Discrete & Computational Geometry
https://doi.org/10.1007/s00454-024-00635-z

Bounds on Polarization Problems on Compact Sets via Mixed Integer Programming

Jan Rolfes[1,2] · **Robert Schüler**[3] · **Marc Christian Zimmermann**[4]

Received: 17 March 2023 / Revised: 23 January 2024 / Accepted: 25 January 2024

Abstract
Finding point configurations, that yield the maximum polarization (Chebyshev constant) is gaining interest in the field of geometric optimization. In the present article, we study the problem of unconstrained maximum polarization on compact sets. In particular, we discuss necessary conditions for local optimality, such as that a locally optimal configuration is always contained in the convex hull of the respective darkest points. Building on this, we propose two sequences of mixed-integer linear programs in order to compute lower and upper bounds on the maximal polarization, where the lower bound is constructive. Moreover, we prove the convergence of these sequences towards the maximal polarization.

Keywords Maximal polarization · Potentials · Mixed integer programming · Geometric optimization

Mathematics Subject Classification 31C20 · 51-08 · 90C11

Editor in Charge: Kenneth Clarkson

Jan Rolfes, Robert Schüler and Marc Christian Zimmermann contributed equally to this work

Jan Rolfes
jrolfes@kth.se

Robert Schüler
robert.schueler2@uni-rostock.de

Marc Christian Zimmermann
marc.christian.zimmermann@gmail.com

1 Department of Data Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany

2 Optimization and System Theory, KTH - Royal Institute of Technology, Lindtstedtsvägen 25, 114 28 Stockholm, Sweden

3 Institute for Mathematics, University of Rostock, Universitätsplatz 1, Rostock, 18051 Rostock, Germany

4 Abteilung Mathematik, Universität zu Köln, Weyertal 86-90, 50931 Cologne, Germany

Published online: 19 March 2024

1 Introduction

Suppose you were given a set A and N lamps you are to place such that the darkest point in A is as bright as possible. In less descriptive terms this max-min problem is known as the maximal polarization problem, which we now state in mathematical language.

Let $A, D \subset \mathbb{R}^n$ be nonempty sets and let $K : A \times D \to \mathbb{R} \cup \{+\infty\}$ be a function bounded from below. An N-point multiset $C \subseteq D$ will be referred to as *point configuration* (*of N points*) and the set of all N-point configurations supported on D will be denoted by $\mathcal{C}$. We assign the discrete K-potential associated with C to every point $p \in A$ as

$$U_{K,A}(p, C) = \sum_{c \in C} K(p, c).$$

To any point configuration we associate its *polarization*

$$P_{K,A}(C) = \inf_{p \in A} U_{K,A}(p, C).$$

It is then natural to consider the *(maximal) polarization problem*:

$$\mathcal{P}_K(A) = \sup_{C \in \mathcal{C}} P_{K,A}(C). \tag{1}$$

For a broader context and overview of this formulation of the polarization problem we refer to the recent monograph [1, CH. 14]. Problems of this kind have been extensively studied. In particular the case of $A = D = S^{n-1}$ being a unit sphere and $K(x, y) = \|x - y\|^{-s}$ being related to a Riesz potential is rich in results on explicit optimal configurations of few points (eg. [2–8]), bounds on maximal polarization (eg. [4, 7]) and asymptotic results (eg. [9–12]). Asymptotic results are also available for more general choices of A, such as rectifiable sets.

Moreover, the polarization problem as stated in (1) is closely related to the well-studied covering problem, i.e. the question, whether A can be covered by N balls of radius $r > 0$. In particular, let $K(x, y) = \mathbb{1}_{[0,r]}(\|x - y\|)$, then, a covering with N balls exists if and only if $1 \leq \mathcal{P}_K(A)$. General discussions of covering problems can be found, for example in the seminal book by Conway and Sloane [13]. For covering problems on compact metric spaces we refer to [14] for an overview, whereas constructive methods have been developed, e.g. in [15] and [16].

In this paper we consider polarization problems of the following kind. The set $A \subset \mathbb{R}^n$ will be a compact set and we will impose no restrictions on the point configurations, i.e. $D = \mathbb{R}^n$. Furthermore, we restrict to functions $K(x, y) = f(\|x - y\|)$ for some continuous strictly monotone decreasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$ and use the notation $U_{f,A}(p, C)$, $P_{f,A}(C)$, $\mathcal{P}_f(A)$. If the subscript parameters are clear from context we omit them.

Under the above assumptions, we therefore consider the optimization problem

$$\mathcal{P}_f(A) = \sup_{C \subset \mathcal{C}} P_{f,A}(C). \tag{2}$$

For explicit computations we choose Gaussians $f(x) = e^{-ax^2}$. These functions appear rather naturally in the context of universal optimality (cf. [17]): Recall that a function $g : (0, \infty) \to \mathbb{R}$ is *completely monotonic* if it is infinitely differentiable and the derivatives satisfy $(-1)^k g^{(k)} \geq 0$ for all k. The functions $g(x) = e^{-\alpha x}$ are completely monotonic and we can write $f(\|x - y\|) = g(\|x - y\|^2)$. In this context functions $f(x) = g(x^2)$ are called completely monotonic functions of squared distance.

A Theorem of Bernstein (cf. [18, Thm. 9.16]) asserts that every completely monotonic function can be written as a convergent integral

$$g(x) = \int_{[0,\infty)} e^{-\alpha x} d\mu(\alpha).$$

From this one obtains that the set of completely monotonic functions of squared distance is the cone spanned by the Gaussians and the constant function $x \mapsto 1$.

In particular the commonly used Riesz potentials can be written in this way.

We fix some more notation for the case that the infimum $P_{f,A}$ is in fact a minimum, i.e. the minimizers of this function are points in A. In this case, any such minimizer will be called a *darkest point* of A. Moreover,

$$\mathrm{Dark}_A(C) = \left\{ p \in A \ : \ \sum_{c \in C} f(\|p - c\|) = P_{f,A}(C) \right\}$$

will be called the *set of darkest points* of C. To explain this wording we invite the reader to recall the interpretation of the problem we gave in the beginning: we center lamps at the points in C which now illuminate A. The polarization of A is then the lowest level of brightness any point in A can have, any point realizing this is a "darkest point".

Note, that requiring A to be compact is rather natural. Indeed if A were unbounded, then the value of the polarization would always tend to $N \cdot \inf f$. If A were not closed, darkest points need not exist. Consider for example A to be the open disc and C only containing the origin. In this case, $P_{f,A}(C)$ is not attained at any point in A.

In Sect. 2 we provide some results connecting a locally optimal configuration to the set of its respective darkest points. Theorem 2.1 states that the points of such a configuration are contained in the convex hull of the darkest points while on the other hand Theorem 2.5 states that the darkest points are located either on the boundary of A or in the interior of the convex hull of the configuration. These restrictions provide necessary conditions for optimality.

In Sect. 3 we investigate mixed-integer approximations of the polarization problem providing upper and lower bounds. These are collected in Theorem 3.5. We then prove that these bounds indeed converge to $\mathcal{P}_f(A)$ in Theorems 3.8 and 3.9.

In Sect. 4 we illustrate capabilities and limitations of the approach on some benchmark instances.

2 Darkest Points and Necessary Conditions

In this section, we investigate structural properties an optimal configuration needs to satisfy in order to potentially falsify the optimality of a given polarization and reduce the search space of optimal configurations.

In particular, we have the following necessary condition that relates local optimality of a configuration to the set of its darkest points:

Theorem 2.1 *If C is a locally optimal solution of (2), then*

$$C \subset \operatorname{conv} \operatorname{Dark}_A(C).$$

Proof Suppose C is a configuration for which we have $c \in C$ such that $c \notin \operatorname{conv}(\operatorname{Dark}_A(C))$. In the following, we discuss how to construct a new configuration C' in an arbitrary neighbourhood of C such that $P(C') > P(C)$. Thus C can not be locally optimal. Since f is continuous, the niveau line

$$S = \{p \in \mathbb{R}^n \,:\, U(p, C) = P(C)\}$$

containing the darkest points is closed and thus $\operatorname{Dark}_A(C) = A \cap S$ is compact. Therefore, $\operatorname{conv}(\operatorname{Dark}_A(C))$ is a compact convex set and we can find a hyperplane $H = \{x \,:\, a^\top x = b\}$ strictly separating this set from c such that $a^\top c < b$. For $\varepsilon > 0$ small enough $c' = c + \varepsilon a$ still satisfies $a^\top c' < b$. We obtain a new configuration $C' = C \cup \{c'\} \setminus \{c\}$. Note, that for every neighbourhood of C, there is a sufficiently small ε such that C' is contained in said neighbourhood. Obviously $|c' - p| < |c - p|$ for all points p in the non-negative halfspace of H. In particular c' is closer to all of the darkest points than c and since f is monotonously decreasing

$$U(p, C') > U(p, C) \geq P(C)$$

for all points p in the non-negative halfspace of H.

It remains to assert this also on the negative halfspace. Since all the darkest points are on the positive side of H, a point $p \in A \cap (H \cup H_-)$ satisfies

$$U(p, C) > P(C).$$

Since $A \cap (H \cup H_-)$ is compact this yields

$$U(p, C) \geq \delta > P(C)$$

for some constant δ. By continuity of f, for ε small enough, we can guarantee that

$$U(p, C') > P(C)$$

for all $p \in A \cap (H \cup H_-)$. Altogether,

$$P(C') = \inf_{p \in A} U(p, C') > P(C).$$

□

The formulated condition is very "unstable" in the following sense:

Proposition 2.2 *Let C be a configuration such that $C \subset \operatorname{conv} \operatorname{Dark}_A(C)$. Let $c \in C$ and $c' \neq c$ and $C' = C \cup \{c'\} \setminus \{c\}$. Then*

1. *$P(C') < P(C)$ and*
2. *$C' \not\subseteq \operatorname{conv} \operatorname{Dark}_A(C')$.*

Proof 1. Consider the hyperplane H with outer normal $c - c'$ through c, oriented such that c' is on the negative side. Since $c \in \operatorname{conv} \operatorname{Dark}_A(C)$ there has to be a darkest point $d \in \operatorname{Dark}_A(C)$ in the non-negative halfspace of H (it might be in H). Then $\|c - d\| < \|c' - d\|$ and by monotonicity $f(\|c - d\|) > f(\|c' - d\|)$. The potentials $U(d, C')$ and $U(d, C)$ differ by $f(\|c' - d\|) - f(\|c - d\|)$, therefore the above implies

$$P(C') \leq U(d, C') < U(d, C) = P(C).$$

2. Suppose $C' \subset \operatorname{conv} \operatorname{Dark}_A(C')$. Then we can apply 1. to C' with the roles of c, c' reversed. But this would give $P(C') < P(C) < P(C')$, which is a contradiction.

□

Optimization methods which only consider single components (like pattern search) or move single configuration points therefore possibly converge to a configuration contained in the convex hull of the darkest points which is not locally optimal. Therefore it seems reasonable to only use optimization methods which are able to move several points at once. Another conclusion is the following, which seems to suggest that the number of optimization variables can be reduced to only $N - 1$ vectors.

Corollary 2.3 *For given points C' with $|C'| = N - 1$ there is* at most *one point c such that $\{c\} \cup C' \subset \operatorname{conv} \operatorname{Dark}_A(\{c\} \cup C')$.*

We can use Theorem 2.1 to study the structure of the darkest points even more. First, we discuss a way to find certificates for $p \notin \operatorname{Dark}_A(C)$. To this end, we recall that the conic hull of a set $S \subseteq \mathbb{R}^n$ is given by $\operatorname{cone}(S) = \bigcap_{K \supset S:\ K \text{ is a convex cone}} C$.

Lemma 2.4 *Let C be a configuration and $p \in \mathbb{R}^n$ be an arbitrary point. Let*

$$N(p, C) = \{p + v \ :\ v \neq 0 \text{ and } v^\top w \geq 0 \text{ for all } w \in \operatorname{cone}\{p - c \ :\ c \in C\}\}.$$

1. *For all $q \in N(p, C)$ we have $U(q, C) < U(p, C)$,*
2. *if $N(p, C) \cap A \neq \emptyset$ then $p \notin \operatorname{Dark}_A(C)$.*

Proof Write $q = p + v \in N(p, C)$ with $v \neq 0$. Then for all $c \in C$ we have

$$|c - (p + v)|^2 = |c - p|^2 + 2(p - c)^\top v + |v|^2 > |c - p|^2.$$

Since f is strictly monotone decreasing, we have $U(q, C) < U(p, C)$. From this, the second claim follows immediately. □

Observe that Lemma 2.4 1 implies, that $N(p, C)$ contains only points whose potential is strictly smaller than the potential at p. If we recall the visualization of the polarization problem as placing light sources C to illuminate A the above definition of $N(p, C)$ of a point p contains only points that are illuminated less than p itself. It is closely related to the idea of a physical shadow as $p + \text{cone}\{p - c'\}$ with $c' \in C$ can be seen as the set of points lying in the shadow thrown by object p with respect to light source c'. In this interpretation $p + \text{cone}\{p - c \ : \ c \in C\}$ resembles the shadow with respect to all light sources simultaneously. Note that $N(p, C)$ is defined by replacing $\text{cone}\{p - c \ : \ c \in C\}$ by its dual cone[1]. With this we prove the following result which further restricts the location of the darkest points:

Theorem 2.5 *Let C be a feasible configuration for (2). Then the points of* $\text{Dark}_A(C)$ *are either in the interior of* $\text{conv}(C)$ *or in δA, i.e.* $\text{Dark}_A(C) \subset \text{int conv}(C) \cup \delta A$. *Moreover, if C is locally optimal for (2), then* $\text{Dark}_A(C) \cap \delta A \neq \emptyset$.

Proof Let $p \in \text{Dark}_A(C)$ and assume $p \notin \text{int conv}(C)$. Furthermore, let $N(p, C)$ be defined as in Lemma 2.4. We can find a hyperplane $H = \{x \in \mathbb{R}^n \ : \ a^\top x = \beta\}$ through p separating C from p, in particular $a^\top c \leq \beta$ for all $c \in C$. Then for all $c \in C$

$$a^\top (p - c) = a^\top p - a^\top c = \beta - a^\top c \geq 0,$$

which shows that $p + \lambda a \in N(p, C)$ for arbitrary $\lambda > 0$. If $p \in \text{int } A$, so is $p + \lambda a$ for λ sufficiently small. Then $A \cap N(p, C) \neq \emptyset$ in contradiction to Lemma 2.4. Thus $p \in \partial A$ as claimed.

In addition, if C is also locally optimal for (2) by Theorem 2.1 we immediately obtain that $C \subset \text{conv Dark}_A(C)$. Now, assume $\text{Dark}_A(C) \cap \delta A = \emptyset$, then as seen above $\text{Dark}_A(C) \subseteq \text{int conv } C$ and we obtain

$$C \subseteq \text{conv Dark}_A(C) \subseteq \text{int conv } C,$$

which is a contradiction since C is finite. □

To summarize, locally optimal configurations C of (2) and its corresponding darkest points $\text{Dark}_A(C)$ share a similar containment property as is illustrated in Fig. 1.

[1] Recall that the dual cone of a convex cone $C \subseteq \mathbb{R}^n$ is the cone $\{v \in \mathbb{R}^n \ : \ v^\top w \geq 0 \text{ for all } w \in C\}$

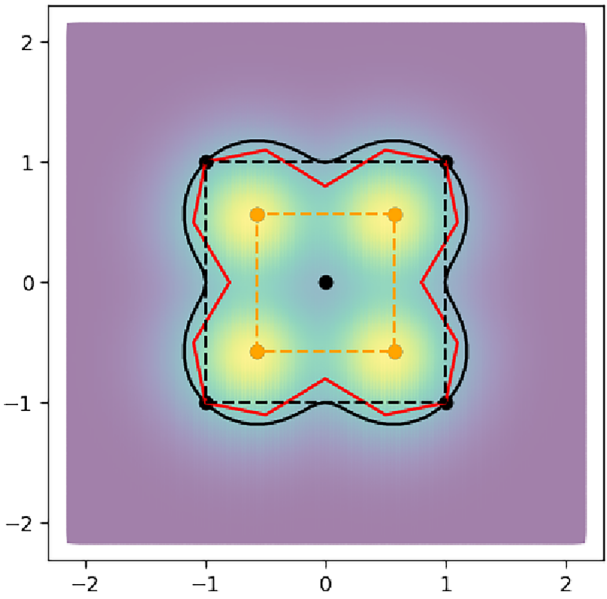

Fig. 1 Illustration of Theorems 2.1 and 2.5. A is depicted in red, Dark_A in black and the configuration C in orange. The dashed lines depict the convex hulls $\operatorname{conv} C$ and $\operatorname{conv} \mathrm{Dark}_A(C)$, whereas the black line depicts all points $p \in \mathbb{R}^2$, such that $U(p, C) = P(C)$

3 An MIP Approach to Polarization

The current section is dedicated to the development of two hierarchies of mixed-integer linear programs (MIP) that approximate the maximal polarization of a compact set A with respect to a monotonically decreasing and continuous function $f : \mathbb{R}_+ \to \mathbb{R}_+$. The MIP, that computes the lower bounds is constructive, i.e. solutions to this MIP are configurations whose polarization is lower bounded by the value of the MIP. The actual polarization of these configurations may very well exceed this lower bound by a significant margin, cf. Figure 3 for some numerical evidence.

First we give an equivalent description of problem (2). For this we observe that by Theorem 2.1 any locally optimal point configuration is necessarily supported on $\operatorname{conv}(A)$. Furthermore we can get rid of the infimum by adding new constraints. The resulting optimization problem is then

$$\begin{aligned} \mathcal{P}_f(A) = \max_{x,C}\ & x \\ & C \in \begin{bmatrix} \operatorname{conv}(A) \\ N \end{bmatrix} \\ & x \le U_{f,A}(p, C) \qquad \text{for all } p \in A, \end{aligned} \tag{3}$$

where $\begin{bmatrix} X \\ N \end{bmatrix}$ describes the set of all multisets of size N with elements in X. It is now clear, that the sup is actually a max, since the feasible region can easily be made

compact by bounding x from below (e.g. $x \geq 0$) without changing the value of the program.

3.1 MIP Hierarchies

We observe that Problem (3) is an optimization problem with finitely many variables (namely x, C), but infinitely many constraints - it is a semiinfinite program (SIP) - and therefore not solvable using standard solvers. In the remainder of this section we introduce two hierarchies of (tractable) MIPs, that approximate $\mathcal{P}(A)$ from above and below (see Theorem 3.5). For this we make use of the following concept of functions which "control" the difference of two values of f.

Definition 3.1 We call a family of functions $g_{c,p} : \mathbb{R}_+ \to \mathbb{R}_+$ for $c \in \text{conv}(A)$, $p \in A$ a *family of control functions* (with respect to f, A) if for all $c \in \text{conv}(A)$, $p \in A$:

1. $g_{c,p}(0) = 0$,
2. $g_{c,p}$ is continuous and non-decreasing,
3. $|f(\|c - p\|) - f(\|c' - p\|)| \leq g_{c,p}(\|c - c'\|)$ for all $c' \in \text{conv}(A)$,
4. $|f(\|c - p\|) - f(\|c - p'\|)| \leq g_{c,p}(\|p - p'\|)$ for all $p \in A$,

where $\|\cdot\|$ denotes the standard Euclidean norm.

Note that f is related to a function K taking two points c, p as arguments: $K(c, p) = f(\|c - p\|)$. A family of control functions allows us to control the way K changes as we vary either c or p.

This control will be an important ingredient of the proof of Theorem 3.5. For continuous functions this is related to bounding the slope of K as can be illustrated by the following example: Suppose the function $K(c, \cdot) = f(\|c - \cdot\|)$ is Lipschitz-continous with Lipschitz constant L for all $p \in A$. Then, $g_{c,p}(\varepsilon) = L \cdot \varepsilon$ is a valid control function for f.

However, applying global Lipschitz-continuity is not a very precise approximation as it ignores local information around specific points c, p. Therefore we provide a more suitable family of control functions.

Proposition 3.2 *For f monotonously decreasing and continuous the following is a family of control functions:*

$$g_{c,p}(\varepsilon) = \max(\hat{g}_{c,p}(\varepsilon), \hat{g}_{c,p}(-\varepsilon)),$$

where

$$\hat{g}_{c,p}(x) = \begin{cases} f(0) - f(\|c - p\|) \text{ if } x < -\|c - p\|, \\ |f(\|c - p\| + x) - f(\|c - p\|)| \text{ otherwise.} \end{cases}$$

Proof We fix c, p and write $g = g_{c,p}$ and $\hat{g} = \hat{g}_{c,p}$. Clearly $g(0) = \hat{g}(0) = 0$. Since f is continuous, so is g.

For $x \in (-\infty, -\|c-p\|)$ the function $\hat{g}(x)$ is constant. For $x \in (-\|c-p\|, 0)$ we have

$$\hat{g}(x) = f(\|c-p\| + x) - f(\|c-p\|),$$

which is decreasing since f is decreasing. For $x \in (0, \infty)$ we have

$$\hat{g}(x) = f(\|c-p\|) - f(\|c-p\| + x),$$

which is increasing since f is decreasing. Overall $g(\varepsilon) = \max(\hat{g}(\varepsilon), \hat{g}(-\varepsilon))$ is an increasing function on $\mathbb{R}_+$.

By symmetry, it is sufficient to prove that g provides an upper bound for $\Delta = |f(\|c-p\|) - f(\|c-p'\|)|$ for all $p' \in \mathrm{conv}(A)$. To this end, we use the triangle inequalities

$$\|c-p\| - \|p'-p\| \le \|c-p'\| \le \|c-p\| + \|p'-p\|$$

and that f is a decreasing function. Then, on the one hand if $\|c-p\| \le \|c-p'\|$, we have

$$\begin{aligned}\Delta &= f(\|c-p\|) - f(\|c-p'\|)\\ &\le f(\|c-p\|) - f(\|c-p\| + \|p'-p\|) = \hat{g}(\|p'-p\|) \le g(\|p'-p\|).\end{aligned}$$

On the other hand, if $\|c-p\| \ge \|c-p'\|$, we obtain

$$\begin{aligned}\Delta &= -f(\|c-p\|) + f(\|c-p'\|)\\ &\le -f(\|c-p\|) + f(\|c-p\| - \|p-p'\|) = \hat{g}(-\|p-p'\|) \le g(\|p-p'\|).\end{aligned}$$

□

For explicit computations we need to discretize two aspects of the problem. Firstly, we discretize the set of possible point configurations. For this we choose a finite sample $\Lambda \subset \mathrm{conv}(A)$ and only optimize over

$$C \in \begin{bmatrix} \Lambda \\ N \end{bmatrix}. \tag{4}$$

Secondly, we replace the infinite number of constraints, parameterized by A, by a finite subcollection. For this we again choose a finite sample $\Gamma \subset A$, and only consider the inequalities

$$x \le U(p, C) \quad \text{for all } p \in \Gamma. \tag{5}$$

However, this naively sampled problem is not necessarily connected to the original problem, since we enforce only a subset of the infinitely many constraints and allow only a finite number of configurations. Either one of these changes would provide

valid bounds but they unfortunately work in different directions. We will now show how to overcome this problem by utilizing the above family of control functions to obtain lower and upper bounds on the original problem.

Let us first consider lower bounds on (3). It is clear that we can restrict the choice of configurations to be supported on a finite sample Λ of $\operatorname{conv}(A)$ as in (4) and obtain a program that computes a lower bound.

Discretizing the constraints is the harder part, since removing constraints lets the maximum grow. The following lemma shows how a slight variation of discretized constraints for some finite sample Γ of A imply the validity of all of the infinitely many original constraints.

Lemma 3.3 *Let $g_{c,p}$ be a family of control functions. Let $\varepsilon > 0$, Λ be an arbitrary finite sample of* $\operatorname{conv}(A)$ *and Γ be an ε-net of A. Furthermore, suppose $x \in \mathbb{R}$, $C \subset \begin{bmatrix} \Lambda \\ N \end{bmatrix}$ satisfy*

$$x \leq \sum_{c \in \Lambda} \mathbb{1}_C(c) \cdot \big(f(\|c - p\|) - g_{c,p}(\varepsilon)\big) \quad \textit{for all } p \in \Gamma.$$

Then

$$x \leq \sum_{c \in \Lambda} \mathbb{1}_C(c) \cdot f(\|c - p\|) \quad \textit{for all } p \in A.$$

Proof Let $p \in A$ be arbitrary and $n(p) = \operatorname{argmin}_{\bar{p} \in \Gamma}\{\|p - \bar{p}\|\}$ denote the closest sample point to $p \in A$. Note that $\|p - n(p)\| < \varepsilon$ since Γ is an ε-net. Then

$$\begin{aligned}
\sum_{c \in \Lambda} \mathbb{1}_C(c) f(\|c - p\|) &= \sum_{c \in \Lambda} \mathbb{1}_C(c) \cdot (f(\|c - p\|) - f(\|c - n(p)\|)) \\
&\quad + \sum_{c \in \Lambda} \mathbb{1}_C(c) \cdot \big(f(\|c - n(p)\|) - g_{c,n(p)}(\varepsilon)\big) \\
&\quad + \sum_{c \in \Lambda} \mathbb{1}_C(c) \cdot g_{c,n(p)}(\varepsilon) \\
&\geq - \sum_{c \in \Lambda} \mathbb{1}_C(c) \cdot g_{c,n(p)}(\|p - n(p)\|) \\
&\quad + x \\
&\quad + \sum_{c \in \Lambda} \mathbb{1}_C(c) \cdot g_{c,n(p)}(\varepsilon),
\end{aligned}$$

which is larger than x since $g_{c,n(p)}$ is non-decreasing and $\|p - n(p)\| < \varepsilon$. □

Conversely, if we consider upper bounds on (3), we now cannot simply choose a finite sample Λ of $\operatorname{conv}(A)$ to approximate the above SIP. Indeed this would restrict the set of feasible solutions of (3) and thereby lower the maximum instead. Again, the following lemma provides a way around this problem using a variation of the constraints.

Lemma 3.4 *Let $g_{c,p}$ be a family of control functions. Let $\varepsilon > 0$ and Λ be an ε-net of* $\mathrm{conv}(A)$. *Furthermore, suppose $C \in \begin{bmatrix} \mathrm{conv}(A) \\ N \end{bmatrix}$ and x satisfy*

$$x \leq U(p, C) = \sum_{c \in C} f(\|c - p\|) \quad \text{for all } p \in \Gamma.$$

Then, there exists a configuration $C' \in \begin{bmatrix} \Lambda \\ N \end{bmatrix}$ such that

$$x \leq \sum_{c \in C'} f(\|c - p\|) + g_{c,p}(\varepsilon) \quad \text{for all } p \in \Gamma.$$

Proof Let $C' = \{\{n(c) \,:\, c \in C\}\}$ where $n(c) = \mathrm{argmin}_{c' \in \Lambda} \|c - c'\|$. Then

$$\begin{aligned} \sum_{c \in C'} f(\|c - p\|) + g_{c,p}(\varepsilon) &= \sum_{c \in C} f(\|n(c) - p\|) - f(\|c - p\|) \\ &\quad + \sum_{c \in C} f(\|c - p\|) + \sum_{c \in C} g_{n(c),p}(\varepsilon) \\ &\geq -\sum_{c \in C} g_{n(c),p}(\|c - n(c)\|) + x + \sum_{c \in C} g_{n(c),p}(\varepsilon) \geq x, \end{aligned}$$

where the last inequality holds since $g_{n(c),p}$ is non-decreasing and $\|c - n(c)\| < \varepsilon$ as Λ is an ε-net of $\mathrm{conv}(A)$. □

Now we can prove the main result of this section.

Theorem 3.5 *Let $\varepsilon_\Lambda, \varepsilon_\Gamma > 0$ and Λ be an ε_Λ-net of* $\mathrm{conv}(A)$ *and Γ be an ε_Γ-net of A. Furthermore, let $g_{c,p}$ be a family of control functions. Then we have the following:*

$$\begin{aligned} &\max\ x && && \text{(6a)} \\ &\quad y \in \{0, \ldots, N\}^\Lambda \\ &\quad \mathbb{1}^\top y = N \\ &\quad x \leq \sum_{c \in \Lambda} y_c \cdot (f(\|c - p\|) - g_{c,p}(\varepsilon_\Gamma)) && \text{for all } p \in \Gamma \\ &\leq \max\ x && && \text{(6b)} \\ &\quad y \in \{0, \ldots, N\}^\Lambda \\ &\quad \mathbb{1}^\top y = N \\ &\quad x \leq \sum_{v \in \Lambda} y_c \cdot f(\|c - p\|) && \text{for all } p \in A \\ &\leq \mathcal{P}(A) && && \text{(6c)} \\ &\leq \max\ x && && \text{(6d)} \end{aligned}$$

$$
\begin{aligned}
& C \in \begin{bmatrix} \operatorname{conv}(A) \\ N \end{bmatrix} \\
& x \leq \sum_{c \in C} f(\|c - p\|) && \text{for all } p \in \Gamma \\
\leq \max\ & x && \text{(6e)} \\
& y \in \{0, \ldots, N\}^{\Lambda} \\
& \mathbb{1}^{\top} y = N \\
& x \leq \sum_{c \in \Lambda} y_c \cdot (f(\|c - p\|) + g_{c,p}(\varepsilon_{\Lambda})) && \text{for all } p \in \Gamma \qquad \text{(6f)}
\end{aligned}
$$

Proof We show, that feasible solutions of the left hand sides are also feasible for the right hand sides with the same objective value justifying the asserted inequalities. First, observe that Lemma 3.3 implies that a feasible solution x, y of (6a) is also feasible for (6b) and the objective values coincide. Next, we consider a feasible solution x, y of (6b) and observe that y encodes a multiset $C \in \begin{bmatrix} \Lambda \\ N \end{bmatrix} \subseteq \begin{bmatrix} \operatorname{conv}(A) \\ N \end{bmatrix}$. Moreover, x, C satisfy the constraints in (2) and with the same objective value x. The next inequality follows rather immediately since (6d) is a relaxation of (2) due to dropping constraints for $p \in A \setminus \Gamma$. Lastly, if x, C is a feasible solution of (6d), we apply Lemma 3.4 to obtain a set $C' \in \begin{bmatrix} \Lambda \\ N \end{bmatrix}$ satisfying the constraints of (6e). Then, by encoding C' through $y \in \{0, \ldots, N\}^{\Lambda}$ with $\mathbb{1}^{\top} y = N$ we obtain a feasible solution to (6e) with the same objectve value x. □

Let us briefly comment on the computational complexity of the mixed-integer programs (6a) and (6e). It is worth noting, that mixed-integer linear programming usually refers to optimization problems that include binary variables, which run significantly faster. We would like to note that the integral variables $y \in \{0, \ldots, N\}^{\Lambda}$ in both (6a) and (6e) can be replaced by $|\Lambda| \cdot \log(N)$ binary variables.

Moreover, in the lower bound of Theorem 3.5 the vector y can be chosen as $y \in \{0, 1\}^{\Lambda}$, which still provides a (potentially worse) lower bound and reduces the number of binary variables significantly. Unfortunately, a similar simplification is not immediately possible for the upper bound. However, we introduce another concept which aims to reduce the computational complexity in a similar fashion in the upper bound case.

Definition 3.6 A finite subset $\Lambda \subset \mathbb{R}^n$ is called an (ε, k)-net of A if

1. $\Lambda \subset A$,
2. For every $p \in A$ there are at least k distinct points $p_1, \ldots, p_k \in \Lambda$ such that $|p_i - p| < \varepsilon$.

Using an $(\varepsilon_{\Lambda}, N)$-net we obtain a hierarchy similar to Theorem 3.5 restricting the possible entries of y to $\{0, 1\}$.

Proposition 3.7 *Let $\varepsilon_\Lambda, \varepsilon_\Gamma > 0$ and Λ be an (ε_Λ, N)-net of* $\operatorname{conv}(A)$ *and Γ be an ε_Γ-net of A. Furthermore, let $g_{c,p}$ be a family of control functions. Then,*

$$
\begin{aligned}
(6d) \leq \max\ & x \\
& y \in \{0,1\}^\Lambda \\
& \mathbb{1}^\top y = N \\
& x \leq \sum_{c\in\Lambda} y_c \cdot (f(\|c-p\|) + g_{c,p}(\varepsilon_\Lambda)) \qquad \text{for all } p \in \Gamma
\end{aligned}
$$

Proof The proof works similar to the proof of Theorem 3.5 by replacing $C' = \{\{n(c) : c \in C\}\}$ in the proof of Lemma 3.4 by a set C' of N distinct points of Λ. This is possible since Λ is an (ε_Λ, N)-net (see Definition 3.6). □

A trivial example of an (ε, N)-net can basically be obtained by a multiset consisting of N copies of an ε-net. However, in practise there are usually solutions that need fewer points, albeit more than a classical ε-net.

3.2 Convergence Results

After establishing upper and lower bounds to $\mathcal{P}(A)$ through the hierarchies presented in Theorem 3.5, we study the quality of these bounds. To this end, we show in this section, that solutions of the bounding problems (6a) and (6e) converge, as $\varepsilon_\Lambda, \varepsilon_\Gamma$ both tend to 0, to a solution of the original problem (3). Both proofs rely in large parts on the proof of Lemma 6.1 in [19], which proves similar convergence for more general semiinfinite programs, but include minor necessary modifications. At first, we focus on the lower bounds, i.e., we show, that (6a) converges to (6b) as $\varepsilon_\Gamma \to 0$:

Theorem 3.8 *Let (ε_k) be a non-negative sequence converging towards* 0*. Furthermore, for every $k \in \mathbb{N}$ choose an ε_k net Γ_k of A. Then, any accumulation point of a sequence $(x_k, y_k)_{k\in\mathbb{N}}$ of optimal solutions of* (6a) *w.r.t. Γ_k and ε_k is an optimal solution of* (6b)*.*

Proof Let $(\bar{x}, \bar{y})$ be an accumulation point of (x_k, y_k). By passing to a subsequence we can assume that $(x_k, y_k) \to (\bar{x}, \bar{y})$ if $k \to \infty$. We are now going to prove, that $(\bar{x}, \bar{y})$ is feasible and in fact optimal for (6b):

Consider an arbitrary $p \in A$ and observe that since Γ_k is an ε_k-net of A, there exists a sequence (p_k) with $p_k \in \Gamma_k$ such that $p_k \to p$ as $k \to \infty$. We observe further, that for all k we have

$$
x_k \leq \sum_{c\in\Lambda} (y_k)_c \cdot (f(\|c-p_k\|) - g_{c,p_k}(\varepsilon_k)) \leq \sum_{c\in\Lambda} (y_k)_c \cdot f(\|c-p_k\|)
$$

and by taking limits

$$
\bar{x} \leq \sum_{c\in\Lambda} \bar{y}_c \cdot f(\|c-p\|).
$$

Hence, $(\bar{x}, \bar{y})$ is feasible for (6b).

Now, let (x, y) be an arbitrary solution to (6b). Since A is compact and $\varepsilon_k > 0$, we know that $g_c = \max_{p \in A} g_{c,p}$ is a continuous, monotonously non-decreasing function with $g_c(0) = 0$. We now observe, that

$$(x - \sum_{c \in \Lambda} y_c \cdot g_c(\varepsilon_k), y)$$

is feasible for (6a) with respect to Γ_k. Since (x_k, y_k) is an optimal solution to (6a), we have $x_k \geq x - \sum_{c \in \Lambda} y_c \cdot g_c(\varepsilon_k)$. Consequently, as $g_c(0) = 0$, in the limit we obtain that $\bar{x} \geq x$. Since x was chosen arbitrarily, we conclude, that $(\bar{x}, \bar{y})$ is indeed optimal for (6b). □

Note that the convergence of (6b) to (6c) as $\varepsilon_\Lambda \to 0$ follows directly since the utility function and f are continuous. Thus, Theorem 3.8 implies the convergence of (6a) to (6c), i.e. the value of (6a) tends to $\mathcal{P}(A)$, as $\varepsilon_\Lambda, \varepsilon_\Gamma \to 0$.

Moreover, with the same arguments, we conclude the convergence of (6d) to (6c) as $\varepsilon_\Gamma \to 0$ and thus only one proof of convergence remains, namely that (6e) converges to (6d) as $\varepsilon_\Lambda \to 0$.

One difficulty of the following theorem is the different kinds of feasible solutions when altering the sample Λ. Feasible solutions of (6e) have the form $y \in \{1, \ldots, N\}^\Lambda$ with $\mathbb{1}^\top y = N$ while feasible solutions of (6d) are N-point multisets supported on $\mathrm{conv}(A)$. Note that these objects do not permit an easy discussion of convergence. However, both notions can be translated into an element $\omega \in (\mathrm{conv}(A))^N$ which is independent of Λ and allows a discussion of convergence. Note that ω can canonically be translated back into a multiset.

Theorem 3.9 *Let (ε_k) be a non-negative sequence converging towards 0. Furthermore, for every $k \in \mathbb{N}$ choose an ε_k-net Λ_k of $\mathrm{conv}(A)$. Let (x_k, y_k) be a sequence of optimal solutions of (6e) w.r.t. Λ_k, ε_k. Identifying each y_k with $\omega_k \in (\mathrm{conv}(A))^N$, any accumulation point $(\bar{x}, \bar{\omega})$ of this sequence corresponds to an optimal solution of (6d) by identification of $\bar{\omega}$ with a multiset.*

Proof The proof is similiar to the proof of Theorem 3.8. Note that, since order of elements is not important for the discussed problems, we can regard to elements of $(\mathrm{conv}(A))^N$ either as tuples or as multisets depending on the context. Suppose (x_k, ω_k) with has an accumulation point $(\bar{x}, \bar{\omega})$. By passing to a subsequence we can assume that $(x_k, \omega_k) \to (\bar{x}, \bar{\omega})$. Consider the continuous function $g_p = \max_{c \in \mathrm{conv}(A)} g_{c,p}$ with $g_p(0) = 0$. Then, we have for all k and $p \in \Gamma$:

$$\begin{aligned} x_k &\leq \sum_{c \in \Lambda_k} (y_k)_c \cdot (f(\|c - p\|) + g_{c,p}(\varepsilon_k)) \\ &\leq \sum_{i=1}^{N} f(\|(\omega_k)_i - p\|) + g_p(\varepsilon_k) \end{aligned}$$

By taking limits we obtain

$$\bar{x} \leq \sum_{i=1}^{N} f(\|\bar{\omega}_i - p\|)$$

for all $p \in \Gamma$. Thus $\bar{x}, \bar{\omega}$ is feasible for (6d).

Now suppose x, ω is an arbitrary solution of (6d). Then by Lemma 3.4 there exists ω'_k such that x, ω'_k is a feasible solution for (6d). Since (x_k, ω_k) is an optimal solution, we have $x_k \geq x$ and by taking limits $\bar{x} \geq x$. Therefore $\bar{x}$ is also optimal for (6e). □

Note, that the proofs of Theorems 3.8, 3.9 still work if we restrict y to be binary as was discussed at the end of Sect. 3.1.

Combining Theorems 3.8 and 3.9, we conclude that by choosing a suitable sequence $(\varepsilon_\Gamma)_k, (\varepsilon_\Lambda)_k$, we can in theory bound the value of $\mathcal{P}(A)$ as tightly as we need. However, solving the respective mixed-integer linear problems in practice will pose a computational challenge.

4 Computational Results

This section presents numerical experiments illustrating the capabilities and limits of the MIP approach presented in this paper. All computations have been performed using Gurobi on a HP DL380 Gen9 server with two Intel(R) Xeon(R) CPU E5-2660v@2.00GHz (each with 14 cores) and 256 GB RAM. We first focus on a simple illustrative example, where A is an equilateral triangle and the size of the configuration is $N = 3$. In addition, we chose $f(x) = e^{-5\|x\|^2}$ for our potential function and $\varepsilon_\Gamma = 0.014, \varepsilon_\Lambda = \varepsilon_\Gamma/3$ as the respective discretization widths of $\Gamma \subseteq A$ and $\Lambda \subseteq \text{conv}(A)$. Lastly, we restrict both, (6a) and (6e) to binary variables $y \in \{0, 1\}^\Lambda$ instead of integral $y \in \{0, \ldots, N\}^\Lambda$ as discussed below Theorem 3.5. Since we expect the resulting configuration to consist of three separate points, this should not significantly impact the quality of the bounds.

We illustrate the configuration given by (6a) in Fig. 2. It was obtained after approximately 10 hours.

We continue by assessing the numerical evidence on the convergence for the above example. To this end, we illustrate the quality of the binary versions of both, (6a) and (6e) for decreasing values of ε_Λ and ε_Γ. Here, the binary variant of (6e) was derived from Proposition 3.7. To be precise, for every $\varepsilon \in \{0.04, 0.038, \ldots, 0.014\}$ we computed the lower bound using $\varepsilon_\Lambda = \varepsilon/3$, $\varepsilon_\Gamma = \varepsilon$ and the upper bound using $\varepsilon_\Gamma = \varepsilon_\Lambda = \varepsilon$. We chose these scalings for a better comparability, since the $(\varepsilon_\Lambda, 3)$-net in the upper-bound case contains more sample points and therefore yields more variables than an $(\varepsilon_\Lambda, 1)$-net. Furthermore, we used scaled versions of the A_2 lattice complemented with additional sample points on the boundary to generate the samples Λ and Γ. This construction ensures that both, Λ and Γ are indeed ε_Λ and ε_Γ-nets respectively. The obtained bounds are visualized in Fig. 3.

It is apparent, that lower values of ε do not always yield better bounds although there is a clearly visible trend to close the gap between the bounds as can be expected

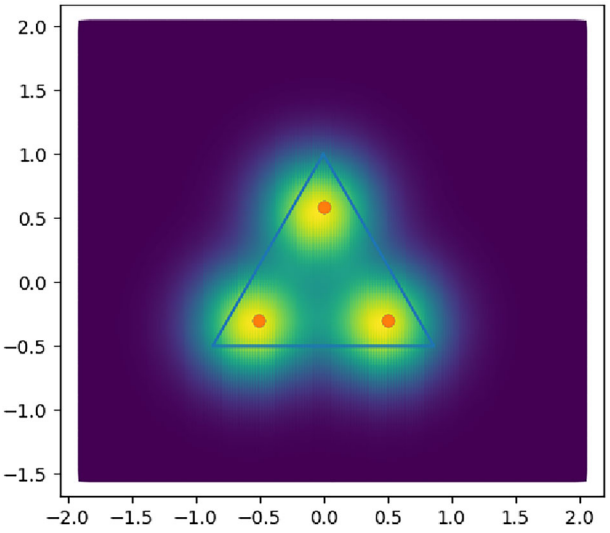

Fig. 2 Optimal configuration for (6a) with $\varepsilon = 0.014$ and a heatmap of the respective f-potential (from dark blue over green to yellow). The points of the configuration are represented by orange circles

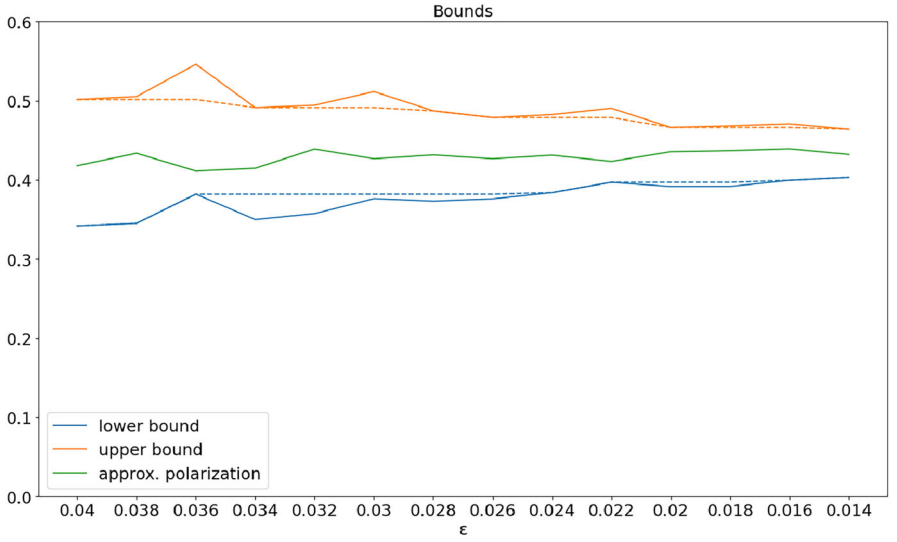

Fig. 3 Upper and lower bounds computed with decreasing values of ε and the respective running optimum (dashed lines) as well as an approximate polarization of the lower bound configuration

from our convergence results established in Theorems 3.8 and 3.9. A drawback of this approach is the computational runtime of the respective MIPs, which vastly increases with the sample size of Γ and Λ from a few seconds if $\varepsilon = 0.04$ to 10 hours for $\varepsilon = 0.014$.

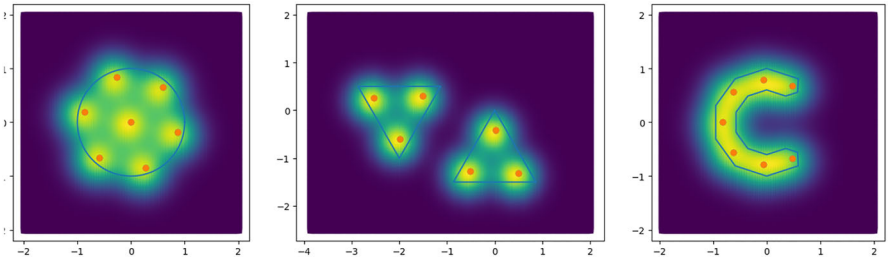

Fig. 4 Optimal configurations of (6a) for different A in orange with a heatmap of the respective potential (from dark blue over green to yellow). The border of the respective shape A is highlighted in blue (from left to right: ball, triangles, non-convex shape)

Table 1 Computational results of polarizations for exemplary shapes

A	Ball	Two twisted triangles	Non-convex shape
N	7	6	7
lower bound	0.391063	0.381283	0.918088
$\varepsilon = \varepsilon_\Lambda = \varepsilon_\Gamma$	0.0625	0.025	0.025
computation time in seconds	6145	1238	2020
upper bound	0.942982	0.506328	1.12719
$\varepsilon = \varepsilon_\Lambda = \varepsilon_\Gamma$	0.0875	0.0375	0.03125
computation time in seconds	6741	858	879
gap	ca. 59%	ca. 25%	ca 19%

As an additional academic example, we use the same approach for different suitable choices of $\varepsilon = \varepsilon_\Lambda = \varepsilon_\Gamma$ and different convex, non-convex or even non-connected A to showcase the wide applicability of our approach. We illustrate the polarizations derived by the binary approximation of our lower bound MIP (6a) in Fig. 4.

Moreover, we briefly summarize the computational results on these additional shapes A in Table 1 below. The respective sample widths were chosen such that the corresponding MIPs could be solved in reasonable time.

We note, that the shape of A significantly impacts the runtime of our MIP approach. It seems that the large symmetry group of the ball may contribute to a larger runtime as good solutions may be found everywhere in the branch-and-bound tree used by solvers such as Gurobi. If true, symmetry reduction techniques may lead to substantial improvements.

5 Outlook

We have seen in Sect. 2 that the location of the darkest points and the location of the points of a locally optimal configuration are intertwined. We suspect that these results can be extended, in particular by utilizing symmetries of A or requiring A to be convex or even a polytope. Furthermore, it would be interesting to extend these results to other choices of D.

However, it is clear that there will be limitations to this approach. Consider for example $A = D = S^{n-1}$ the unit sphere, where the convexity condition of Theorem 2.1 is not applicable. However, using different techniques, information on the set of darkest points for certain configurations on S^{n-1} has been obtained, e.g. for regular simplices [8, Thm. 2.4] and for m-stiff and strongly m-sharp configurations in [20, Thms. 4.3 and 4.5] extending previous results in [2, 21].

In this paper, we have not dealt with explicit computations of locally or globally optimal point configurations, even on simple sets such as n-gons or the unit ball. However, numerical experiments suggest that such configurations show some structure and we hope that extensions of the results in Sect. 2 can be utilized to obtain proof of optimality for some configurations. Here, we would like to highlight one result in this direction we are aware of, namely that for certain Riesz potentials of modest decay and with A being chosen as the closed d-dimensional unit ball, the optimal point configuration consists of N copies of the origin (see [1, Thm. 14.2.6]). We were able to observe similar effects in numerical experiments on regular polytopes.

The MIP hierarchies presented in Sect. 3 give provable upper and lower bounds converging to the optimal solution. However, unsurprisingly computing these bounds for sufficiently fine samples is very time consuming since MIP is NP-complete. A natural question is, whether well known techniques from mathematical programming - such as convex relaxations, inner approximations, column generation or local refinement, that speed up the computations can be utilized to achieve results for finer samples. However, most of these techniques only provide approximations of the discussed MIP hierarchies, which might limit the gain achieved through the finer samples.

Moreover, it might be helpful to carefully fit the choice of the samples to the specific instance of the problem. For example, if one has a conjecture for an optimal configuration and/or the correct location of the darkest points, this information can be fitted into the samples while retaining the ε-net property of the samples. Furthermore, these ideas might provide a way to use our bounds for analytic proofs of optimality in highly structured situations.

Acknowledgements The authors like to thank Frank Vallentin for useful suggestions. M.C.Z. is partially supported "Spectral bounds in extremal discrete geometry" (project number 414898050) funded by the DFG.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

1. Borodachov, S.V., Hardin, D.P., Saff, E.B.: Discrete Energy on Rectifiable Sets. Springer Monographs in Mathematics. Springer, New York (2019). https://books.google.se/books?id=Eie7DwAAQBAJ
2. Stolarsky, K.: The sum of the distances to certain pointsets on the unit circle. Pac. J. Math. **59**(1), 241–251 (1975)
3. Ambrus, G.: Analytic and probabilistic problems in discrete geometry. PhD thesis, University College London (2009)
4. Ambrus, G., Ball, K.M., Erdélyi, T.: Chebyshev constants for the unit circle. Bull. Lond. Math. Soc. **45**(2), 236–248 (2013). https://doi.org/10.1112/blms/bds082
5. Nikolov, N., Rafailov, R.: On the sum of powered distances to certain sets of points on the circle. Pac. J. Math. **253**(1), 157–168 (2011)
6. Hardin, D.P., Kendall, A.P., Saff, E.B.: Polarization optimality of equally spaced points on the circle for discrete potentials. Discrete Comput. Geom. **50**, 236–243 (2013)
7. Erdélyi, T., Saff, E.B.: Riesz polarization inequalities in higher dimensions. J. Approx. Theory **171**, 128–147 (2013)
8. Borodachov, S.: Polarization problem on a higher-dimensional sphere for a simplex. Discrete Comput. Geom. **67**, 525–542 (2022). https://doi.org/10.1007/s00454-021-00308-1
9. Borodachov, S.V., Bosuwan, N.: Asymptotics of discrete Riesz d-polarization on subsets of d-dimensional manifolds. Potential Anal. **41**(1), 35–49 (2014)
10. Borodachov, S., Hardin, D., Reznikov, A., Saff, E.: Optimal discrete measures for Riesz potentials. Trans. Am. Math. Soc. **370**(10), 6973–6993 (2018)
11. Hardin, D.P., Petrache, M., Saff, E.B.: Unconstrained polarization (chebyshev) problems: basic properties and Riesz kernel asymptotics. Potential Anal. 1–44 (2020)
12. Anderson, A., Reznikov, A., Vlasiuk, O., White, E.: Polarization and covering on sets of low smoothness. Adv. Math. **410**, 108720 (2022)
13. Conway, J.H., Sloane, N.J.A.: Sphere Packings, Lattices and Groups. Grundlehren der mathematischen Wissenschaften. Springer, New York (1998)
14. Naszódi, M.: In: Ambrus, G., Bárány, I., Böröczky, K.J., Fejes Tóth, G., Pach, J. (eds.) Flavors of Translative Coverings, pp. 335–358. Springer, Berlin (2018). https://doi.org/10.1007/978-3-662-57413-3_14
15. Naszódi, M.: On some covering problems in geometry. Proc. Am. Math. Soc. **144**(8), 3555–3562 (2016). https://doi.org/10.1090/proc/12992
16. Rolfes, J.H., Vallentin, F.: Covering compact metric spaces greedily. Acta Mathematica Hungarica **155**, 130–140 (2017)
17. Cohn, H., Kumar, A.: Universally optimal distribution of points on spheres. J. Am. Math. Soc. **20**(1), 99–148 (2007)
18. Simon, B.: Convexity: An Analytic Viewpoint, vol. 187. Cambridge University Press, Cambridge (2011)
19. Shapiro, A.: Semi-infinite programming, duality, discretization and optimality conditions. Optimization **58**(2), 133–161 (2009)
20. Borodachov, S.: Absolute Minima of Potentials of a Certain Class of Spherical Designs (2022)
21. Stolarsky, K.: The sum of the distances to n points on a sphere. Pac. J. Math. **57**(2), 563–573 (1975)