Optimal Coreset for Gaussian Kernel Density Estimation

Wai Ming Tai ✉
University of Chicago, IL, USA

Abstract

Given a point set $P \subset \mathbb{R}^d$, the kernel density estimate of P is defined as

$$\overline{\mathcal{G}}_P(x) = \frac{1}{|P|}\sum_{p\in P} e^{-\|x-p\|^2}$$

for any $x \in \mathbb{R}^d$. We study how to construct a small subset Q of P such that the kernel density estimate of P is approximated by the kernel density estimate of Q. This subset Q is called a coreset. The main technique in this work is constructing a ± 1 coloring on the point set P by discrepancy theory and we leverage Banaszczyk's Theorem. When $d > 1$ is a constant, our construction gives a coreset of size $O\left(\frac{1}{\varepsilon}\right)$ as opposed to the best-known result of $O\left(\frac{1}{\varepsilon}\sqrt{\log\frac{1}{\varepsilon}}\right)$. It is the first result to give a breakthrough on the barrier of $\sqrt{\log}$ factor even when $d = 2$.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms

Keywords and phrases Discrepancy Theory, Kernel Density Estimation, Coreset

Digital Object Identifier 10.4230/LIPIcs.SoCG.2022.63

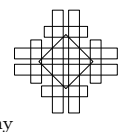
Related Version *Full Version*: https://arxiv.org/abs/2007.08031

1 Introduction

Kernel density estimation is a non-parametric way to estimate a probability distribution. Given a point set $P \subset \mathbb{R}^d$, the kernel density estimate (KDE) of P smooths out P to a continuous function [35, 36]. More precisely, given a point set $P \subset \mathbb{R}^d$ and a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, KDE is defined as the function $\overline{\mathcal{G}}_P(x) = \frac{1}{|P|}\sum_{p\in P} K(x,p)$ for any $x \in \mathbb{R}^d$. Here, the point x is called a *query*. One common example of kernel K is the Gaussian kernel, which is $K(x,y) = e^{-\|x-y\|^2}$ for any $x, y \in \mathbb{R}^d$, and it is the main focus of this paper. A wide range of application includes outlier detection [41], clustering [33], topological data analysis [32, 10], spatial anomaly detection [1, 18], statistical hypothesis test [17] and other [19, 23].

Generally speaking, the techniques using kernels are called *kernel methods*, in which KDE is the central role in these techniques. Kernel methods are prevalent in machine learning and statistics and often involve optimization problems. Optimization problems are generally hard in the sense that solving them usually has a super-linear or even an exponential dependence on the input's size in its running time. Therefore, reducing the size of the input will be desirable. A straightforward way to achieve this is extracting a small subset Q of the input P. This paper will study the construction of the subset Q such that $\overline{\mathcal{G}}_Q$ approximates $\overline{\mathcal{G}}_P$.

Classically, statisticians concern about different types of error such as L_1-error [14] or L_2-error [35, 36]. However, there are multiple modern applications that require L_∞-error such as preserving classification margin [34], density estimation [40], topology [32] and hypothesis test on distributions [17]. For example, in topological data analysis, we might want to study the persistent homology of a super-level set of a kernel density estimate. In this case, L_∞-error plays an important role here since a small perturbation could cause a significant change in its persistence diagram. Formally, we would like to solve the following problem.



Given a point set $P \subset \mathbb{R}^d$ and $\varepsilon > 0$, we construct a subset Q of P such that

$$\sup_{x \in \mathbb{R}^d} |\overline{\mathcal{G}}_P(x) - \overline{\mathcal{G}}_Q(x)| = \sup_{x \in \mathbb{R}^d} |\frac{1}{|P|} \sum_{p \in P} e^{-\|x-p\|^2} - \frac{1}{|Q|} \sum_{q \in Q} e^{-\|x-q\|^2}| \leq \varepsilon.$$

Then, how small can the size of Q, $|Q|$, be?

We call this subset Q an *ε-coreset*.

1.1 Known results

We now discuss some previous results for the size of an ε-coreset.

Josh et al. [20] showed that random sampling can achieve the size of $O(\frac{d}{\varepsilon^2})$. They investigated the VC-dimension of the super-level sets of a kernel and analyzed that the sample size can be bounded by it. In particular, the super-level sets of the Gaussian kernel are balls in $\mathbb{R}^d$. It reduces the problem to bounding the sample size of the range space of balls.

Lopaz-Paz et al. [24] later proved that the size of the coreset can be reduced to $O(\frac{1}{\varepsilon^2})$ by random sampling. They studied the reproducing kernel Hilbert space (RKHS) associated with a positive-definite kernel [3, 39, 38]. Note that the Gaussian kernel is a positive-definite kernel. In RKHS, one can bound the L_∞-error between two KDEs of point sets P and Q by the kernel distance of P and Q. They showed that the sample size of $O(\frac{1}{\varepsilon^2})$ is sufficient to bound the kernel distance.

Other than random sampling, Lacoste-Julien et al. [22] showed a greedy approach can also achieve the size of $O(\frac{1}{\varepsilon^2})$. They applied Frank-Wolfe algorithm [13, 15] in RKHS to bound the error of the kernel distance.

Note that all of the above results have a factor of $\frac{1}{\varepsilon^2}$. Josh et al. [20] first showed that a sub-$O(\frac{1}{\varepsilon^2})$ result can be obtained by reducing the problem to constructing an ε-approximation for the range space of balls [26]. They assumed that d is constant. For the case of $d = 1$, their result gives the size of $O(\frac{1}{\varepsilon})$.

Later, Phillips [29] improved the result to $O((\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon})^{\frac{d}{d+2}})$ for constant d via geometric matching. It is based on the discrepancy approach. Namely, they construct a ± 1 coloring on the point set, recursively drop the points colored -1 and construct another ± 1 coloring on the points colored $+1$. We will discuss it in more detail below. Notably, for the case of $d = 2$, their bound is $O(\frac{1}{\varepsilon}\sqrt{\log \frac{1}{\varepsilon}})$ which is nearly-optimal (as a preview, the optimal bound is $\Omega(\frac{1}{\varepsilon})$) and is the first nearly-linear result for the case of $d > 1$.

Recently, Phillips and Tai [30] further improved the size of a coreset to $O(\frac{1}{\varepsilon} \log^d \frac{1}{\varepsilon})$ for constant d. It is also based on the discrepancy approach. They exploited the fact that the Gaussian kernel is multiplicatively separable. It implies that the Gaussian kernel can be rewritten as the weighted average of a family of axis-parallel boxes in $\mathbb{R}^d$. Finally, they reduced the problem to Tusnády's problem [6, 2].

Also, Phillips and Tai [31] proved a nearly-optimal result of $O(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log \frac{1}{\varepsilon}})$ shortly after that. It is also based on the discrepancy approach. They observed that the underlying structure of the positive-definite kernel allows us to bound the norm of the vectors and apply the lemma in [27], which used Banaszczyk's Theorem [4, 5]. Recall that the Gaussian kernel is a positive-definite kernel.

Except for the upper bound, there are some results on the lower bound for the size of an ε-coreset. Phillips [29] provided the first lower bound for the size of a coreset. They proved a lower bound of $\Omega(\frac{1}{\varepsilon})$ by giving an example that all points are spread out. When assuming

$d > \frac{1}{\varepsilon^2}$, Phillips and Tai [30] gave another example that forms a simplex and showed a lower bound of $\Omega(\frac{1}{\varepsilon^2})$. Later, Phillips and Tai [31] combined the techniques of the above two results and showed the lower bound of $\Omega(\frac{\sqrt{d}}{\varepsilon})$.

There are other conditional bounds for this problem. We suggest the readers refer to [31] for a more extensive review. Recently, Karnin and Liberty [21] defined the notion of Class Discrepancy which governs the coreset-complexity of different families of functions. Specifically, for analytic functions of squared distances (such as the Gaussian kernel), their analysis gives a discrepancy bound $D_m = O(\frac{\sqrt{d}}{m})$ which gives a coreset of size $O(\frac{\sqrt{d}}{\varepsilon})$. Their approach also used the discrepancy technique or, more precisely, Banaszczyk's Theorem [4, 5]. Unfortunately, their analysis requires *both* the point set P and the query x lie in a ball of a fixed radius R. Therefore, their result has a dependence on R. Strictly speaking, their result is not comparable to ours. It is not clear how to remove this assumption of R based on their result. Also, the lower bound constructions in [29, 31] rely on the fact that P is in an unbounded region and hence it is not clear how their result is comparable to the existing lower results.

1.2 Related works

In computational geometry, an ε-approximation is the approximation of a general set by a smaller subset. Given a set S and a collection $\mathcal{C}$ of subsets of S, a subset $A \subset S$ is called an ε-approximation if $|\frac{|T|}{|S|} - \frac{|T \cap A|}{|A|}| \leq \varepsilon$ for all $T \in \mathcal{C}$. The pair $(S, \mathcal{C})$ is called a set system (also known as a range space or a hypergraph). One can rewrite the above guarantee as $|\frac{1}{|S|}\sum_{x \in S} \mathbb{1}_T(x) - \frac{1}{|A|}\sum_{x \in A} \mathbb{1}_T(x)| \leq \varepsilon$ where $\mathbb{1}_T$ is the indicator function of set T. If we replace this indicator function by a kernel such as the Gaussian kernel, it is the same as our ε-coreset. There is a rich history on the construction of an ε-approximation [11, 26]. One notable method is discrepancy theory, which is also our main technique. There is a wide range of techniques employed in this field. In the early 1980s, Beck devised the technique of partial coloring [7], and later a refinement of this technique called entropy method was introduced by Spencer [37]. The entropy method is first used to solve the famous "six standard deviations" theorem: given a set system of n points and n subsets, there is a coloring of discrepancy at most $6\sqrt{n}$. In contrast, random coloring gives the discrepancy of $O(\sqrt{n \log n})$. A more geometric example in discrepancy theory is Tusnády's problem. It states that, given point set P of size n in $\mathbb{R}^d$, construct a ± 1 coloring σ on P such that the discrepancy $\min_\sigma \max_R |\sum_{P \cap R} \sigma(p)|$ is minimized where $\max_R$ is over all axis-parallel boxes R. One previous approach of our ε-coreset problem reduces the problem to Tusnády's problem.

On the topic of approximating KDE, Fast Gauss Transform [16] is a method to preprocess the input point set such that the computation of KDE at a query is faster than the brute-force approach. The idea in this method is expanding the Gaussian kernel by Hermite polynomials and truncating the expansion. Assuming that the data set lies inside a bounded region, the query time in this method is poly-logarithmic of n for constant dimension d. Also, Charikar et al. [9] studied the problem of designing a data structure that preprocesses the input to answer a KDE query in a faster time. They used locality-sensitive hashing to perform their data structure. However, the guarantee they obtained is a relative error, while ours is an additive error. More precisely, given a point set $P \subset \mathbb{R}^d$, Charikar et al. designed a data structure such that, for any query $x' \in \mathbb{R}^d$, the algorithm answers the value $\overline{\mathcal{G}}_P(x') = \sum_{p \in P} e^{-\|x'-p\|^2}$ within $(1+\varepsilon)$-relative error. Also, the query time of their data structure is sublinear of n.

1.3 Our result

We construct an ε-coreset and bound the size of the ε-coreset via discrepancy theory. Roughly speaking, we construct a ± 1 coloring on our point set such that its discrepancy is small. Then, we drop the points colored -1 and recursively construct a ± 1 coloring on the points colored $+1$. Eventually, the remaining point set is the desired coreset. A famous theorem in discrepancy theory is Banaszczyk's Theorem [4, 5]. We will use Banaszczyk's Theorem to construct a coloring and prove the discrepancy is small by induction. To the best of our knowledge, this induction analysis combining with Banaszczyk's Theorem has not been seen in discrepancy theory before. In the constant dimensional space, we carefully study the structure of the Gaussian kernel and it allows us to construct an ε-coreset of size $O(1/\varepsilon)$. Our result is the first result to break the barrier of $\sqrt{\log}$ factor even when $d = 2$.

▶ **Theorem 1.** *Suppose $P \subset \mathbb{R}^d$ a point set of size n. Let $\overline{\mathcal{G}}_P$ be the Gaussian kernel density estimate of P, i.e. $\overline{\mathcal{G}}_P(x) = \frac{1}{|P|}\sum_{p\in P} e^{-\|x-p\|^2}$ for any $x \in \mathbb{R}^d$. For a fixed constant d, there is an algorithm that constructs a subset $Q \subset P$ of size $O(\frac{1}{\varepsilon})$ such that $\sup_{x\in\mathbb{R}^d} |\overline{\mathcal{G}}_P(x) - \overline{\mathcal{G}}_Q(x)| < \varepsilon$ and has a polynomial running time in n.*

Even if $d = 1$, the best known result is $O(1/\varepsilon)$ by [20, 30], which is optimal. Their approach is to reduce the problem to Tusnády's problem. A trivial solution of Tusnády's problem (and hence our problem) is: sort P and assign ± 1 on each point alternately. However, it is not clear that how this simple solution can be generalized to the higher dimensional case. Our algorithm gives a non-trivial perspective even though the optimal result was achieved previously.

2 Preliminaries

Our approach for constructing a coreset relies on discrepancy theory, which is a similar technique in range counting coreset [12, 28, 8]. We first introduce an equivalent problem (up to a constant factor) as follows.

Given a point set $P \subset \mathbb{R}^d$, what is the smallest quantity of $\sup_{x\in\mathbb{R}^d} |\sum_{p\in P} \sigma(p)e^{-\|x-p\|^2}|$ over all σ in the set of colorings from P to $\{-1, +1\}$?

Now, one can intuitively view the equivalence in the following way. If we rewrite the objective as:

$$\frac{1}{|P|}|\sum_{p\in P}\sigma(p)e^{-\|x-p\|^2}| = |\frac{1}{|P|}\sum_{p\in P} e^{-\|x-p\|^2} - \frac{1}{|P|/2}\sum_{p\in P_+} e^{-\|x-p\|^2}|$$

where $P_+ \subset P$ is the set of points that is assigned $+1$, then we can apply the halving technique [12, 28] which recursively invokes the coloring algorithm and retains the points assigned $+1$ until the subset of the desired size remains. Note that there is no guarantee that half of the points are assigned $+1$, while the other half is assigned -1. However, we can handle this issue by some standard techniques [26] or see our proof for details.

Also, we define the following notations. Given a point set $P \subset \mathbb{R}^d$, a coloring $\sigma : P \to \{-1, +1\}$ and a point $x \in S$, we define the signed discrepancy $\mathcal{D}_{P,\sigma}(x)$ as

$$\mathcal{D}_{P,\sigma}(x) = \sum_{p\in P}\sigma(p)e^{-\|x-p\|^2}$$

It is worth noting that we expect $|\mathcal{D}_{P,\sigma}(x)| < O(1)$ in order to construct an ε-coreset of size $O(\frac{1}{\varepsilon})$ via this halving technique.

An important result in discrepancy theory is Banaszczyk's Theorem [4].

▶ **Theorem 2** (Banaszczyk's Theorem [4]). *Suppose we are given a convex body $K \subset \mathbb{R}^m$ of the Gaussian measure at least $\frac{1}{2}$ and n vectors $v^{(1)}, v^{(2)}, \ldots, v^{(n)} \in \mathbb{R}^m$ of norm at most 1, there is a coloring $\sigma : [n] \to \{-1, +1\}$ such that the vector $\sum_{i=1}^{n} \sigma(i)v^{(i)} \in cK = \{c \cdot y \mid y \in K\}$. Here, c is an absolute constant and the Gaussian measure of a convex body K is defined as $\int_{x \in K} \frac{1}{(2\pi)^{d/2}} e^{-\|x\|^2/2} dx$.*

The original proof of this theorem is non-constructive. Bansal et al. [5] proved that there is an efficient algorithm to construct the coloring in Banaszczyk's Theorem. Moreover, assuming $m < n$, the running time is $O(n^{\omega+1})$ where ω is the exponent of matrix multiplication.

▶ **Theorem 3** (Constructive version of Banaszczyk's Theorem [5]). *Suppose we are given n vectors $v^{(1)}, \ldots, v^{(n)} \in \mathbb{R}^m$ of norm at most 1, there is an efficient randomized algorithm that constructs a coloring σ on P with the following guarantee: there are two absolute constants C', C'' such that, for any unit vector $\theta \in \mathbb{R}^m$ and $\alpha > 0$, we have*

$$\boldsymbol{Pr}[|\langle \theta, X \rangle| > \alpha] < C' e^{-C''\alpha^2}$$

where X is the random variable of $\sum_{i=1}^{n} \sigma(i)v^{(i)}$. The probability in the above statement is distributed over all ± 1 colorings.

Finally, we introduce a useful theorem which is Markov Brother's Inequality.

▶ **Theorem 4** (Markov Brother's Inequality [25]). *Let $\mathcal{P}(x)$ be a polynomial of degree ρ. Then,*

$$\sup_{x \in [0,1]} |\mathcal{P}'(x)| \leq 2\rho^2 \sup_{x \in [0,1]} |\mathcal{P}(x)|$$

Here, $\mathcal{P}'$ is the derivative of $\mathcal{P}$.

3 Proof overview

As we mentioned before, our equivalent problem statement suggests that we need to construct a ± 1 coloring on the input point set such that the absolute value of the signed discrepancy at all points is small. In this section, we will give an overview on how we construct the coloring and how it gives us the desired guarantees.

For exposition purposes, we illustrate the idea for the case of $d = 1$ even though previous results [20, 30] showed this case is trivial. Recall that our problem definition is: given a point set $P \subset \mathbb{R}$ of size n, construct a ± 1 coloring σ on P such that the absolute value of the signed discrepancy

$$|\mathcal{D}_{P,\sigma}(x)| = |\sum_{p \in P} \sigma(p) e^{-(x-p)^2}|$$

is bounded from above by a constant for *all* $x \in \mathbb{R}$.

Some general observations. We first make some observations. Note that $\mathcal{D}_{P,\sigma}$ is a smooth function of x that the slope at any x is bounded. It means that if $|\mathcal{D}_{P,\sigma}(x_0)|$ is small for some point x_0 then $|\mathcal{D}_{P,\sigma}(y)|$ is also small for any point y at a neighborhood of x_0. Another observation is that $\mathcal{D}_{P,\sigma}$ is basically a linear combination of Gaussians and hence $|\mathcal{D}_{P,\sigma}(x)|$ is small for any x that is far away from all points in P.

Combining these two observations, if we lay down a grid on $\mathbb{R}$ and consider the grid points that is not too far away from P, then we only need to construct a coloring σ such that $|\mathcal{D}_{P,\sigma}(x)|$ is small for all x in a *finite* set and it implies that $|\mathcal{D}_{P,\sigma}(x)|$ is small for all $x \in \mathbb{R}$. It is crucial because we preview that our algorithm for constructing the coloring σ is a randomized algorithm and the size of the finite set controls the number of events when we apply the union bound. Note that these observations hold for *any* coloring.

Techniques from the previous result. Now, we make the above observations more quantitative. Since the slope of each Gaussian at any point is bounded by $O(1)$ and there are n Gaussians in $\mathcal{D}_{P,\sigma}$, by triangle inequality, the absolute value of the slope of $\mathcal{D}_{P,\sigma}$ at any point is bounded by $O(n)$. Hence, if $|\mathcal{D}_{P,\sigma}(x_0)|$ is bounded by α for any point x_0 for any α then $|\mathcal{D}_{P,\sigma}(y)|$ is bounded by $\alpha + O(1)$ for all y that $|x_0 - y| < O(1/n)$. Also, Gaussians decay exponentially and hence $|\mathcal{D}_{P,\sigma}(x)| < O(1)$ for any x that $|x - p| > \Omega(\sqrt{\log n})$ for all $p \in P$.

If a coloring σ satisfies that

$$|\mathcal{D}_{P,\sigma}(x)| < \alpha \text{ for } \textit{any } x \in \mathbb{R} \text{ with probability at least } 1 - O(e^{-\Omega(\alpha^2)}) \text{ for any } \alpha \tag{1}$$

then it implies, by union bound, this coloring σ satisfies that

$$|\mathcal{D}_{P,\sigma}(x)| < \alpha + O(1) \text{ for } \textit{all } x \in \mathbb{R} \text{ with probability at least } 1 - N \cdot O(e^{-\Omega(\alpha^2)})$$

where N is the number of grid points that are in the grid of cell width $\Omega(1/n)$ and lie around some point in P within a radius of $O(\sqrt{\log n})$. The number N is bounded by $O(n^2\sqrt{\log n})$ because for each point $p \in P$ there are $O(\sqrt{\log n}/(1/n)) = O(n\sqrt{\log n})$ grid points around p within a radius of $O(\sqrt{\log n})$ and there are n points in P. By setting $\alpha = O(\sqrt{\log n})$, we have

$$|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log n}) \text{ for } \textit{all } x \in \mathbb{R} \text{ with probability at least } 1 - 1/10$$

if we manage to construct a coloring σ satisfying (1). Phillips and Tai [31] managed to construct such coloring σ by Banaszczyk's Theorem and proved their result. Namely, a coloring satisfying (1) is construct-able.

Attempts to improve the result. We have seen how to show $|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log n})$. There is still a gap from showing $|\mathcal{D}_{P,\sigma}(x)| < O(1)$. We observe that the above argument aims at minimizing α such that the total failure probability $Ne^{-\Omega(\alpha^2)}$ is bounded by a constant. If we manage to make the factor N smaller, it helps setting α smaller and hence we can improve the result.

Recall that $N = O(n^2\sqrt{\log n}) = O(n \cdot n\sqrt{\log n})$ and the first factor n comes from the fact that P has n points and these n points could be widely spread out. Namely, we need at most n neighborhoods to cover all relevant grid points. What if all points in P lie inside a bounded region say $[-1, 1]$? In this case, we just need to consider *one* neighborhood to cover all relevant grid points. Nonetheless, we do not assume that they are in a bounded region and we take care of it in the following way. We partition $\mathbb{R}$ into infinitely many bounded regions (say $\dots, [-3, -1], [-1, 1], [1, 3], \dots$) and assign each point in P to its corresponding region. Then, we construct a coloring on the points in each bounded region and each coloring is constructed independently. By triangle inequality, we have

$$|\mathcal{D}_{P,\sigma}(x)| \le \sum |\mathcal{D}_{P_i,\sigma_i}(x)| \tag{2}$$

where each $P_i \subset P$ is the set of points in the same bounded region and σ_i is the coloring σ restricted on P_i.

If we manage to construct the colorings σ_i satisfying (1) then we will end up getting $|\mathcal{D}_{P,\sigma}(x)| < n_0 \cdot O(\alpha)$ where n_0 is the number of bounded regions that contain at least one point in P. However, n_0 can be as large as $O(n)$. To address this issue, we take the advantage of the assumption that all points in P_i are in a bounded region (say $[-1, 1]$). Since all points in P_i are in $[-1, 1]$ now and Gaussians decay exponentially, intuitively we should be able to construct a coloring σ_i that

$$|\mathcal{D}_{P_i,\sigma_i}(x)| < \alpha e^{-\frac{2}{3}x^2} \text{ for } \textit{any } x \in \mathbb{R} \text{ with probability at least } 1 - O(e^{-\Omega(\alpha^2)}) \text{ for any } \alpha$$

if a coloring satisfying (1) is construct-able. It is because we can rewrite $|\mathcal{D}_{P_i,\sigma_i}(x)|$ as

$$|\mathcal{D}_{P_i,\sigma_i}(x)| = |\sum_{p\in P_i} \sigma_i(p)e^{-(x-p)^2}| = e^{-\frac{2}{3}x^2} \cdot |\sum_{p\in P_i} \sigma_i(p)e^{2p^2}e^{-(\frac{1}{\sqrt{3}}x-\sqrt{3}p)^2}| \tag{3}$$

and the expression in the RHS has a form similar to $\mathcal{D}_{P_i,\sigma_i}(x)$. The constant $\frac{2}{3}$ in the factor $e^{-\frac{2}{3}x^2}$ can be any constant between 0 and 1. The extra factor $e^{-\frac{2}{3}x^2}$ is crucial: when we plug the bound $\alpha e^{-\frac{2}{3}x^2}$ into (2), $|\mathcal{D}_{P,\sigma}(x)|$ is bounded by $O(1)\cdot O(\alpha)$ instead of $n_0 \cdot O(\alpha)$.

One minor issue here is that the failure probability is accumulated when we ensure all σ_i have the desired discrepancy. We fix this issue by turning the construction of each σ_i into a Las Vegas Algorithm. Namely, we check if each σ_i satisfies the desired discrepancy and repeat the construction if not.

Now, *if* we manage to construct a coloring σ such that: given $P \subset [-1,1]$,

$$|\mathcal{D}_{P,\sigma}(x)| < \alpha e^{-\frac{2}{3}x^2} \text{ for } \textit{any } x \in \mathbb{R} \text{ with probability at least } 1 - O(e^{-\Omega(\alpha^2)}) \text{ for any } \alpha \tag{4}$$

then we only need to consider *one* neighborhood to cover all relevant grid points when applying the union bound. We also preview here that (4) is the only property a coloring needs to show our result. From now on, we assume $P \subset [-1,1]$. Even though (4) (the properties of the coloring σ we are looking for) is slightly different than (1) (what we stated in the beginning) because of the extra factor $e^{-\frac{2}{3}x^2}$, we can still perform a similar argument to prove that

$$|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log n})e^{-\frac{2}{3}x^2} \text{ for } \textit{all } x \in \mathbb{R} \text{ with probability at least } 1 - 1/10 \tag{5}$$

by arguing the slope of $\mathcal{D}_{P,\sigma}(x)$ is bounded by $O(n)e^{-\frac{2}{3}x^2}$ for any $x \in \mathbb{R}$.

Reusing the guarantees for $\mathcal{D}_{P,\sigma}$. Now, we look at the second factor $n\sqrt{\log n}$ in N. It turns out that we are not going to make this factor smaller. Instead, we will look at what guarantees this factor can give us and reuse these guarantees.

We further split $n\sqrt{\log n}$ into two parts: n and $\sqrt{\log n}$. Recall that the first part n comes from the configuration that the cell width of the grid is $\Omega(1/n)$ and the second part $\sqrt{\log n}$ comes from the configuration that we need to consider the neighborhood of radius $O(\sqrt{\log n})$ to cover all relevant grid points. However, we set up these two configurations *without taking σ into consideration.* As we mentioned before, if we have a coloring σ satisfying (4) then we have (5). Can we reuse this guarantee and exploit the coloring σ? To answer this question, we first investigate the term $|\mathcal{D}_{P,\sigma}(x) - \mathcal{D}_{P,\sigma}(y)|$ for any $x, y \in \mathbb{R}$ and, by exploiting the structure of the Gaussians, we can prove

$$|\frac{\mathcal{D}_{P,\sigma}(x) - \mathcal{D}_{P,\sigma}(y)}{x-y}| < O(|\xi|)\cdot|\mathcal{D}_{P,\sigma}(\xi)| \tag{6}$$

for any $x \neq y$ where ξ is in between x and y. *The takeaway from this inequality is the slope of $\mathcal{D}_{P,\sigma}$ is bounded by $\mathcal{D}_{P,\sigma}$ itself.* It is how we can reuse our guarantees.

If we plug our guarantee (5) into (6), we can show that the slope of $\mathcal{D}_{P,\sigma}(x)$ for this σ is bounded by $O(\sqrt{\log n \log\log n})e^{-\frac{2}{3}x^2}$ for any x within a radius of $O(\sqrt{\log\log n})$. For x that lies beyond a radius of $\Omega(\sqrt{\log\log n})$, we have

$$|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log n})e^{-\frac{2}{3}x^2} < \frac{O(\sqrt{\log n})}{\Omega(\sqrt{\log n})}e^{-\frac{1}{3}x^2} < O(1)e^{-\frac{1}{3}x^2} < O(\sqrt{\log\log n})e^{-\frac{1}{3}x^2} \tag{7}$$

Note that the constant in the exponent becomes $\frac{1}{3}$ and it can be any constant smaller than $\frac{2}{3}$. If we have a coloring σ satisfying *additionally* that $|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log\log n})e^{-\frac{2}{3}x^2}$ for *all* x in the set of grid points that are in the grid of cell width $\Omega(1/\sqrt{\log n})$ (instead of $\Omega(1/n)$) and bounded within a radius of $O(\sqrt{\log\log n})$ (instead of $O(\sqrt{\log n})$), then we have

$$|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log\log n})e^{-\frac{1}{3}x^2} \text{ for } all\ x \in \mathbb{R}.$$

There is a caveat: to ensure the coloring σ satisfies the additional properties, we have to include more events in the union bound when invoking (4). In other words, the failure probability is now larger than 1/10. Nonetheless, we improved the previous result to $|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log\log n})e^{-\frac{1}{3}x^2}$.

Hints of using induction. From the improvement we just made, it gives us a hint to refine the quality of our result by induction. One may notice the following pattern. Suppose we have a coloring σ satisfying

$$|\mathcal{D}_{P,\sigma}(x)| < \beta e^{-\kappa x^2} \text{ for } all\ x \in \mathbb{R} \tag{8}$$

for some β where κ is any constant between 0 and 1 (like 2/3 before). Let S be the set of grid points that are in the grid of cell width $\Omega(1/\beta)$ and lie within a radius of $O(\sqrt{\log\beta})$. Note that $|S| = O(\beta\sqrt{\log\beta})$. If this coloring σ also satisfies that

$$|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log\beta})e^{-\kappa x^2} \text{ for } all\ x \in S \tag{9}$$

then we can modify the previous argument in the following way. From (8) and (6), we have the absolute value of the slope of $\mathcal{D}_{P,\sigma}$ at any point within a radius of $O(\sqrt{\log\beta})$ is bounded by $O(\beta\sqrt{\log\beta})e^{-\kappa x^2}$. From an argument similar to (7), we also have $|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log\beta})e^{-\kappa' x^2}$ for all x that lies beyond a radius of $\Omega(\sqrt{\log\beta})$ where κ' is any constant between 0 and κ (like 1/3 before). We combine them with (9) and it implies

$$|\mathcal{D}_{P,\sigma}(x)| < O(\sqrt{\log\beta})e^{-\kappa' x^2} \text{ for } all\ x \in \mathbb{R}. \tag{10}$$

If we take (5) as the base step and the implication from (8) to (10) as the inductive step, we should expect

$$|\mathcal{D}_{P,\sigma}(x)| < O(1)e^{-\frac{1}{3}x^2} \text{ for } all\ x \in \mathbb{R}.$$

after $O(\log^* n)$ inductive steps.

As we mentioned before, we also need to keep track of the failure probability and the exponent κ in the factor $e^{-\kappa x^2}$. We first deal with the failure probability. In each inductive step, we need extra guarantees on the set of grid points of a smaller size (i.e. (9) when invoking (4)). Hence, the total failure probability is a sum of $O(\log^* n)$ failure probabilities in each inductive step. We can set these $O(\log^* n)$ failure probabilities to be a geometric sequence such that the total failure probability is a constant. The other issue is the exponent. We can again make this exponent decrease from 2/3 to 1/3 geometrically as it proceeds in the inductive steps. In each inductive step, we need to set α in (4) larger than what we stated earlier accordingly when invoking (4) in the union bound. Nonetheless, we eventually prove that $|\mathcal{D}_{P,\sigma}(x)| < O(1)e^{-\frac{1}{3}x^2}$ for *all* $x \in \mathbb{R}$ with probability 1/2.

Construction of the coloring. It all boils down to the problem of how to construct a coloring σ satisfying (4). Namely, given a point set $P \subset [-1, 1]$,

$$|\mathcal{D}_{P,\sigma}(x)| < \alpha e^{-\frac{2}{3}x^2} \text{ for } \textit{any } x \in \mathbb{R} \text{ with probability at least } 1 - O(e^{-\Omega(\alpha^2)}) \text{ for any } \alpha.$$

We introduced Banaszczyk's Theorem (Theorem 3) before and if we can rewrite (4) as the inner product form shown in Theorem 3 then we can apply the algorithm in Theorem 3. As we mentioned in (3), we first rewrite

$$|\mathcal{D}_{P,\sigma}(x)| = |\sum_{p\in P} \sigma(p) e^{-(x-p)^2}| = e^{-\frac{2}{3}x^2} \cdot |\sum_{p\in P} \sigma(p) e^{2p^2} e^{-(\frac{1}{\sqrt{3}}x - \sqrt{3}p)^2}|.$$

and hence we can ease the notation by dropping the factor $e^{-\frac{2}{3}x^2}$. Namely, we need a coloring σ such that, given a point set $P \subset [-1, 1]$,

$$|\sum_{p\in P} \sigma(p) e^{2p^2} e^{-(\frac{1}{\sqrt{3}}x - \sqrt{3}p)^2}| < \alpha \text{ for } \textit{any } x \in \mathbb{R}$$

with probability at least $1 - O(e^{-\Omega(\alpha^2)})$ for any α. Since the Gaussian kernel is a positive-definite kernel, it implies that the term $e^{-(\frac{1}{\sqrt{3}}x - \sqrt{3}p)^2}$ can be rewritten as $\langle u^{(\frac{1}{\sqrt{3}}x)}, u^{(\sqrt{3}p)} \rangle$ where $u^{(\cdot)}$ is a vector such that $\langle u^{(s)}, u^{(t)} \rangle = e^{-(s-t)^2}$ for any $s, t \in \mathbb{R}$. It is worth noting that $\|u^{(s)}\|^2 = \langle u^{(s)}, u^{(s)} \rangle = e^{-(s-s)^2} = 1$ for any $s \in \mathbb{R}$. Hence, we further rewrite (4) as: given a point set $P \subset [-1, 1]$,

$$|\langle u^{(\frac{1}{\sqrt{3}}x)}, \Sigma \rangle| < \alpha \text{ for } \textit{any } x \in \mathbb{R} \text{ with probability at least } 1 - O(e^{-\Omega(\alpha^2)}) \text{ for any } \alpha$$

where $\Sigma = \sum_{p\in P} \sigma(p) e^{2p^2} u^{(\sqrt{3}p)}$. It is the inner product form we are looking for in order to apply the algorithm in Theorem 3. Recall that the norms of the input vectors and the query vectors in Banaszczyk's Theorem are required to be not larger than 1. We check that the norm of the query vector $\|u^{(\frac{1}{\sqrt{3}}x)}\| = 1$ and the norm of the input vector $\|e^{2p^2} u^{(\sqrt{3}p)}\| = O(1)$ since we assume that $P \subset [-1, 1]$. Karnin and Liberty [21] assumed *both* the point set P and the query x lie within a constant radius because their result stops short of handling the norms of these vectors when using Banaszczyk's Theorem. If we take $\frac{e^{2p^2} u^{(\sqrt{3}p)}}{\|e^{2p^2} u^{(\sqrt{3}p)}\|}$ as the input vectors, we can apply the algorithm in Theorem 3 to construct the desired coloring.

4 Proofs

In this section, we will show how to construct an ε-coreset via discrepancy theory. From now on, we assume that d is a constant. The log function in this paper is base e. Also, we define the following notations. Let $\mathsf{Grid}_d(\gamma) \subset \mathbb{R}^d$ be an infinite lattice grid of cell width γ, i.e. $\{(\gamma i_1, \ldots, \gamma i_d) \mid i_1, \ldots, i_d \text{ are integers}\}$. Denote $B^d_\infty(r) = \{x \mid |x_j| < r \text{ for } j = 1, \ldots, d\}$ to be a ℓ_∞-ball of radius r. We define a decreasing sequence n_i in the following way: $n_0 = \log^2 n$, $n_1 = \sqrt{3 \log n} + 3$ and $n_{i+1} = \sqrt{3 \cdot 2^{\ell(n)-i} \log n_i}$ for $i = 1, \ldots, \ell(n) - 1$. Here, $\ell(n) + 3$ is the smallest integer k that $\mathrm{ilog}(k, n) < 0$ where $\mathrm{ilog}(k, n) = \log \cdots \log n$ (there are k log functions) and it is easy to see that $\ell(n) = O(\log^* n)$. For $i = 0, \ldots, \ell(n) - 1$, denote $S_i = \mathsf{Grid}_d(\frac{1}{C_0 n_i}) \cap [-n_{i+1}, n_{i+1}]^d = \mathsf{Grid}_d(\frac{1}{C_0 n_i}) \cap B^d_\infty(n_{i+1})$ where C_0 is a sufficiently large constant. Namely, S_i is a bounded lattice grid and its size is at most $(2C_0 n_i n_{i+1})^d$. Note that S_i may be interpreted as a subset of S_0 but, for clarity, we still view them as different sets. Throughout this section, the absolute constants C_0, C_1, C are unchanged. Also, C is larger than C_1 and C_1 is larger than C_0.

4.1 Useful lemmas

Before we go into the main proof, we first present some important observations.

▶ **Lemma 5.** *Suppose $P \subset B^d_\infty(1)$ be a point set of size n and σ is a coloring on P. Then, we have*

$$\sup_{x \in B^d_\infty(\sqrt{3\log n}+3)} |\sum_{p\in P} \sigma(p) e^{2\|p\|^2} e^{-\frac{1}{3}\|x-3p\|^2}| \le 4 \cdot \sup_{s\in S_0} |\sum_{p\in P} \sigma(p) e^{2\|p\|^2} e^{-\frac{1}{3}\|s-3p\|^2}| + 7$$

where $S_0 = \mathsf{Grid}_d(w) \cap [-\sqrt{3\log n} - 3, \sqrt{3\log n} + 3]^d = \mathsf{Grid}_d(w) \cap B^d_\infty(\sqrt{3\log n} + 3)$ with $w = \frac{1}{C_0 \log^2 n}$. Here, $\mathsf{Grid}_d(\gamma) = \{(\gamma i_1, \ldots, \gamma i_d) \mid i_1, \ldots, i_d \text{ are integers}\} \subset \mathbb{R}^d$ is an infinite lattice grid.

The main technique in Lemma 5 is expanding the expression by Taylor expansion. Then, by truncating the Taylor expansion with a finite number of terms, one can bound the derivatives of the expression by using Markov Brother's inequality (Theorem 4). Since the width of the grid cell in S_0 depends on the number of terms in the Taylor expansion, we need to argue that a small number of terms suffices to bound the error.

▶ **Lemma 6.** *Given a coloring σ. For any $x, s \in \mathbb{R}^d$ such that $|x_j| \le |s_j|$ for all $j = 1, 2, \ldots, d$, we have*

$$\begin{aligned}
&|\sum_{p\in P} \sigma(p) e^{-\|x-p\|^2} - \sum_{p\in P} \sigma(p) e^{-\|s-p\|^2}| \\
&\qquad \le (\|s\|^2 - \|x\|^2)|\sum_{p\in P} \sigma(p) e^{-\|x-p\|^2}| + 2\sum_{j=1}^{d} |s_j - x_j| \cdot |\sum_{p\in P} \sigma(p) e^{-\|\xi^{(j)}-p\|^2}|
\end{aligned}$$

where $\xi^{(j)} = (x_1, \ldots, x_{j-1}, \xi_j, s_{j+1}, \ldots, s_d)$ for some ξ_j between $|x_j|$ and $|s_j|$.

In the inductive step, the main observation is the absolute difference of the discrepancy objective at two different points, $|\mathcal{D}_{P,\sigma}(x) - \mathcal{D}_{P,\sigma}(y)|$, can be bounded by the discrepancy objective itself. Lemma 6 is the lemma providing the key inequality to perform the inductive steps.

Finally, we also show the asymptotic bound of the recurrence equation n_i in Lemma 7.

▶ **Lemma 7.** *Let $n_0 = \log^2 n$, $n_1 = \sqrt{3\log n} + 3$ and $n_{i+1} = \sqrt{3 \cdot 2^{\ell(n)-i} \log n_i}$ for $i = 1, \ldots, \ell(n) - 1$. Then, $n_{\ell(n)} = O(1)$. Recall that $\ell(n) + 3$ is the smallest integer k that $\mathrm{ilog}(k, n) < 0$.*

4.2 Base step

Recall that the definition of $\mathcal{D}_{P,\sigma}(x)$ is $\sum_{p\in P} \sigma(p) e^{-\|x-p\|^2}$. Lemma 8 shows that *if* a coloring σ satisfies that $|\mathcal{D}_{P,\sigma}(x)|$ is small for all x in a finite subset (which is a grid) of $\mathbb{R}^d$, then the coloring σ also satisfies that $|\mathcal{D}_{P,\sigma}(x)|$ is small for all $x \in \mathbb{R}^d$. Note that we still haven't provided the detail on *how* to find such coloring and we will do it in the full algorithm.

▶ **Lemma 8.** *Suppose $P \subset B^d_\infty(1)$. Given a coloring σ such that, for all $s' \in S_0 = \mathsf{Grid}_d(w) \cup B^d_\infty(n_1)$ where $w = \frac{1}{C_0 \log^2 n} = \frac{1}{C_0 n_0}$ is the same w shown in Lemma 5 and $n_1 = \sqrt{3\log n} + 3$,*

$$|\mathcal{D}_{P,\sigma}(s')| = |\sum_{p\in P} \sigma(p) e^{-\|s'-p\|^2}| < C_1 n_1 e^{-\frac{2}{3}\|s'\|^2}$$

Then, we have, for all $x \in \mathbb{R}^d$,

$$|\mathcal{D}_{P,\sigma}(x)| = |\sum_{p\in P} \sigma(p)e^{-\|x-p\|^2}| < Cn_1 e^{-\frac{2}{3}\|x\|^2}.$$

Here, C, C_1, C_0 *are sufficiently large constant depending on* d *only.*

We make a short remark here. One might notice that Lemma 8 states $w = \frac{1}{C_0 \log^2 n}$ while it is sufficient to set $w = \Omega(\frac{1}{n})$ as suggested in Section 3. As we mentioned before, our final algorithm is a Las Vegas algorithm and hence we need to check if the output coloring has the desired discrepancy. We check it by enumerating the relevant grid points and computing the discrepancy at them. Making w larger reduces the size of the grid and hence improves the running time. Nonetheless, $w = \frac{1}{C_0 \log^2 n} = \Omega(\frac{1}{n})$ and hence it doesn't change the logic.

4.3 Inductive step

Lemma 9 suggests that *if* a coloring σ satisfies that $|\mathcal{D}_{P,\sigma}(x)|$ is small for all $x \in \mathbb{R}^d$, then the coloring σ satisfies that the absolute difference $|\mathcal{D}_{P,\sigma}(x) - \mathcal{D}_{P,\sigma}(s)|$ is also small for any two close points x, s within a certain region. It is achieved by the observation that the slope of $\mathcal{D}_{P,\sigma}$ can be bounded by $\mathcal{D}_{P,\sigma}$ itself in magnitude.

▶ **Lemma 9.** *Suppose* $P \subset B^d_\infty(1)$. *Let* $D_i = C \cdot \frac{5}{4}(1 - \frac{1}{5^i})$ *and* $I_i = \frac{1}{3} + \frac{1}{3}(1 - \frac{1}{2^{\ell(n)-i}})$. *Given a coloring* σ *such that, for all* $x \in \mathbb{R}^d$,

$$|\mathcal{D}_{P,\sigma}(x)| = |\sum_{p\in P} \sigma(p)e^{-\|x-p\|^2}| < D_i \cdot n_i e^{-I_i\|x\|^2}.$$

If $x \in B^d_\infty(n_{i+1})$, *then*

$$|\mathcal{D}_{P,\sigma}(x) - \mathcal{D}_{P,\sigma}(s)| = |\sum_{p\in P} \sigma(p)e^{-\|x-p\|^2} - \sum_{p\in P} \sigma(p)e^{-\|s-p\|^2}| \le \frac{1}{5} D_i \cdot n_{i+1} e^{-I_i\|x\|^2}.$$

where $s \in S_i$ *is the closest point to* x *that* $|s_j| > |x_j|$ *for all* $j = 1, 2, \ldots, d$.

Similar to Lemma 8, Lemma 10 shows that *if* a coloring σ satisfies that $|\mathcal{D}_{P,\sigma}(x)|$ is small for all x in a finite subset (which is a grid) of $\mathbb{R}^d$, then the coloring σ also satisfies that $|\mathcal{D}_{P,\sigma}(x)|$ is small for all $x \in \mathbb{R}^d$. The only difference is that we can take the advantage of the discrepancy guarantee from the previous iterations.

▶ **Lemma 10.** *Suppose* $P \subset B^d_\infty(1)$. *Recall that* $D_i = C \cdot \frac{5}{4}(1 - \frac{1}{5^i})$ *and* $I_i = \frac{1}{3} + \frac{1}{3}(1 - \frac{1}{2^{\ell(n)-i}})$ *which is the same definition as in Lemma 9. Given a coloring* σ *such that, for all* $s' \in S_i$,

$$|\sum_{p\in P} \sigma(p)e^{-\|s'-p\|^2}| < C_1 n_{i+1} e^{-\frac{2}{3}\|s'\|^2}$$

and, for all $x \in \mathbb{R}^d$,

$$|\sum_{p\in P} \sigma(p)e^{-\|x-p\|^2}| \le D_i \cdot n_i e^{-I_i\|x\|^2}.$$

Then, we have, for all $x \in \mathbb{R}^d$,

$$|\sum_{p\in P} \sigma(p)e^{-\|x-p\|^2}| < D_{i+1} \cdot n_{i+1} e^{-I_{i+1}\|x\|^2}.$$

Here, C, C_1 *are sufficiently large constants.*

4.4 Full algorithm

For now, we still assume that $P \subset B^d_\infty(1)$. Now, we can apply the algorithm in Theorem 3 to construct our coloring σ that produces a low discrepancy, $|\mathcal{D}_{P,\sigma}(x)|$, for all $x \in \mathbb{R}^d$. Recall that $\ell(n)+3$ is the smallest integer k that $\mathrm{ilog}(k,n) < 0$. Also, we defined n_i before such that $n_0 = \log^2 n$, $n_1 = \sqrt{3\log n}+3$ and $n_{i+1} = \sqrt{3 \cdot 2^{\ell(n)-i}\log n_i}$.

▶ **Lemma 11.** *Assuming $P \subset B^d_\infty(1)$. Given a set of vectors V_P defined as follows.*

$$V_P = \{\frac{1}{\sqrt{1+e^{4d}}}\begin{pmatrix}1\\ v^{(p)}e^{2\|p\|^2}\end{pmatrix} \mid p \in P\}$$

such that $\langle v^{(p)}, v^{(q)}\rangle = e^{-3\|p-q\|^2}$ for any $p,q \in P$. Then, by taking V_P as the input, the algorithm in Theorem 3 constructs a coloring σ on P such that

$$|\sum_{p\in P}\sigma(p)e^{-\|x-p\|^2}| < C \cdot \frac{5}{4} \cdot n_{\ell(n)}e^{-\frac{1}{3}\|x\|^2}$$

for all $x \in \mathbb{R}^d$ and $|\sum_{p\in P}\sigma(p)| \le C$ with probability at least $\frac{1}{2}$.

Recall that we eventually would like to use the halving technique to construct our ε-coreset. To use the halving technique, we need to ensure that half of the points in P are $+1$ and the other half are -1. In Lemma 11, the 1s concatenated on top of the vectors $v^{(p)}e^{2\|p\|^2}$ in V_P ensure the coloring has the above property.

▶ **Lemma 12.** *Assuming $P \subset B^d_\infty(1)$. There is an efficient algorithm that constructs a coloring σ such that $|\sum_{p\in P}\sigma(p)e^{-\|x-p\|^2}| = O(n_{\ell(n)}e^{-\frac{1}{3}\|x\|^2})$ for all $x \in \mathbb{R}^d$ and half of points are assigned $+1$ with probability at least $\frac{1}{2}$.*

Algorithm 1 Construction of the coloring.

input: a point set $P \subset \mathbb{R}^d$

1: initialize $Q_g = \emptyset$ for all $g \in \mathsf{Grid}_d(2)$
2: **for** each $p \in P$ **do**
3: insert p into Q_g where $g \in \mathsf{Grid}_d(2)$ is the closest point to p.
4: **for** each non-empty Q_g **do**
5: construct a collection V_g of vector $\{\frac{1}{\sqrt{1+e^{4d}}}\begin{pmatrix}1\\ v^{(p)}e^{2\|p\|^2}\end{pmatrix} \mid p \in Q_g\}$ such that $\langle v^{(p)}, v^{(q)}\rangle = e^{-3\|p-q\|^2}$ for any $p,q \in Q_g$
6: use V_g as the input and run the algorithm in Theorem 3 to obtain a coloring σ_g on Q_g
7: check if σ_g satisfies the conditions in Lemma 8 and Lemma 10 and repeat line 6 if not
8: flip the color of any points such that half of points in Q_g are colored $+1$.
9: **return** a coloring $\sigma : P \to \{-1,+1\}$ such that $\sigma(p) = \sigma_g(p)$ when $p \in Q_g$

We can now remove the assumption of $P \subset B^d_\infty(1)$. Algorithm 1 is a Las Vegas algorithm that constructs a coloring on the input point set P. We can now show how to construct a coloring such that the discrepancy is small. Recall that we defined $\mathsf{Grid}_d(\gamma) = \{(\gamma i_1, \ldots, \gamma i_d) \mid i_1, \ldots, i_d \text{ are integers}\} \subset \mathbb{R}^d$ to be an infinite lattice grid. The idea of Algorithm 1 is that we first decompose the entire $\mathbb{R}^d$ into infinitely many ℓ_∞-balls of radius 1. Then, we partition our input P such that each point $p \in P$ lies in some ℓ_∞-ball. For each non-empty ℓ_∞-ball, run the algorithm in Theorem 3 to construct a coloring with the desired discrepancy by Lemma 12. Finally, we argue that there is an extra constant factor in the final discrepancy.

▶ **Lemma 13.** *Suppose $P \subset \mathbb{R}^d$ be a point set of size n. Then, Algorithm 1 constructs a coloring σ on P efficiently such that $\sup_{x\in\mathbb{R}^d} |\sum_{p\in P} \sigma(p)e^{-\|x-p\|^2}| = O(1)$ and half of the points in P are colored $+1$.*

One can first perform random sampling [24] before running Algorithm 1 such that the input size $n = O(\frac{1}{\varepsilon^2})$. Finally, by the standard halving technique, we have the following theorem.

▶ **Theorem 14** (Restated Theorem 1). *Suppose $P \subset \mathbb{R}^d$ be a point set of size n. Let $\overline{\mathcal{G}}_P$ be the Gaussian kernel density estimate of P, i.e. $\overline{\mathcal{G}}_P(x) = \frac{1}{|P|}\sum_{p\in P} e^{-\|x-p\|^2}$ for any $x \in \mathbb{R}^d$. For a fixed constant d, there is an algorithm that constructs a subset $Q \subset P$ of size $O(\frac{1}{\varepsilon})$ such that $\sup_{x\in\mathbb{R}^d} |\overline{\mathcal{G}}_P(x) - \overline{\mathcal{G}}_Q(x)| < \varepsilon$ and has a polynomial running time in n.*

5 Conclusion and discussion

In this paper, we studied the question of constructing coresets for kernel density estimates. We proved that the Gaussian kernel has an ε-coreset of the optimal size $O(1/\varepsilon)$ when d is a constant. This coreset can be constructed efficiently. We leveraged Banaszczyk's Theorem to construct a coloring such that the kernel discrepancy is small. Then, we constructed an ε-coreset of the desired size via the halving technique.

Some open problems in discrepancy theory, such as Tusnády's Problem, have an issue that an extra factor shows up when we generalize the result from the case of $d = 1$ to the case of larger d. A previous result of our problem is reducing our problem to Tusnády's Problem. It turns out that, if $d = 1$, the trivial solution gives the optimal result. Unfortunately, it cannot be generalized to the higher dimensional case. Our new induction analysis combining with Banaszczyk's Theorem provides a non-trivial perspective even when $d = 1$. Hence, it might open up a possibility of improving the results on these open problems.

Even though the Gaussian kernel is a major class of kernels in most of applications, it would be interesting to investigate similar results on other kernel settings such as the Laplace kernel. Our approach exploits the properties of the Gaussian kernel such as factoring out the $e^{-\Omega(\|x\|^2)}$ factor while maintaining the positive-definiteness. Generalizing the result to a broader class of kernels might require deeper understandings of the properties that other kernels share with the Gaussian kernel.

In some applications, the input data might be in the high dimensional space. Our result assumes that d is a constant. Note that one of the previous results is sub-optimal in terms of ε but is optimal in terms of d. The dependence on d in our result is exponential which we might want to avoid in the high dimensional case. Hence, improving the dependence on d to polynomial is also interesting because it would be more practical in some applications.

References

1. Pankaj K Agarwal, Sariel Har-Peled, Haim Kaplan, and Micha Sharir. Union of random minkowski sums and network vulnerability analysis. *Discrete & Computational Geometry*, 52(3):551–582, 2014.
2. Christoph Aistleitner, Dmitriy Bilyk, and Aleksandar Nikolov. Tusnády's problem, the transference principle, and non-uniform qmc sampling. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 169–180. Springer, 2016.
3. Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.
4. Wojciech Banaszczyk. Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998.

5 Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The gram-schmidt walk: a cure for the banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–597, 2018.

6 Nikhil Bansal and Shashwat Garg. Algorithmic discrepancy beyond partial coloring. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 914–926, 2017.

7 József Beck. Roth's estimate of the discrepancy of integer sequences is nearly sharp. *Combinatorica*, 1(4):319–325, 1981.

8 Jon Louis Bentley and James B Saxe. Decomposable searching problems i: Static-to-dynamic transformation. *J. algorithms*, 1(4):301–358, 1980.

9 Moses Charikar, Michael Kapralov, Navid Nouri, and Paris Siminelakis. Kernel density estimation through density constrained near neighbor search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 172–183. IEEE, 2020.

10 Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *The Journal of Machine Learning Research*, 18(1):5845–5884, 2017.

11 Bernard Chazelle. *The discrepancy method: randomness and complexity*. Cambridge University Press, 2001.

12 Bernard Chazelle and Jiřı Matoušek. On linear-time deterministic algorithms for optimization problems in fixed dimension. *Journal of Algorithms*, 21(3):579–597, 1996.

13 Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):1–30, 2010.

14 Luc Devroye and László Györfi. *Nonparametric Density Estimation: The L_1 View*. Wiley, 1984.

15 Bernd Gärtner and Martin Jaggi. Coresets for polytope distance. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 33–42, 2009.

16 Leslie Greengard and John Strain. The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

17 Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

18 Mingxuan Han, Michael Matheny, and Jeff M Phillips. The kernel spatial scan statistic. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 349–358, 2019.

19 Phillips Jeff and Tai Wai Ming. The gaussiansketch for almost relative error kernel distance. In *International Conference on Randomization and Computation (RANDOM)*, 2020.

20 Sarang Joshi, Raj Varma Kommaraji, Jeff M Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *Proceedings of the twenty-seventh annual symposium on Computational geometry*, pages 47–56, 2011.

21 Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Conference on Learning Theory*, pages 1975–1993, 2019.

22 Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pages 544–552, 2015.

23 Jasper CH Lee, Jerry Li, Christopher Musco, Jeff M Phillips, and Wai Ming Tai. Finding an approximate mode of a kernel density estimate. In *29th Annual European Symposium on Algorithms (ESA 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

24 David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

25 AA Markov. On a question of di mendeleev, zap. *Petersburg Akad. Nauk*, 62:1–24, 1889.

26 Jiri Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 2009.

27 Jiří Matoušek, Aleksandar Nikolov, and Kunal Talwar. Factorization norms and hereditary discrepancy. *International Mathematics Research Notices*, 2020(3):751–780, 2020.

28 Jeff M Phillips. Algorithms for ε-approximations of terrains. In *International Colloquium on Automata, Languages, and Programming*, pages 447–458. Springer, 2008.

29 Jeff M Phillips. ε-samples for kernels. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1622–1632. SIAM, 2013.

30 Jeff M Phillips and Wai Ming Tai. Improved coresets for kernel density estimates. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2718–2727. SIAM, 2018.

31 Jeff M Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates. *Discrete & Computational Geometry*, pages 1–21, 2019.

32 Jeff M Phillips, Bei Wang, and Yan Zheng. Geometric inference on kernel density estimates. In *31st International Symposium on Computational Geometry (SoCG 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

33 Alessandro Rinaldo, Larry Wasserman, et al. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.

34 Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

35 David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

36 Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

37 Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985.

38 Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

39 Grace Wahba et al. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

40 Yan Zheng and Jeff M Phillips. L_∞ error and bandwidth selection for kernel density estimates of large data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1533–1542, 2015.

41 Shaofeng Zou, Yingbin Liang, H Vincent Poor, and Xinghua Shi. Unsupervised nonparametric anomaly detection: A kernel method. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 836–841. IEEE, 2014.