


Diversity Maximization in Doubling Metrics


Alfonso Cevallos

Swiss Federal Institute of Technology (ETH), Switzerland
alfonso.cevallos@ifor.math.ethz.ch

 <https://orcid.org/0000-0001-8622-5830>

Friedrich Eisenbrand¹

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
friedrich.eisenbrand@epfl.ch

 <https://orcid.org/0000-0001-7928-1076>

Sarah Morell²

Technische Universität Berlin (TU Berlin), Germany
morell@math.tu-berlin.de

Abstract

Diversity maximization is an important geometric optimization problem with many applications in recommender systems, machine learning or search engines among others. A typical diversification problem is as follows: Given a finite metric space (X, d) and a parameter $k \in \mathbb{N}$, find a subset of k elements of X that has maximum diversity. There are many functions that measure diversity. One of the most popular measures, called *remote-clique*, is the sum of the pairwise distances of the chosen elements. In this paper, we present novel results on three widely used diversity measures: Remote-clique, remote-star and remote-bipartition.

Our main result are polynomial time approximation schemes for these three diversification problems under the assumption that the metric space is doubling. This setting has been discussed in the recent literature. The existence of such a PTAS however was left open.

Our results also hold in the setting where the distances are raised to a fixed power $q \geq 1$, giving rise to more variants of diversity functions, similar in spirit to the variations of clustering problems depending on the power applied to the pairwise distances. Finally, we provide a proof of NP-hardness for remote-clique with squared distances in doubling metric spaces.

2012 ACM Subject Classification Theory of computation \rightarrow Facility location and clustering

Keywords and phrases Remote-clique, remote-star, remote-bipartition, doubling dimension, grid rounding, ε -nets, polynomial time approximation scheme, facility location, information retrieval

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2018.33

Related Version A full version of the paper is available at [8], <https://arxiv.org/abs/1809.09521>.

1 Introduction

A *dispersion* or *diversity maximization* problem is as follows: Given a ground set X and a natural number $k \in \mathbb{N}$, find a subset $S \subseteq X$ among those of cardinality k that maximizes a certain *diversity function* $\text{div}(S)$.

¹ The second author acknowledges support from the Swiss National Science Foundation grant 163071, “Convexity, geometry of numbers, and the complexity of integer programming”.

² Work conducted while the third author was affiliated to EPFL, Switzerland.



While diversity maximization has been of interest in the algorithms and operations research community for some time already, see e.g. [11, 5, 25, 20], the problem received considerable attention in the recent literature regarding information retrieval, recommender systems, machine learning and data mining, see e.g. [28, 29, 23, 24, 1].

Distances used in these applications may be metric or non-metric. However, most popular distances either are metric or correspond to the q -th power of metric distances for some $q > 1$. The cosine distance for a set $X \subseteq \mathbb{R}^d \setminus \{0\}$, for example, is a popular non-metric measure of dissimilarity for text documents [26], which can be interpreted as the squared Euclidean distance of the input vectors, after scaling all vectors to unit length.

In this paper we focus on three popular diversity functions over metric spaces, see e.g. [11, 5, 2, 21, 17, 7, 6, 4]. In particular, for a given n -point metric space (X, d) , a constant $q \in \mathbb{R}_{\geq 1}$ and a parameter $k \in \mathbb{Z}$ with $2 \leq k \leq n$, we consider the family of problems

$$\max_{T \subseteq X, |T|=k} \text{div}^q(T),$$

where $\text{div}^q(T)$ corresponds to one of the following three diversity functions:

- *Remote-clique*: $\text{cl}^q(T) := \sum_{\{u,v\} \in \binom{T}{2}} d^q(u,v) = \frac{1}{2} \sum_{u,v \in T} d^q(u,v)$.
- *Remote-star*: $\text{st}^q(T) := \min_{z \in T} \sum_{u \in T \setminus \{z\}} d^q(z,u)$.
- *Remote-bipartition*: $\text{bp}^q(T) := \min_{L \subseteq T, |L|=\lfloor |T|/2 \rfloor} \sum_{\ell \in L, r \in T \setminus L} d^q(\ell, r)$.

Here, $d^q(u, v)$ is the q -th power of the distance between u and v . In the literature, these problems have been mainly considered for $q = 1$ to which we refer as *standard* remote-clique, remote-star and remote-bipartition respectively.

In the present work, we present polynomial time approximation schemes for the generalized versions ($q \geq 1$) of the remote-clique, remote-star and remote-bipartition problems in the case where the metric space is *doubling*. The latter is a general and robust class of metric spaces that have low intrinsic dimension. We provide a proper definition in Section 2.

Contributions of this paper

Suppose that (X, d) is a metric space of bounded doubling dimension D and that the power $q \geq 1$ is fixed. In this setting, our main results are as follows:

- i) We show that there exist polynomial time approximation schemes (PTAS) for the remote-clique, remote-star and remote-bipartition problems. In other words, for each $\varepsilon > 0$ and for each of the three diversity functions $\text{cl}^q(T)$, $\text{st}^q(T)$ and $\text{bp}^q(T)$, there exists a polynomial time algorithm that computes a k -subset of X whose diversity is at least $(1 - \varepsilon)$ times the diversity of the optimal set. We prove this result by means of a single and very simple algorithm that identifies a cluster which is then rounded, while all points outside of the cluster have to be in the optimal solution.
- ii) For the standard ($q = 1$) remote-clique problem we refine our generic algorithm into a fast PTAS that runs in time $O(n(k + \varepsilon^{-D})) + (\varepsilon^{-1} \log k)^{O(\varepsilon^{-D})} \cdot k$.
- iii) For the remote-bipartition problem, our algorithm assumes access to a polynomial time oracle that, for any k -set T , returns the value of $\text{bp}^q(T)$. For $q = 1$, this corresponds to the metric min-bisection problem, known to be NP-hard and admitting a PTAS [16]. We generalize this last result and provide a PTAS for min-bisection over doubling metric spaces for *any* constant $q \geq 1$, thus validating our main result.

■ **Table 1** Current best approximation ratios and hardness results for remote-clique, remote-star and remote-bipartition with a highlight on our results. The sign † indicated that the result assumes hardness of the planted-clique problem.

Problem	Distance class	Unbounded dimension		Fixed (doubling) dimension	
		Approx.	Hardness	Approx.	Hardness
clique, $q = 1$	Metric	1/2 [20, 5]	1/2 + ε † [6]	PTAS (Thm. 4)	–
	ℓ_1 and ℓ_2	PTAS [9, 10]	NP-hard [9]	PTAS [15, 9, 10]	–
clique, $q = 2$	Euclidean	PTAS [9, 10]	NP-hard [9]	PTAS [9, 10]	NP-hard (Thm. 8)
star, $q = 1$	Metric	1/2 [11]	1/2 + ε † [8]	PTAS (Thm. 4)	–
bipartition, $q = 1$	Metric	1/3 [11]	1/2 + ε † [8]	PTAS (Thm. 4)	–
3 problems, any const. $q \geq 1$	Metric	–	$2^{-q} + \varepsilon$ † [8]	PTAS (Thm. 4)	NP-hard (Thm. 8)

- iv) We provide the first NP-hardness proof for remote-clique in fixed doubling dimension. More precisely, we prove that the version of remote-clique with squared Euclidean distances in \mathbb{R}^3 is NP-hard.

Related work

For the standard case $q = 1$ and for general metrics, Chandra and Halldórsson [11] provided a thorough study of several diversity problems, including remote-clique, remote-star and remote-bipartition. They observed that all three problems are NP-hard by reductions from the CLIQUE-problem and provided a $\frac{1}{2}$ -factor and a $\frac{1}{3}$ -factor approximation algorithm for remote-star and remote-bipartition respectively. Several approximation algorithms are known for remote-clique as well [25, 20, 5] with the current best factor being $\frac{1}{2}$.

► **Remark.** Borodin et al. [6] proved that the approximation factor of $\frac{1}{2}$ is best possible for standard remote-clique over general metrics under the assumption that the *planted-clique problem* [3] is hard. In the full version we prove that, under the same assumption and for any $q \geq 1$, neither remote-clique, remote-star nor remote-bipartition admits a constant approximation factor higher than 2^{-q} . Thus, none of the three problems nor their generalizations for $q \geq 1$ admits a PTAS over general metrics.

In terms of relevant special cases for standard remote-clique, Ravi et al. [25] provided an efficient exact algorithm for instances over the real line, and a factor of $\frac{2}{\pi}$ over the Euclidean plane. Later on, Fekete and Meijer [15] provided the first PTAS for this problem for fixed-dimensional ℓ_1 distances, and an improved factor of $\frac{\sqrt{2}}{2}$ over the Euclidean plane. Very recently, Cevallos et al. [9, 10] provided PTASs over ℓ_1 and ℓ_2 distances of unbounded dimension as well as for distances of *negative type*, a class that contains some popular non-metric distances including the cosine distance. We remark however that the running times of all previously mentioned PTASs [15, 9, 10] have a dependence on n given by high-degree polynomials (in the worst case) and thus are not suited for large data sets.

For remote-star and remote-bipartition, to the best of the authors' knowledge there were no previous results in the literature on improved approximability for any fixed-dimensional setting, nor for other non-trivial special settings beyond general metrics. Moreover, there was no proof of NP-hardness for any of the three problems in a fixed-dimensional setting. In particular, showing NP-hardness of a fixed-dimensional geometric version of remote-clique was left as an open problem in [15].

Further related results and implications

In applications of diversity maximization in the area of information retrieval, common challenges come from the fact that the data sets are very large and/or are naturally embedded in a high dimensional vector space. There is active research in dimensionality reduction techniques, see [13] for a survey. It has also been remarked that in many scenarios such as human motion data and face recognition, data points have a hidden intrinsic dimension that is very low and independent from the ambient dimension, and there are ongoing efforts to develop algorithms and data structures that exploit this fact, see [27, 22, 14, 18]. One of the most common and theoretically robust notions of intrinsic dimension is precisely the doubling dimension. We remark that our algorithm does not need to embed the input points into a vector space (of low dimension or otherwise) and does not require knowledge of the doubling dimension, as this parameter only plays a role in the run-time analysis.

A sensible approach when dealing with very large data sets is to perform a *core-set reduction* of the input as a pre-processing step. This procedure quickly filters through the input points and discards most of them, leaving only a small subset – the core-set – that is guaranteed to contain a near-optimal solution. There are several recent results on core-set reductions for standard ($q = 1$) dispersion problems, see [21, 2, 7]. In particular, Ceccarello et al. [7] recently presented a PTAS-preserving reduction (resulting in an arbitrarily small deterioration of the approximation factor) for all three problems in doubling metric spaces, with the existence of a PTAS left open. Their construction allows for our algorithm to run in a machine of restricted memory and adapts it to streaming and distributed models of computation. Besides showing that a PTAS exists, we can also combine our results with theirs. We refer the interested reader to the previously mentioned references and limit ourselves to remark a direct consequence of Theorem 4 and [7, Theorems 3 and 9].

► **Corollary 1.** *For $q = 1$ and any constant $\varepsilon > 0$, our three diversity problems over metric spaces of constant doubling dimension D admit $(1 - \varepsilon)$ -approximations that execute as single-pass and 2-pass streaming algorithms, in space $O(\varepsilon^{-D}k^2)$ and $O(\varepsilon^{-D}k)$ respectively.*

Organization of the paper. In Section 2, we provide some needed notation and background techniques. Section 3 presents our general algorithm (Theorem 4) and Section 4 is dedicated to the NP-hardness result (Theorem 8). Due to space constraints, the description of the faster PTAS for standard remote-clique and the PTAS for the generalized min-bisection problem as well as the proofs of some lemmas have been deferred to the full version of this paper [8].

2 Preliminaries

A (*finite*) *metric space* is a tuple (X, d) , where X is a finite set and $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$ is a symmetric distance function that satisfies the triangle inequality with $d(u, u) = 0$ for each point $u \in X$. For a point $u \in X$ and a parameter $r \in \mathbb{R}_{\geq 0}$, the *ball centered at u of radius r* is defined as $B(u, r) := \{v \in X : d(u, v) \leq r\}$. The *doubling dimension* of (X, d) is the smallest $D \in \mathbb{R}_{\geq 0}$ such that any ball in X can be covered by at most 2^D balls of half its radius. In other words, for each $u \in X$ and $r > 0$, there exist points $v_1, \dots, v_t \in X$ with $t \leq 2^D$ such that $B(u, r) \subseteq \cup_{i=1}^t B(v_i, r/2)$. A family of metric spaces is *doubling* if their doubling dimensions are bounded by a constant. It is well known that all metric spaces induced by a normed vector space of bounded dimension are doubling.

We rely on the standard *cell-decomposition* technique and *grid-rounding*, see [19]. We assume without loss of generality that the diameter of (X, d) , i.e. the largest distance between two points, is 1. For a parameter $\delta > 0$, the following greedy procedure partitions X into

cells of radius δ . Initially, define all points in X to be white. While there exist white points, pick one that we call u , color it red and assign all white points $v \in X$ with $d(u, v) \leq \delta$ to u and color them blue. A cell is now comprised of a red point, declared to be the cell center, and all the blue points assigned to it. Grid rounding means to *move* or *round* each point to its respective cell center. This incurs a location error of at most δ for each point.

How many cells and thus different points does this algorithm produce? If (X, d) is of constant doubling dimension D , a direct consequence of the definition of D is that for any parameters r and ρ in $\mathbb{R}_{>0}$, a ball of radius r can be covered by at most $(2/\rho)^D$ balls of radius ρr . Since X is contained in a ball of radius 1, the number of cells produced is bounded by $(4/\delta)^D$. Indeed, X can be covered by $(4/\delta)^D$ balls of radius $\delta/2$ and each such ball contains at most one cell center since, by construction, the distance between any two cell centers is strictly larger than δ . Notice that this procedure executes in time $O((\# \text{ cells}) \cdot |X|)$ and that it requires no knowledge of the value of the doubling dimension D .

The following two lemmas correspond respectively to standard inequalities used for powers of metric distances and to trivial relations among our three diversity functions, see also [12]. Their proofs are deferred to the full version.

► **Lemma 2.** Fix a constant $q \geq 1$. For any three points $u, v, w \in X$ one has

$$d^q(u, w) \leq 2^{q-1} [d^q(u, v) + d^q(v, w)] \quad \text{or equivalently} \quad (1)$$

$$d^q(u, v) \geq 2^{-(q-1)} d^q(u, w) - d^q(v, w). \quad (2)$$

For any numbers $x, y \in \mathbb{R}_{\geq 0}$ and $0 \leq \varepsilon \leq 1$,

$$(x + \varepsilon y)^q \leq x^q + 2^q \varepsilon \cdot \max\{x^q, y^q\}. \quad (3)$$

► **Lemma 3.** Fix a constant $q \geq 1$. For any k -set $T \subseteq X$,

$$\frac{k}{2} \cdot \text{st}^q(T) \leq \text{cl}^q(T) \leq 2^{q-1} k \cdot \text{st}^q(T) \quad \text{and} \quad (4)$$

$$\frac{2(k-1)}{k} \cdot \text{bp}^q(T) \leq \text{cl}^q(T) \leq (2^q + 1) \cdot \text{bp}^q(T) \quad (\text{assuming that } k \text{ is even}). \quad (5)$$

Whenever we deal with remote-bipartition, we assume for simplicity that k is even – all our results can easily be extended to the odd case, up to a change in constants by a factor $2^{O(q)}$. Therefore, the diversity functions correspond to the sum of $\binom{k}{2}$, $(k-1)$ and $k^2/4$ terms, respectively for remote-clique, remote-star and remote-bipartition. Consequently, for each function div^q and for a given instance, we fix an optimal k -set denoted by OPT_{div^q} and define its *average optimal value* Δ_{div^q} as follows:

- $\Delta_{\text{cl}^q} := \text{cl}^q(OPT_{\text{cl}^q}) / \binom{k}{2}$,
- $\Delta_{\text{st}^q} := \text{st}^q(OPT_{\text{st}^q}) / (k-1)$,
- $\Delta_{\text{bp}^q} := \text{bp}^q(OPT_{\text{bp}^q}) / (k^2/4)$.

Whenever the diversity function div^q is clear from context, or for general statements on all three functions, we use OPT and Δ as short-hands for OPT_{div^q} and Δ_{div^q} respectively.

► **Remark.** It directly follows from Lemma 3 that for a common metric space and common parameters $q \geq 1$ and k , the average optimal values Δ_{cl^q} , Δ_{st^q} and Δ_{bp^q} are all just a constant away from each other (a constant $2^{O(q)}$ that is independent of n and k). We heavily use this property linking our three problems in the proof of our key structural result (Theorem 5). A similar result does not extend to other common diversity maximization problems such as remote-edge, remote-tree and remote-cycle, see [11] for definitions. This seems to be a bottleneck for possibly adapting our approach to those problems.

3 A PTAS for all three diversity problems

We now come to our main result which is the following theorem.

► **Theorem 4.** *For any constant $q \in \mathbb{R}_{\geq 1}$, the q -th power versions of the remote-clique, remote-star and remote-bipartition problems admit PTASs over doubling metric spaces.*

Let us fix a constant error parameter $\varepsilon > 0$. Our algorithm is based on grid rounding. However, if we think about the case $q = 1$, a direct implementation of this technique requires a cell decomposition of radius $O(\varepsilon \cdot \Delta)$, which is manageable only if Δ is large enough with respect to the diameter. Otherwise, the number of cells produced may be super-constant in n . Hence, a difficult instance is one where Δ is very small, which intuitively occurs only in the degenerate case where most of the input points are densely clustered in a small region, with very few points outside of it. The algorithmic idea is thus to partition the input points into a *main cluster* and a collection of *outliers*, and treat these sets differently.

3.1 Key structural result

We identify in any instance a main cluster containing most of the input points. This cluster corresponds to a ball with a radius that is bounded with respect to $\Delta^{1/q}$. Thanks to the nature of the diversity functions, we can guarantee that *all outliers are contained in OPT*.

► **Theorem 5.** *Fix a constant $q \geq 1$. For each diversity function div^q in $\{\text{cl}^q, \text{st}^q, \text{bp}^q\}$ and a fixed optimal k -set $\text{OPT}_{\text{div}^q} \subseteq X$, there is a point $z_0 = z_0(\text{div}^q)$ in $\text{OPT}_{\text{div}^q}$ so that*

$$X \setminus B(z_0, c_{\text{div}^q}(\Delta_{\text{div}^q})^{1/q}) \subseteq \text{OPT}_{\text{div}^q},$$

where $c_{\text{cl}^q} = 2$, $c_{\text{st}^q} = 4$, and $c_{\text{bp}^q} = 6$.

Proof. For each function div^q in $\{\text{cl}^q, \text{st}^q, \text{bp}^q\}$, let $z_0 = z_0(\text{div}^q)$ be the center of the minimum weight spanning star in $\text{OPT}_{\text{div}^q}$ so that $\text{st}^q(\text{OPT}_{\text{div}^q}) = \sum_{u \in \text{OPT}_{\text{div}^q}} d^q(z_0, u)$. Consider a point $s = s(\text{div}^q)$ outside of the ball $B(z_0, c_{\text{div}^q}(\Delta_{\text{div}^q})^{1/q})$, i.e.

$$d^q(z_0, s) > (c_{\text{div}^q})^q \cdot \Delta_{\text{div}^q}. \quad (6)$$

Assume that s is not in $\text{OPT}_{\text{div}^q}$ and define the k -set $\text{OPT}'_{\text{div}^q} := \text{OPT}_{\text{div}^q} \cup \{s\} \setminus \{z_0\}$. We will show for each diversity function that $\text{div}^q(\text{OPT}'_{\text{div}^q}) > \text{div}^q(\text{OPT}_{\text{div}^q})$, thus contradicting the optimality of $\text{OPT}_{\text{div}^q}$. To simplify notation in the remainder of the proof, we make the corresponding function clear from context and remove the subscripts div^q .

For remote-clique, we have

$$\begin{aligned} \text{cl}^q(\text{OPT}') - \text{cl}^q(\text{OPT}) &= \sum_{u \in \text{OPT} \setminus \{z_0\}} [d^q(s, u) - d^q(z_0, u)] \\ &\geq \sum_{u \in \text{OPT} \setminus \{z_0\}} [2^{-(q-1)} d^q(z_0, s) - 2d^q(z_0, u)] && \text{(by (2))} \\ &= \frac{k-1}{2^{q-1}} d^q(z_0, s) - 2 \cdot \text{st}^q(\text{OPT}) && \text{(by choice of } z_0) \\ &> \frac{k-1}{2^{q-1}} (2^q \Delta) - 2 \cdot \frac{2}{k} \cdot \text{cl}^q(\text{OPT}) && \text{(by (6) and (4))} \\ &= 2(k-1)\Delta - 2(k-1)\Delta = 0 && \text{(by def. of } \Delta). \end{aligned}$$

For remote-star, let z be the center of the minimum weight spanning star in OPT' so that $\text{st}^q(\text{OPT}') = d^q(z, s) + \sum_{u \in \text{OPT}' \setminus \{z_0\}} d^q(z, u)$. We claim that

$$d^q(z_0, z) \leq 2^q \Delta, \tag{7}$$

as otherwise we obtain

$$\begin{aligned} \text{st}^q(\text{OPT}) + \text{st}^q(\text{OPT}') &= d^q(z, s) + \sum_{u \in \text{OPT}' \setminus \{z_0\}} [d^q(z_0, u) + d^q(z, u)] \\ &\geq 2^{-(q-1)} \sum_{u \in \text{OPT}' \setminus \{z_0\}} d^q(z_0, z) && \text{(by (1))} \\ &> \frac{k-1}{2^{q-1}} (2^q \Delta) = 2(k-1)\Delta = 2 \cdot \text{st}^q(\text{OPT}) && \text{(negating (7)).} \end{aligned}$$

Inequality (7) implies in particular that $z \neq s$, hence $z \in \text{OPT}$. Notice by the minimality of the remote-star function that $\text{st}^q(\text{OPT}) \leq \sum_{u \in \text{OPT}} d^q(z, u)$. By inequalities (2), (6) and (7), we obtain

$$\begin{aligned} \text{st}^q(\text{OPT}') - \text{st}^q(\text{OPT}) &\geq \sum_{u \in \text{OPT}'} d^q(z, u) - \sum_{u \in \text{OPT}} d^q(z, u) = d^q(z, s) - d^q(z, z_0) \\ &\geq 2^{-(q-1)} d^q(z_0, s) - 2d^q(z_0, z) \\ &> 2^{-(q-1)} (4^q \Delta) - 2(2^q \Delta) = 0. \end{aligned}$$

For remote-bipartition, let $\text{OPT}' = L' \cup R$ be the minimum weight bipartition of OPT' so that $\text{bp}^q(\text{OPT}') = \sum_{\ell \in L', r \in R} d^q(\ell, r)$. Assume without loss of generality that $s \in L'$. We claim that

$$\sum_{r \in R} d^q(z_0, r) \leq \frac{2^q + 1}{2} k \Delta, \tag{8}$$

as otherwise we obtain

$$\begin{aligned} \text{bp}^q(\text{OPT}) &\geq \frac{1}{2^q + 1} \text{cl}^q(\text{OPT}) \geq \frac{k}{2(2^q + 1)} \text{st}^q(\text{OPT}) && \text{(by (5) and (4))} \\ &= \frac{k}{2(2^q + 1)} \sum_{u \in \text{OPT}} d^q(z_0, u) \geq \frac{k}{2(2^q + 1)} \sum_{r \in R} d^q(z_0, r) && \text{(as } R \subseteq \text{OPT)} \\ &> \frac{k}{2(2^q + 1)} \cdot \frac{2^q + 1}{2} k \Delta = \frac{k^2}{4} \Delta = \text{bp}^q(\text{OPT}) && \text{(negating (8)).} \end{aligned}$$

Define $L := L' \cup \{z_0\} \setminus \{s\}$ and notice that $L \cup R = \text{OPT}$. By the minimality of the remote-bipartition function, $\text{bp}^q(\text{OPT}) \leq \sum_{\ell \in L} \sum_{r \in R} d^q(\ell, r)$. Hence,

$$\begin{aligned} \text{bp}^q(\text{OPT}') - \text{bp}^q(\text{OPT}) &\geq \sum_{\ell \in L'} \sum_{r \in R} d^q(\ell, r) - \sum_{\ell \in L} \sum_{r \in R} d^q(\ell, r) \\ &= \sum_{r \in R} [d^q(s, r) - d^q(z_0, r)] \\ &\geq \sum_{r \in R} [2^{-(q-1)} d^q(z_0, s) - 2d^q(z_0, r)] && \text{(by (2))} \\ &> \frac{|R|}{2^{q-1}} (6^q \Delta) - 2 \sum_{r \in R} d^q(z_0, r) && \text{(by (6))} \\ &\geq 3^q k \cdot \Delta - (2^q + 1)k \cdot \Delta \geq 0. && \text{(by (8)).} \end{aligned}$$

This completes the proof of the theorem. ◀

3.2 The algorithm

For any diversity function and a fixed optimal k -set, we refer to the ball $B := B(z_0, c\Delta^{1/q})$ defined in Theorem 5 as the *main cluster* and to z_0 as the *instance center*. Our algorithm consists of two phases: Finding the main cluster B and performing grid rounding on B . We remark that for a well-dispersed instance, B may well contain all input points. In that case, our algorithm amounts to a direct application of the grid rounding procedure.

Finding the main cluster

There are several possible ways to (approximately) find B . For simplicity, we present a naive approach based on exhaustive search. A smarter technique is described in the full version, where we provide a more refined algorithm for standard remote-clique.

Assuming without loss of generality that the instance diameter is 1, we obtain for each diversity function the bounds $1/k^2 \leq \Delta^{1/q} \leq 1$. Hence, by performing $O(\log k)$ trials, we can “guess” the value of $\Delta^{1/q}$ up to a constant factor arbitrarily close to one, which means that for any constant $\lambda > 0$, we can find an estimate Δ' so that $(1 - \lambda)\Delta^{1/q} \leq \Delta' \leq \Delta^{1/q}$. Similarly, by trying out all n input points, we can “guess” the instance center z_0 . For each one of these guesses, we perform the second phase (described in the next paragraph) and output the best k -set found over all trials. To simplify our exposition, we assume in what follows that we have found $\Delta^{1/q}$ and z_0 (and thus B) exactly. Our analysis can be adapted to any constant-factor estimation of $\Delta^{1/q}$, as it is enough to find a slightly larger ball B' containing B and to slightly change the value of constant c . More precisely, if we have an estimate Δ' so that $(1 - \lambda)\Delta^{1/q} \leq \Delta' \leq \Delta^{1/q}$ and we set $c' := \frac{c}{1-\lambda}$, then $B' := B(z_0, c'\Delta'^{1/q})$ is guaranteed to contain B and hence all points outside of B' are in OPT.

Rounding the cluster

We now assume that we have found the main cluster B (see the previous paragraph). For a constant $\delta > 0$ to be defined later, with $1/\delta = \Theta(2^q/\varepsilon)$, we perform a cell decomposition of radius $\delta\Delta^{1/q}$ over B . As the radius of ball B is $c\Delta^{1/q}$, this decomposition produces at most $(4 \cdot \frac{c\Delta^{1/q}}{\delta\Delta^{1/q}})^D = (4c/\delta)^D = O(2^q/\varepsilon)^D$ cells, i.e. constantly many cells. Let $\pi : B \rightarrow B$ be the function that maps each point to its cell center. For notational convenience, we extend this into a function $\pi : X \rightarrow X$ by applying the identity on $X \setminus B =: \bar{B}$ (and thinking of each point in \bar{B} as the center of its own cell). Finally, for any set $T \subseteq X$, we denote by $\hat{\pi}(T)$ the multiset over set $\pi(T)$ having multiplicities $|\pi^{-1}(u) \cap T|$ for each $u \in \pi(T)$.

Next, we perform exhaustive search to find a k -set T in X with the property that

$$\text{div}^q(\hat{\pi}(T)) \geq \text{div}^q(\hat{\pi}(\text{OPT})). \quad (9)$$

This can be done in polynomial time as follows: We try out all multisets in $\hat{\pi}(X)$ that a) contain \bar{B} and b) have cardinality k counting multiplicities. Then, we keep the multiset with largest diversity and return any k -set T that is a pre-image of this multiset. Clearly, this search considers only $k^{O(2^q/\varepsilon)^D}$ multisets and is bound to consider $\hat{\pi}(\text{OPT})$.

As mentioned in the introduction, our algorithm assumes access to a polynomial-time oracle that, for any k -set T , returns the value of $\text{div}^q(T)$ or a $(1 + \varepsilon)$ -factor estimate of it which is sufficient for our purposes. The use of this estimate produces a corresponding small deterioration in our final approximation guarantee, but for simplicity we ignore this in the remainder. No exact efficient algorithm is known to compute $\text{bp}^q(T)$ for a given k -set T . However, we provide a PTAS for this problem in the full version.

3.3 Analysis

What is the approximation guarantee of our algorithm? By an application of inequality (3), our cell decomposition gives the following guarantee for each pair of points.

► **Lemma 6.** *Let $\pi : X \rightarrow X$ be a map such that $d(u, \pi(u)) \leq \delta \Delta^{1/q}$ for each u in X . Then, for any pair of points $u, v \in X$,*

$$|d^q(u, v) - d^q(\pi(u), \pi(v))| \leq 2^{q+1} \delta \cdot (\Delta + \min\{d^q(u, v), d^q(\pi(u), \pi(v))\}).$$

Proof. We consider two cases. If $d(u, v) \leq d(\pi(u), \pi(v))$, we have by hypothesis

$$\begin{aligned} d^q(\pi(u), \pi(v)) &\leq [d(\pi(u), u) + d(u, v) + d(v, \pi(v))]^q \leq [d(u, v) + 2\delta \Delta^{1/q}]^q \\ &\leq d^q(u, v) + 2^{q+1} \delta \cdot \max\{\Delta, d^q(u, v)\} \leq d^q(u, v) + 2^{q+1} \delta \cdot (\Delta + d^q(u, v)), \end{aligned}$$

where we used inequality (3) in the second line. This proves the claim.

Similarly, if $d(\pi(u), \pi(v)) < d(u, v)$, then

$$d^q(u, v) \leq d^q(\pi(u), \pi(v)) + 2^{q+1} \delta \cdot (\Delta + d^q(\pi(u), \pi(v))),$$

which again proves the claim. ◀

Lemma 6, together with the definition of Δ , implies the following result whose proof is deferred to the full version.

► **Lemma 7.** *Let $\pi : X \rightarrow X$ be a map such that $d(u, \pi(u)) \leq \delta \Delta^{1/q}$ for each u in X . Then, for each one of our three diversity functions and for each k -set $T \subseteq X$,*

$$|\operatorname{div}^q(T) - \operatorname{div}^q(\hat{\pi}(T))| \leq 2^{q+1} \delta \cdot [\operatorname{div}^q(\operatorname{OPT}) + \operatorname{div}^q(T)] \leq 2^{q+2} \delta \cdot \operatorname{div}^q(\operatorname{OPT}).$$

Applying the previous lemma twice as well as inequality (9) once, we conclude that

$$\begin{aligned} \operatorname{div}^q(T) &\geq \operatorname{div}^q(\hat{\pi}(T)) - 2^{q+2} \delta \cdot \operatorname{div}^q(\operatorname{OPT}) \geq \operatorname{div}^q(\hat{\pi}(\operatorname{OPT})) - 2^{q+2} \delta \cdot \operatorname{div}^q(\operatorname{OPT}) \\ &\geq \operatorname{div}^q(\operatorname{OPT}) - 2^{q+3} \delta \cdot \operatorname{div}^q(\operatorname{OPT}) = (1 - 2^{q+3} \delta) \cdot \operatorname{div}^q(\operatorname{OPT}). \end{aligned}$$

Hence, in order to achieve an approximation factor of $1 - \varepsilon$, it suffices to select $\delta := \varepsilon/2^{q+3}$. The number of cells produced by the cell decomposition is thus bounded by $(2^{q+5}c/\varepsilon)^D = O(2^q/\varepsilon)^D$. This completes the analysis of our algorithm and the proof of Theorem 4.

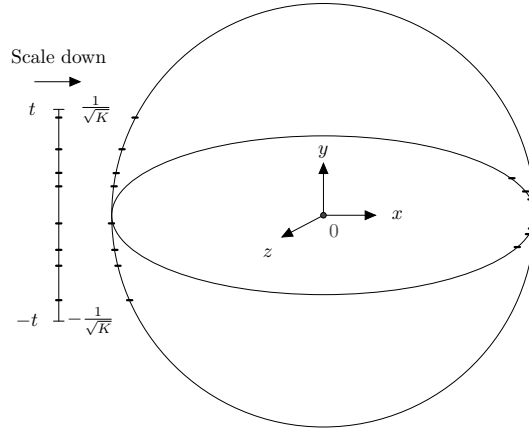
4 Proof of NP-hardness

In this section, we present the first proof of NP-hardness for any of the three diversity problems in fixed dimension (in fact, the only other diversity maximization problem known to be NP-hard in a fixed-dimensional setting is remote-edge [30]). In particular, we prove NP-hardness for the squared distances ($q = 2$) version of remote-clique in the case where all input points are *unit vectors* in the Euclidean space \mathbb{R}^3 , i.e. $X \subseteq \mathbb{S}^2$.

► **Theorem 8.** *The squared distances version ($q = 2$) of the remote-clique problem is NP-hard over the three-dimensional Euclidean space.*

We remark that squared Euclidean distances over unit vectors correspond precisely to the popular cosine distances, hence the case considered is highly relevant.

For a k -set $T \subseteq \mathbb{S}^2$ with Euclidean distances, the function $\operatorname{cl}^2(T) := \sum_{\{u,v\} \in \binom{T}{2}} d^2(u, v)$ has very particular geometric properties related to the concept of *centroid*. The centroid of a



■ **Figure 1** Reduction from K -SUM to remote-clique with $q = 2$, $|X| = 2|M|$ and $k = 2K$.

k -set T is defined as $z_T := \frac{1}{k} \sum_{u \in T} u$. It represents the coordinate-wise average of the points in T . The following result greatly simplifies the computation of function $\text{cl}^2(T)$ in terms of the centroid. We state it for a general dimension D even though we only use it for the case $D = 3$. Its proof is deferred to the full version.

► **Lemma 9.** For a k -set $T \subseteq \mathbb{S}^{D-1} \subseteq \mathbb{R}^D$ with centroid $z_T := \frac{1}{k} \sum_{u \in T} u$,

$$\text{cl}^2(T) = k^2 \cdot (1 - \|z_T\|^2).$$

We present a reduction from the K -SUM problem which is known to be NP-hard: Given a set M of integer numbers in the range $[-t, t]$ for some threshold t and a positive integer K , determine whether there is a K -set $S \subseteq M$ that sums to zero. Given such an instance of K -SUM, we define the following instance $X \subseteq \mathbb{S}^2$ of remote-clique with $q = 2$, $|X| = 2|M|$ and $k = 2K$, see Figure 1. For each $m \in M$, set $m' := \frac{m}{t\sqrt{K}}$ and define

$$X := \left\{ \ell_m := (-\sqrt{1 - m'^2}, m', 0)^\top : m \in M \right\} \cup \left\{ r_m := (\sqrt{1 - m'^2}, 0, m')^\top : m \in M \right\}.$$

Due to the scaling down by a factor of $\frac{1}{t\sqrt{K}}$, the y - and z -components of all points in X are upper bounded by $\frac{1}{\sqrt{K}}$ in absolute value, while their x -components are lower bounded by $\sqrt{1 - \frac{1}{K}}$ in absolute value. The points are thus tightly clustered around one of the two antipodal points $\pm(1, 0, 0)$, and X is partitioned into a *left cluster* and a *right cluster*.

From Lemma 9, it is clear that solving this instance of remote-clique is equivalent to finding the k -set whose centroid is closest to the origin. Hence, the proof of Theorem 8 is complete once we show the following claim.

► **Lemma 10.** If M has a K -set S with zero sum, then X has a k -set T with centroid $z_T = 0$. Otherwise, for every k -set $T \subseteq X$ we have $\|z_T\| \geq \frac{1}{2tK^{3/2}}$.

Proof. Suppose that M has a K -set S with zero sum and define the k -set $T := \{\ell_m, r_m : m \in S\} \subseteq X$. Recall that its centroid z_T corresponds to the component-wise average of the points in T , so we analyze these components separately. In z , all points of T on the left cluster are zero and those on the right cluster have a zero sum, so $(z_T)_z = 0$. In y , all points

of T on the right cluster are zero and those on the left cluster have a zero sum, so $(z_T)_y = 0$. And in x , each point ℓ_m of T on the left cluster is canceled out by its paired point r_m on the right cluster, so $(z_T)_x = 0$. Therefore, $z_T = 0$.

Finally, we prove the contrapositive of the second statement, i.e. we assume that there is a k -set $T \subseteq X$ with $\|z_T\| < \frac{1}{2tK^{3/2}}$. The set T must contain exactly K points in the left cluster and K points in the right cluster. Indeed, if T had at most $K - 1$ points in the left cluster, then the x -component of its centroid would give

$$(z_T)_x \geq (K - 1)(-1) + (K + 1)\sqrt{1 - \frac{1}{K}} \geq -(K - 1) + (K + 1)\left(1 - \frac{1}{K}\right) = 1 - \frac{1}{K},$$

and hence $\|z_T\| \geq |(z_T)_x| \geq 1 - \frac{1}{K} > \frac{1}{2tK^{3/2}}$ for $K \geq 2$ and $t \geq 1$, leading to a contradiction.

Let $T = L \cup R$ be the corresponding (balanced) bipartition of T given by the left and right clusters. Each of L and R must correspond to a K -set of M with zero sum. Otherwise, without loss of generality L corresponds to a K -set S of M with sum at least 1, but then

$$(z_T)_y = \frac{1}{2K} \sum_{m \in S} m' = \frac{1}{2tK^{3/2}} \sum_{m \in S} m \geq \frac{1}{2tK^{3/2}}$$

and thus $\|z_T\| \geq |(z_T)_y| \geq \frac{1}{2tK^{3/2}}$, again a contradiction. This completes the proof. ◀

References

- 1 Z. Abbassi, V. S. Mirrokni, and M. Thakur. Diversity maximization under matroid constraints. In *19th Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 32–40. ACM, 2013.
- 2 S. Aghamolaei, M. Farhadi, and H. Zarrabi-Zadeh. Diversity Maximization via Composable Coresets. In *27th Canadian Conference on Computational Geometry (CCCG)*, page 43, 2015.
- 3 N. Alon, S. Arora, R. Manokaran, D. Moshkovitz, and O. Weinstein. Inapproximability of densest κ -subgraph from average case hardness. *Unpublished manuscript*, 2011.
- 4 A. Bhaskara, M. Ghadiri, V. Mirrokni, and O. Svensson. Linear relaxations for finding diverse elements in metric spaces. In *Advances in Neural Information Processing Systems*, pages 4098–4106, 2016.
- 5 B. Birnbaum and K. J. Goldman. An improved analysis for a greedy remote-clique algorithm using factor-revealing LPs. *Algorithmica*, 55(1):42–59, 2009.
- 6 A. Borodin, H. C. Lee, and Y. Ye. Max-sum diversification, monotone submodular functions and dynamic updates. In *Proceedings of the 31st Symposium on Principles of Database Systems*, pages 155–166, 2012.
- 7 M. Ceccarello, A. Pietracaprina, G. Pucci, and E. Upfal. MapReduce and streaming algorithms for diversity maximization in metric spaces of bounded doubling dimension. *Proceedings of the VLDB Endowment*, 10(5):469–480, 2017.
- 8 A. Cevallos, F. Eisenbrand, and S. Morell. Diversity maximization in doubling metrics. *arXiv preprint*, 2018. [arXiv:1809.09521](https://arxiv.org/abs/1809.09521).
- 9 A. Cevallos, F. Eisenbrand, and R. Zenklusen. Max-Sum Diversity via Convex Programming. In *32nd Annual Symposium on Computational Geometry (SoCG)*, pages 26:1–26:14, 2016.
- 10 A. Cevallos, F. Eisenbrand, and R. Zenklusen. Local Search for Max-Sum Diversification. In *28th Symposium on Discrete Algorithms (SODA)*, pages 130–142. SIAM, 2017.
- 11 B. Chandra and M. M. Halldórsson. Approximation algorithms for dispersion problems. *Journal of algorithms*, 38(2):438–465, 2001.

- 12 V. Cohen-Addad, P. N. Klein, and C. Mathieu. Local search yields approximation schemes for k -means and k -median in Euclidean and minor-free metrics. In *57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 353–364. IEEE, 2016.
- 13 J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: survey, insights, and generalizations. *Journal of Machine Learning Research*, 16(1):2859–2900, 2015.
- 14 S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the 40th Symposium on Theory of Computing*, pages 537–546. ACM, 2008.
- 15 S. P. Fekete and H. Meijer. Maximum dispersion and geometric maximum weight cliques. *Algorithmica*, 38(3):501–511, 2004.
- 16 W. Fernandez de la Vega, M. Karpinski, and C. Kenyon. A Polynomial Time Approximation Scheme for Metric MIN-BISECTION. *Electronic Colloquium on Computational Complexity (ECCC)*, pages 1–12, 2002.
- 17 S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *18th International Conference on World Wide Web (WWW)*, pages 381–390. ACM, 2009.
- 18 L. A. Gottlieb and R. Krauthgamer. A nonlinear approach to dimension reduction. *Discrete & Computational Geometry*, 54(2):291–315, 2015.
- 19 S. Har-Peled. *Geometric approximation algorithms*, volume 173. American mathematical society Boston, 2011.
- 20 R. Hassin, S. Rubinstein, and A. Tamir. Approximation algorithms for maximum dispersion. *Operations Research Letters*, 21(3):133–137, 1997.
- 21 P. Indyk, S. Mahabadi, M. Mahdian, and V. S. Mirrokni. Composable core-sets for diversity and coverage maximization. In *33rd ACM Symposium on Principles of Database Systems*, pages 100–108, 2014.
- 22 P. Indyk and A. Naor. Nearest-neighbor-preserving embeddings. *ACM Transactions on Algorithms (TALG)*, 3(3):31, 2007.
- 23 L. Qin, J. X. Yu, and L. Chang. Diversifying top- k results. *Proceedings of the VLDB Endowment*, 5(11):1124–1135, 2012.
- 24 F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *29th SIGIR Conference on Research and Development in Information Retrieval*, pages 691–692. ACM, 2006.
- 25 S. S. Ravi, D. J. Rosenkrantz, and G. K. Tayi. Heuristic and special case algorithms for dispersion problems. *Operations Research*, 42(2):299–310, 1994.
- 26 A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.
- 27 J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- 28 N. Vasconcelos. Feature selection by maximum marginal diversity. In *Advances in Neural Information Processing Systems*, pages 1375–1382, 2003.
- 29 M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina, and V. J. Tsotras. On query result diversification. In *27th International Conference on Data Engineering (ICDE)*, pages 1163–1174. IEEE, 2011.
- 30 D.W. Wang and Y.S. Kuo. A study on two geometric location problems. *Information processing letters*, 28(6):281–286, 1988.