since any two centers c_1, c_2 such that both do not belong to one of the balls have cost strictly larger than $\mathsf{OPT}(S,2)$. Assume without loss of generality that $m_1^*(S) \in B_1, m_2^*(S) \in B_2$. Thus, by Lemma 2.3.7, assuming $\gamma \leq \frac{\alpha-2}{3}$, $m_1^*(S), m_2^*(S)$ partitions S into S_1, S_2 and we have,

$$\mathsf{OPT}(S,2) = \mathrm{cost}(S, \{m_1^*(S), m_2^*(S)\}) = \sum_{x \in S_1} \|x - m_1^*(S)\|_p + \sum_{x \in S_2} \|x - m_2^*(S)\|_p .$$

If we assume that one of the 2-medians of S *isn't* a 1-median of S_1 or S_2, then the above cost can be strictly improved, resulting in a contradiction to the optimality of m_1^*, m_2^*. ■

Finally, for $k = 2$, we choose $\gamma = \frac{3k}{\eta} + 1$ and $\alpha = \frac{9k}{\eta} + 5$ which satisfies all conditions encountered above, and we can prove our main theorem.

Proof of Theorem 2.3.1. By Lemmas 2.3.2 and 2.3.4, we have that with probability at least $\frac{3}{4}$, each optimal cluster was sampled at least $\frac{2}{3}\eta\frac{s}{k}$ times, and S_i serves as a weak coreset for A_i^*. Assuming this event occurs, and using Lemma 2.3.11, every optimal 2-medians of S can be viewed as two 1-medians of S_1, S_2 separately, i.e. $\mathsf{OPT}(S,2) = \mathrm{cost}(S, \{m(S_1), m(S_2)\})$. Finally, by Corollary 2.3.8, we know that $m(S_1), m(S_2)$ separates A into A_1^*, A_2^*. Overall,

$$\begin{aligned}
\mathrm{cost}(A, \{m_1^*(S), m_2^*(S)\}) &= \mathrm{cost}(A, \{m(S_1), m(S_2)\}) && \text{Lemma 2.3.11} \\
&= \sum_{i=1}^{2} \sum_{x \in A_i} \|x - m(S_i)\|_p && \text{Corollary 2.3.8} \\
&\leq \sum_{i=1}^{2} (1+\varepsilon) \sum_{x \in A_i^*} \|x - m(A_i^*)\|_p && \text{weak } (\varepsilon,1)\text{-coreset} \\
&= (1+\varepsilon)\,\mathrm{cost}(A, \{m(A_1^*), m(A_2^*)\}).
\end{aligned}$$

■

2.4 Discussion

It is important to extend these results to $(1+\varepsilon)$-approximate medians of S, since the results in this chapter currently only apply to the optimal medians of the sample S. Furthermore, the inherent power of uniform sampling is still unresolved. We showed that uniform sampling can produce weak $(\varepsilon,1)$-coresets in the metric spaces $(\mathbb{R}^d, \ell_p)$ for $p \in \{1,2\}$. Thus, it is straightforward to ask whether this technique also applies to other metric spaces.

The analysis of the weak (ε,k)-coreset result for $k = 2$ only require that α is large enough, namely $\alpha \geq \frac{9k}{\eta} + 5$. It is emphasized that both the size of the sample $|S|$, and the approximation factor ε are independent of α. Hence, one can ask what happen if we use sample size $|S|$ or an approximation factor that depends also on the separability parameter. Can we improve the analysis in such case? Can we decouple the dependency of α in k and η? It is also interesting to understand whether we can prove a similar result for inputs that only satisfy the η-balanced property, or at least reduce the dependency in the constants of α?

Finally, an obvious direction for future work is to generalize the current proof of weak $(\varepsilon,2)$-coreset for general $k > 2$. We remark that if the input satisfy an additional structural property, it is indeed possible. In short, consider a ball of radius γR_i around each optimal median $m(A_i^*)$, and by writing the radii of the optimal clusters of A in decreasing magnitude, i.e. $R_1 \geq R_2 \geq \cdots \geq R_k$, and assume that the input satisfies $\frac{R_k}{R_1} \geq \mu$ for some constant μ. Then, it is possible to obtain a weak (ε,k)-coreset, for $\alpha \geq \frac{9k}{\eta\mu} + 5$ by using only uniform sampling and with the same size guarantee.

Chapter 3

From Rank Aggregation to Clustering Permutations

In this chapter we consider the k-rank-aggregation problem, that is the generalization of the rank-aggregation to a k-median objective. In Section 3.1 we devise a strong (ε, k)-coreset construction (Definition 1.0.1) for an input set $A \subseteq \mathcal{P}_N$. It uses a known construction of [BJKW21] and an embedding of the permutations into ℓ_2^2 [JKS16]. This coreset is later used as a component of a PTAS for the k-rank-aggregation problem for two special cases. The first one, presented in Section 3.2, is a PTAS for small (e.g. constant) k. This FPT-PTAS in k enumerates over all possible partitions of a strong (ε, k)-coreset of the input A. If ε is treated as a constant, this results in running time that is polynomial in both n and N and is exponential in k. The second case applies for a restricted sets of inputs that satisfy the β-distributed condition (Definition 1.1.6), and is examined in Section 3.3. We propose a variation of an algorithm from [ABS10], that produces a $(1+\varepsilon)$-factor approximate solution in time that is polynomial in n, N, k where ε, β are treated as constants. In contrast to the algorithm of [ABS10], our modification allows the selection of ambient points as centers and is achieved by using a a strong $(\varepsilon, 1)$-coreset for each cluster with high cost separately.

Preliminaries. The input permutation set is denoted by $A = \{\sigma_1, \ldots, \sigma_n\} \subseteq \mathcal{P}_N$, and it describes a set of n voters, that each ranks N different candidates. We use $\sigma[i]$ to denote the location of candidate i in the ranking σ. A naive computation of the Kendall-tau distance takes $O(N^2)$ time, but we will assume access to a distance oracle that computes the Kendall-tau distance between two permutations in $O(N \log N)$ time (for example, by using merge-sort). This can be further improved to $O(N\sqrt{\log N})$ time by [CP10], though not crucial to our application since it is the same up to poly-logarithmic factors.
Now, we define an embedding of the Kendall-tau distance into $(\mathbb{R}^{\binom{N}{2}}, \ell_2^2)$ semi-metric space, and show this embedding is an isometric embedding.

Definition 3.0.1 (Kemeny Embedding [JKS16])**.** The Kemeny embedding is a mapping of permutations over N items into a binary vector of length $\binom{N}{2}$. It is defined by $\psi : \mathcal{P}_N \to \{0,1\}^{\binom{N}{2}}$, where for every $\sigma \in \mathcal{P}_N$, and every $i < j \in [N]$, it holds

$$\psi(\sigma)[i,j] = \begin{cases} 0 & \text{if } \sigma[i] < \sigma[j] \\ 1 & \text{if } \sigma[j] < \sigma[i] \end{cases}$$

Claim 3.0.2 (Kemeny Embedding is Isometric)**.** *For every two permutations $\sigma, \rho \in \mathcal{P}_N$,*

$$K_\tau(\sigma, \rho) = \|\psi(\sigma) - \psi(\rho)\|_2^2$$

Proof. Let $\sigma, \rho \in \mathcal{P}_N$. By the definition of Kendall-tau,

$$K_\tau(\sigma, \rho) = \sum_{i<j \in [N]} \mathbb{1}\left\{i, j \text{ in opposite order in } \sigma \text{ and } \rho\right\}.$$

If σ and ρ disagree on the order of a pair of items i, j, then $\psi(\sigma)[i, j] \neq \psi(\rho)[i, j]$. Hence,

$$\|\psi(\sigma) - \psi(\rho)\|_2^2 = \sum_{i<j \in [N]} \mathbb{1}\left\{\psi(\sigma)[i, j] \neq \psi(\rho)[i, j]\right\}^2$$

Conclude that ψ is an isometric embedding into ℓ_2^2. ■

We remark that ψ and its inverse ψ^{-1} (for every point in $\psi(\mathcal{P}_n)$) are efficiently computable in $O(N^2)$ time. Additionally, this embedding can also be seen as an isometric embedding into ℓ_1, but the ℓ_2^2 point of view will be more useful later for our coreset construction.

Additionally, we assume access to the 1-rank-aggregation PTAS from [KS07, Sch12]. We denote an invocation of this PTAS on a set $A \subseteq \mathcal{P}_N$ with precision parameter ε by $F(A, \varepsilon)$, and we additionally use the notation $F((A, W), \varepsilon)$ in cases where A is a weighted set of permutations. Whether we use the non-weighted or weighted case should be clear from the context. Extending the PTAS to weighted instances is important since we will be applyig it on strong coresets, which are weighted sets. It is performed by adding the weights of the rankings to the edge-weighted tournament graph that the PTAS uses internally. The only requirement from this tournament is that the edges weights satisfy that $w_{u,v} + w_{v,u} = 1$ for every pair $u, v \in [N], u \neq v$. Indeed the requirement holds since the weight of each edge (u, v) is defined as $w_{u,v} = \frac{1}{W}\sum_{\sigma \in A} w_\sigma \mathbb{1}\left[\sigma(u) < \sigma(v)\right]$ where $W = \sum_{\sigma \in A} w_\sigma$. The PTAS running time is according to its derandomized version (See Corollary 7.3 by [Sch12]), hence, given a set of n rankings, it produces a $(1+\varepsilon)$-approximation for their Kemeny-aggregation, in time

$$T_{rank}(n, N, \varepsilon) = O\left(\frac{N^3 \log N}{\varepsilon}\right) + N 2^{\tilde{O}(1/\varepsilon^{12})} + O(nN^2). \tag{3.1}$$

3.1 Strong Coresets in Permutations Space

This section presents a strong (ε, k)-coreset construction for a set A in the $(\mathcal{P}_N, K_\tau)$ metric space. The main components are the isometric Kemeny embedding ψ (Definition 3.0.1) and a coreset construction of dimension-independent size introduced by [BJKW21]. Generally, the algorithm in [BJKW21] computes a coreset for Euclidean (k, z)-clustering whose goal is to find a center set $\mathcal{C}$ of size k that minimizes the objective $\sum_{x \in A} D(x, \mathcal{C})^z$. For $z = 2$, the objective is also known as the k-means problem and their construction only needs to use the distance D between the points. The coreset size is denoted by $s(k, \varepsilon) = \tilde{O}(\varepsilon^{-4} k^2 \log \frac{1}{\delta})$ where δ is the failure probability, and the running time is denoted by $T(n, k, d, \varepsilon) = \tilde{O}(nkd)$ where d is the dimension of the input. For more details, we refer the reader to Corollary 4.18 in [BJKW21, Section 4.3].

Remark. Coreset for k-means of dimension-independent size was also previously designed by [BFL16, FSS20]. We use the subsequent work of [BJKW21] to have simple bounds on the weights of the coreset points as seen in Lemma 3.1.3.

The following algorithm constructs a strong (ε, k)-coreset for a set of permutations A with respect to the k-rank-aggregation objective. In step 1, it invokes the algorithm from [BJKW21] in a black-box manner, with $z = 2$ and the square root of the Kendall-tau distance. This algorithm

returns a k-means coreset for A with respect to the distance $D(\sigma_1, \sigma_2) = \sqrt{K_\tau(\sigma_1, \sigma_2)}$ for every two $\sigma_1, \sigma_2 \in \mathcal{P}(N)$.

Algorithm 1: Coreset Construction in $(\mathcal{P}_N, K_\tau)$

Input: $A = \{\sigma_1, \ldots, \sigma_n\}$, k, $\varepsilon \in (0,1)$, $\delta \in (0,1)$

1 compute a strong (ε, k)-coreset $S \subseteq A$ with weights $w : S \to \mathbb{R}_{\geq 0}$ with respect to the k-means objective using and distance $D(\cdot,\cdot) = \sqrt{K_\tau(\cdot,\cdot)}$ as described above ;

2 return (S, w)

Theorem 3.1.1. *Let $\varepsilon, \delta \in (0,1)$, and let $A \subseteq \mathcal{P}_N$ be a k-rank-aggregation instance. Then, with probability at least $1-\delta$, the above algorithm returns a strong (ε, k)-coreset of A of size $|S| = \tilde{O}(\varepsilon^{-4} k^2 \log \frac{1}{\delta})$ and in time $T_{coreset}(n, k, \varepsilon, d) = \tilde{O}(nkN)$.*

Proof. Let $\varepsilon \in (0,1), \delta \in (0,1)$, and an integer k. Let $S \subset A$ with weights $w : S \to \mathbb{R}_{\geq 0}$ the output of the algorithm, and let $\Pi \subset \mathcal{P}_N$ a set of size k the candidate k-centers rankings. Consider, for the sake of analysis, the embedding ψ (Definition 3.0.1). In Claim 3.0.2, it is shown that ψ is an isometric embedding, hence, the distances are equal under this mapping, i.e. for every two $\sigma_1, \sigma_2 \in \mathcal{P}(N)$, $D(\sigma_1, \sigma_2) = \sqrt{K_\tau(\sigma_1, \sigma_2)} = \|\psi(\sigma_1) - \psi(\sigma_2)\|_2$. Since the algorithm of [BJKW21] succeeds in finding a coreset for the k-means objective with probability at least $1-\delta$, Algorithm 1 also succeed with the same probability. Overall,

$$\begin{aligned}
\sum_{\sigma \in S} w(\sigma) K_\tau(\sigma, \Pi) &= \sum_{\sigma \in S} w(\sigma) \|\psi(\sigma) - \psi(\Pi)\|_2^2 && \text{Claim 3.0.2} \\
&\in (1 \pm \varepsilon) \sum_{\sigma \in A} \|\psi(\sigma) - \psi(\Pi)\|_2^2 && \text{coreset } S \text{ correctness} \\
&= (1 \pm \varepsilon) \sum_{\sigma \in A} K_\tau(\sigma, \Pi). && \text{Claim 3.0.2}
\end{aligned}$$

Clearly, the size of S is $|S| = s(k, \varepsilon) = \tilde{O}(\varepsilon^{-4} k^2 \log \frac{1}{\delta})$, and the running time is comprised of using the coreset construction of [BJKW21], where the distance computation time is $O(N \log N)$. Overall, applying their analysis produce a running time of $\tilde{O}(nkN)$. ∎

For the rest of the thesis, we will use the above algorithm to construct coresets under the Kendall-tau metric space. We will need an additional property of this coreset, captured by the following lemma. Proving it requires understanding the inner workings of the coreset construction introduced by [BJKW21]. We thus briefly state its steps. Their algorithm (algorithm 1 in [BJKW21], denoted by $\mathcal{B}$) performs $t = \log^\star n$ iterations of an importance sampling based coreset construction (algorithm 2 in [BJKW21], denoted by $\mathcal{B}'$). In every iteration $i \in [1, \log^\star n]$, algorithm $\mathcal{B}'$ is invoked with S_{i-1} as input (for $i = 1$, $S_0 = A$), and it outputs a smaller subset S_i that functions as a strong (ε, k)-coreset of the input set A. This technique is called *iterative size reduction*, and it succeeds with probability at least $1-\delta$ (Theorem 3.1 in [BJKW21]). The following is an immediate consequence from the proof of Theorem 3.1 by [BJKW21], and subsequently we state the required lemma.

Corollary 3.1.2. *Assuming the whole iterative size reduction described above succeeds, then for every integer k, at the end of each iteration $i \in [1, \log^\star n]$, the set S_i is a strong (ε, k)-coreset of the input set A. Hence, by the strong coreset properties, at the end of each iteration i, the total weight of S_i is bounded by*

$$w_{S_i} := \sum_{x \in S_i} w_{S_i}(x) \in (1 \pm \varepsilon) n.$$

Lemma 3.1.3. *For $k \geq 1$, if Algorithm 1 outputs a strong (ε, k)-coreset $S \subseteq A$, then the weight of every $x \in S$ is bounded. Specifically, for every $x \in S$, $w_S(x) \in \left[c_1 \frac{1}{|S|}, c_2 \frac{|A|}{|S|} k\right]$ for some constants c_1, c_2.*

Proof. This proof shows general properties of the weights obtained for the points in the coreset from [BJKW21]. Since our coreset construction is using it in a black-box manner, these properties apply to our coreset as well. Thus, this proof deals with the actual construction algorithm from [BJKW21] and leverage its iterative nature.

Consider some step i of algorithm $\mathcal{B}$ described above. Denote by S_{i-1} the input to $\mathcal{B}'$ and S_i its output. We now show that the weights of S_i are bounded in $\left[c_1 \frac{1}{|S_i|}, c_2 \frac{|A|}{|S_i|} k\right]$ for some constants c_1, c_2. In particular, to complete the proof, observe that it is also true for the last iteration, where $S_i = S$.

The proof will follow by bounding the sensitivities and probabilities of each point and thus bounding the weights. We analyze the sensitivities, probabilities and the weights of points in the coreset S_{i-1} that was output at some step $i-1 \in [0, \log^\star n]$ of $\mathcal{B}$. It is emphasized that S_i was calculated by a single step of the iterative size reduction, which is an invocation of $\mathcal{B}'$.

First, we aim at bounding the sensitivities, and we remember that for every $x \in S_{i-1}$, its sensitivity is defined by

$$\sigma_x = \sigma_x^{apx} \coloneqq w_{S_{i-1}}(x) \cdot \left(\frac{K_\tau(x, C^{apx})}{\text{cost}(S_{i-1}, C^{apx})} + \frac{1}{w_{S_{i-1}}(C^{apx}(x))} \right).$$

Using the same notation as defined by [BJKW21], where C^{apx} is an approximate set of centers for the k-clusters problem on S_{i-1}. Additionally, denote by $C^{apx}(x)$ the set of all points in S_{i-1} that are assigned to the same cluster as x, i.e. have the same nearest center as x. Naive bounds on the above sensitivity produce

$$\sigma_x \in w_{S_{i-1}}(x) \cdot \left[\frac{1}{w_{S_{i-1}}(C^{apx}(x))}, 1 + \frac{1}{w_{S_{i-1}}(C^{apx}(x))} \right].$$

Using Corollary 3.1.2, we can bound $w_{S_{i-1}}(C^{apx}(x))$ by

$$w_{S_{i-1}}(C^{apx}(x)) = \sum_{y \in C^{apx}(x)} w_{S_{i-1}}(y) \in [w_{S_{i-1}}(x), (1+\varepsilon)n].$$

Therefore, for $\varepsilon < 1$,

$$\sigma_x \in w_{S_{i-1}}(x) \cdot \left[\frac{1}{(1+\varepsilon)n}, \left(1 + \frac{1}{w_{S_{i-1}}(x)}\right) \right] \in w_{S_{i-1}}(x) \cdot \left[\frac{1}{2n}, O(1) \right] \tag{3.2}$$

Next, we would like to bound the total sensitivity $\sigma_{S_{i-1}} \coloneqq \sum_{x \in S_{i-1}} \sigma_x$. For the upper bound, we use Lemma 3.4 of [BJKW21], which shows that $\sigma_{S_{i-1}} \leq 1 + \alpha k$ for $\alpha = O(1)$. To lower bound, we use Equation (3.2) for every σ_x, and Corollary 3.1.2, and we have for $\varepsilon < \frac{1}{2}$,

$$\sigma_{S_{i-1}} \geq \sum_{x \in S_{i-1}} w_{S_{i-1}}(x) \cdot \frac{1}{2n} \geq \frac{1}{2n} \cdot (1-\varepsilon)n \geq \frac{1}{4} = \Omega(1).$$

Overall, the total sensitivity is bounded in

$$\sigma_{S_{i-1}} \in [\Omega(1), O(k)]. \tag{3.3}$$

To bound the sampling probability of each point $x \in S_{i-1}$, we use its definition $p_x \coloneqq \frac{\sigma_x}{\sigma_{S_{i-1}}}$. For every $x \in S_{i-1}$, by Equations (3.2) and (3.3),

$$p_x \in w_{S_{i-1}}(x) \cdot \left[\Omega\left(\frac{1}{nk}\right), O(1) \right]. \tag{3.4}$$

Finally, since $w_{S_i}(x) \coloneqq \frac{w_{S_{i-1}}(x)}{p_x |S|}$, we have that the weights are bounded as well. For each $x \in S_i$, it holds,

$$w_{S_i}(x) \in \left[\Omega\left(\frac{1}{|S_i|}\right), O\left(\frac{nk}{|S_i|}\right) \right].$$

■

3.2 FPT-PTAS for k-Rank-Aggregation

In this section, we present an FPT-PTAS in k for k-rank-aggregation. The algorithm has a similar spirit to the one presented in [BJKW21, Section 5.1]. To achieve this, we combine two components as follows. Given A, the algorithm computes a strong (ε, k)-coreset S of A by Algorithm 1, of size $\|S\|_0 = \tilde{O}(\varepsilon^{-4}k^2 \log \frac{1}{\delta})$ (Theorem 3.1.1). Then, it considers every possible partition (by exhaustive search) of S, and for each partition it computes a $(1+\varepsilon)$-approximate center for every set in the partition using the known PTAS for 1-rank-aggregation (see Preliminaries). This procedure outputs a $(1+\varepsilon)$-approximate solution for the k-rank-aggregation problem in time that is polynomial in n and N when ε is treated as a constant, but exponential in k.

Algorithm 2: FPT-PTAS for k-rank-aggregation

Input : $A = \{\sigma_1, \ldots, \sigma_m\} \subseteq \mathcal{P}_N$, k, K_τ, F, $\varepsilon \in (0,1)$, $\delta \in (0,1)$
Output: Coreset S

1 use Algorithm 1 to compute a strong $(0.25\varepsilon, k)$-coreset S ;
2 for each partition $S_1, \ldots, S_k$ of S compute a $(1+0.1\varepsilon)$-approximate center $\hat{\pi}_i \leftarrow F((S_i, W_i), 0.1\varepsilon)$ for each set S_i with weights W_i, and save the lowest cost partition $\hat{S}_1, \ldots, \hat{S}_k$ with $\hat{\Pi} = \{\hat{\pi}_1, \ldots, \hat{\pi}_k\}$;
3 return the best set of approximate k-centers $\hat{\Pi}$;

Theorem 3.2.1. *Let $\varepsilon, \delta \in (0,1)$, and let $A \subseteq \mathcal{P}_N$ be a k-rank-aggregation instance. Then, with probability at least $1-\delta$, the above algorithm returns a set of k-centers, denoted by $\hat{\Pi}$, s.t.*

$$\mathrm{cost}(A, \hat{\Pi}) \leq (1+\varepsilon)\mathsf{OPT}(A, k),$$

and runs in time

$$O\left(k^{\tilde{O}(\varepsilon^{-4}k^2 \log \frac{1}{\delta})} \cdot \left(\frac{N^3 \log N}{\varepsilon} + N2^{\tilde{O}(1/\varepsilon^{12})} + nN^2\right)\right).$$

Proof. Let $\varepsilon \in (0,1), \delta \in (0,1)$, and a k-rank-aggregation instance $A \subseteq \mathcal{P}_N$. Also, write $\varepsilon_1 = 0.25\varepsilon$ and $\varepsilon_2 = 0.1\varepsilon$.

First, the algorithm success probability and running time are analyzed. Since we are using the deterministic version of the 1-rank-aggregation PTAS, the above algorithm success depends only on the event that the coreset construction succeeds, which happens with probability at least $1-\delta$. The running time of the algorithm depends on steps 1 and 2. The first is the coreset construction, which takes time $T_{coreset}(n, k, \varepsilon, d) = \tilde{O}(nkN^2)$ (Theorem 3.1.1). The second step is an exhaustive search over all possible partitions of the coreset with k invocations of $F((S_i, W_i), \varepsilon_2)$ for each partition (once for every set in the partition). It require $O\left(k^{\|S\|_0} \cdot T_{rank}(n, N, \varepsilon)\right)$ time. Overall, the running time is

$$O\Big(nkN^2 + k^{\|S\|_0} \cdot T_{rank}(n, N, \varepsilon)\Big).$$

Using $\|S\|_0 = \tilde{O}(\varepsilon^{-4}k^2 \log \frac{1}{\delta})$ and $T_{rank}(n, N, \varepsilon) = O\left(\frac{N^3 \log N}{\varepsilon}\right) + n2^{\tilde{O}(1/\varepsilon^{12})} + O(nN^2)$ (Equation 3.1) yields the required running time.

Now, assuming S is indeed a strong $(0.25\varepsilon, k)$-coreset, we can analyze the approximation factor given by the above algorithm. We will need three different sets of k-centers for the proof, and for simplicity, we ignore the weights of the different partitions of S. Let $\hat{\Pi} = \{\hat{\pi}_1, \ldots, \hat{\pi}_k\} \subset \mathcal{P}_N$ the k candidate centers returned by Algorithm 2, and denote its corresponding partition of S by $\hat{S}_1, \ldots, \hat{S}_k$. We also use denote the optimal partition of S by $S_1^*, \ldots, S_k^*$, and their optimal k-centers by $\Pi^* = \{\pi_1^*, \ldots, \pi_k^*\}$. Finally, for the intricate part of the proof, denote with

$\hat{\Pi}^* = \{\hat{\pi}_1^*, \ldots, \hat{\pi}_k^*\}$ the *approximate* centers of each set in the *optimal* partition. That is, for every $i \in [k]$, $\hat{\pi}_i^*$ is the approximate permutation returned by $F(S_i^*, \varepsilon_2)$. Then,

$$\begin{aligned}
\text{cost}(A, \hat{\Pi}) &\leq \frac{1}{1-\varepsilon_1}\text{cost}(S, \hat{\Pi}) && \text{strong } (\varepsilon_1, k)\text{-coreset} \\
&= \frac{1}{1-\varepsilon_1}\sum_{i=1}^{k}\sum_{x\in\hat{S}_i} w(x)K_\tau(x, \hat{\pi}_i) \\
&\leq \frac{1}{1-\varepsilon_1}\sum_{i=1}^{k}\sum_{x\in S_i^*} w(x)K_\tau(x, \hat{\pi}_i^*) && \text{exhaustive search} \\
&\leq \frac{1}{1-\varepsilon_1}\sum_{i=1}^{k}\sum_{x\in S_i^*} (1+\varepsilon_2)w(x)K_\tau(x, \pi_i^*) && \hat{\pi}_i^* = F(S_i^*, \varepsilon_2) \\
&= \frac{(1+\varepsilon_2)}{1-\varepsilon_1}\text{cost}(S, \Pi^*) \\
&\leq \frac{(1+\varepsilon_2)}{1-\varepsilon_1}(1+\varepsilon_1)\text{cost}(A, \Pi^*) && \text{strong } (\varepsilon_1, k)\text{-coreset} \\
&\leq (1+\varepsilon)\text{cost}(X, \Pi^*) && \varepsilon_1 = 0.25\varepsilon, \varepsilon_2 = 0.1\varepsilon
\end{aligned}$$

We remark that the difficult step is the second inequality, that can be justified by remembering that the algorithm is performing an exhaustive search over all k-partitions of the coreset S. Hence, even the optimal partition $S_1^*, \ldots, S_k^*$ with its approximate centers $\hat{\Pi}^*$ have higher cost then the chosen partition $\hat{\Pi}$. ■

Remark. This approach should work for any metric space that has an efficient $(1+\varepsilon)$-approximation for the 1-median problem, even in the presence of a very large or possibly infinite ambient space.

3.3 PTAS for β-Distributed k-Rank-Aggregation

This section presents a $(1+\varepsilon)$-approximation scheme for the k-rank-aggregation problem with inputs that satisfy the β-distributed property (Definition 1.1.6). The algorithm and its proof are adapted from [ABS10], where a PTAS for the β-distributed k-median problem in any *finite* metric space is shown. In contrast to [ABS10], where each center must be an input point, we consider a large metric space ($N!$ points), where the center points can be taken from the ambient space. To achieve our result, we again leverage the coreset construction from Section 3.1, this time only for $k = 1$, and the PTAS for 1-rank-aggregation. The idea is to perform an exhaustive search and guess a small coreset (along with a corresponding set of weights) for every large-cost cluster. Then, for each such coreset, the algorithm finds an approximate center using the 1-rank-aggregation PTAS. The running time of our adapted algorithm is higher by the time it takes to perform an exhaustive search for strong $(\varepsilon, 1)$-coreset with a correct set of weights for each large-cost cluster multiplied by the time it takes to find an approximate center of each set.

Definitions and Notation. We denote by $\mathcal{C}^* = \{C_1^*, \ldots, C_k^*\}$ a set of optimal clusters, with optimal centers $c_1^*, \ldots, c_k^*$. The output of the 1-rank-aggregation PTAS will be denoted by $\hat{c}$. Additionally, we use the following definitions that appeared in [ABS10].

1. A cluster C_i^* is called *cheap* if its optimal cost is bounded, i.e. $\mathsf{OPT}_i \leq \frac{\beta\varepsilon\mathsf{OPT}}{32}$, and called *expensive* otherwise. Notice that there are at most $\frac{32}{\beta\varepsilon}$ expensive clusters.

2. The *inner ring* of a cluster C_i^* is the set $C_i^*(core) := \{x \in C_i^* : K_\tau(x, c_i^*) \leq \frac{\beta\mathsf{OPT}}{8|C_i^*|}\}$.

Denote by $B(x,r) = \{y \in A : K_\tau(x,y) \leq r\}$ a ball of radius r around a point x. The algorithm will maintain a list $\mathcal{Q}$ of candidate sets for the clusters cores and we denote by $K_\tau(x,\mathcal{Q}) = \min_{T\in Q; y\in T} K_\tau(x,y)$ the distance of a point x from the list. A component T from the list $\mathcal{Q}$ is called *good* if it contains an inner ring point of some optimal cheap cluster C_i^*, and called *bad* otherwise.

Algorithm for k-Rank-Aggregation with β-Distributed Inputs

The algorithm pseudo code is now presented, and as in the paper by [ABS10], we assume for simplicity the algorithm knows the optimal cost, denoted by $\mathsf{OPT} = \mathsf{OPT}(A,k)$ for clarity. Otherwise, it can perform a doubling technique starting from a known approximation.

The first procedure, describing the inner loop, is handling only the cheap clusters. It receives as input the permutations set $A = \{\sigma_1, \ldots \sigma_n\} \subseteq \mathcal{P}_N$, a set $\mathcal{Q}_{init}$ of approximate centers for guessed coresets of expensive clusters (will later be chosen by exhaustive search performed by an external loop), the separability parameter β and the 1-rank-aggregation PTAS function F. We emphasize that only the centers retrieving stage is changed compared to the algorithm by [ABS10]. It computes an approximate center in line 13 instead of an exact center.

Algorithm 3: β-Distributed k-Rank-Aggregation - Inner Loop

Input: A, K_τ, k, $\varepsilon \in (0,1)$, $\mathcal{Q}_{init}$, β, F

Initialization Stage:

1 Set $\mathcal{Q} \leftarrow \mathcal{Q}_{init}$;

Population Stage:

2 **for** $s = n, n-1, n-2, \ldots, 1$ **do**

3 set $r = \frac{\beta \mathsf{OPT}}{4s}$;

4 remove any point x s.t. $K_\tau(x,\mathcal{Q}) < 2r$;

5 **for** *every pair of remaining points* a, b **do**

6 **if** $K_\tau(a,b) \leq r \wedge |B(a,r)| > \frac{s}{2} \wedge |B(b,r)| > \frac{s}{2}$ **then**

7 connect a and b ;

8 **for** *every connected component* T *of size* $> \frac{s}{2}$ **do**

9 add T to Q ;

10 define $B(T) = \{x : \exists y \in T \text{ s.t. } K_\tau(x,y) \leq 2r\}$;

11 remove the points of $B(T)$ from the instance ;

Centers-Retrieving Stage:

12 **for** *every choice of* k *components* $T_1, \ldots, T_k$ *out of* $\mathcal{Q}$ **do**

13 set $\hat{c}_i = F(T_i \cup B(T_i), \varepsilon)$ for every $i \in [k]$;

14 partition all n points according to the nearest point among $\{\hat{c}_1, \ldots, \hat{c}_k\}$;

15 if a clustering of cost at most $(1 + O(\varepsilon))\mathsf{OPT}$ is found, output these k centers and halt ;

Now that we know how to handle the case where there are only cheap clusters, we need to perform an exhaustive search over at most $\frac{32}{\beta\varepsilon}$ possible expensive clusters. The external loop is the source of the expensive clusters optional centers that are given as input to the inner loop in the list $\mathcal{Q}_{init}$. It is infeasible to try all possible points for an expensive cluster. Hence, we need to shrink the data size, and we thus resort to using strong $(\varepsilon,1)$-coresets (proved to exist in Section 3.1). We claim that if we try all possible sets of points, and all possible integers that serve as powers for the rounded weights, then in one of the iterations, the algorithm will obtain a set of points and a corresponding rounded weights that represent a correct strong $(\varepsilon,1)$-coreset for an expensive cluster. The algorithm chooses $t = \tilde{O}\left(\varepsilon^{-4}\right)$, which is the size of strong $(\varepsilon,1)$-coreset,

chosen by setting $k = 1$ in the size of the strong (ε, k)-coreset from Theorem 3.1.1.

Algorithm 4: β-Distributed k-Rank-Aggregation - External Loop

Input: $A = \{\sigma_1, \ldots \sigma_m\} \subseteq \mathcal{P}_N$, k, K_τ, $\varepsilon \in (0,1)$, β, F

1 **for** $\ell = 0, 1, \ldots, \frac{32}{\beta\varepsilon} - 1$ **do**

2 Set $t = \tilde{O}(\frac{1}{\varepsilon^4})$ and $I = \left[-\frac{1}{\varepsilon}\log\frac{t}{c_1}, \frac{1}{\varepsilon}\log\frac{c_2 n}{t}\right]$; ▷ `coreset size and weights bounds`

3 **for** *every choice of* ℓt *points from* A *and* ℓt *integers in* I **do**

4 **for** *every* $i \in [\ell], j \in [t]$ **do**

5 denote by $p(i,j) \in A$ each guessed point and by $q(i,j) \in I$ each guessed integer ;

6 set weight $w(i,j) = (1+\varepsilon)^{q(i,j)}$;

7 **for** $i = 1, 2, .., \ell$ **do**

8 Set $\hat{c}_i \leftarrow F((\{p(i,1), \ldots, p(i,t)\}, \{w(i,1), \ldots, w(i,t)\}), \varepsilon)$;

9 Run algorithm 3 with $\mathcal{Q}_{init} = \{\{\hat{c}_1\}, \{\hat{c}_2\}, \ldots, \{\hat{c}_\ell\}\}$;

10 **if** *the algorithm returns a clustering of cost at most* $(1 + O(\varepsilon))\mathsf{OPT}$ **then**

11 output it and halt ;

The modified algorithm uses ideas from both the k-median and k-means algorithms of [ABS10] and their analysis, with three main changes. In line 13 of the inner loop, and line 8 of the external loop, we use a PTAS F for finding an approximate center c_i (with respect to the algorithm in [ABS10]) instead of finding an exact center. Also, we increase the size of the sample t, to use the fact there exists a strong $(\varepsilon, 1)$-coreset for each optimal cluster. The analysis below is adapting the proof of [ABS10]. We show that the approximation factor is still $(1 + O(\varepsilon))$, and that the running time is increased, according to the time it takes F to find an approximate center and the additional time it takes to guess coreset points and approximate weights.

Analysis

The following theorem applies to the above modified algorithm, and its proof is very similar to the proof from [ABS10] (Section 4, Theorem 4.6). The changes are due to the use of strong coresets and $(1+\varepsilon)$-approximate centers.

Theorem 3.3.1. *Let $A \subseteq \mathcal{P}_N$ be a k-rank-aggregation instance that satisfies the β-distributed property (Definition 1.1.6). Then, Algorithm 4 outputs k centers whose cost is at most*

$$(1 + O(\varepsilon))\,\mathsf{OPT}(A, k)$$

in time

$$O\left(\left(\frac{n}{\varepsilon}\right)^{O\left(\varepsilon^{-5}\beta^{-1}\right)}\left(\left(k + \frac{1}{\varepsilon\beta}\right)\left(\frac{N^3 \log N}{\varepsilon} + N 2^{\tilde{O}\left(\varepsilon^{-12}\right)} + N^2 n\right) + k^{O\left(\frac{1}{\beta}\right)} n k N^2\right)\right).$$

For completeness, we cite here three technical lemmas from [ABS10] that are used in the proof below. They are still true, with unchanged proof, since they refer to the population stage of Algorithm 3, which is identical to [ABS10].

Lemma 3.3.2 (Claim 4.3 in [ABS10])**.** *Let C_i^* be any cluster in the target clustering. By stage $s = |C_i^*|$, the algorithm adds to $\mathcal{Q}$ a component T that contains a point from the inner ring of C_i^*. That is, the algorithm adds a good component with respect to C_i^*.*

Lemma 3.3.3 (Claim 4.4 in [ABS10]). *Let T be a good component added to $\mathcal{Q}$, i.e., it contains an inner ring point from cluster C_i^*. Then:*

1. *$\forall x \in T$, it holds $K_\tau(x, c_i^*) \leq \frac{\beta \mathsf{OPT}}{2|C_i^*|}$.*
2. *$C_i^*(core) \subseteq T \cup B(t) \subseteq C_i^*$.*
3. *$\forall T' \in \mathcal{Q}$ s.t. $T' \neq T$, $C_i^*(core) \cap T' = \emptyset$.*

Lemma 3.3.4 (Claim 4.5 in [ABS10]). *The list $\mathcal{Q}$ contains less than $\frac{16}{3\beta}$ bad components at the end of the population stage.*

In addition, we also need the following result regarding the discretization process for the coresets weights that will be used later in the analysis. Since the weights are taken from a polynomially-bounded continuous range (Lemma 3.1.3), the algorithm performs discretization of the range into a logarithmic set of weights and performs an exhaustive search over all possibilities in polynomial time. The following corollary guarantee that the discretization process only incur an extra factor of $(1+\varepsilon)$ to the cost.

Corollary 3.3.5. *Let $S \subseteq A$ be a strong $(\varepsilon, 1)$-coreset of size $s = |S|$. Denote its set of weights by $w^*(x_1), \ldots, w^*(x_s)$. For every $x \in S$, define $\hat{w}(x) = (1+\varepsilon)^{q(x)}$ where $q(x)$ is the smallest integer s.t. $\hat{w}(x) \geq w^*(x)$. Then, for every point c that serves as center,*

$$\sum_{x \in S} \hat{w}(x_i) K_\tau(x, c) \leq (1+\varepsilon) \sum_{x \in S} w^*(x) K_\tau(x, c)$$

Additionally, using Lemma 3.1.3 and the above corollary, we have that the range of powers used by the algorithm is large enough.

Corollary 3.3.6. *Let $S \subseteq A$ be a coreset constructed by Algorithm 1. By Lemma 3.1.3, we know that every $x \in S$, has bounded weight $w_S(x) \in \left[c_1 \frac{1}{|S|}, c_2 \frac{n}{|S|} k\right]$. Thus, considering the discrete set of weights defined by the discretization process from Corollary 3.3.5, we conclude that there are $O\left(\frac{1}{\varepsilon}(\log n + \log |S|)\right)$ possibilities for rounding up each weight of $x \in S$ into the nearest power of $(1+\varepsilon)$. Simply by choosing the power as integer in $q(x) \in \left[-\frac{1}{\varepsilon} \log \frac{|S|}{c_1}, \frac{1}{\varepsilon} \log \frac{c_2 n}{|S|}\right]$ for some constants c_1, c_2.*

Proof of Theorem 3.3.1. First, we deal with the approximation factor of the returned set of centers. Let $\varepsilon < \frac{1}{2}$. Using Lemma 3.3.3 and the fact we are trying all possible coresets for the expensive clusters, it follows that there exists some choice of k components of $\mathcal{Q}$, s.t. all components are either an approximate center of a strong coreset of expensive cluster or a good component of a cheap cluster. Therefore, each component represents a single, unique, cluster of the optimal clustering $\mathcal{C}^*$. We fix this choice of k components, and show that for the optimal clustering, replacing the true centers $\{c_1^*, c_2^*, \ldots, c_k^*\}$ with the approximate centers $\{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_k\}$ that the algorithm outputs, increases the cost by at most $(1+\varepsilon')$ factor, for $\varepsilon' = 12\varepsilon$. Consider each optimal cluster separately and fix some optimal cluster C_i^* with optimal center c_i^* of the optimal clustering. Denote by OPT_i the optimal cost of C_i^* using an optimal c_i^*.

There are two options. The first one is that C_i^* is an expensive cluster. In this case, denote by $S_i = \{x_1, \ldots, x_t\}$ a correct guess for the set of points of a strong $(\varepsilon, 1)$-coreset of C_i^*. Also, denote a correct set of corresponding weights with $W^* = \{w^*(x_1), \ldots, w^*(x_t)\}$, with a corresponding rounded set of weights $\hat{W}_i = \{\hat{w}(x_1), \ldots, \hat{w}(x_t)\}$ (as outlined in the discretization process of Corollary 3.3.5). Clearly, since Algorithm 4 performs exhaustive search, then at some iteration it encounters the correct set S_i and the correct set of discrete weights $\hat{W}_i$. We consider now this iteration of the algorithm, and write by c_i the optimal center for $(S_i, \hat{W}_i)$, and denote by $\{\hat{c}_i\}$

the singleton set that has the $(1+\varepsilon)$-approximate center computed by $F((S_i, \hat{W}_i), \varepsilon)$ that was given as input from the external loop. Hence,

$$\begin{aligned}
\sum_{x \in C_i^*} K_\tau(x, \hat{c}_i) &\leq \sum_{x \in S_i} \frac{1}{1-\varepsilon} w^*(x) K_\tau(x, \hat{c}_i) && \varepsilon\text{-coreset} \\
&\leq \frac{1}{1-\varepsilon} \sum_{x \in S_i} \hat{w}(x) K_\tau(x, \hat{c}_i) && \text{rounded up weights} \\
&\leq \frac{1+\varepsilon}{1-\varepsilon} \sum_{x \in S_i} \hat{w}(x) K_\tau(x, c_i) && \varepsilon\text{-approx} \\
&\leq \frac{(1+\varepsilon)^2}{1-\varepsilon} \sum_{x \in S_i} w^*(x) K_\tau(x, c_i^*) && \text{optimality \& Corollary 3.3.5} \\
&\leq (1+12\varepsilon) \sum_{x \in C_i^*} K_\tau(x, c_i^*) = (1+O(\varepsilon))\mathsf{OPT}_i. && \varepsilon\text{-coreset}
\end{aligned}$$

Note that c_i is the optimal center for S_i with the weights $\hat{W}$, hence we need to use its optimality and only then Corollary 3.3.5.

The second option is that C_i^* is a cheap cluster. Then, we denote by T the good component corresponding to it (one such component must exists by Lemma 3.3.2). We break the cost of C_i^* using $\hat{c}_i$, the approximate point picked in step 13 by the Algorithm 3 to serve as center, into two parts,

$$\sum_{x \in C_i^*} K_\tau(x, \hat{c}_i) = \sum_{x \in T \cup B(T)} K_\tau(x, \hat{c}_i) + \sum_{x \in C_i^* \setminus (T \cup B(T))} K_\tau(x, \hat{c}_i).$$

Consider each term separately. The first term, is bounded by $(1+\varepsilon)\sum_{x \in T \cup B(T)} K_\tau(x, c_i^*)$ since $\hat{c}_i$ is a $(1+\varepsilon)$-approximate computed on the set $T \cup B(T)$, and thus is a $(1+\varepsilon)$ "competitive" against every other center including c_i^*. Bounding the second term is the intricate part of the proof. Consider a point $x \in C_i^* \setminus (T \cup B(T))$. Since $C_i^*(core) \subseteq T \cup B(T)$, we know that $x \notin C_i^*(core)$, and therefore,

$$d(x, c_i^*) > \frac{\beta \mathsf{OPT}}{8|C_i^*|}. \tag{3.5}$$

We show that

$$d(\hat{c}_i, c_i^*) \leq \frac{\varepsilon\beta \mathsf{OPT}}{4|C_i^*|}. \tag{3.6}$$

Define the set $G = \{y \in C_i^* : K_\tau(y, c_i^*) \leq \frac{\varepsilon\beta\mathsf{OPT}}{16|C_i^*|}\}$. Notice that $G \subseteq C_i^*(core) \subseteq T \cup B(T)$ and that, using Markov's inequality, $|G| \geq \frac{1}{2}|C_i^*|$ (otherwise $\mathsf{OPT}_i \geq \frac{\varepsilon\beta\mathsf{OPT}}{32}$ which contradicts the fact that C_i^* is cheap). Assume towards contradiction that $K_\tau(\hat{c}_i, c_i^*) > \frac{\varepsilon\beta\mathsf{OPT}}{4|C_i^*|}$. Then,

$$\begin{aligned}
\sum_{x \in T \cup B(T)} K_\tau(x, \hat{c}_i) &\geq \sum_{x \in (T \cup B(T)) \cap G} K_\tau(\hat{c}_i, c_i^*) - K_\tau(x, c_i^*) \\
&> \frac{1}{2}|C_i^*| \left(\frac{\varepsilon\beta\mathsf{OPT}}{4|C_i^*|} - \frac{\varepsilon\beta\mathsf{OPT}}{16|C_i^*|} \right) \\
&\geq 2\frac{\varepsilon\beta\mathsf{OPT}}{32} \geq 2\mathsf{OPT}_i.
\end{aligned}$$

However, for $\varepsilon < 1$, and since c_i is the optimal center of $T \cup B(T)$, we get a contradiction,

$$\begin{aligned}
\sum_{x \in T \cup B(T)} K_\tau(x, \hat{c}_i) &< 2 \cdot \sum_{x \in T \cup B(T)} K_\tau(x, c_i) \leq 2 \cdot \sum_{x \in T \cup B(T)} K_\tau(x, c_i^*) \\
&\leq 2 \cdot \sum_{x \in C_i^*} K_\tau(x, c_i^*) = 2\mathsf{OPT}_i.
\end{aligned}$$

Conclude, using triangle inequality, and the combination of (3.5) and (3.6) that:

$$\begin{aligned}\sum_{x\in C_i^*\setminus(T\cup B(T))} K_\tau(x,\hat{c}_i) &\leq \sum_{x\in C_i^*\setminus(T\cup B(T))} [K_\tau(x,c_i^*)+K_\tau(\hat{c}_i,c_i^*)]\\ &\leq \sum_{x\in C_i^*\setminus(T\cup B(T))} [K_\tau(x,c_i^*)+2\varepsilon K_\tau(x,c_i^*)]\\ &\leq (1+2\varepsilon)\sum_{x\in C_i^*\setminus(T\cup B(T))} K_\tau(x,c_i^*).\end{aligned}$$

Overall, for a cheap cluster C_i^*,

$$\sum_{x\in C_i^*} K_\tau(x,\hat{c}_i) \leq (1+2\varepsilon)\sum_{x\in C_i^*} K_\tau(x,c_i^*) = (1+O(\varepsilon))\mathsf{OPT}_i.$$

Finally, summing over all k-centers output by the algorithm produces the required approximation.

To analyze the running time, we highlight that it is infeasible to find the best center, hence we find an approximate center in time given by $T_{rank}(n,N,\varepsilon)$ as seen in Equation (3.1). Additionally, one need to take into account the additional running time to guess a correct candidate coresets along with a correct set of weights. We also remark that computing the distance between two permutations is naively performed in time in $O(N\log N)$ (discussed above). Therefore, we need at most $O(n^2N\log N)$ time to compute the distances between all pairs of points beforehand and store these distances in an $O(n^2)$-size data structure. The reason we do it is to reduce the computational cost for each distance lookup between two points. From now on, we ignore this computation, and assume $O(1)$ running time for each distance query. Finally, adding up the relevant running times for each step in our algorithm as follows.

1. The population stage in Algorithm 3 is unchanged, and thus takes $O(n^3)$ time as before. For every $s\in[n]$, and every point $x\in A$, we need to find how many of the remaining points are in $B(x,r)$.

2. By Lemmas 3.3.3 and 3.3.4, $\mathcal{Q}$ contains a single good component for each optimal cluster, and at most $O\left(\frac{1}{\beta}\right)$ bad components. Hence, finding a candidate center for all components in $\mathcal{Q}$ takes $|\mathcal{Q}|T_{rank}(n,N,\varepsilon) = \left(k+O\left(\frac{1}{\beta}\right)\right)T_{rank}(n,N,\varepsilon)$ time. Furthermore, trying every possible k components in $\mathcal{Q}$ in the centers-retrieving stage takes $\binom{k+O(\frac{1}{\beta})}{k} \leq k^{O(\frac{1}{\beta})}$ attempts.

3. Evaluating the cost of a possible clustering takes $O(nkN\log N)$ since we need to compute the distance between all n points and the k possible centers and take the nearest neighbour. Each such distance computation takes $O(N\log N)$ time.

4. The external loop is performing an exhaustive search for strong coresets and their weights for at most $\frac{32}{\beta\varepsilon}$ possible expensive clusters. Thus, assuming a coreset of size $t=\tilde{O}\left(\frac{1}{\varepsilon^4}\right)$, and a set of $O\left(\frac{1}{\varepsilon}(\log n+\log t)\right)$ possible discrete weights (Corollary 3.3.6), there are $\sum_{\ell=0}^{\frac{32}{\beta\varepsilon}}\binom{n}{\ell t}\left(\frac{1}{\varepsilon}(\log n+\log t)\right)^{\ell t}$ executions of the inner loop. This can be bounded by

$$O\left(\frac{1}{\beta\varepsilon}\binom{n}{\frac{t}{\beta\varepsilon}}(\log n+\log t)^{O(\frac{t}{\beta\varepsilon})}\right).$$

In addition, for each such coreset, we use the 1-rank-aggregation PTAS to find an approximate center, thus we need to multiply by $T_{rank}(t,N,\varepsilon)$.

Overall, running time of the entire algorithm is at most

$$O\left(\frac{1}{\beta\varepsilon}\binom{n}{\frac{t}{\beta\varepsilon}}\left(\frac{1}{\varepsilon}\left(\log n+\log t\right)\right)^{O(\frac{t}{\beta\varepsilon})}\cdot \left(n^3+\left(k+O\left(\frac{1}{\beta}\right)\right)T_{rank}(n,N,\varepsilon)+k^{O\left(\frac{1}{\beta}\right)}nkN^2+\frac{T_{rank}(t,N,\varepsilon)}{\beta\varepsilon}\right)\right).$$

Simplifying using $t=\varepsilon^{-4}$ produces the following expression,

$$O\left(\frac{1}{\beta\varepsilon}\left(\frac{2}{\varepsilon}n\log n\right)^{O\left(\frac{1}{\beta\varepsilon^5}\right)}\left(\left(k+O\left(\frac{1}{\varepsilon\beta}\right)\right)T_{rank}(n,N,\varepsilon)+k^{O(\frac{1}{\beta})}nkN^2\right)\right).$$

Finally,

$$O\left(\left(\frac{n}{\varepsilon}\right)^{O\left(\frac{1}{\beta\varepsilon^5}\right)}\left(\left(k+\frac{1}{\varepsilon\beta}\right)T_{rank}(n,N,\varepsilon)+k^{O(\frac{1}{\beta})}nkN^2\right)\right).$$

Setting $T_{rank}(n,N,\varepsilon)$ using Equation (3.1) concludes that the overall time is polynomial in both n, N and k when treating ε and β as fixed constants.

■

3.4 Discussion

In the beginning of this chapter we showed that there exists a strong (ε,k)-coreset for $(\mathcal{P}_N,K_\tau)$. It is interesting to understand if coresets exists for other ranking distances or partial rankings as defined for example in [FKS03, FKM+04, FKM+06, Ail10, BBGH15]. If such small strong coresets exists, then one can use the PTAS for bucketed top-r rankings from [MM20] to obtain a $(1+\varepsilon)$-approximation for k-rank-aggregation on bucket lists in a similar manner. Additionally, we would like to understand if it is possible to obtain a PTAS for k-rank-aggregation without the β-distributed separability assumption, e.g., by leveraging the structure of the permutation space $(\mathcal{P}_N,K_\tau)$.

We stress that the results shown in this chapter are demonstrated using the metric space of permutations and Kendall-tau distance, mainly because they fulfill two requirements. First, in $(\mathcal{P}_N,K_\tau)$ there exists a strong (ε,k)-coreset that is independent of both the number of points n and their "dimension" N. Second, there is a polynomial-time deterministic $(1+\varepsilon)$-approximation algorithm for the corresponding 1-median problem. Specifically, our proofs did not use any specific feature of $(\mathcal{P}_N,K_\tau)$, hence, to use our framework in a different large, possibly infinite, metric space, one just needs to show these two requirements hold, and then the results of this part of the thesis will apply to the new metric space. To conclude, in a metric space satisfying these two conditions, our algorithm is a PTAS for k-median on inputs satisfying the β-distributed property.

Bibliography

[ABD09] Margareta Ackerman and Shai Ben-David. Clusterability: A theoretical study. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 1–8. PMLR, 2009. URL: `https://proceedings.mlr.press/v5/ackerman09a.html`.

[ABS10] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a PTAS for k-median and k-means clustering. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 309–318, 2010. `doi:10.1109/FOCS.2010.36`.

[ABS12] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Inf. Process. Lett.*, 112(1-2):49–54, 2012. `doi:10.1016/j.ipl.2011.10.006`.

[Ach03] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. Special Issue on PODS 2001. `doi:10.1016/S0022-0000(03)00025-4`.

[ACM+20] Pankaj K. Agarwal, Hsien-Chih Chang, Kamesh Munagala, Erin Taylor, and Emo Welzl. Clustering Under Perturbation Stability in Near-Linear Time. In *40th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2020)*, pages 8:1–8:16, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. `doi:10.4230/LIPIcs.FSTTCS.2020.8`.

[ACN08] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5), 2008. `doi:10.1145/1411509.1411513`.

[AGK+04] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004. `doi:10.1137/S0097539702416402`.

[AHPV04] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004. `doi:10.1145/1008731.1008736`.

[Ail10] Nir Ailon. Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica*, 57:284–300, 06 2010. `doi:10.1007/s00453-008-9211-1`.

[ANFSW19] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and Euclidean k-median by primal-dual algorithms. *SIAM Journal on Computing*, 49(4):FOCS17–97, 2019.

[ARR98] Sanjeev Arora, Prabhakar Raghavant, and Satish Rao. Approximation schemes for Euclidean k-medians and related problems. *Proceedings of the Annual ACM Symposium on Theory of Computing*, pages 106–113, 1998. doi:10.1145/276698.276718.

[BBD05] Therese Biedl, Franz J. Brandenburg, and Xiaotie Deng. Crossings and permutations. In *Proceedings of the 13th International Conference on Graph Drawing*, GD'05, page 1–12, Berlin, Heidelberg, 2005. Springer-Verlag. doi:10.1007/11618058_1.

[BBG09] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, page 1068–1077. Society for Industrial and Applied Mathematics, 2009.

[BBG+19] Georg Bachmeier, Felix Brandt, Christian Geist, Paul Harrenstein, Keyvan Kardel, Dominik Peters, and Hans Georg Seedig. k-majority digraphs and the hardness of voting with a constant number of voters. *Journal of Computer and System Sciences*, 105:130–157, 2019. doi:10.1016/j.jcss.2019.04.005.

[BBGH15] Christian Bachmaier, Franz J. Brandenburg, Andreas Gleißner, and Andreas Hofmeier. On the hardness of maximum rank aggregation problems. *Journal of Discrete Algorithms*, 31:2–13, 2015. 24th International Workshop on Combinatorial Algorithms (IWOCA 2013). doi:https://doi.org/10.1016/j.jda.2014.10.002.

[BBH+20] Daniel N. Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 569–579. PMLR, 2020. URL: http://proceedings.mlr.press/v119/baker20a.html.

[Ben15] Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. *CoRR*, abs/1501.00437, 2015. URL: http://arxiv.org/abs/1501.00437, arXiv:1501.00437.

[BFL16] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016. arXiv:1612.00889.

[BJKW21] Vladimir Braverman, Shaofeng H-C Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2679–2696. SIAM, 2021.

[BL12] Yonatan Bilu and Nathan Linial. Are stable instances easy? *Comb. Probab. Comput.*, 21(5):643–660, 2012. doi:10.1017/S0963548312000193.

[BL16] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102–155, 2016.

[Bor84] JC de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784)*, 1784.

[BPR+15] Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, page 737–756, 2015.

[BR14] Shalev Ben-David and Lev Reyzin. Data stability in clustering: A closer look. *Theoretical Computer Science*, 558:51–61, 2014.

[BS80] Jon Louis Bentley and James B Saxe. Decomposable searching problems I. static-to-dynamic transformation. *Journal of Algorithms*, 1(4):301–358, 1980. doi:10.1016/0196-6774(80)90015-2.

[BTT89a] J. Bartholdi, C. A. Tovey, and M. A. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6(2):157–165, 1989. URL: http://www.jstor.org/stable/41105913.

[BTT89b] J. J. Bartholdi, C. A. Tovey, and M. A. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989. URL: http://www.jstor.org/stable/41105918.

[CFR10] Don Coppersmith, Lisa K. Fleischer, and Atri Rurda. Ordering by weighted number of wins gives a good ranking for weighted tournaments. *ACM Trans. Algorithms*, 6(3), July 2010. doi:10.1145/1798596.1798608.

[CFS19] Vincent Cohen-Addad, Andreas Emil Feldmann, and David Saulpic. Near-linear time approximations schemes for clustering in doubling metrics. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 540–559, 2019. doi:10.1109/FOCS.2019.00041.

[CG99] Moses Charikar and Sudipto Guha. Improved combinatorial algorithms for the facility location and k-median problems. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 378–388. IEEE, 1999.

[CGTS02] Moses Charikar, Sudipto Guha, Eva Tardos, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences*, 65(1):129–149, 2002. doi:10.1006/jcss.2002.1882.

[Che09] Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009. doi:10.1137/070699007.

[CKL21] Vincent Cohen-Addad, C. S. Karthik, and Euiwoong Lee. On approximability of clustering problems without candidate centers. In *Proceedings of the Thirty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '21, page 2635–2648, 2021.

[CKM19] Vincent Cohen-Addad, Philip N Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics. *SIAM Journal on Computing*, 48(2):644–667, 2019.

[CP10] Timothy M. Chan and Mihai Pătraşcu. Counting inversions, offline orthogonal range counting, and related problems. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, page 161–173, 2010.

[CS04] Artur Czumaj and Christian Sohler. Sublinear-time approximation for clustering via random sampling. In *Automata, Languages and Programming*, pages 396–407. Springer Berlin Heidelberg, 2004.

[CSS21] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 169–182, 2021. doi:10.1145/3406325.3451022.

[DG77] Persi Diaconis and R. L. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268, 1977. URL: http://www.jstor.org/stable/2984804.

[DG03] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. doi:10.1002/rsa.10073.

[DKNS01] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 613–622. ACM, 2001. doi:10.1145/371920.372165.

[DLS12] Amit Daniely, Nati Linial, and Michael E. Saks. Clustering is difficult only when it does not matter. *CoRR*, 2012. URL: http://arxiv.org/abs/1205.4891.

[Fel20] Dan Feldman. Core-sets: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(1), 2020. doi:10.1002/widm.1335.

[FKM+04] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. Comparing and aggregating rankings with ties. In *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, page 47–58, 2004. doi:10.1145/1055558.1055568.

[FKM+06] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. Comparing partial rankings. *SIAM J. Discrete Math.*, 20:628–648, 01 2006. doi:10.1137/05063088X.

[FKS03] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003. doi:10.1137/S0895480102412856.

[FL11] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, page 569–578, 2011. doi:10.1145/1993636.1993712.

[FMS07] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proceedings of the Twenty-Third Annual Symposium on Computational Geometry*, SCG '07, page 11–18. ACM, 2007. doi:10.1145/1247069.1247072.

[FRS19] Zachary Friggstad, Mohsen Rezapour, and Mohammad R Salavatipour. Local search yields a PTAS for k-means in doubling metrics. *SIAM Journal on Computing*, 48(2):452–480, 2019.

[FSS20] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020. doi:10.1137/18M1209854.

[GI03] Venkatesan Guruswami and Piotr Indyk. Embeddings and non-approximability of geometric problems. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 537–538, 2003.

[GK99] Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *Journal of Algorithms*, 31(1):228–248, 1999. doi:10.1006/jagm.1998.0993.

[HJLW18] Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 814–825, 2018. doi:10.1109/FOCS.2018.00082.

[HPK07] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete and Computational Geometry*, 37(1):3–19, 2007. doi:10.1007/s00454-006-1271-x.

[HPM04] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '04, page 291–300, 2004. doi:10.1145/1007352.1007400.

[HV20] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in Euclidean spaces: Importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 1416–1429, 2020. doi:10.1145/3357713.3384296.

[Ind99] Piotr Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 428–434, 1999.

[JKS16] Yunlong Jiao, Anna Korba, and Eric Sibony. Controlling the distance to a Kemeny consensus without computing it. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2971–2980. PMLR, 2016. URL: http://proceedings.mlr.press/v48/korba16.html.

[JL84] William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space 26. *Contemporary mathematics*, 26, 1984.

[JMS02] Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, page 731–740, 2002. doi:10.1145/509907.510012.

[JV01] Kamal Jain and Vijay V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. *J. ACM*, 48(2):274–296, 2001. doi:10.1145/375827.375845.

[Kem59] John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959. URL: http://www.jstor.org/stable/20026529.

[Ken38] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. URL: http://www.jstor.org/stable/2332226.

[KPR00] Madhukar R Korupolu, C Greg Plaxton, and Rajmohan Rajaraman. Analysis of a local search heuristic for facility location problems. *Journal of algorithms*, 37(1):146–188, 2000.

[KS07] Claire Kenyon-Mathieu and Warren Schudy. How to rank with few errors. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, page 95–103, 2007. doi:10.1145/1250790.1250806.

[LS10] Michael Langberg and Leonard J. Schulman. Universal epsilon-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, page 598–607. Society for Industrial and Applied Mathematics, 2010.

[LS13] Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 901–910, 2013. doi:10.1145/2488608.2488723.

[LV92] Jyh-Han Lin and Jeffrey Scott Vitter. Approximation algorithms for geometric median problems. *Information Processing Letters*, 44(5):245–249, 1992. doi:https://doi.org/10.1016/0020-0190(92)90208-D.

[McG18] Andrew McGregor. Advanced Algorithms. https://people.cs.umass.edu/~mcgregor/711S18/vectors-1.pdf, 2018. [Online].

[MHP18] Robin Milosz, Sylvie Hamel, and Adeline Pierrot. Median of 3 permutations, 3-cycles and 3-hitting set problem. In *Combinatorial Algorithms*, pages 224–236. Springer International Publishing, 2018.

[MM20] Claire Mathieu and Simon Mauras. How to aggregate top-lists: Approximation algorithms via scores and average ranks. In *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '20, page 2810–2822. Society for Industrial and Applied Mathematics, 2020.

[MMR19] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 1027–1038. ACM, 2019. doi:10.1145/3313276.3316350.

[MOP01] Nina Mishra, Dan Oblinger, and Leonard Pitt. Sublinear time approximate clustering. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, page 439–447. Society for Industrial and Applied Mathematics, 2001.

[MOP04] Adam Meyerson, Liadan O'callaghan, and Serge Plotkin. A k-median algorithm with running time independent of data size. *Machine Learning*, 56(1):61–87, 2004.

[MP92] Arakaparampil M Mathai and Serge B Provost. *Quadratic forms in random variables: theory and applications*. Dekker, 1992.

[MP04] Ramgopal R Mettu and C Greg Plaxton. Optimal time bounds for approximate clustering. *Machine Learning*, 56(1):35–60, 2004.

[MS18] Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *KI-Künstliche Intelligenz*, 32(1):37–53, 2018.

[ORSS13] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of Lloyd-type methods for the k-means problem. *J. ACM*, 59(6), January 2013. doi:10.1145/2395116.2395117.

[Phi16] Jeff M. Phillips. Coresets and sketches. *CoRR*, abs/1601.00617, 2016. arXiv:1601.00617.

[Sch12] Warren Schudy. *Approximation Schemes for Inferring Rankings and Clusterings from Pairwise Data*. PhD thesis, Brown University, 2012.

[SW18] Christian Sohler and David P. Woodruff. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 802–813, 2018. doi:10.1109/FOCS.2018.00081.

[Tho05] Mikkel Thorup. Quick k-median, k-center, and facility location for sparse graphs. *SIAM Journal on Computing*, 34(2):405–432, 2005.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi:10.1017/9781108231596.

[YL78] H Peyton Young and Arthur Levenglick. A consistent extension of Condorcet's election principle. *SIAM Journal on applied Mathematics*, 35(2):285–300, 1978.

[You88] H. P. Young. Condorcet's theory of voting. *American Political Science Review*, 82(4):1231–1244, 1988. doi:10.2307/1961757.

[ZW07] Anke Zuylen and David Williamson. Deterministic algorithms for rank aggregation and other ranking and clustering problems. *WAOA*, 4927:260–273, 10 2007. doi:10.1007/978-3-540-77918-6_21.

Appendix A

Concentration of the JL Sum

A.1 Background

Given a matrix $G \in \mathbb{R}^{d\times D}$ whose entries are standard normal independently and identically distributed, $N(0,1)$, defining $L = \frac{1}{\sqrt{d}}G$ creates the standard JL-type dimension reduction linear transform. In this self-contained appendix, we suggest a new concentration bound for $\sum_{i=1}^{n}\|Lx_i\|_2^z \in (1\pm\varepsilon)\sum_{i=1}^{n}\|x_i\|_2^z$ for $z \in \{1,2\}$ where the target dimension is $d = \tilde{O}(\frac{1}{\varepsilon^2})$ (Theorems A.2.1, A.3.1). Theorem A.3.1 also offers an alternative proof for Lemma 2.1.9 with Gaussian entries JL transform. We emphasize that with an extra factor $O(\log n)$ in d, we can have this concentration bound immediately from the standard JL lemma, since the norm of every x_i is preserved up to a factor of $(1\pm\varepsilon)$. The main idea here, is that if we consider of the entire sum, then we can conserve on the usual factor of $\log n$ for the target dimension.

We heavily rely on the probability book by Roman Vershynin ([Ver18], §2.5-3.1). For completeness, the relevant results are cited in this section.

Definition A.1.1 (Sub-gaussian random variable, Definition 2.5.6 [Ver18])**.** A random variable X that satisfies the following property, for some absolute constant K_1, is called *sub-gaussian random variable*.

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-t^2/K_1^2\right) \quad \text{for all } t \geq 0.$$

The *sub-gaussian norm* of X, denoted $\|X\|_{\psi_2}$, is defined by

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$$

The sub-gaussian norm is a useful notion, since it is possible to write the definition, using some absolute constant c, as follows,

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-ct^2/\|X\|_{\psi_2}^2\right) \quad \text{for all } t \geq 0. \tag{A.1}$$

Using known properties of normal random variables of we have the following lemma.

Lemma A.1.2 (Gaussian is sub-gaussian, Exercise 2.5.8(a) [Ver18])**.** *$X \sim N(0,\sigma^2)$ is a sub-gaussian random variable with $\|X\|_{\psi_2} \leq C\sigma$ where C is an absolute constant.*

Definition A.1.3 (Sub-exponential random variable, Definition 2.7.5 [Ver18])**.** A random variable X that satisfies the following property, for some absolute constant K_1, is called *sub-exponential random variable*.

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-t/K_1\right) \quad \text{for all } t \geq 0.$$

The *sub-exponential norm* of X, denoted $\|X\|_{\psi_1}$, is defined by

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}\exp(X/t) \leq 2\}$$

Remark. Both the sub-gaussian and sub-exponential definition have multiple equivalent definitions. This definition was chosen for readability and concreteness.

The following is a relation between the sub-exponential and sub-gaussian random variables.

Lemma A.1.4 (Sub-exponential is sub-gaussian squared, Lemma 2.7.6 [Ver18])**.** *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover, $\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$.*

Using the above, we immediately obtain the following corollary.

Corollary A.1.5 (Gaussian squared is sub-exponential)**.** *If $X \sim N(0, \sigma^2)$, then X^2 is a sub-exponential random variable where $\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2 \leq C^2\sigma^2$. For example, for $U \sim \chi^2(1)$, it holds that $\|U\|_{\psi_1} \leq C^2$.*

Lastly, we will need the following two concentration bounds for sub-exponential and sub-gaussian random variables.

Lemma A.1.6 (Bernstein's inequality, Theorem 2.8.2 [Ver18])**.** *Let $X_1, X_2, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, and writing $K = \max_{i\in[N]} \|X_i\|_{\psi_1}$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right] \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right)$$

Lemma A.1.7 (Norm concentration with general variance, generalization of Theorem 3.1.1 [Ver18])**.** *Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a random vector with independent, sub-gaussian coordinates X_i that satisfy $\mathbb{E}X_i^2 = \sigma^2$. Then*

$$\left\|\|X\|_2 - \sqrt{d}\sigma\right\|_{\psi_2} \leq \frac{cK^2}{\sigma}$$

where $K = \max_i \|X_i\|_{\psi_2}$ and c is a constant.

Proof. Let $Y = \frac{1}{\sigma}X \in \mathbb{R}^d$. Then $Y_i = \frac{1}{\sigma}X_i$ and $\mathbb{E}\left[Y_i^2\right] = \frac{1}{\sigma^2}\mathbb{E}\left[X_i^2\right] = 1$. Hence, we can use Theorem 3.1.1 from [Ver18] w.r.t to random vector with unit-variance coordinates

$$\left\|\|Y\|_2 - \sqrt{d}\right\|_{\psi_2} \leq c\tilde{K}^2.$$

Where $\tilde{K} = \max_i \|Y_i\|_{\psi_2} = \frac{1}{\sigma}\max_i \|X_i\|_{\psi_2} = \frac{1}{\sigma}K$.
Finally, after rearranging:

$$\left\|\frac{1}{\sigma}\|X\|_2 - \sqrt{d}\right\|_{\psi_2} \leq c\tilde{K}^2 \Rightarrow \left\|\|X\|_2 - \sigma\sqrt{d}\right\|_{\psi_2} \leq \frac{cK^2}{\sigma}.$$

■

A.2 Concentration Bounds for ℓ_2^2 Sum

We begin with the ℓ_2^2 case.

Theorem A.2.1. *Let $G \in \mathbb{R}^{d\times D}$ be a matrix whose entries are independent standard normal $N(0,1)$, and define $L = \frac{1}{\sqrt{d}}G$. Then, for sufficiently large $d = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$, for every $x_1, \ldots, x_n \in \mathbb{R}^D$, with probability at least $1-\delta$,*

$$\sum_{\ell=1}^{n}\|Lx_\ell\|_2^2 \in (1\pm\varepsilon)\sum_{\ell=1}^{n}\|x_\ell\|_2^2.$$

Proof. First, we consider a single row of G, denoted $g \sim N(0, I)$ s.t. $g \in \mathbb{R}^D$. Then $\forall \ell \in [n]$ we have $\langle g, x_\ell \rangle = \sum_{j=1}^D g_j x_{\ell,j}$. Then, summing over all $x \in A$:

$$Q(g) \coloneqq \sum_{\ell=1}^n \langle g, x_\ell \rangle^2 = \sum_{\ell=1}^n g^T x_\ell x_\ell^T g = g^T \left(\sum_{\ell=1}^n x_\ell x_\ell^T \right) g.$$

Writing $A \coloneqq \sum_{\ell=1}^n x_\ell x_\ell^T$, we have the following quadratic forms

$$Q(g) = g^T A g = \sum_{j=1}^D \sum_{k=1}^D a_{j,k} g_j g_k.$$

Observe that

$$\mathbb{E}\left[Q(g)\right] = \sum_{j=1}^D \sum_{k=1}^D a_{j,k} \mathbb{E}[g_j g_k] = \sum_{j=1}^D a_{j,j} = \mathrm{Tr}(A) = \sum_{\ell=1}^n \|x_\ell\|_2^2 . \tag{A.2}$$

As shown in [MP92, Chapter 3], we can decompose the quadratic form of $Q(g)$ using a spectral decomposition of A, denoted by $A = P^T \Lambda P$, to obtain a sum of independent chi-square with 1-degree of freedom random variables as follows,

$$Q(g) = \sum_{j=1}^D \lambda_j U_j^2$$

where $U = Pg \in \mathbb{R}^D$, thus $U_j^2 \sim \chi^2(1)$ for any $j \in [D]$. Now, using Corollary A.1.5, we can deduce that $\left\| U_j^2 \right\|_{\psi_1} \leq C^2$ and since $Q(g)$ is a sum of independent chi-square random variables,

$$\|Q(g)\|_{\psi_1}^2 \leq O\left(\sum_{j=1}^D \left\| \lambda_j U_j^2 \right\|_{\psi_1}^2 \right) \leq O\left(\sum_{j=1}^D \lambda_j^2 \right) = O\left(\mathrm{Tr}(A^2) \right).$$

Now, we are ready to consider $G \in \mathbb{R}^{d \times D}$ where each row acts as g from above, i.e., we have d independent copies of g, resulting with

$$\sum_{\ell=1}^n \|Gx_\ell\|_2^2 = \sum_{\ell=1}^n \sum_{i=1}^d \langle g_i, x_\ell \rangle^2 = \sum_{i=1}^d Q(g_i). \tag{A.3}$$

By centering the R.Vs and using Bernstein's inequality A.1.6 with d independent R.Vs:

$$\begin{aligned}
\mathbb{P}\left[\left| \sum_{i=1}^d Q(g_i) - d\mathbb{E}[Q(g_i)] \right| \geq t \right] &\leq 2 \exp\left(-O\left(\min\left\{ \frac{t^2}{\sum_{i=1}^d \|Q(g_i)\|_{\psi_1}^2}, \frac{t}{\max_i \|Q(g_i)\|_{\psi_1}} \right\} \right) \right) \\
&\leq 2 \exp\left(-O\left(\min\left\{ \frac{t^2}{d\,\mathrm{Tr}(A^2)}, \frac{t}{\sqrt{\mathrm{Tr}(A^2)}} \right\} \right) \right).
\end{aligned}$$

Choosing $t = \varepsilon d \cdot \mathrm{Tr}(A) = \varepsilon d \sum_{\ell=1}^n \|x_\ell\|_2^2$ and using $L = \frac{1}{\sqrt{d}} G$ and Equations (A.2), (A.3) we obtain,

$$\begin{aligned}
\mathbb{P}\left[\left| \sum_{\ell=1}^n \|Lx_\ell\|_2^2 - \sum_{\ell=1}^n \|x_\ell\|_2^2 \right| \geq \varepsilon \sum_{\ell=1}^n \|x_\ell\|_2^2 \right] &\leq 2 \exp\left(-O\left(\min\left\{ \frac{\varepsilon^2 d^2\,\mathrm{Tr}(A)^2}{d\,\mathrm{Tr}(A^2)}, \frac{\varepsilon d\,\mathrm{Tr}(A)}{\sqrt{\mathrm{Tr}(A^2)}} \right\} \right) \right) \\
&\leq 2 \exp\left(-O\left(\min\left\{ d\varepsilon^2, d\varepsilon \right\} \right) \right).
\end{aligned}$$

The last inequality is obtained by observing that A is a positive semi-definite matrix and therefore $\mathrm{Tr}(A) \geq \sqrt{\mathrm{Tr}(A^2)}$ and $\mathrm{Tr}(A)^2 \geq \mathrm{Tr}(A^2)$. Finally, choosing $d = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ for large enough constant concludes that with probability at at least $1-\delta$,

$$\sum_{\ell=1}^{n} \|Lx_\ell\|_2^2 \in (1\pm\varepsilon)\sum_{\ell=1}^{n}\|x_\ell\|_2^2 .$$

■

A.3 Concentration Bounds for ℓ_2 Sum

Here, we provide an additional proof for Lemma 2.1.9, using sub-gaussian properties of the JL transform L. It is emphasized that the proof require entries that are normally distributed. However, it should be possible to extend it to every sub-gaussian entries and thus to any JL transform. Additionally, the target dimension here conserve on the additional factor of $\log\frac{1}{\varepsilon}$ that appears in Lemma 2.1.9.

Theorem A.3.1. *Let $G \in \mathbb{R}^{d\times D}$ be a matrix whose entries are independent standard normal $N(0,1)$, and define $L = \frac{1}{\sqrt{d}}G$. Then, for sufficiently large $d = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$, for every $x_1,\ldots,x_n \in \mathbb{R}^D$, with probability at least $1-\delta$,*

$$\sum_{x\in A}\|Lx\|_2 \in (1\pm\varepsilon)\sum_{x\in A}\|x\|_2 .$$

Proof. Let $G \in \mathbb{R}^{d\times D}$ a random matrix where each entry is independently drawn according to the standard normal distribution $G_{i,j} \sim N(0,1)$. Hence, by Lemma A.1.2, $\|G_{i,j}\|_{\psi_2} \leq C$. Next, we want to understand how does a single summand $\|Gx\|_2$ behaves. Notice that $(Gx)_i = \sum_{j=1}^{D} G_{i,j}x_j$ is a just a linear combination of independent gaussians, thus $(Gx)_i \sim N(0,\|x\|_2^2)$, i.e. $\mathbb{E}\left[(Gx)_i^2\right] = \|x\|_2^2$, which implies again by Lemma A.1.2 that $\|(Gx)_i\|_{\psi_2} \leq C\|x\|_2$. Overall, the entries of Gx are independent sub-gaussian entries with constant (assuming data is fixed) second moment $\sigma^2 := \|x\|_2^2$ and we can use Lemma A.1.7 for isotropic, non-unit variance. The lemma implies that each summand of $\|Gx\|_2 - \sqrt{d}\|x\|_2$ is a sub-gaussian random variable with the sub-gaussian norm,

$$\left\|\|Gx\|_2 - \sqrt{d}\|x\|_2\right\|_{\psi_2} \leq \frac{O(1)K^2}{\|x\|_2} = O(1)\cdot\|x\|_2$$

since $K = \max_i\|(Gx)_i\|_{\psi_2} = C\|x\|_2$. Now, summing over n dependent sub-gaussians and using triangle inequality,

$$\sigma := \left\|\sum_{x\in A}\left[\|Gx\|_2 - \sqrt{d}\|x\|_2\right]\right\|_{\psi_2}^2 \leq \left(\sum_{x\in A}\left\|\|Gx\|_2 - \sqrt{d}\|x\|_2\right\|_{\psi_2}\right)^2 \leq O\left(\sum_{x\in A}\|x\|_2\right)^2 .$$

To claim that the sum is also a sub-gaussian random variable, we can use Equation (A.1), to obtain,

$$\mathbb{P}\left[\left|\sum_{x\in A}\left[\|Gx\|_2 - \sqrt{d}\|x\|_2\right]\right| \geq t\right] \leq 2\exp\left(-O(1)\cdot\frac{t^2}{\sigma}\right).$$

Taking $t = \varepsilon\sqrt{d}\sum_{x\in A}\|x\|_2$ concludes,

$$\mathbb{P}\left[\sum_{x\in A}\|Gx\|_2 \notin (1\pm\varepsilon)\sqrt{d}\sum_{x\in A}\|x\|_2\right] \leq 2\exp\left(-O(1)\cdot\frac{\varepsilon^2 d\left(\sum_{x\in A}\|x\|_2\right)^2}{\left(\sum_{x\in A}\|x\|_2\right)^2}\right) = 2\exp\left(-O(1)\cdot\varepsilon^2 d\right).$$

Overall, choosing $d = O(\varepsilon^{-2}\log(\frac{2}{\delta}))$ for large enough constant yields that required result. ■