Applications of Discrepancy Theory to Machine Learning

Rainie Heck

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Thomas Rothvoss, Chair
Rekha Thomas
Cynthia Vinzant

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

Applications of Discrepancy Theory to Machine Learning

Rainie Heck

Chair of the Supervisory Committee:
Dr. Thomas Rothvoss
Departments of Mathematics and Computer Science

In the combinatorial discrepancy theory problem, one is given a base set $[n]$ and a collection of subsets $S_1, ..., S_m \subseteq [n]$ and asked to color the elements of $[n]$ so that each set S_i is as balanced as possible. This simple set-system based question has spawned a multitude of generalizations and found many recent applications in various areas of machine learning. In this dissertation, we introduce the discrepancy problem and its geometric generalization, the vector balancing problem, and then prove two sets of results about applications of the discrepancy problem to machine learning. To conclude, we prove a more abstract result about the vector balancing constant for zonotopes. The first application to machine learning–coresets for kernel density estimators–gives both improved bounds over existing results for a variety of applications of interest, as well as a new chaining-based technique that allows for a more data-driven approach to the problem. The second application–to quantization of neural networks–is a new application of discrepancy theory that provides improvements over existing algorithmic approaches to the problem. Finally, our results for vector balancing for zonotopes address and nearly resolves an open conjecture, leaving only a $\log \log \log d$ gap.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Thomas Rothvoss, my advisor and the Chair of my supervisory committee, for his support and collaboration in research during my PhD, and for helping me and supporting me in designing a wonderful PhD experience. I am also thankful to my supervisory committee, Dr. Maria Meila, Dr. Rekha Thomas, and Dr. Cynthia Vinzant, for their support during my comprehensive exam and final exam, as well as for research support, guidance, and help extended outside of their work on my committee. Many thanks to Dr. Sara Billey and Dr. Tatiana Toro, whose support and advice during the early part of my graduate career was invaluable. Many thanks also to my friends and family for supporting me along the way. Finally, special thanks to the National Science Foundation Graduate Research Fellowship program and the Department of Defense SMART program for the financial and career support throughout my doctoral studies.

Chapter 1

INTRODUCTION

1.1 Introduction to Discrepancy Theory and the Discrepancy Method

Discrepancy Theory, a subfield of combinatorics, assigns the following task: given a ground set $[n] := \{1, ..., n\}$ and a collection of subsets $\mathcal{S} = \{S_1, ..., S_m\} \subseteq 2^{[n]}$, color each element of $[n]$ either red or blue so that each set $S_i \in \mathcal{S}$ is as balanced (in terms of color) as possible. To make this question rigorous, we will consider *colorings* of the elements in $[n]$, denoted by maps $\chi : [n] \to \{\pm 1\}$. Then given a fixed coloring χ of $[n]$, we denote $\chi(S_i) := \sum_{j \in S_i} \chi(j)$ for each $i \in [m]$ and define the *discrepancy* of $\mathcal{S}$ under χ as

$$\text{disc}_\chi(\mathcal{S}) = \max_{i \in [m]} |\chi(S_i)|.$$

Then we can define the discrepancy of the whole set system $\mathcal{S}$ as the discrepancy of the optimally chosen coloring, i.e.

$$\text{disc}(\mathcal{S}) = \min_{\chi:[n]\to\{\pm1\}} \max_{i \in [m]} |\chi(S_i)| = \min_{\chi:[n]\to\{\pm1\}} \text{disc}_\chi(\mathcal{S}). \tag{1.1}$$

We expect bounds on $\text{disc}(\mathcal{S})$ to depend on the number of elements n, the number of sets m, and the structure of the set system itself. A bound based on purely random colorings can be obtained fairly immediately by applying the following standard theorem (see for example [1]).

Theorem 1.1.1 (Chernov bound)**.** *Take independent random variables $X_1, ..., X_k$ such that $\Pr[X_i = 1] = \Pr[X_i = -1] = 1/2$. Then for any $\lambda \geq 0$,*

$$\Pr\left[\left|\sum_{i=1}^{k} X_k\right| > \lambda\sqrt{k}\right] < 2e^{-\lambda^2/2}.$$

Applying Theorem 1.1.1 for $\lambda := \sqrt{2\log(4m)}$ to the discrepancy of each set S_i together with the union bound shows that with probability at least $1/2$, the coloring gives a discrepancy of $O(\sqrt{n \log m})$. One of the most ground-breaking theorems in this area, colloquially known as Spencer's Theorem, successfully removes the $\log m$ term.

Theorem 1.1.2 (Spencer's Theorem [66])**.** *For any set system $\mathcal{S}$ with $m \geq n$ sets on n elements, one has*

$$\operatorname{disc}(\mathcal{S}) = O(\sqrt{n \log(2m/n)}).$$

In particular, for $m = O(n)$, $\operatorname{disc}(\mathcal{S}) = O(\sqrt{n})$.

Spencer's original proof was based on a very non-constructive approach using the pigeonhole principle; in his original paper he demonstrated that finding a constructive solution would be extremely challenging. Bansal [10] and Lovett and Meka [44] were nevertheless able to successfully prove a constructive version of Spencer's Theorem. We state Lovett and Meka's version below, which was the first to give bounds matching that of Theorem 1.1.2.

Theorem 1.1.3. *For any set system $\mathcal{S}$ with $|\mathcal{S}| = m$ on base set $[n]$, there exists a randomized algorithm running in time $\widetilde{O}((n+m)^3)$[1] that with probability at least $1/2$ computes a coloring $\chi : [n] \to \{\pm 1\}$ such that*

$$\operatorname{disc}_\chi(\mathcal{S}) \leq K\sqrt{n \cdot \log_2(m/n)}$$

for $K > 0$ a universal constant.

For several conjectured generalizations of Spencer's Theorem, see Section 4.1. In order to prove his theorem, Spencer introduced a more general linear-algebraic formulation of the combinatorial discrepancy problem. This generalization, called the *matrix discrepancy* problem, can be formulated as follows: given a matrix $A \in [-1,1]^{m \times n}$, find signs $x \in \{\pm 1\}^n$ to minimize $\|Ax\|_\infty$. The *discrepancy* of A is $\operatorname{disc}(A) := \min_{x \in \{\pm 1\}^n} \|Ax\|_\infty$. If we represent the set system $\mathcal{S}$ by its incidence matrix $A \in \{0,1\}^{m \times n}$, then we immediately

[1] $\widetilde{O}$ suppresses polylogarithmic factors.

obtain $\text{disc}(A) = \text{disc}(\mathcal{S})$; however, this generalization also captures many other problems of interest, including a further geometric generalization introduced in Section 1.2.

The matrix discrepancy problem has drawn significant interest from machine learning applications, to the extent that such approaches have been dubbed the *discrepancy method*. For a summary of this method, see the excellent book of Chazelle [20]. Broadly speaking, the mathematical content of these applications involves analyzing the combinatorial and/or geometric structures introduced by collections of matrices derived from specific machine learning problems. We will describe two such approaches in Chapters 2 and 3; in particular, the problem described in Chapter 3, quantization of neural networks, is a new application of the discrepancy method, introduced in our paper [6]. For a few other examples of applications of the discrepancy method see [53] for a connection to differential privacy and [35] for the best-known approximation algorithm for the Bin Packing problem. Algorithmic solutions to discrepancy problems–such as Theorem 1.1.3–have been extremely important in the discrepancy method, as from the standpoint of applications, computational feasibility is essential.

1.2 A Geometric Version of Discrepancy Theory: Vector Balancing

In this section we introduce an interesting geometric generalization of the discrepancy theory problem, called the vector balancing problem. Given $n \in \mathbb{N}$ and two norms $\|\cdot\|_a, \|\cdot\|_b$ on $\mathbb{R}^d$, the vector balancing problem asks the following: given a collection of vectors $v_1, ..., v_n$ such that $\|v_i\|_a \leq 1$ for all $i \in [n]$, find an assignment of signs $\varepsilon \in \{\pm 1\}^n$ to minimize

$$\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_b.$$

The *n-vector balancing constant* of $\|\cdot\|_a$ and $\|\cdot\|_b$ is the best bound that holds for any selection of n such vectors:

$$\text{vb}_n(\|\cdot\|_a, \|\cdot\|_b) := \max_{v_1,...,v_n:\ \|v_i\|_a \leq 1} \min_{\varepsilon \in \{\pm 1\}^n} \|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_b.$$

We can also define the *vector balancing constant*, where one may choose any number of vectors:

$$\mathrm{vb}(\|\cdot\|_a, \|\cdot\|_b) := \sup_{n\in\mathbb{N}} \max_{v_1,\dots,v_n:\ \|v_i\|_a\le 1} \min_{\varepsilon\in\{\pm 1\}^n} \|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_b.$$

One may also define these quantities in an equivalent, more geometric way, using the Minkowski norm: for a given symmetric convex body $K \subseteq \mathbb{R}^d$, one obtains the corresponding Minkowski norm on $\mathbb{R}^d$:

$$\|x\|_K := \min\{\lambda \ge 0 : x \in \lambda K\}.$$

Then one can equivalently state the above definitions in terms of symmetric convex bodies $K, Q \subseteq \mathbb{R}^d$: the n-vector balancing constant of K and Q is

$$\mathrm{vb}_n(K,Q) := \inf\{r \ge 0 :\ \forall v_1,\dots,v_n \in K,\ \exists \varepsilon \in \{\pm 1\}^n \text{ s.t. } \varepsilon_1 v_1 + \cdots + \varepsilon_n v_n \in rQ\},$$

and similarly the vector balancing constant is

$$\mathrm{vb}(K,Q) := \inf\{r \ge 0 :\ \forall n \in \mathbb{N},\ \forall v_1,\dots,v_n \in K,\ \exists \varepsilon \in \{\pm 1\}^n \text{ s.t. } \varepsilon_1 v_1 + \cdots + \varepsilon_n v_n \in rQ\}.$$

The following surprising result of Lovasz, Spencer, and Vesztergombi relates these two definitions [45].

Theorem 1.2.1. *For any symmetric convex $K, Q \subseteq \mathbb{R}^d$,*

$$\mathrm{vb}(K,Q) \le 2 \cdot \mathrm{vb}_d(K,Q).$$

In practice, this means that it suffices to consider the case where one has $O(d)$ vectors.

To see that the vector balancing problem is a generalization of the matrix discrepancy problem, fix the maximum norm $\|\cdot\|_\infty$ on $\mathbb{R}^d$ and consider the task of balancing vectors $v_1,\dots,v_n \in B_\infty^d$. Defining a matrix $M \in \mathbb{R}^{d\times n}$ with columns $v_1,\dots,v_n$, we see that indeed $M \in [-1,1]^{d\times n}$, and that for a fixed assignment of signs $\varepsilon \in \{\pm 1\}^n$,

$$\|M\varepsilon\|_\infty = \|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_\infty, \tag{1.2}$$

which directly recovers the vector balancing problem $\mathrm{vb}(B_\infty^d, B_\infty^d)$. Using this example, we can interpret Spencer's theorem as a vector balancing result.

Theorem 1.2.2. *For $n, d \in \mathbb{N}$ with $n \leq d$,*

$$\mathrm{vb}_n(B_\infty^d, B_\infty^d) = \Theta\Big(\sqrt{n \log(2d/n)}\Big). \tag{1.3}$$

Further combining with Theorem 1.2.1 shows that in fact $\mathrm{vb}(B_\infty^d, B_\infty^d) = \Theta(\sqrt{d})$ (a matching lower bound can be explicitly constructed with Hadamard matrices).

Many results exist for a wide variety of vector balancing settings; we highlight a few that will be of particular use for our work and then close the section by mentioning other related results of interest. The primary vector balancing tool that will be of use in our work (see Chapter 2 and Chapter 4 in particular) will be the following theorem of Banaszczyk [7] and its corresponding algorithmic version [9]. By γ_d we denote the standard Gaussian measure on $\mathbb{R}^d$; see Section 4.1.3 for more details.

Theorem 1.2.3. *Let $K \subseteq \mathbb{R}^d$ be any convex body with $\gamma_d(K) \geq 1/2$. Given any vectors $v_1, ..., v_n \in B_2^d$, there exist signs $\varepsilon \in \{\pm 1\}^n$ so that*

$$\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n \in CK,$$

for $C > 0$ a universal constant.

Similar to the story of Spencer's theorem, for a long time this result was purely non-constructive; the first complete proof of a constructive version is due to Bansal, Dadush, Garg, and Lovett [9]. We will state a version of the algorithmic result based on the theory of subgaussian random variables in Appendix A. One aspect of Banaszczyk's theorem that is particularly interesting is that it produces a coloring without relying on the *partial coloring method*, first introduced by Beck in [13] and used by Spencer [66]. The partial coloring method constructs a coloring vector $\{\pm 1\}^n$ by iteratively coloring a constant fraction of the remaining coordinates until a full coloring has been obtained. To this end, we make the following definition: a vector $x \in [-1, 1]^n$ is a *good partial coloring* if

$$|\{j \in [n] : \ x_j \in \{-1, 1\}\}| \geq n/2.$$

The following result, which we will use in Chapter 4, shows a connection between the existence of good partial colorings and Gaussian lower measure bounds on sections of convex bodies [58].

Theorem 1.2.4. *For any $\alpha > 0$, there is a constant $c := c(\alpha) > 0$ and a randomized polynomial time algorithm that for a symmetric convex body $K \subseteq \mathbb{R}^n$, a $2n/3$-dimensional subspace $F \subseteq \mathbb{R}^n$ with $\gamma_F(K \cap F) \geq e^{-\alpha n}$, and a shift $y \in (-1, 1)^n$, finds $x \in c \cdot K \cap F$ so that $x + y$ is a good partial coloring.*

In addition to the results mentioned above, many other interesting results are known for various vector balancing problems, of which we list several. It is well-known that $\mathrm{vb}(B_2^d, B_2^d) = \sqrt{d}$ [62, 19, 68]. This result, along with Spencer's Theorem, were generalized to a colorful setting in [2]. It is also well-known that $\mathrm{vb}(B_1^d, B_1^d) = \Theta(d)$, and Beck and Fiala proved that $\mathrm{vb}(B_1^d, B_\infty^d) \leq 2$. The long-standing open Komlós conjecture posits that $\mathrm{vb}(B_2^d, B_\infty^d) = O(1)$; the strongest known result is $O(\sqrt{\log d})$, implied by Banaszczyk's theorem [7]. There are also many interesting online versions of the vector balancing problem, as well as related combinatorial games [67, 40]. Finally, for a summary of classic vector balancing results, see [30]. Many classical vector balancing results are surveyed by Giannopoulos [30]. Vector balancing in the plane was studied by Swanepoel [70] and Lund, Magazinov [46]. Online versions of vector balancing and related combinatorial games were considered by Spencer in [67, 69]. Various anti-balancing questions were discussed by Banaszczyk [8] and Ambrus, González Merino [3].

1.3 Summary of Topics to Be Covered

To conclude Chapter 1, we summarize the work that will be presented in this dissertation. In Chapter 2, we cover improved bounds for the problem of finding coresets for kernel density estimators, an application of the discrepancy method. In particular, we develop a new technique based on *chaining* (see Appendix A for more details) that allows dataset specific considerations not possible using previous techniques, and we show improved bounds for a

wide variety of kernel functions. The complete work can be found in [17]. In Chapter 3, we describe a new application of the discrepancy method to the problem of quantizing weights for neural networks. In this dissertation we focus on the theoretical portion of the work; for more details on the applications and experimental results, see [6]. In Chapter 4, we present a more abstract, geometric result in vector balancing that generalizes Spencer's Theorem to the setting of *zonotopes* [34]. In addition, in order to prove the existence of good partial colorings for zonotopes, we prove a lower bound on the Gaussian measure of sections of zonotopes that generalizes a similar result for the cube [77], and is of independent interest. Finally, in Chapter 5 we provide a brief summary of the results obtained and a few directions for future research.

Chapter 2

ε-CORESETS FOR KERNEL DENSITY ESTIMATORS

In this chapter we develop a new *chaining*-based technique using the discrepancy method to tackle the problem of the *coreset complexity* of *kernel density estimators* from machine learning. In addition to giving improved bounds on the coreset complexity of a wide class of kernel functions, our technique allows one to give bounds dependent on the geomertry of the data, offering a strong advantage for practical applications. In Section 2.1 we introduce the application of interest, coresets for kernel density estimators, followed by an overview of how the discrepancy method can be applied to this problem in Section 2.2. This is followed by a summary of existing results in Section 2.3 and the statement of our main results in Section 2.4. We complete the chapter with the proofs of our main results and a summary of their applications of interest in Sections 2.5 and 2.6, respectively.

2.1 Kernel Density Estimation and the Coresets Problem

Kernel density estimators provide a non-parametric method for estimating probability distributions. In contrast to parametric machine learning models, in which a set of training data is used to determine an optimal parameter vector which can be used to predict future outcomes without the training data, for non-parametric methods the number of parameters of the model has the capability to grow as the size of the dataset grows. Our paper will focus on the task of kernel density estimation, in which one estimates a probability distribution using kernel functions of the form $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. Kernels are often defined with a *bandwidth parameter* $1/\alpha$ that controls the width of the kernel function.

Given a probability distribution ρ and a collection of points $X = \{x_1, ..., x_n\} \sim \rho$ sampled independently, it is well known that for certain well-behaved kernel functions K, the

distribution ρ can be approximated very well by the *kernel density estimator* (KDE)

$$\mathrm{KDE}_X(y) = \frac{1}{n}\sum_{i\in[n]} K(x_i, y).$$

In particular, it is known that under certain conditions on the kernel function, KDE_X approximates ρ at the minimax optimal rate as $|X| \to \infty$ [76].

Although this is an elegant theoretical result, in practice it is computationally inefficient to store and make computations with an arbitrarily large number of data points n. One solution to reduce the computational complexity is to use an *ε-coreset* for a kernel density estimator.

Definition 2.1.1 (KDE ε-coreset)**.** *For fixed $\varepsilon > 0$, kernel function $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and data set $X \subseteq \mathcal{D}$, an* ε-coreset *for K is a subset $Q \subseteq X$ so that*

$$\|\mathrm{KDE}_X(y) - \mathrm{KDE}_Q(y)\|_\infty = \sup_{y\in\mathcal{D}} \left| \frac{1}{|X|}\sum_{x\in X} K(x,y) - \frac{1}{|Q|}\sum_{q\in Q} K(q,y) \right| \leq \varepsilon.$$

We will say that the coreset complexity of a kernel function K is the minimum possible size of a ε-coreset Q for K.

In general, coreset complexity bounds will depend on ε and the dimension d of the kernel domain, and they will be independent of the size of the set X. These bounds are also often independent of the choice of $X \subseteq \mathcal{D}$, although several previous results and several of our results give an explicit dependence on X that may allow improvement over existing bounds for sufficiently nice data sets. In particular, several of our bounds will depend on the radius of the set X. For more details about non-parametric methods and kernel density estimation, see for example [76].

2.2 The Discrepancy Method for Coresets for KDEs

One powerful method for proving bounds on the coreset complexity of kernel functions is the *discrepancy approach*. It has also been used in [55, 72, 57, 39] and is based on a method for computing range counting coresets [21, 57, 15, 56]. Following the notational conventions of [55], we make the following definition.

Definition 2.2.1 (Kernel Discrepancy)**.** *Given a data set $X \subseteq \mathcal{D}$, a kernel $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and a coloring $\beta \in \{\pm 1\}^X$, the kernel discrepancy at a point $y \in \mathcal{D}$ is defined as*

$$\mathrm{disc}_K(X, \beta, y) := \Big| \sum_{x \in X} \beta(x) K(x, y) \Big|.$$

The kernel discrepancy can then be defined as

$$\mathrm{disc}_K(n) = \max_{\substack{X \subseteq \mathcal{D}: \\ |X| = n}} \min_{\beta \in \{\pm 1\}^X} \max_{y \in \mathcal{D}} \mathrm{disc}_K(X, \beta, y).$$

We will also use the notation $\mathrm{disc}_K(X)$ to denote the kernel discrepancy with respect to a fixed data set X.

We can interpret Definition 2.2.1 in terms of both matrix discrepancy (see Section 1.1) and vector balancing (see Section 1.2). For both it will be necessary to make the simplifying assumption that the domain $\mathcal{D}$ of the kernel function is finite; this assumption can be realized by discretizing the domain using a sufficiently fine epsilon net. Such a tool is used in [55], and one of the main advantages of our chaining-based technique will be removing this restriction. Assuming that $|\mathcal{D}| = d$, then we note that defining the *kernel matrix* $A^K \in \mathbb{R}^{d \times n}$ with $A^K_{ij} = K(y_i, x_j)$ for $i \in [d]$, $j \in [n]$,

$$\mathrm{disc}_K(n) = \mathrm{disc}(A^K). \tag{2.1}$$

Similarly, for each data point x_j, $j \in [n]$, we can define a vector $v_j \in \mathbb{R}^d$ with entries $v_j^i = K(y_i, x_j)$, and then ask to balance $v_1, ..., v_n$ in the maximum norm to obtain $\mathrm{disc}_K(X)$. In these interpretations, we are then in practice studying the geometric and analytic properties of vectors and/or matrices defined using a variety of kernel functions.

In order to obtain useful results about the discrepancy of such "kernel matrices", it is necessary to make a few basic assumptions about the underlying kernels. One such assumption is of significant importance, and so we introduce it here.

Definition 2.2.2. *A kernel $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is said to be* positive definite *if given any selection of points $x_1, ..., x_m \in \mathcal{D}$, the Gram matrix G given by $G_{ij} = K(x_i, x_j)$ is positive definite.*

Positive definiteness has played a variety of roles in previous approaches to this problem: for example, in [55] Phillips and Tai use the decomposition property of positive definite matrices to write A_K as $A_K = B^T B$ and then leverage this decomposition to bound the matrix discrepancy. Our result relies instead on the theory of reproducing kernel Hilbert spaces, described by the following famous theorem [4].

Theorem 2.2.1 (Moore-Aronszajn, 1950)**.** *Let T be a set and K a positive definite function on $T \times T$. Then there is a map $\phi : T \to \mathcal{H}_K$ to a unique corresponding Hilbert space $\mathcal{H}_K$ so that for any $s, t \in T$,*

$$K(s,t) = \langle \phi(s), \phi(t) \rangle_{\mathcal{H}_K}.$$

The Hilbert space $\mathcal{H}_K$ is called the *reproducing kernel Hilbert space* (RKHS) associated to K. Using this theorem, one can make the following definition.

Definition 2.2.3. *The kernel distance associated to a kernel function K is the function*

$$D_K(x,y) := \|\phi(x) - \phi(y)\|_{\mathcal{H}_K}.$$

Here $\|h\|_{\mathcal{H}_K} = \sqrt{\langle h, h \rangle_{\mathcal{H}_K}}$ is the Euclidean norm in $\mathcal{H}_K$. In general, it is only true that D_K is a pseudometric on the domain of K; however, this is all we will need for our proofs. Almost all of the kernels commonly used in machine learning applications are positive definite.

We close this section by explaining how bounds on kernel discrepancy can be leveraged to obtain bounds on the coreset complexity for a given kernel K. The strategy is the following, often called the "halving trick" [21, 15, 56]: we construct a coreset of X by iteratively removing half of the points in X, and we select which half of the points are removed by creating colorings $\beta \in \{\pm 1\}^X$ minimizing the kernel discrepancy and then removing those points assigned -1 (in principle, there is no reason to expect that exactly half of the points are assigned $+1$, and half -1, but there are standard techniques to overcome this challenge [49]). Indeed, supposing that we have an optimal choice of signs $\beta \in \{\pm 1\}^X$ such that

$$\sup_{y \in \mathcal{D}} \left| \sum_{x \in X} \beta(x) K(x,y) \right| \leq f(n),$$

then we simply note that, letting X^+ be the set of points assigned $+1$ and X^- be the set of points assigned -1, then under the assumption that $|X^-| = |X|/2$, for any $y \in \mathcal{D}$,

$$\begin{aligned}\frac{1}{|X|}\sum_{x\in X}\beta(x)K(x,y) &= \frac{1}{|X|}\sum_{x\in X^+}K(x,y) - \frac{1}{|X|}\sum_{x\in X^-}K(x,y)\\ &= \frac{1}{|X|}\sum_{x\in X}K(x,y) - \frac{1}{|X|/2}\sum_{x\in X^-}K(x,y).\end{aligned} \tag{2.2}$$

Taking a supremum over $y \in \mathcal{D}$, the final line of (2.2) is exactly $\mathrm{KDE}_X(y) - \mathrm{KDE}_{X^-}(y)$. Thus, iterating this procedure t times, and denoting the resulting set at iteration s by X_s (with $X_0 := X$), we find that

$$\|\mathrm{KDE}_X - \mathrm{KDE}_{X_t}\|_\infty \leq \sum_{s\in[t]}\|\mathrm{KDE}_{X_{s-1}} - \mathrm{KDE}_{X_s}\|_\infty \leq \sum_{s\in[t]}\frac{2^{s-1}}{n}f\big(n/2^{s-1}\big).$$

Assuming that the function f grows sufficiently slowly[1], this sum will be dominated by the final term, which allows us to calculate the size of a coreset yielding error at most ε. Based on this connection, our proofs will focus on bounding the quantity $\mathrm{disc}_K(n)$ for different kernels K (or in some cases $\mathrm{disc}_K(X)$, when we want to account for the geometry of the data set X), and then the "halving trick" can easily be used to determine the corresponding size of the coreset thus obtained.

2.3 History of the Coresets for KDEs Problem

The problem of coreset complexity for a wide variety of kernel functions has been studied for several decades. Early approaches focused on random samples [64, 61, 37], but more recently analytic approaches and new algorithms have been discovered, leading to much stronger bounds. We provide a brief summary of previous approaches and results.

Joshi et al. [38] used a sampling technique to prove a bound of $O((1/\varepsilon^2)(d + \log(1/\delta)))$ on the coreset complexity of any centrally symmetric, non-increasing kernel, where δ is the probability of failure. Fasy et al. [27] used sampling to prove a different bound of $O((d/\varepsilon^2)\log(d\Delta/\varepsilon\delta))$, where $\Delta := \alpha \sup_{x,x'\in X}\|x - x'\|_2$ is the diameter of the data set. This

[1]It suffices if $f(n) \leq n^c$ for some fixed constant $0 < c < 1$.

result may improve upon that of Joshi et al. in the case that $K(x, x) > 1$, and it also applies to the broader class of Lipschitz kernels.

The following collection of results applies to the collection of *characteristic kernels*, a subset of positive definite kernels that satisfy the additional property that the mapping ϕ_K into the associated RKHS is injective, which implies that the induced kernel distance D_K is a metric. This class contains many, though not all, positive definite kernels, with a notable exception being the exponential kernel (see Theorem 2.6.1). Again using random sampling, Lopaz-Paz et al. [43, 51] gave a simpler bound of $O((1/\varepsilon^2)\log(1/\delta))$.

Improved results were proved using an iterative technique called *kernel herding* introduced by Chen et al. [22] to solve a closely related problem called *kernel mean approximation*, which is shown to upper bound kernel density approximation in the case of reproducing kernels [22, 65]. Chen et al. proved a bound of $O(1/(\varepsilon r_X))$, where r_X is the largest radius of a ball centered at $\frac{1}{|X|}\sum_{x\in X}\phi_K(x) \in \mathcal{H}_K$ that is completely contained in $\text{conv}\{\phi(x) : \ x \in X\}$. This paper claimed that r_X is always a constant greater than 0; however, Bach et al. [5] gave a new interpretation of the algorithm and argued that although r_X is arbitrarily small for continuous distributions, the constant $1/r_X$ is finite when X is finite. Their interpretation provided a bound of $(1/r_X^2)\log(1/\varepsilon)$. Bach et al. also provided a bound of $O(1/\varepsilon^2)$ in the case of *weighted coreset complexity*, where points in X can be assigned a non-negative weight, and this result was later improved to the setting of unweighted coresets [41].

Harvey and Samadi [32] applied the kernel herding technique to an even more general problem called *general mean approximation in* $\mathbb{R}^d$ to provide bounds on the coreset complexity of order $O((1/\varepsilon)\sqrt{n}\log^{2.5}(n))$, where $n = |X|$. The dependence on n is introduced by the worst case outcome of manipulating r_X using affine scaling. Locoste-Julien et al. [41] showed that actually one can take $n = O(1/\varepsilon^2)$ which improves the bound to $O((1/\varepsilon^2)\log^{2.5}(1/\varepsilon))$.

The next collection of bounds applies to *Lipschitz kernels*, that is, kernels where we can bound the *Lipschitz factor* $C_K := \max_{x,y,z\in\mathcal{D}} \frac{\|K(z,x)-K(z,y)\|_2}{\|x-y\|_2}$ of K. In the case that C_K is a small constant, as is the case for most kernels of interest, it is easy to see that taking a $2\varepsilon/(C_K\sqrt{d})$-net G_ε over the domain of K and mapping each point $x \in X$ to the closest point

in G_ε (with multiplicity) to obtain X_{G_ε}, we find that $\sup_{y\in\mathcal{D}} |\mathrm{KDE}_X(y) - \mathrm{KDE}_{X_{G_\varepsilon}}(y)| \leq \varepsilon$. Cortes and Scott's work on the sparse kernel mean problem [24] combined with the discretization argument above implies a bound of $O((\Delta/\varepsilon)^d)$ on the coreset-complexity, in the case that Δ is bounded.

Phillips [57]	$O((\alpha/\varepsilon)^{2d/(d+2)} \log^{d/(d+2)}(\alpha/\varepsilon))$	Lipschitz, PD
Phillips and Tai [55]	$\sqrt{d \log n}$	Lipschitz, PD
Tai [72]	$2^{O(d)}$	Gaussian
New	$2^{O(d)}\sqrt{\log\log n}$	Laplacian
New	$O(\sqrt{d \log(\mathrm{radius}X + \log n)})$	Gaussian, Laplacian
New	$O(\sqrt{d \log(2\max\{\alpha, 1\})})$	Exponential, JS, Hellinger

Table 2.1: Results from the Discrepancy Approach

Phillips [57] first used the discrepancy method to show a bound of

$$O((\alpha/\varepsilon)^{2d/(d+2)} \log^{d/(d+2)}(\alpha/\varepsilon))$$

for kernels with a Lipchitz factor $C_K = O(\alpha)$. Using a sorting argument, Phillips also showed in this paper that for $d = 1$ one can achieve a coreset of size $O(1/\varepsilon)$, matching a tight lower bound. Note that in general the coreset complexity is always bounded below by $O(1/\varepsilon)$, as can be seen by taking $O(1/\varepsilon)$ points that are spread far apart [57]. Phillips and Tai [55] improved these results by combining the discretization approach with the discrepancy approach to give a bound of $O((1/\varepsilon)\sqrt{d \log(1/\varepsilon)})$ for any positive definite, Lipschitz kernel with *bounded influence*, a restriction similar to the impact radius condition that we will define. This result applies to a very wide class of kernels, including all of the kernels that we will discuss in our paper, and also the sinc kernel, for which no earlier non-trivial bounds were known. They also provide a lower bound of $\Omega(\sqrt{d}/\varepsilon)$ for $d \in [2, 1/\varepsilon^2)$, and a tight lower bound of $O(1/\varepsilon^2)$ in the case that $d \geq 1/\varepsilon^2$, for all shift and rotation invariant kernels that are somewhere-steep. Tai [72] later proved that for d constant, the Gaussian kernel

has coreset complexity $O(1/\varepsilon)$, matching the optimal lower bound in terms of ε. This result supresses an exponential dependence on d, but is still interesting for small-dimensional data sets.

Finally, a related but not directly comparable result due to Karnin and Liberty [39] applies to kernels that are analytic functions of the dot product and satisfy the very strong condition that $\sup_{x,x' \in \mathcal{D}} \|x - x'\|_2 \leq R_K$, where R_K is a fixed constant determined by the kernel K. In this setting, they show a bound of $O(\sqrt{d}/\varepsilon)$ on the coreset complexity. However, as we will see, for kernels such as the Gaussian or Laplacian kernel, one can only assume that $\sup_{x,x' \in \mathcal{D}} \|x - x'\|_2 \leq n \log n$, where $n = |X|$, and thus this result does not apply. Their result can however be interpreted to give a bound of $O(\alpha \exp(\alpha)\sqrt{d}/\varepsilon)$ on the coreset complexity of the exponential kernel, as in this case the domain of the kernel function does have constant diameter.

2.4 Our Results

In order to state our main results, we will need to introduce the following notation related to kernel density estimation.

Definition 2.4.1 (Impact Radius)**.** *Given a kernel $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ for $\mathcal{D} \subseteq \mathbb{R}^d$, we define the impact radius of K as*

$$r_K(n) := \inf \big\{ r \geq 0 : \ \|x - y\|_2 \geq r \implies |K(x, y)| \leq 1/n \ \ \forall x, y \in \mathcal{D} \big\}.$$

Definition 2.4.2 (Query Space)**.** *Given a kernel $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ and a data set $X \subseteq \mathcal{D}$, we define the query space of K with respect to X as*

$$Q = \Big\{ y \in \mathcal{D} : \ \exists x \in X \ \text{s.t.} \ \|x - y\|_2 \leq r_K(|X|) \Big\} = \mathcal{D} \cap \Big(\bigcup_{x \in X} (x + r_K(|X|) B_2^d) \Big).$$

Note that in general both the impact radius and the query space may depend explicitly on the bandwidth parameter α, but this dependence often cancels out, making the bounds obtained independent of α. One notable exception where this cancellation does not occur is

when the domain $\mathcal{D}$ of the kernel is compact, for example for the exponential, Hellinger, and Jensen-Shannon kernels; we return to this idea in Section 2.6.

We also note that in the event that the query space Q given by a particular data set is disconnected, it suffices to consider the largest connected component of Q. To see this, note that by the definition of impact radius, the query points and data points from distinct connected components do not interact, thus we can apply the same bound to each connected component independently. For the proof of our results, we will assume that Q is connected.

Our two main results will be in terms of the size of the query space and the impact radius of the kernel, respectively. One of the key challenges in earlier applications of the discrepancy approach was that the domain $\mathcal{D}$ is often $\mathbb{R}^d$, the sphere S^d, the standard $(d-1)$-dimensional simplex Δ^d, or some other potentially unbounded and/or uncountably infinite space. These domains make bounding the discrepancy challenging, as probabilistic techniques are often used, and the size of these spaces make the union bound ineffective. The following lemma shows how the query space and impact radius can simplify this problem to at least ensure that the domain is bounded for kernels with sufficiently nice decay properties.

Lemma 2.4.1. *Let $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ be a kernel, $X \subseteq \mathcal{D}$ a data set, and Q the query space associated to K and X. Then*

$$\operatorname{disc}_K(n) \leq \operatorname{disc}_{K|_Q}(n) + O(1).$$

Proof. To prove the lemma, it suffices to show that for any $y \in \mathcal{D} \setminus Q$, $\operatorname{disc}_K(X, \beta, y) = O(1)$ for all $\beta \in \{\pm 1\}^X$. But this follows immediately, as we know by the definition of Q that for any $y \in \mathcal{D} \setminus Q$ and $x \in X$, $|K(x, y)| \leq 1/n$. Thus

$$\operatorname{disc}_K(X, \beta, y) = \Big| \sum_{x \in X} \beta(x) K(x, y) \Big| \leq \sum_{x \in X} |K(x, y)| \leq 1. \qquad \square$$

Earlier approaches using the discrepancy approach [55, 72] needed an even stronger version of Lemma 2.4.1 for Lipschitz kernels that also ensured that the query space could be made finite up to an $O(1)$ error using the Lipschitz constant of the kernel and a sufficiently

small ε-net. This lemma provides an extra factor of $\sqrt{\log n}$ in many results that we will avoid by using *chaining* [79], a multi-step construction of ε-nets (see Appendix A for more details and related results).

Note that in general, for an arbitrary data set, the query space can have volume up to $n\mathrm{Vol}_d(r_K B_2^d)$, as it is possible that the data points are well spread out and thus the impact radii of data points do not intersect. This observation is why [39] cannot be applied to the general Gaussian kernel, for example.

Our first result gives a discrepancy bound that depends explicitly on the choice of the data set $X \subseteq \mathbb{R}^d$.

Theorem 2.4.1 (H. and Rothvoss [17])**.** *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel with bandwidth parameter $1/\alpha$ and $X \subseteq \mathbb{R}^d$ be a dataset. Denote the query space of (K, X) by Q, and define $R = R(X)$ so that $Q \subseteq RB_2^d$. If K satisfies the following properties:*

(i) K is positive definite;

(ii) K satisfies $K(x, x) = 1$ for all $x \in Q$;

(iii) $K = \kappa(\alpha\|x - y\|_2)$, where $\kappa : \mathbb{R}_{\geq 0} \to [-1, 1]$ is strictly decreasing and continuous; and

(iv) The following bound holds, where $[0, b_K)$ is the domain of the integrand:

$$\int_0^{b_K} \sqrt{-\ln\left[\kappa^{-1}\left(\frac{2 - r^2}{2}\right)\right]}\,\mathrm{d}r = O(1);$$

then

$$\mathrm{disc}_K(X) = O(\sqrt{d \log(2 \max\{R\alpha, 1\})}).$$

Moreover, there is a randomized algorithm that on input of X and K, finds an according coloring $\beta \in \{\pm 1\}^X$ in polynomial time.

We will see in Section 2.6 that Theorem 2.4.1 can give strong improvements on the current best known bounds on the coreset complexity for the Gaussian and Laplacian kernels in the case that the data set X has bounded diameter. In particular, we have the following corollary.

Corollary 2.4.1. *Let K be a kernel satisfying the conditions of Theorem 2.4.1, and let $X \subseteq \mathbb{R}^d$ be any data set. Then*

$$\mathrm{disc}_K(X, n) \leq O(\sqrt{d \log(\mathrm{radius}(X) + \kappa^{-1}(1/n))}).$$

Note that because of the discretization necessary to prove Phillip and Tai's bounds [55], even in the case where the data set is bounded uniformly, the $\sqrt{\log n}$ factor in their discrepancy bound cannot be removed.

In the case where we do not want to account for the geometry of the data set X, we present the following stronger bound in the case that the dimension d is taken to be constant.

Theorem 2.4.2 (H. and Rothvoss [17])**.** *Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel with bandwidth parameter $1/\alpha$ and $X \subseteq \mathbb{R}^d$ a data set. If K satisfies the following properties:*

(i) K is positive definite;

(ii) K satisfies $K(x, x) = 1$ for all $x \in Q$;

(iii) $K = \kappa(\alpha\|x - y\|_2)$, where $\kappa : \mathbb{R}_{\geq 0} \to [-1, 1]$ is strictly decreasing and continuous; and

(iv) The following bound holds:

$$\int_0^{b_K} \sqrt{-\ln\left[\kappa^{-1}\left(\frac{2 - r^2}{2}\right)\right]} \mathrm{d}r = O(1);$$

then

$$\mathrm{disc}_K(n) = 2^{O(d)} \sqrt{\log(\kappa^{-1}(1/n))}.$$

Moreover, there is a randomized algorithm that on input of X and K, finds an according coloring $\beta \in \{\pm 1\}^X$ in polynomial time.

We will see in Section 2.6 that Theorem 2.4.2 yields significantly stronger bounds than [55] for several important kernels in machine learning, assuming the data is sufficiently low dimensional.

2.5 Our Proofs

For the proofs of Theorems 2.4.1 and 2.4.2 we will need the following results about covering numbers. Let $N(A, B)$ denote the number of copies of body B needed to cover A.

Lemma 2.5.1. *For any convex set $K \subseteq \mathbb{R}^d$ and $r > 0$, one has*

$$N(K, rB_2^d) \leq 2^d \frac{\mathrm{Vol}_d(K + \frac{r}{2}B_2^d)}{\mathrm{Vol}_d(rB_2^d)}.$$

Lemma 2.5.2. *For any symmetric convex body $P \subseteq \mathbb{R}^d$ and $r > 0$, one has*

$$N(P, rP) \leq \left(1 + \frac{2}{r}\right)^d.$$

We begin with the proof of Theorem 2.4.1.

Proof of Theorem 2.4.1. We may assume that $R\alpha \geq 1$, otherwise replace R by $\frac{1}{\alpha}$. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $X \subseteq \mathbb{R}^d$ satisfy conditions (i)-(iv) outlined in Theorem 2.4.1. As K is positive definite, there exists an RKHS $\mathcal{H}_K$ and a map $\phi : \mathbb{R}^d \to \mathcal{H}_K$ so that

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_K} \quad \forall x, y \in \mathbb{R}^d.$$

We apply the Gram-Schmidt walk from Theorem A.0.1 to the vectors $\phi(x)$ for $x \in X$, noting that by condition (ii), $\|\phi(x)\|_{\mathcal{H}_K} = 1$ for each $x \in X$. The algorithm yields a distribution $\mathcal{P}$ over $\{\pm 1\}^X$ so that the random variable $\Sigma := \sum_{x \in X} \beta(x)\phi(x)$, with $\beta \sim \mathcal{P}$, is $O(1)$-subgaussian. In particular, as we know that $\|\phi(y)\|_{\mathcal{H}_K} = 1$ for any $y \in Q$, the (mean zero) random variable

$$\Sigma_y := \big\langle \Sigma, \phi(y) \big\rangle = \sum_{x \in X} \beta(x) K(x, y) = \mathrm{disc}_K(X, \beta, y)$$

is $O(1)$-subgaussian. We will apply Dudley's integral inequality (see Theorem A.0.3) to the mean zero process $(\Sigma_y)_{y \in Q}$ with the pseudometric $D_K(x, y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_K}$, the kernel distance defined in Section 2.2. To see why D_K satisfies the condition of Theorem A.0.3, note that for any $y, q \in Q$,

$$\mathrm{Var}[\Sigma_y - \Sigma_q]^{1/2} = \mathbb{E}[\big\langle \Sigma, \phi(y) - \phi(q) \big\rangle^2]^{1/2} \lesssim \|\phi(y) - \phi(q)\|_{\mathcal{H}_K} = D_K(y, q).$$

Thus by Dudley's integral inequality,

$$\mathbb{E}\ \mathrm{disc}_{K|_Q}(X) = \mathbb{E} \sup_{y \in Q} |\Sigma_y| \lesssim \int_0^{\mathrm{diam}(D_K)} \sqrt{\log \mathcal{N}(Q, D_K, r)}\ \mathrm{d}r + \|\Sigma_{y_0}\|_{\psi_2} \tag{2.3}$$

for any fixed $y_0 \in Q$. Here $\|\Sigma_{y_0}\|_{\psi_2} = O(1)$, and so we may ignore this lower order term. To estimate $\mathcal{N}(Q, D_K, r)$, we use conditions (ii) and (iii) to note that

$$D_K(q,y) = \sqrt{2 - 2\kappa(\alpha\|q-y\|_2)} \le r \quad \Longleftrightarrow \quad \|q-y\|_2 \le \frac{1}{\alpha}\kappa^{-1}\left(\frac{2-r^2}{2}\right),$$

where we use that κ is strictly decreasing. From this calculation we conclude that

$$\mathcal{N}(Q, D_K, r) = \mathcal{N}\left(Q, \|\cdot\|_2, \frac{1}{\alpha}\kappa^{-1}\left(\frac{2-r^2}{2}\right)\right).$$

Taking $c := \frac{1}{\alpha}\kappa^{-1}\left(\frac{2-r^2}{2}\right)$ for a moment, we can bound the quantity on the right using Lemma 2.5.1:

$$\begin{aligned}
\mathcal{N}\left(Q, \|\cdot\|_2, c\right) &= N(Q,\ cB_2^d) \\
&\le N(RB_2^d, cB_2^d) \\
&\le 2^d \frac{\mathrm{Vol}_d((R + \frac{c}{2})B_2^d)}{\mathrm{Vol}_d(cB_2^d)} \\
&\le \left(\frac{4R}{c}\right)^d,
\end{aligned} \tag{2.4}$$

where we use Lemma 2.5.1 and $c \le R$ (if $c > R$ then the covering number is trivially 1). Using this bound in (2.3), we find

$$\begin{aligned}
\mathbb{E}\ \mathrm{disc}_{K|_Q}(X) &\lesssim \int_0^{\mathrm{diam}(D_K)} \sqrt{\log\left[\left(\frac{4R\alpha}{\kappa^{-1}\left(\frac{2-r^2}{2}\right)}\right)^d\right]}\ \mathrm{d}r \\
&\lesssim \sqrt{d}\int_0^{\mathrm{diam}(D_K)} \sqrt{\log\left[\frac{4R\alpha}{\kappa^{-1}\left(\frac{2-r^2}{2}\right)}\right]}\ \mathrm{d}r \\
&\lesssim \sqrt{d\log(2R\alpha)} + \sqrt{d}\int_0^{\mathrm{diam}(D_K)} \sqrt{\log\left[\frac{1}{\kappa^{-1}\left(\frac{2-r^2}{2}\right)}\right]}\ \mathrm{d}r \\
&\lesssim \sqrt{d\log(2R\alpha)} + O(\sqrt{d}).
\end{aligned} \tag{2.5}$$

To justify these calculations, note that in the event that the quantity inside the radical is negative, it means that

$$4R\alpha \leq \kappa^{-1}\Big(\frac{2-r^2}{2}\Big),$$

and hence by (2.4) we know that $\mathcal{N}(Q, D_K, r) = 2^{O(d)}$. As $\operatorname{diam}(D_K) \leq 2$, the domain of r values for which this occurs has length ≤ 2, so we can simply restrict the domain to this point while losing at most an additive $O(\sqrt{d})$. Note that as κ is assumed strictly decreasing and continuous, $\kappa^{-1}\big(\frac{2-r^2}{2}\big)$ is an increasing, continuous function of r. Thus in particular $\kappa^{-1}\big(\frac{2-r^2}{2}\big) \to 0$ as $r \to 0$, based on our assumption that $K(x,x) = 1$ for all $x \in \mathbb{R}^d$. Finally, by Lemma 2.4.1,

$$\mathbb{E}\ \operatorname{disc}_K(X) \leq \mathbb{E}\ \operatorname{disc}_{K|_Q}(X) + O(1) \lesssim \sqrt{d\log(2\alpha R)}. \qquad \square$$

Next we prove Theorem 2.4.2, which gives improved bounds independent of the geometry of the data set X.

Proof of Theorem 2.4.2. Fix a kernel K satisfying conditions (i)-(iv) with impact radius r_K, and let $X \subseteq \mathbb{R}^d$. Fix the associated query space Q as given by Definition 2.4.2. As we assume that K is positive definite and satisfies $K(x,x) = 1$ for all $x \in Q$, there exists a map $\phi : \mathbb{R}^d \to \mathcal{H}_K$, where $\mathcal{H}_K$ is a RKHS such that $K(x,y) = \langle \phi(x), \phi(y) \rangle$ for all $x, y \in \mathbb{R}^d$, and $\|\phi(x)\|_{\mathcal{H}_K} = 1$ for all $x \in \mathbb{R}^d$. We first find a maximal set of points $q_1, \ldots, q_m \in Q$ such that $\|q_i - q_j\|_2 \geq r_K$ for any $i, j \in [m]$. We then partition the space Q into disjoint cells

$$Q := R_1 \dot\cup \cdots \dot\cup R_m$$

so that

$$R_i \subseteq q_i + r_K B_2^d \tag{2.6}$$

for each $i \in [m]$. These cells also partition the set X into smaller sets $X_i := X \cap R_i$. We begin with the set X_1; because the conditions in Theorems 2.4.1 and 2.4.2 are the same, by identical arguments to the proof of Theorem 2.4.1 we obtain an $O(1)$-subgaussian random variable

$\Sigma^1 := \sum_{x\in X_1} \beta^{(1)}(x)\phi(x)$, where $\beta^{(1)} \sim \{\pm 1\}^{X_1}$ is drawn according to the Gram-Schmidt walk. Note that then for any $y \in Q$, as $\|\phi(y)\|_{\mathcal{H}_K} = 1$, the random variable

$$\Sigma^1(y) := \langle \phi(y), \Sigma^1\rangle = \Big\langle \phi(y), \sum_{x\in X_1} \beta^{(1)}(x)\cdot\phi(x)\Big\rangle$$

is also $O(1)$-subgaussian.

Repeating this argument, for each $j \in [m]$, we now have the collection of (mean zero) $O(1)$-subgaussian random variables $\{\Sigma^j(y) : \ y \in R_j\}$. We fix $j \in [m]$; then by our assumption in (2.6), our query space R_j has volume at most $\mathrm{Vol}_d(r_K B_2^d)$. By a calculation analogous to that in (2.4), we have that for all $r > 0$,

$$\mathcal{N}(R_j, D_K, r) \le \max\Big\{\Big(\frac{4\alpha r_K}{r}\Big)^d, 1\Big\}. \tag{2.7}$$

Here we note that because of our assumption that $K(x,y) = \kappa(\alpha\|x-y\|_2)$ and the definition of $r_K(n)$, we have for x, y with $\|x-y\|_2 = r_K(n)$,

$$K(x,y) = \kappa(\alpha r_K(n)) \le 1/n \implies \alpha r_K(n) \le \kappa^{-1}(1/n) \implies r_K(n) \le \frac{1}{\alpha}\kappa^{-1}(1/n). \tag{2.8}$$

Thus the quantity αr_K appearing in (2.7) can actually be bounded by a term independent of α. Note that by assumption (iii), κ^{-1} is a decreasing function. Thus by the same analysis in the proof of Theorem 2.4.2, we have

$$\mathrm{disc}_{K|_{R_j}}(X_1) \lesssim \int_0^{\mathrm{diam}(D_K|_{R_j})} \sqrt{\log \mathcal{N}(R_j, D_K, r)}\ \mathrm{d}r = O(\sqrt{d\log\kappa^{-1}(1/n)}).$$

Applying Dudley's concentration inequality (see Theorem A.0.4) with $u := C_0\sqrt{d}$ yields that

$$\mathbb{P}\Big\{\sup_{y\in R_j} |\Sigma^1(y)| \ge C\cdot C_0\sqrt{d}\Big\} \le 2e^{-dC_0^2}, \tag{2.9}$$

where C is the constant from Theorem A.0.4 and $C_0 > 0$.

Moreover, by standard packing arguments, this probability is in fact 0 for all but at most $2^{O(d)}$ choices of $j \in [m]$, as any R_j not adjacent to (or the same as) R_1 will have distance $\Omega(r_K)$ from any point $x \in X_1$, hence by the definition of the impact radius r_K and the bound

in (2.8), for any such R_j the above probability is 0. We apply the union bound over the $2^{O(d)}$ cells where (2.9) could fail and obtain that for C_0 large enough,

$$\begin{aligned}\mathbb{P}\Big\{\forall j \in [m] : \sup_{y \in R_j} |\Sigma^1(y)| \geq C \cdot C_0\sqrt{d \log r_K}\Big\} &= \mathbb{P}\Big\{\sup_{y \in Q} |\Sigma^1(y)| \geq C_0\sqrt{d \log r_K}\Big\} \\ &\leq 2^{O(d)} \cdot 2(\kappa^{-1}(1/n))^{-C_0^2 d} < 1.\end{aligned} \tag{2.10}$$

Thus with positive probability, $\mathrm{disc}_K(X_1) \leq C \cdot C_0\sqrt{d \log \kappa^{-1}(1/n)}$. We then fix such an outcome and repeat this construction independently for $R_2, ..., R_m$, at each step repeating the Gram-Schmidt algorithm until we get the outcome as in (2.10). After repeating this construction m times, we have a choice of signs $\beta = (\beta^{(1)}, ..., \beta^{(m)}) \in \{\pm 1\}^X$ so that by the triangle inequality

$$\mathrm{disc}_K(X) \leq \sum_{i \in [m]} \mathrm{disc}_K(X_i). \tag{2.11}$$

For our purposes this bound is too weak; however, because the function $K = \kappa(\|x - y\|_2)$ is symmetric in x and y, by the same packing argument we made above, we know that for each fixed point $y \in Q$, only $2^{O(d)}$ terms in the summand in (2.11) can contribute a discrepancy larger than $O(1)$. Thus we conclude that

$$\mathrm{disc}_K(X) \leq 2^{O(d)}\sqrt{d \log \kappa^{-1}(1/n)}.$$

As $X \subseteq \mathbb{R}^d$ was arbitrary, the bound follows. □

2.6 Applications to Kernels of Interest

In this section we highlight the applications of Theorems 2.4.1 and 2.4.2 to several kernels of interest in machine learning, beginning with two of the most commonly used kernels in machine learning: the Gaussian and Laplacian kernels. The Gaussian kernel is defined by

$$K_G : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \quad K_G(x, y) = \exp(-\alpha^2\|x - y\|_2^2),$$

and the Laplacian kernel is defined by

$$K_L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \quad K_L(x, y) = \exp(-\alpha\|x - y\|_2),$$

where $1/\alpha > 0$ is the bandwidth parameter of the kernels. Currently the best known bound for the kernel discrepancy of both the Gaussian and the Laplacian kernels in arbitrary dimension d is $O(\sqrt{d \log n})$, and because these bounds rely on taking $1/n$-nets of the query space, even if the given data set X satisfies nice properties such as boundedness, the $\sqrt{\log n}$ term remains. However, Theorem 2.4.1 allows us to give significant improvements on this $O(\sqrt{d \log n})$ bound given such a data set.

Corollary 2.6.1. *For the Gaussian kernel K_G with any bandwidth parameter $\alpha > 0$ and any data set $X \subseteq RB_2^d$ for any fixed constant $R > 0$, we have*

$$\operatorname{disc}_{K_G}(X) = O(\sqrt{d \log \log n}).$$

In particular, one can find (in randomized polynomial time) a coreset for X of size

$$O\left(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log \log \frac{1}{\varepsilon}}\right).$$

Proof. We first check the conditions of Theorem 2.4.1 for the Gaussian kernel. It is well-known that K_G is positive definite, and clearly $K_G(x, x) = e^0 = 1$ for any $x \in \mathbb{R}^d$. Taking $\kappa_G(z) := e^{-z^2}$ gives us $K_G(x, y) = \kappa_G(\alpha\|x - y\|_2)$, and κ is strictly decreasing on $\mathbb{R}^+$; finally,

$$\sqrt{-\log\left[\kappa^{-1}\left(\tfrac{2-r^2}{2}\right)\right]} \lesssim \sqrt{\log\left[\frac{1}{\log\left(\frac{2}{2-r^2}\right)}\right]}$$

is dominated by $O(1/r^2)$ for $r \in \mathbb{R}^+$, hence is integrable on its domain. The query space Q is given by

$$Q := \bigcup_{x \in X} B_2^d(x, r_{K_G}) \subseteq O(r_{K_G})B_2^d,$$

as we assume that $\operatorname{radius}(X) = O(1)$. Thus Theorem 2.4.1 yields

$$\operatorname{disc}_{K_G}(X) \lesssim O(\sqrt{d \log R\alpha}) = O(\sqrt{d \log \log n})$$

by the fact that $R = O(r_K)$ and the same observation as in the proof of Theorem 2.4.2 that

$$\alpha r_{K_G} \leq \kappa_G^{-1}(1/n) = \sqrt{\log n}.$$

To see the bound on the size of a minimal coreset, we will apply the "halving trick" shown in Section 2.2. The above argument shows a bound of $f(n) = O(\sqrt{d \log\log n})$ on the kernel discrepancy; iterating the discrepancy calculation t times and removing the points colored -1 at each step, we obtain the following bound after t iterations:

$$\|\mathrm{KDE}_X - \mathrm{KDE}_{X_t}\|_\infty \lesssim \sum_{s\in[t]} \frac{2^{s-1}}{n}\sqrt{d \log\log \tfrac{n}{2^{s-1}}}.$$

This sum is dominated by the t^{th} term, so we can repeat this calculation until the size of the remaining set of data points $m := n/2^{s-1}$ satisfies

$$m\sqrt{d \log\log m} = \varepsilon,$$

which occurs for $m = O\big(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log\log \frac{1}{\varepsilon}}\big)$, yielding the bound. □

This exact technique can always be used to provide coreset bounds, assuming that the term $f(n)$ grows significantly slowly with n, which it always will for our applications. Thus we will not repeat this calculation for the proofs of the remaining theorems. Note that an identical argument for the discrepancy yields the proof of Corollary 2.4.1. However, in general if the diameter of the set X depends on the number of data points, one has to be more careful to obtain bounds on the coreset complexity, as this dependence may change as we remove data points.

The following theorem follows from essentially identical arguments to those in the previous proof, noting that the Laplacian kernel can be written as $K_L(x, y) = \kappa_L(\alpha\|x - y\|_2)$, with $\kappa_L(z) = e^{-z}$.

Corollary 2.6.2. *For the Laplacian kernel K_L with any $\alpha > 0$ and any data set $X \subseteq RB_2^d$ for any fixed constant $R > 0$ we have*

$$\mathrm{disc}_{K_L}(X) = O(\sqrt{d \log\log n}).$$

In particular, one can find (in randomized polynomial time) a coreset for X of size

$$O\left(\frac{\sqrt{d}}{\varepsilon}\sqrt{\log\log \frac{1}{\varepsilon}}\right).$$

Thus in the interest of applications, we can invoke useful properties of the data set X in order to obtain better bounds, something that was not possible due to the discretization technique used in [55]. Karnin and Liberty [39] provide bounds assuming that the entire query space lies in a bounded region (for a particular constant depending on the kernel), but our assumption is much weaker: we are only interested in the geometry of the data set, not the query space.

In the case when our data is sufficiently low dimensional that a $2^{O(d)}$ factor is not prohibitively large, i.e. for d constant, Theorem 2.4.2 yields the following improvement on the current best known bound for the Laplacian kernel [55].

Corollary 2.6.3. *For the Laplacian kernel K_L with any $\alpha > 0$, we have*

$$\operatorname{disc}_{K_L}(n) = 2^{O(d)}\sqrt{\log\log n}.$$

In particular, given any data set $X \subseteq \mathbb{R}^d$, we can find (in randomized polynomial time) a coreset of size

$$\frac{2^{O(d)}}{\varepsilon}\sqrt{\log\log\frac{1}{\varepsilon}}.$$

Proof. The verification that the Laplacian kernel satisfies conditions (i)-(iv) of Theorem 2.4.2 is essentially identical to showing that the Gaussian kernel satisfies these properties, so we refer back to the proof of Corollary 2.6.1. As mentioned previously, the Laplacian kernel can be written as $K_L(x, y) = \kappa_L(\alpha\|x - y\|_2)$, where $\kappa_L(z) = e^{-z}$, and applying Theorem 2.4.2 then yields the bound above. □

We note that the same bound follows for the Gaussian kernel as well, though a recent paper of Tai [72] gives a bound of $O(1)$ on the discrepancy of the Gaussian kernel for d constant. The technique Tai used for the Gaussian does not appear to generalize to other kernels, whereas our technique applies to a broader class of kernels that also includes the Laplacian kernel.

We can also obtain improved bounds for the exponential and Hellinger kernels by applying Theorem 2.4.1.

Theorem 2.6.1. *The exponential kernel*

$$K_e : S^d \times S^d \to \mathbb{R}, \quad K_e(x, y) = \exp(-\alpha(1 - \langle x, y \rangle)),$$

has discrepancy satisfying

$$\mathrm{disc}_{K_e}(n) = O(\sqrt{d \log(2 \max\{\alpha, 1\})}).$$

Further, the exponential kernel has coreset complexity

$$O(\sqrt{d \log(2 \max\{\alpha, 1\})}/\varepsilon),$$

and such a coreset can be constructed in randomized polynomial time.

Proof of Theorem 2.6.1. First we note that $x, y \in S^d$, hence in particular $\|x\|_2 = \|y\|_2 = 1$, so we can re-write the exponential kernel as

$$K_e(x, y) = \exp\Big(-\frac{\alpha}{2} \|x - y\|_2^2 \Big) = K_G|_{S^d},$$

from which we see that K_e satisfies the conditions of Theorem 2.4.1. As the query space is $Q = S^d \subset B_2^d$, we can apply Theorem 2.4.1 with $R = 1$. □

This result gives a doubly exponential improvement in terms of dependence on the bandwidth parameter from Karnin and Liberty's result [39].

Theorem 2.6.2. *The Hellinger kernel*

$$K_H : \Delta^d \times \Delta^d \to \mathbb{R}, \quad K_H(x, y) = \exp\left(-\alpha \sum_{i=1}^{d} \left(\sqrt{x} - \sqrt{y} \right) \right),$$

has discrepancy satisfying

$$\mathrm{disc}_{K_H}(n) = O(\sqrt{d \log(2 \max\{\alpha, 1\})})$$

Further, the Hellinger kernel has coreset complexity

$$O(\sqrt{d \log(2 \max\{\alpha, 1\})}/\varepsilon),$$

and such a coreset can be constructed in randomized polynomial time.

Theorem 2.6.2 will follow from showing that the Hellinger kernel has discrepancy bounded by that of the exponential kernel.

Lemma 2.6.1. *The Hellinger kernel K_H has discrepancy at most that of the exponential kernel K_e, i.e.*

$$\mathrm{disc}_{K_H}(n) \leq \mathrm{disc}_{K_e}(n).$$

Proof of Lemma 2.6.1. Given a data set $X \subseteq \Delta_d$ of size n, we make the transformation

$$f : \mathbb{R}^d \to \mathbb{R}^d, \quad (x_1, \ldots, x_d) \to (\sqrt{x_1}, \ldots, \sqrt{x_d}).$$

Here we make two observations: first, f maps Δ^d into S^d, and second, for $x, y \in S^d$, and

$$K_H(x, y) = \exp\Big(- \alpha \sum_{i \in [d]} (\sqrt{x_i} - \sqrt{y_i})^2 \Big) = \exp(-\alpha \|f(x) - f(y)\|_2^2) = K_e(f(x), f(y)).$$

By the first observation, we can apply our bound on the exponential kernel discrepancy of the set $\{f(x) : \ x \in X\} \subseteq S^d$ to find signs $\beta \in \{\pm 1\}^X$ so that

$$\begin{aligned} \sup_{y \in S^d} \Big| \sum_{x \in X} \beta(x) K_e(f(x), y) \Big| &= \sup_{z \in \Delta^d} \Big| \sum_{x \in X} \beta(x) K_e(f(x), f(z)) \Big| \\ &= \sup_{z \in \Delta^d} \Big| \sum_{x \in X} \beta(x) K_H(x, z) \Big| \\ &= \mathrm{disc}_{K_H}(X) \end{aligned}$$

satisfies the bound of Theorem 2.6.1. Thus we conclude that

$$\mathrm{disc}_{K_H}(X) \leq \mathrm{disc}_{K_e}(X),$$

and as X was an arbitrary data set, the lemma follows.

□

Finally, we prove the following bound for a specific kernel that does not directly satisfy the conditions of Theorems 2.4.1 and 2.4.2: the Jensen-Shannon (JS) Kernel.

Theorem 2.6.3. *The Jensen-Shannon (JS) kernel*

$$K_{JS} : \Delta_d \times \Delta_d \to \mathbb{R}, \quad K_{JS}(x,y) = \exp\left(-\alpha\big(H\big(\tfrac{x+y}{2}\big) - \tfrac{H(x)+H(y)}{2}\big)\right),$$

where $H(x) = -\sum_{i\in[d]} x_i \log x_i$ *is the Shannon entropy function, has discrepancy satisfying*

$$\mathrm{disc}_{K_{JS}}(n) = O(\sqrt{d\log(2\max\{\alpha,1\})}).$$

Further, the JS Kernel has coreset complexity

$$O(\sqrt{d\log(2\max\{\alpha,1\})}/\varepsilon),$$

and such a coreset can be constructed in randomized polynomial time.

To prove this result, we will need the following lemmas bounding the one-dimensional entropy function, which we will denote by $h(x) := x\ln(\frac{1}{x}) = -x\ln(x)$.

Lemma 2.6.2. *For* $a, \delta \geq 0$ *one has* $\big|2h(a+\delta) - (h(a) + h(a+2\delta))\big| \leq 3\delta$.

Proof. By continuity of h it suffices to prove the inequality for the case of $a > 0$. We have

$$\begin{aligned}
& \big|2h(a+\delta) - h(a) - h(a+2\delta)\big| \\
= \;& \big|2(a+\delta)\ln(a+\delta) - a\ln(a) - (a+2\delta)\ln(a+2\delta)\big| \\
= \;& \left|2(a+\delta)\cdot\left(\ln(a) + \ln\left(1+\frac{\delta}{a}\right)\right) - a\ln(a) - (a+2\delta)\cdot\left(\ln(a) + \ln\left(1+\frac{2\delta}{a}\right)\right)\right| \\
= \;& \left|2(a+\delta)\ln\left(1+\frac{\delta}{a}\right) - (a+2\delta)\ln\left(1+\frac{2\delta}{a}\right)\right| \\
\overset{\text{triangle ineq.}}{\leq} \;& a\cdot\underbrace{\left|2\ln\left(1+\frac{\delta}{a}\right) - \ln\left(1+\frac{2\delta}{a}\right)\right|}_{\leq \delta/a} + 2\delta\cdot\underbrace{\left|\ln\left(1+\frac{\delta}{a}\right) - \ln\left(1+\frac{2\delta}{a}\right)\right|}_{\leq 1} \leq 3\delta
\end{aligned}$$

Here one can verify that $|2\ln(1+z) - \ln(1+2z)| \leq z$ for all $z \geq 0$ which we use to bound the left term. To bound the right term we use that $|\ln(1+z) - \ln(1+2z)| \leq 1$ for all $z \geq 0$. □

Lemma 2.6.3. *For* $a, b \geq 0$ *one has* $|h(\frac{a+b}{2}) - \frac{h(a)+h(b)}{2}| \leq \frac{3}{4}|a-b|$.

Proof. Suppose $b \geq a$. Applying Lemma 2.6.2 with a and $\delta := \frac{b-a}{2}$ gives

$$\left|2h\Big(\frac{a+b}{2}\Big) - (h(a)+h(b))\right| \leq 3 \cdot \frac{b-a}{2} = \tfrac{3}{2}|a-b|.$$

The claim follows after dividing both sides by 2. □

Proof of Theorem 2.6.3. As the JS kernel is positive definite, there exists a map $\phi : \Delta^d \to \mathcal{H}_{K_{JS}}$, where $\mathcal{H}_{K_{JS}}$ is a RKHS for K_{JS}, i.e. for any choice of $x, y \in \Delta_d$, $K_{JS}(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_{K_{JS}}}$. Then as $\|\phi(x)\|_{\mathcal{H}_{K_{JS}}} = e^0 = 1$ for any choice of $x \in \Delta^d$, we can apply the Gram-Schmidt walk to the collection of vectors $\{\phi(x)\}_{x \in \Delta_d}$ exactly as in the proofs of Theorems 2.4.1 and 2.4.2 to obtain a distribution $\mathcal{D}$ over $\{\pm 1\}^X$ and a corresponding family of $O(1)$-subgaussian random variables

$$\Sigma_y := \mathrm{disc}_{K_{JS}}(X, \beta, y), \quad y \in \Delta_d,\ \beta \sim \mathcal{D}.$$

Then applying Dudley's integral inequality (see Theorem A.0.3) to this collection of random variables with the pseudometric

$$D_{K_{JS}}(x,y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_{K_{JS}}} = \sqrt{2 - 2K_{JS}(x,y)},$$

we find that

$$\mathbb{E}\ \mathrm{disc}_{K_{JS}}(X) \lesssim \int_0^{\mathrm{diam}(D_{K_{JS}})} \sqrt{\log \mathcal{N}(\Delta_d, D_{K_{JS}}, r)}\ \mathrm{d}r. \tag{2.12}$$

We focus on bounding the term $\mathcal{N}(\Delta_d, D_{K_{JS}}, r)$, for which we will use Lemma 2.6.3. First, note that for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, breaking the d-dimensional entropy function H down component-wise as $H(x) = \sum_{i \in [d]} h(x_i)$:

$$D_{K_{JS}}(x,y) \leq r \quad \Longleftrightarrow \quad \sum_{i \in [d]} \left(h\left(\frac{x_i + y_i}{2}\right) - \frac{h(x_i) + h(y_i)}{2} \right) \leq \frac{1}{\alpha} \log\left(\tfrac{2}{2-r^2}\right). \tag{2.13}$$

By Lemma 2.6.3, we know that for each $i \in [d]$,

$$h\left(\frac{x_i + y_i}{2}\right) - \frac{h(x_i) + h(y_i)}{2} \leq |x_i - y_i|,$$

hence

$$\sum_{i\in[d]}\left(h\left(\frac{x_i+y_i}{2}\right)-\frac{h(x_i)+h(y_i)}{2}\right)\leq\sum_{i\in[d]}|x_i-y_i|=\|x-y\|_1.$$

In particular, by (2.13), $\|x-y\|_1\leq\frac{1}{\alpha}\log\left(\frac{2}{2-r^2}\right)=:c$ implies that $D_{K_{JS}}(x,y)\leq r$, so we have the bound

$$\mathcal{N}(\Delta_d,D_{K_{JS}},r)\leq\mathcal{N}(\Delta_d,\|\cdot\|_1,c)\leq N(B_1^d,cB_1^d)\leq\left(1+\frac{2}{c}\right)^d\leq\left(\frac{3}{c}\right)^d$$

for $c\leq 1$ using Lemma 2.5.2 (and if $c>1$, the covering number is 1 anyway).

Returning to (2.12), we conclude

$$\mathbb{E}\ \mathrm{disc}_{K_{JS}}(X)\lesssim\int_0^{\mathrm{diam}(D_K)}\sqrt{\log\left[\left(\frac{3\alpha}{\log\frac{2}{2-r^2}}\right)^d\right]}\,\mathrm{d}r\lesssim\sqrt{d\log(2\max\{1,\alpha\})},$$

by the same argument as in the proofs of Theorems 2.4.1 and 2.4.2. □

These results greatly improves on the current best bound of $O(\sqrt{d\log n})$ for the discrepancy of the JS and Hellinger kernels, in particular by dropping all dependence on the size n of the data set. We note that as these kernels are not shift or rotation invariant, they are not known to satisfy any lower bounds other than $O(1/\varepsilon)$, which holds for all kernels.

Chapter 3

QUANTIZATION OF NEURAL NETWORKS

In this chapter we prove a new application of the discrepancy method to the problem of quantizing neural networks. We introduce the problem in Section 3.1, as well as providing some background on previous results. In Section 3.2 we introduce the rigorous mathematical framework for our setting and introduce the discrepancy method approach to this problem. Finally, in Section 3.3 we state and prove our main theoretical result

3.1 Motivation and History of the Problem

Quantization is a technique developed to compress large deep learning models, in particular for our purposes large language models (LLMs) for faster training and/or inference. In order to give a basic introduction to quantization, we will introduce the following notations related to machine learning models. As our approach focuses on post-training compression approaches, we will assume that we are given a neural network $\mathcal{N}_w$ with pre-trained weight vector $w \in \mathbb{R}^n$, trained with loss function $f(w; s)$, where $s \in \mathbb{R}^d$ is an input sample from some given data distribution $\mathcal{D}_{data}$. Post-Training Quantization (PTQ) consists of two steps: determining a good low-bit complexity representation for the weights w, which we will call the *quantization grid*, and rounding the original weights to values in the quantization grid. The problem of constructing good low-bit representations has been explored extensively (see [63, 75, 26]); however, we work in the setting of rounding optimally to any given quantization grid, which is under-explored. The primary rounding methods in the literature to date are Round-to-Nearest (RTN), in which one rounds each weight to the nearest point in the quantization grid, and which is typically used as a simple baseline, and GPTQ (see [33, 29, 28]), a data dependent algorithm.

We now define the mathematical problem a bit more formally. Our goal is to perturb the original weights $w \in \mathbb{R}^n$ to values in a given quantization grid without increasing the loss $f(w; s)$ too much, over any potential sample $s \sim \mathcal{D}_{data}$. In order to avoid changing the original model weights too much, we impose the constraint that each weight can only be rounded up or down. Geometrically, we can visualize the set of allowed quantization points as the vertices of an n-dimensional hypercube H centered at w. Supposing we have determined perturbed weights $\hat{w} \in \mathbb{R}^n$, we can define the resulting change in the loss function for a given sample s as $\Delta f = f(\hat{w}; s) - f(w; s)$ and approximate Δf with a first order Taylor expansion: $\delta f \approx \langle \nabla_w f(w; s), \hat{w} - w \rangle$. While it is the case that the average gradients of a pre-trained model are close to zero (see prior work based on this assumption [52, 33]), per-sample gradients can be large, and experimental evidence shows that the first order term is a good approximation to δf (see our paper [6] for more details and examples). Following this logic, our goal is to to choose $\hat{w}$ so that $\langle \nabla_w f(w; s), \hat{w} - w \rangle \approx 0$ for s sampled from the data distribution $\mathcal{D}_{data}$.

A simple linear-algebraic argument shows that if we are given m independent samples $s_1, ..., s_m \sim \mathcal{D}_{data}$, for $m \leq n$, we can choose $\hat{w}$ so that $\langle \nabla_w f(w; s), \hat{w} - w \rangle = 0$. This equation determines an affine subspace V of dimension $n - m$, and the intersection $V \cap H$ is a convex polytope K, each vertex of which contains at least $n - m$ fully rounded parameters. In terms of our application of interest, the number of parameters n will always be much larger than the number of samples m, hence choosing an arbitrary vertex for any given sample gives an almost fully rounded solution. This solution is guaranteed to be optimal for the given samples, but we still have no guarantee that it will generalize to unseen samples from the data distribution $\mathcal{D}_{data}$.

We note that empirically, the distribution of gradients $g = \nabla_w f(w; s)$ for $s \sim \mathcal{D}_{data}$ is approximately low-rank (that is, the covariance matrix of gradients $\Sigma = \mathbb{E}_{s \sim \mathcal{D}_{data}}[gg^T]$, where $g = \nabla_w f(w; s)$, is approximately low-rank). Specifically, we will prove the following: if the eigenvalues of the covariance matrix of gradients Σ decay polynomially fast, then given $m = \text{poly}(\log n/\varepsilon)$ samples $s_1, ..., s_m \sim \mathcal{D}_{data}$, there is a randomized algorithm to find

quantized weights $\hat{w}$ with $n - m$ weights rounded such that $\mathbb{E}_{s\sim\mathcal{D}_{data}}[|\Delta f|] \leq \varepsilon$.

The theoretical guarantees that we will demonstrate in section 3.2 are based on tools and techniques from discrepancy theory, in particular the algorithmic version of Spencer's Theorem due to Lovett and Meka (see Theorem 1.1.3). The only other work to connect discrepancy theory to quantization is [47], though their method is quite different from ours and based on a discrepancy-motivated approach for quantizing the output of a single neuron. On the other hand, our approach applies to entire neural networks.

The practical algorithm, called *DiscQuant*, that is applied in the experimental sections of our paper is based on a slightly different algorithm, informed by our theoretical results. Experimental results show that DiscQuant provides significant improvements over GPTQ and RTN over multiple benchmarks on real-world models. See [6] for more details on the applied side of the work.

3.2 Theoretical Results

In this section we formalize the discrepancy-based theory of quantization and prove our main theoretical result, stated informally in Section 3.1. Suppose we are given $f(w; s)$, the loss function of a pre-trained neural network $\mathcal{N}_w$ with weights $w \in \mathbb{R}^n$ on an input sample s, the data distribution $\mathcal{D}_{data}$, and a scalar quantization grid $\mathcal{Q} = Q_1 \times \cdots \times Q_n$, where $Q_j \subset \mathbb{R}$ is a finite set of quantization points that can be used to quantize the j^{th} parameter. Here we note that in general the quantization grid can depend on w, see for example the Block Scaling method [28], but to simplify notation we ignore this dependence. By scalar quantization, we mean that each parameter can be independently rounded to a finite set of available values, as opposed to vector quantization, in which a group of d variables are rounded together to one of a finite set of quantization points in $\mathbb{R}^d$ (see for example [75, 26, 78]).

To re-state our goal, we would like to find a rounding $\hat{w} \in \mathcal{Q}$ of the original weights w satisfying $f(\hat{w}; s) \approx f(w; s)$ for $s \sim \mathcal{D}_{data}$. We restrict the rounding of a parameter $\hat{w}_j$ to either

$$w_j^{up} := \min\{z \geq w_j : \ z \in Q_j\} \quad \text{or} \quad w_j^{down} := \max\{z \leq w_j : \ z \in Q_j\};$$

in the case that $w_j < \min Q_j$ or $w_j > \max Q_j$, we set $w_j^{up} = w_j^{down}$ to either $\min Q_j$ or $\max Q_j$, respectively. We consider the Taylor expansion

$$\Delta f = f(\hat{w}; s) - f(w; s) = \langle \nabla_w f(w; s), \hat{w} - w \rangle + (\hat{w} - w)^T \nabla_w^2 f(w; s)(\hat{w} - w) + \cdots$$

As long as the quantization grid $\mathcal{Q}$ is sufficiently fine, and given our assumption that $\hat{w}_j \in \{w_j^{up}, w_j^{down}\}$, $\|\hat{w} - w\|$ is small enough that we can restrict consideration to the first and second order terms. In our work we will use that the linear term is dominant, an assumption that was justified empirically. Assuming that we are given $s_1, ..., s_m \sim \mathcal{D}_{data}$ sampled independently, with $m \ll n$, we can identify two questions of interest.

Question 3.2.1 (Bounding Empirical Error)**.** *Can one find $\hat{w} \in \mathcal{Q}$ such that (i) $\hat{w}_j \in \{w_j^{down}, w_j^{up}\}$ for each $j \in [n]$ and (ii) $\langle \nabla_w f(w; s_i), \hat{w} - w \rangle \approx 0$ for each $i \in [m]$?*

Question 3.2.2 (Bounding Generalization Error)**.** *Given a $\hat{w}$ satisfying the conditions of Question 3.2.1, will it generalize to the true data distribution, i.e. will*

$$\langle \nabla_w f(w; s), \hat{w} - w \rangle \approx 0 \quad \textit{for } s \sim \mathcal{D}_{data}?$$

If so, how many samples are needed for such a guarantee?

We first address Question 3.2.1; to begin, we assume the quantization grid is uniform (i.e. $w_i^{up} - w_i^{down} = \delta$ for all $i \in [n]$ and some $\delta > 0$), an assumption we will later explain how to remove. Define a new vector of parameters $x \in [0, 1]^n$ such that $w^x = w^{down} + \delta x$, so that w_i^x is an interpolation between w_i^{down} and w_i^{up}, with $w_i = w_i^{down}$ for $x_i = 0$ and $w_i = w_i^{up}$ for $x_i = 1$. Choosing $y \in [0, 1]^n$ so that $w^y = w$, then we can re-write the first order term of interest in terms of x as

$$\langle \nabla_w f(w; s_i), w^x - w \rangle = \langle \nabla_w f(w; s_i), w^x - w^y \rangle = \delta \langle \nabla_w f(w; s_i), x - y \rangle \quad \forall i \in [m]. \tag{3.1}$$

Defining $M \in \mathbb{R}^{m \times n}$ with i^{th} row given by $\nabla_w f(w; s_i)$, we can succinctly write all linear constraints in equation 3.1 as $M(x - y)$. In this language, our goal is to find a fully integral $\hat{x} \in \{0, 1\}^n$ such that $M(\hat{x} - y) = 0$. These linear constraints define an affine subspace

$V = \{x \in \mathbb{R}^n : \ Mx = My\}$ of dimension at least $n - m$; then $K := [0,1]^n \cap V$ is a non-empty (as $y \in K$) convex polytope, hence it is a well known result (see for example [31]) that any vertex of K contains at least $n - m$ integral coordinates. As we assume that there are far more parameters than samples (i.e. $n \gg m$), any vertex of K is almost fully integral and exactly satisfies all of the constraints in (3.1). In general one could further apply techniques from discrepancy theory to obtain a fully integral solution, however for practical applications of interest to our work, a vertex of K combined with RTN for the remaining fractional parameters is more than sufficient.

To conclude our discussion of Question 3.2.1, we show how to reduce the case of a non-uniform quantization grid to the setting described above by defining a slightly more complicated matrix. In this setting, we introduce a new parameter vector $z \in [0,1]^n$ and define

$$w^z = w^{down} \ \odot \ (1 - z) + w^{up} \ \odot \ z,$$

where $\odot$ denotes the component-wise product. Thus each component w_i^z interpolates between w_i^{down} and w_i^{up}, as in the uniform setting. Defining $y \in [0,1]^n$ so that $w^y = w$, then in this setting we can rewrite the constraints from (3.1) as

$$\begin{aligned}\langle \nabla_w f(w; s_i), w^z - w \rangle &= \langle \nabla_w f(w; s_i), w^z - w^y \rangle \\ &= \langle \nabla_w f(w; s_i), (w^{up} - w^{down}) \odot (z - y) \rangle \\ &= \langle \nabla_w f(w; s_i) \odot (w^{up} - w^{down}), z - y \rangle.\end{aligned}$$

Then similar to the uniform case, we can define an $m \times n$ matrix M with i^{th} row given by $\nabla_w f(w; s_i) \odot (w^{up} - w^{down})$, and the linear constraints can be written as $M(z - y)$.

We now turn to Question 3.2.2. Again, for simplicity we will work in the setting of uniform quantization grids, noting that by the above argument these results apply to non-uniform grids as well (assuming that the necessary conditions hold on the associated gradient matrix M). Supposing that we have chosen $\hat{w}$ as in Question 3.2.1 so that $\hat{w} - w$ is approximately[1]

[1] $\hat{w} - w$ is approximately orthogonal because there are at most m coordinates that are still to be rounded.

orthogonal to the sample gradients $\nabla_w f(w; s_i)$ for $i \in [m]$, a priori there is no reason to expect that $\hat{w} - w$ will also be orthogonal to unseen gradients $\nabla_w f(w; s)$ for samples $s \sim \mathcal{D}_{data}$. In order for such generalization to occur, we will need to add the additional (empirically justified) assumption that the covariance matrix of the distribution of sample gradients,

$$\Sigma = \mathbb{E}_{s \sim \mathcal{D}_{data}}[gg^T], \quad \text{for} \quad g = \nabla_w f(w; s),$$

has eigenvalues decaying quite quickly. Our model assumption will be that $\lambda_k \leq \lambda_1 / k^\alpha$ for some $\alpha > 1$; note that we may assume $\alpha > 1$ as

$$\mathbb{E}_{s \sim \mathcal{D}_{data}}[\|g\|^2] = \mathbb{E}_{s \sim \mathcal{D}_{data}}[\mathrm{Tr}(gg^T)] = \mathrm{Tr}(\mathbb{E}_{s \sim \mathcal{D}_{data}}[gg^T]) = \mathrm{Tr}(\Sigma) = \sum_{i=1}^{n} \lambda_i.$$

It is well-known that gradients of a pre-trained model have constant norm on most samples, hence $\sum_{i=1}^{n} \lambda_i = O(1)$, and the decay coefficient α must be at least 1.

With this additional spectral assumption on the covariance matrix, it is reasonable to expect generalization, as we will explain in Section 3.3; however, in general it is quite difficult to find such a generalizing solution, as any deterministic algorithm choosing a vertex of K is unlikely to generalize. In the next section we will give theoretical guarantees on a randomized rounding algorithm based on techniques from discrepancy theory. The final assumption that we will need (again, empirically verified) is a certain notion of well-behavedness of the gradient distribution.

Definition 3.2.1 (β-Reasonable Distributions [54])**.** *For a parameter $\beta \geq 1$, a random vector $X \in \mathbb{R}^n$ is* β-reasonable *if*

$$\mathbb{E}[\langle X, \theta \rangle^4] \leq \beta \cdot \mathbb{E}[\langle X, \theta \rangle^2]^2.$$

For a few simple examples, note that $x \sim \{-1, 1\}^n$ and any Gaussian $X \sim N(0, \Sigma)$ are $O(1)$-reasonable.

3.3 Proof of Main Theoretical Result

In this section we will prove our main theoretical result, which gives a constructive argument for finding a generalizable solution under appropriate assumptions on the data distribution.

Theorem 3.3.1 (Backurs, Chee, H., Li, Kulkarni, Rotvhoss, and Sivakanth [6])**.** *Let $\alpha > 1$ and $\beta \geq 1$ be constants and let $1 \leq m \leq n/16$. Let $\mathcal{D}$ be a β-reasonable distribution with unknown covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ whose Eigenvalues satisfy $\lambda_k \leq \lambda_1/k^\alpha$ for $k \in [n]$. Then there is a randomized polynomial time algorithm that given any $y \in [0,1]^n$ and m independent samples $g_1, ..., g_m \sim \mathcal{D}$, produces an $x \in [0,1]^n$ with probability at least 0.99 such that*

(i) $|\text{frac}(x)| \leq 16m$*, and*

(ii) $\langle \Sigma, (x-y)(x-y)^T \rangle \lesssim_\alpha \log(n/m) \cdot F_\alpha(m,n)$*, where*

$$F_\alpha(m,n) := \begin{cases} m^{1-\alpha} & \text{if } 1 < \alpha < 3/2 \\ \frac{\log(n)}{\sqrt{m}} & \text{if } \alpha = 3/2 \\ \frac{1}{\sqrt{m}} & \text{if } \alpha > 3/2. \end{cases}$$

Stated more simply, we obtain (with high probability) a solution $x \in [0,1]^n$ such that all but $O(m)$ parameters in x are fully rounded and

$$\mathbb{E}_{g \sim \mathcal{D}}[\langle g, x-y \rangle^2] = (x-y)^T \Sigma (x-y) \lesssim_{\alpha,\beta} \lambda_1 m^{-\min\{1/2, \alpha-1\}} (\log n)^2. \tag{3.2}$$

Note that this theorem proves the statement given in Section 3.1, as we take the distribution $\mathcal{D}$ to be that of the gradients, and note that the guarantee on the expectation in (3.2) exactly bounds the generalization error introduced in Question 3.2.2. The theorem will be proved via an analysis of a rounding algorithm that is based on the Lovett-Meka algorithm (see Theorem 1.1.3). To begin, we briefly describe the Lovett-Meka algorithm: given a point $y \in [0,1]^n$, vectors $v_1, ..., v_m \in B_2^n$, and parameters $c_j \geq 0$ so that $\sum_{j=1}^m \exp(-c_j^2/16) \leq n/16$,

then in randomized polynomial time we find a point $x \in [0,1]^n$ with at least half integral coordinates and so that $|\langle v_j, x-y\rangle| \leq c_j$ for each $j \in [m]$. The algorithm works as follows: x is the outcome of a random walk that starts at y and iteratively takes Gaussian steps scaled by a parameter $\delta > 0$. After hitting a constraint (which is of the form $x_i = 0, x_i = 1$, or $|\langle v_j, x-y\rangle| = C_j$), all future Gaussian updates are taken orthogonal to the corresponding normal vectors. Geometrically, this can be visualized as a random Gaussian walk inside of a polytope which–upon hitting a constraint of the polytope–continues inside of the given face. Lovett and Meka's result shows that if the updates are performed $O(1/\delta^2)$ times, the walk will cover enough distance so that on average $\Omega(n)$ coordinates will be integer.

We will apply this strategy with the parameters $c_j = 0$, as we are simply interested in finding a vertex. However, we will need a few additional properties of the outcome of the Lovett-Meka algorithm that are not explicitly stated in the literature.

Theorem 3.3.2 (Derived from [44])**.** *Let $g_1, ..., g_m \in \mathbb{R}^n$ be any vectors with $m \leq n/16$ and let $y \in [0,1]^n$. Then in polynomial time one can compute a sample $x \sim \mathcal{D} := \mathcal{D}_{LM}(g_1, ..., g_m, y)$ so that*

(i) One has $x \in [0,1]^n$, and with probability at least $1/10$

$$|\{j \in [n] : \ x_j \in \{0,1\}\}| \geq n/2.$$

(ii) For any vector $\theta \in \mathbb{R}^n$ one has

$$\mathbb{E}_{x\sim\mathcal{D}}[\langle \theta, x-y\rangle^2] \leq O(\|\theta\|_2^2).$$

(iii) For any symmetric matrix $M \in \mathbb{R}^{n\times n}$ one has $\mathbb{E}[\langle M, (x-y)(x-y)^T\rangle] \leq O(\|M\|_{\mathcal{S}(1)})$.

Here $\|\cdot\|_{\mathcal{S}(1)}$ denotes the Schatten 1-norm (equivalently the nuclear norm or the trace norm) of M, which is defined as the sum of its singular values.

Proof. Property (i) is explicit in [44]. To prove (ii), note that the outcome of the random walk can be written in the form

$$x = y + \delta \sum_{t=1}^{O(1/\delta^2)} u_t \quad \text{where} \quad u_t \sim N(0, \Sigma_t).$$

Here $0 \preceq \Sigma_t \preceq I_n$, and we note that as each covariance matrix Σ_t may depend on the outcome of $u_1, ..., u_{t-1}$, $x - y$ is not Gaussian. However, it is a Martingale, and since at each step t one has $\mathbb{E}[\langle u_t, \theta \rangle] = 0$ and $\mathbb{E}[\langle u_t, \theta \rangle^2] \leq O(\|\theta\|_2^2)$, we still have that the variance satisfies

$$\mathbb{E}[\Big\langle \delta \sum_{t=1}^{O(1/\delta^2)} u_t, \theta \Big\rangle^2],$$

showing (ii). To see why (iii) holds, note that (ii) can also be stated as $\mathbb{E}_{x \sim \mathcal{D}}[(x-y)(x-y)^T] \preceq O(1) \cdot I_n$, hence

$$\begin{aligned} \mathbb{E}[\langle M, (x-y)(x-y)^T \rangle] &= \langle M, \mathbb{E}[(x-y)(x-y)^T] \rangle \\ &\leq \|M\|_{\mathcal{S}(1)} \cdot \|\mathbb{E}[(x-y)(x-y)^T]\|_{op} \\ &\leq O(\|M\|_{\mathcal{S}(1)}). \end{aligned}$$

□

We now state the rounding algorithm that we will use:

Lovett-Meka Rounding Algorithm
Input: Weight vector $y \in [0,1]^n$ and parameter m
Output: Rounded vector x
(1) Sample $g_1, \ldots, g_m \sim \mathcal{D}$. Initialize $x^{(0)} := y$
(2) FOR $t = 1$ TO ∞ DO
(3) IF $\lvert\text{frac}(x^{(t-1)})\rvert \leq 16m$ then return $x^{(t-1)}$
(4) Set $x^{(t)} := \mathcal{D}_{LM}(g_1, \ldots, g_m, x^{(t-1)})$

The proof of Theorem 3.3.1, which rests on an analysis of Algorithm 3.3, requires understanding the Schatten 1-norm distance between the *covariance estimator* $\frac{1}{m}\sum_{j=1}^{m} g_j g_j^T$ and the actual covariance matrix. To this end we formulate the following matrix concentration bound, the proof of which we defer to the end of this section.

Lemma 3.3.1. *Let $\alpha > 1$, $\beta \geq 1$, and let $\mathcal{D}$ be a β-reasonable distribution with covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ with Eigenvalues satisfying $\lambda_k \leq 1/k^\alpha$ for all $k \in [n]$. Let $g_1, ..., g_m \sim \mathcal{D}$ be independent samples and let $X^{(\ell)} := \frac{1}{m} g_\ell g_\ell^T$ and $X := \sum_{\ell=1}^{m} X^{(\ell)}$. Then*

$$\mathbb{E}[\|X - \Sigma\|_{\mathcal{S}(1)}] \lesssim_{\alpha,\beta} F_\alpha(m, n),$$

where $F_\alpha(m, n)$ is as defined in Theorem 3.3.1.

Equipped with these tools, we prove our main theoretical result.

Proof of Theorem 3.3.1. Let $x^{(t^*)}$ be the vector returned by Algorithm 3.3. For notational convenience, we define $x^{(t)} := x^{(t^*)}$ for all $t > t^*$. We will call iteration t *good* if either $|\mathrm{frac}(x^{(t-1)})| \leq 16m$ or if $|\mathrm{frac}(x^{(t)})| \leq \frac{1}{2}|\mathrm{frac}(x^{(t-1)})|$. If iteration t is not good, we repeat the iteration until it is good, which we know occurs with probability at least $1/10$, independent of previous outcomes. Thus by standard Chernov bounds, with probability at least 0.99 there are at least $\log(n/m)$ good iterations in the first $T := C' \log(n/m)$ iterations, for $C' > 0$ a sufficiently large constant. Having acheived $\log(n/m)$ good iterations, $|\mathrm{frac}(x^{(T)})| \leq 16m$, and the discrepancy is

$$\mathbb{E}[\langle \Sigma, (x^{(T)} - y)(x^{(T)} - y)^T \rangle] \leq \sum_{t=1}^{T} \mathbb{E}[\langle \Sigma, (x^{(t)} - x^{(t-1)})(x^{(t)} - x^{(t-1)})^T \rangle] \lesssim_{\alpha,\beta} T \cdot F_\alpha(m, n),$$

proving the theorem. □

As promised, we conclude the section with the proof of Lemma 3.3.1.

Proof of Lemma 3.3.1. We will first prove the case $1 < \alpha < \frac{3}{2}$ and then discuss how a modification of the proof shows the remaining two cases. Because the claim is invariant

under changing basis, we can assume Σ is diagonal with Eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, i.e. $\Sigma_{ii} = \lambda_i$ for $i \in [n]$.

The first step of the proof is to bound the variance terms entry-wise: we show that for each $i, j \in [n]$, one has $\mathbb{E}[|X_{ij} - \Sigma_{ij}|^2] \lesssim_\beta \frac{\lambda_i \lambda_j}{m}$. To this end, we note that $\mathbb{E}[X] = \Sigma$ and $\mathbb{E}[X^{(\ell)}] = \frac{1}{m}\Sigma$. Thus

$$\begin{aligned}
\mathbb{E}[|X_{ij} - \Sigma_{ij}|^2] &= \mathrm{Var}[X_{ij}] \\
&= \sum_{\ell=1}^{m} \mathrm{Var}[X_{ij}^{(\ell)}] \\
&= \frac{1}{m}\mathbb{E}_{h\sim\mathcal{D}}[|h_i h_j - \Sigma_{ij}|^2] && (a-b)^2 \leq 2a^2 + 2b^2 \\
&\leq \frac{2}{m}(\mathbb{E}_{h\sim\mathcal{D}}[h_i^2 h_j^2] + \Sigma_{ij}^2) \\
&\leq \frac{2}{m}(\mathbb{E}_{h\sim\mathcal{D}}[h_i^4]^{1/2}\mathbb{E}_{h\sim\mathcal{D}}[h_j^4]^{1/2} + \lambda_i\lambda_j) && \text{Cauchy Schwarz + } \Sigma \text{ diagonal} \\
&\leq \frac{2}{m}(\beta^{1/2}\mathbb{E}_{h\sim\mathcal{D}}[h_i^2] \cdot \beta^{1/2}\mathbb{E}_{h\sim\mathcal{D}}[h_j^2] + \lambda_i\lambda_j) && \mathcal{D} \text{ is } \beta\text{-reasonable.}
\end{aligned}$$

For the next step of the proof, consider $J_\ell := \{i \in [n] : \ 2^{\ell-1} \leq i < 2^\ell\}$. Note that $|J_\ell| \leq 2^\ell$, and the sum of the eigenvalues in each block satisfies

$$\sum_{i\in J_\ell} \lambda_i \lesssim 2^\ell \cdot (2^\ell)^{-\alpha} = (2^\ell)^{1-\alpha}$$

by our decay assumption on the Eigenvalues of Σ. We will prove the lemma by using the triangle inequality to bound

$$\mathbb{E}[\|X - \Sigma\|_{\mathcal{S}(1)}] \leq 2\sum_{\ell\geq 1}\sum_{k\geq\ell} \mathbb{E}[\|X_{J_\ell,J_k} - \Sigma_{J_\ell,J_k}\|_{\mathcal{S}(1)}], \tag{3.3}$$

where X_{J_ℓ,J_k} is the $|J_\ell| \times |J_k|$ submatrix of X indexed by rows in J_ℓ and columns in J_k. We will use different estimates of the summands on the right hand side based on both parameter regime and diagonal vs off diagonal. These bounds are described in the following series of cases; however, first we prove the following claim which be useful for several regimes.

Claim 3.3.1. *Let $\ell \leq k$ and denote $Y := X_{J_\ell,J_k} - \Sigma_{J_\ell,J_k}$; then*

$$\mathbb{E}[\|Y\|_{\mathcal{S}(1)}] \lesssim \sqrt{\frac{r}{m}} \cdot 2^{\frac{\ell+k}{2}},$$

assuming that $\mathrm{rank}(Y) \le r$ *for any outcome of* Y*.*

Proof of Claim 3.3.1. Recall the following standard norm equivalence result:

$$\|A\|_{\mathcal{S}(1)} \le \sqrt{\mathrm{rank}(A)} \cdot \|A\|_F.$$

Then for all $\ell \le k$ we have

$$\begin{aligned}
\mathbb{E}[\|Y\|_{\mathcal{S}(1)}] &\le \sqrt{r}\mathbb{E}[\|Y\|_F] \\
&\le \sqrt{r} \cdot \mathbb{E}[\|Y\|_F^2]^{1/2} && \text{Jensen's Inequality} \\
&\lesssim_\beta \sqrt{r} \cdot \Big(\frac{1}{m}\big(\sum_{i\in J_\ell}\lambda_i\big)\big(\sum_{j\in J_k}\lambda_j\big)\Big)^{1/2} && \text{Variance bound} \\
&\lesssim \sqrt{r} \cdot \sqrt{\frac{1}{m} \cdot (2^\ell)^{1-\alpha} \cdot (2^k)^{1-\alpha}} \\
&= \sqrt{\frac{r}{m}} \cdot 2^{\frac{\ell+k}{2}(1-\alpha)}.
\end{aligned}$$

□

Case 1: Off-Diagonal Blocks: First we consider the contribution of the off-diagonal blocks to (3.3). Noting that $\mathrm{rank}(X_{J_\ell,J_k}) \le \min\{m, 2^\ell\}$,

$$\begin{aligned}
\sum_{\ell\ge 1}\sum_{k>\ell}\mathbb{E}[\|X_{J_\ell,J_k} - \Sigma_{J_\ell,J_k}\|_{\mathcal{S}(1)}] &= \sum_{\ell\ge 1}\sum_{k>\ell}\mathbb{E}[\|X_{J_\ell,J_k}\|_{\mathcal{S}(1)}] && \Sigma_{J_\ell,J_k} = 0 \\
&\le \sum_{\ell\ge 1}\sum_{k>\ell}\frac{\sqrt{\min\{m,2^\ell\}}}{\sqrt{m}} \cdot 2^{\frac{\ell+k}{2}(1-\alpha)} && \text{Claim 3.3.1} \\
&\lesssim_\alpha \sum_{\ell\ge 1}\min\{1, \sqrt{2^\ell/m}\} \cdot (2^\ell)^{1-\alpha} \\
&\lesssim_\alpha m^{1-\alpha},
\end{aligned}$$

where the last line follows from the fact that $f(z) = \sqrt{z}\cdot z^{1-\alpha}$ is monotonically increasing and $f(z) = z^{1-\alpha}$ is monotonically decreasing for $1 < \alpha < 3/2$, thus the $m = 2^\ell$ term dominates the sum.

Case 2: Diagonal Blocks with Small Indices: for this case we consider the range of small indices, i.e. blocks with $2^\ell \leq m$; then by the bound

$$\operatorname{rank}(X_{J_\ell,J_\ell} - \Sigma_{J_\ell,J_\ell}) \leq |J_\ell| \leq 2^\ell,$$

we conclude by Claim 3.3.1 that

$$\sum_{\ell:\ 2^\ell \leq m} \mathbb{E}[\|X_{J_\ell,J_\ell} - \Sigma_{J_\ell,J_\ell}\|_{\mathcal{S}(1)}] \leq \sum_{\ell:\ 2^\ell \leq m} \sqrt{\frac{2^\ell}{m}} \cdot 2^{\ell(1-\alpha)} \lesssim m^{1-\alpha}.$$

Again we use that in this regime of α, $f(z) = \sqrt{z} \cdot z^{1-\alpha}$ is monotonically increasing to conlude that the last summand, $2^\ell = m$, dominates the sum.

Case 3: Diagonal Blocks with Large Indices: Once we are in the setting $2^\ell > m$, we can ignore concentration provided by the randomness and bound

$$\begin{aligned}
\sum_{\ell:2^\ell > m} \mathbb{E}[\|X_{J_\ell,J_\ell} - \Sigma_{J_\ell,J_\ell}\|_{\mathcal{S}(1)}] &\leq \sum_{\ell:\ 2^\ell > m} (\mathbb{E}[\|X_{J_\ell,J_\ell}\|_{\mathcal{S}(1)}] + \|\Sigma_{J_\ell,J_\ell}\|_{\mathcal{S}(1)}) \\
&= \sum_{\ell:\ 2^\ell > m} (\mathbb{E}[\operatorname{Tr}[X_{J_\ell,J_\ell}]] + \operatorname{Tr}[\Sigma_{J_\ell,J_\ell}]) \\
&= \sum_{j=m}^{n} (\underbrace{\mathbb{E}[X_{jj}]}_{=\Sigma_{jj}} + \Sigma_{jj}) \\
&\lesssim \sum_{j \geq m} \frac{1}{j^\alpha} \lesssim m^{1-\alpha}.
\end{aligned}$$

In this calculation, we have used positive semi-definiteness of X_{J_ℓ,J_ℓ} and Σ_{J_ℓ,J_ℓ}, as well as the decay assumptions on the eigenvalues. This case concludes the argument for $1 < \alpha < 3/2$.

Now consider $\alpha = 3/2$: then $\sqrt{2^\ell/m} \cdot (2^\ell)^{1-\alpha} \leq 1/\sqrt{m}$, so Case 1 (off-diagonal blocks) is bounded by $\log(n)/\sqrt{m}$. For $\alpha = 3/2$, Cases 2 and 3 can be merged: by Claim 3.3.1,

$$\sum_{\ell \geq 1} \mathbb{E}[\|X_{J_\ell,J_\ell} - \Sigma_{J_\ell,J_\ell}\|_{\mathcal{S}(1)}] \leq \sum_{\ell \geq 1} \sqrt{\frac{2^\ell}{m}} \cdot 2^{\ell(1-\alpha)} \lesssim \frac{\log(n)}{\sqrt{m}}. \tag{3.4}$$

In the remaining case that $\alpha > 3/2$, then in Case 1 and in the merged Cases 2 and 3 shown in (3.4), the first term $\ell = 1$ dominates the sums and so we can omit the $\log(n)$ term. This completes the proof. □

Chapter 4

VECTOR BALANCING FOR ZONOTOPES

In this chapter we prove new bounds for the vector balancing constant of zonotopes. We also prove a lower bound on the Gaussian measure of sections of zonotopes. In Section 4.1 we introduce the problem and its history. In Section 4.2 we introduce an appropriate notion of normalization of zonotopes and prove a few corresponding properties, including approximation and decomposition properties that will be essential to our proof. Sections 4.3 and 4.4 contain the main details of our proof: the lower bound for the Gaussian measure of sections of normalized zonotopes and the upper bounds on the vector balancing constants of zonotopes, respectively.

4.1 History of the Problem

A *zonotope* is the linear image of a cube. For a matrix $A \in \mathbb{R}^{m\times d}$, with $m \geq d$,

$$K = \Big\{ \sum_{i=1}^{m} y_i A_i : \ y \in [-1,1]^m \Big\} = A^T B_\infty^m \subseteq \mathbb{R}^d$$

is a d-dimensional zonotope with m segments. The cube B_∞^d is the simplest example of a d-dimensional zonotope, and it is known that for all $p \geq 2$, the ℓ_p ball B_p^d is the (Hausdorff) limit of a sequence of zonotopes, called a *zonoid* [16]. Our result addresses the following conjecture of Schechtman:

Conjecture 4.1.1. *For any zonotope $K \subseteq \mathbb{R}^d$,* $\mathrm{vb}(K, K) = O(\sqrt{d})$.

Conjecture 4.1.1 is a proposed geometric generalization of Spencer's theorem (see Theorem 1.2.2), which shows that Conjecture 4.1.1 holds for the simplest example of a zonotope: the cube. At least two other extensions of Spencer's Theorem exist in the literature, which

we briefly describe here. The first, the Matrix Spencer conjecture, is an analytic generalization popularized by Zouzias [80] and Meka [50]; it posits that for any symmetric matrices $A_1, ..., A_n \in \mathbb{R}^{n \times n}$ with all eigenvalues in $[-1, 1]$, there are signs $x \in \{\pm 1\}^n$ so that the maximum singular value of $\sum_{i=1}^n x_i A_i$ is at most $O(\sqrt{n})$. This problem reduces to Spencer's theorem by considering only diagonal matrices, and taking the diagonal of each matrix to be the vectors to be balanced. Standard matrix concentration bounds show that a random coloring gives an $O(\sqrt{n \log n})$ bound, and the conjectured $O(\sqrt{n})$ bound holds for matrices that are block-diagonal with constant-sized blocks [25] and for matrices with rank $O(\sqrt{n})$ [36]. Improved matrix concentration bounds allow one to weaken this assumption to rank at most $n/\log^3(n)$.

The second generaliztion, called the Beck-Fiala conjecture, is a combinatorial generalization of Spencer's theorem. It considers set systems in which every element is in at most t sets, and posits that the discrepancy of such set systems is $O(\sqrt{t})$. In the same paper where this conjecture was originally stated, Beck and Fiala used the linear algebra method to prove a discrepancy bound of $2t$ [14]; the strongest known bounds are $O(\sqrt{t \log n})$ [7] (see [9] for the algorithmic version) and $2t - \log^* t$ [18]. The Beck-Fiala conjecture can be even further generalized to obtain the Komlós conjecture, which hypothesizes that $\mathrm{vb}(B_2^d, B_\infty^d) = O(1)$. In this direction the strongest known result shows that $\mathrm{vb}(B_2^d, B_\infty^d) = O(\sqrt{\log d})$ [7].

Conjecture 4.1.1 is closely related to the problem of *sparsification of zonotopes*; the strongest result in this direction is stated below [73].

Theorem 4.1.1. *For any zonotope $K \subseteq \mathbb{R}^d$ and $0 < \varepsilon \leq 1/2$, there is a zonotope Q with at most $O(\frac{d}{\varepsilon^2} \log d)$ segments so that $Q \subseteq K \subseteq (1 + \varepsilon)Q$.*

It is further conjectured that an $O_\varepsilon(d)$ bound holds, which would imply Conjecture 4.1.1 after combining an $O(d)$ sparsification with Spencer's Theorem on the underlying cube. The same argument combined with Theorem 4.1.1 implies $\mathrm{vb}(K, K) = O(\sqrt{d \log \log d})$. Our result improves on this bound:

Theorem 4.1.2 (H., Reis, and Rothvoss [34]). *For any zonotope $K \subseteq \mathbb{R}^d$,*

$$\mathrm{vb}(K, K) = O(\sqrt{d} \log\log\log d).$$

Moreover for any $v_1, ..., v_n \in K$, one can find in randomized polynomial time a coloring $x \in \{-1, 1\}^n$ with $\| \sum_{i=1}^n x_i v_i \|_K = O(\sqrt{d} \log\log\log d)$.

Because Theorem 4.1.2 is invariant under linear transformations, we will define an appropriate normalized position for zonotopes; see Section 4.2 for more details. Our main technical result, and the key ingredient of our proof, is the following tight lower bound on the Gaussian measure of sections of any normalized zonotope.

Theorem 4.1.3 (H., Reis, and Rothvoss [34]). *For any normalized zonotope $K \subseteq \mathbb{R}^d$, any subspace $H \subseteq \mathbb{R}^d$ with $n := \dim(H)$, and any $t \geq 1$, one has*

$$\gamma_H(t \cdot C \cdot K \cap H) \geq \exp(-e^{-t^2/2} \cdot n),$$

where $C > 0$ is a universal constant.

Theorem 4.1.3 can also be used to balance vectors between different normalized zonotopes.

Theorem 4.1.4 (H., Reis, and Rothvoss [34]). *For any normalized zonotopes $K, Q \subseteq \mathbb{R}^d$,*

$$\mathrm{vb}(K, Q) = O(\sqrt{d \log d}).$$

Moreover, for any $v_1, ..., v_n \in K$, one can find in randomized polynomial time a coloring $x \in \{-1, 1\}^n$ such that $\| \sum_{i=1}^n x_i v_i \|_Q = O(\sqrt{d \log \min\{d, n\}})$.

In Section 4.4 we will explain why the case of non-matching zonotopes is more challenging than that of matching zonotopes.

4.2 Normalized Zonotopes

In this section we define an appropriate notion of normalization of zonotopes for our purposes and show that one can always approximate a given zonotope by a normalized zonotope while

only losing a constant factor in the vector balancing constant. We also show that normalized zonotopes have a nice decomposition property. We begin by defining approximately regular matrices and normalized zonotopes.

Definition 4.2.1. *A matrix $A \in \mathbb{R}^{m\times d}$ is* approximately regular *if:*

(i) The columns $A^1, ..., A^d$ are orthonormal.

(ii) The rows satisfy $\|A_i\|_2 \leq 2\sqrt{\frac{d}{m}}$ for all $i = 1, ..., m$.

A zonotope $K \subseteq \mathbb{R}^d$ is normalized *if there exists an approximately regular matrix $A \in \mathbb{R}^{m\times d}$ so that $K = \sqrt{\frac{d}{m}} A^T B_\infty^m$.*

Note that by the choice of scaling, any cube B_∞^d is normalized and any d-dimensional zonotope has volume and radius comparable to that of B_∞^d, independent of the number of segments m. To this end, we have the following lemma.

Lemma 4.2.1. *Any normalized zonotope $K \subseteq \mathbb{R}^d$ satisfies $K \subseteq \sqrt{d} B_2^d$.*

Proof. As K is normalized, $K = \sqrt{\frac{d}{m}} A^T B_\infty^m$, where $A \in \mathbb{R}^{m\times d}$ is approximately regular. Then $A^T A = I_d$ by orthonormality of the columns of A, hence $\|A\|_{op} = \|A^T A\|_{op}^{1/2} = 1$. For any $x \in K$, there exists $y \in B_\infty^m$ such that $x = \sqrt{\frac{d}{m}} A^T y$ so that

$$\|x\|_2 = \sqrt{\tfrac{d}{m}} \|A^T y\|_2 \leq \sqrt{\tfrac{d}{m}} \|A^T\|_{op} \cdot \|y\|_2 \leq \sqrt{d}. \tag{4.1}$$

□

We now show the following approximation property for normalized zonotopes.

Lemma 4.2.2. *For any full-dimensional zonotope $K = A^T B_\infty^m \subseteq \mathbb{R}^d$, there is a normalized zonotope $\tilde{K}$ and an invertible linear map T so that*

$$\frac{4}{5}\tilde{K} \subseteq T(K) \subseteq \tilde{K}.$$

In particular,

$$\frac{4}{5}\mathrm{vb}(\widetilde{K}, \widetilde{K}) \leq \mathrm{vb}(K, K) \leq \frac{5}{4}\mathrm{vb}(\widetilde{K}, \widetilde{K}).$$

One main ingredient of the proof of Lemma 4.2.2 will be the existence of *Lewis weights* [23].

Theorem 4.2.1. *Given a matrix $A \in \mathbb{R}^{m\times d}$, there exists a unique vector $\overline{w} \in \mathbb{R}^m_{>0}$ so that for all $i \in [m]$, one has*

$$\overline{w}_i^{-2} A_i^T (A^T \overline{W}^{-1} A)^{-1} A_i = 1,$$

where $\overline{W} := \mathrm{diag}(\overline{w})$. Moreover, $\mathrm{Tr}(\overline{W}) \leq d$, with equality for full rank A.

Proof of Lemma 4.2.2. Let $K = A^T B_\infty^m$ be a full-dimensional zonotope with $A \in \mathbb{R}^{m\times d}$, and let $\overline{W}$ be the diagonal matrix defined by the Lewis weights of A. Scale $\overline{W}$ by an appropriately large constant $D > 0$ so that $W := D\overline{W}$ satisfies that $w_i := W_{i,i} \geq 1$ for all i. Define a matrix $B := A(A^T W^{-1} A)^{-1/2} \in \mathbb{R}^{m\times d}$, and define a second matrix $\widetilde{A}$ by replacing each row B_i of B by $\lfloor w_i \rfloor$ copies of $w_i^{-1} B_i$ and, in the case that $\{w_i\} \neq 0$, one copy of $\{w_i\}^{1/2} w_i^{-1} b_i$. Under this construction the new matrix $\widetilde{A}$ contains $m' := \sum_{i=1}^m \lceil w_i \rceil$ many rows. Define the linear transform

$$T : \mathbb{R}^d \to \mathbb{R}^d, \;\; T(K) = \sqrt{\tfrac{d}{m'}} (A^T W^{-1} A)^{-1/2} K = \sqrt{\tfrac{d}{m'}} \cdot B^T B_\infty^m$$

and $\widetilde{K} := \sqrt{\frac{d}{m'}} \widetilde{A}^T B_\infty^{m'}$. Note that by construction of $\widetilde{A}$, $B^T B_\infty^m = \widetilde{A}^T B_\infty^{m'}$. We first show that $\widetilde{K}$ is normalized by showing that $\widetilde{A}$ is approximately regular. For orthonormality of the columns, note that

$$(\widetilde{A}^T \widetilde{A})_{j,k} = \sum_{i=1}^m w_i^{-2} (\lfloor w_i \rfloor + \{w_i\}) \cdot B_{i,j} B_{i,k} = \sum_{i=1}^m w_i^{-1} \cdot B_{i,j} B_{i,k},$$

and thus by construction of B,

$$\widetilde{A}^T \widetilde{A} = B^T W^{-1} B = (A^T W^{-1} A)^{-1/2} A^T W^{-1} A (A^T W^{-1} A)^{-1/2} = I_d.$$

To see boundedness of the row norms, note that for each $i \in [m']$, there exists a $j \in [m]$ so that either $A_i = w_j^{-1} B_j$ or $\{w_j\}^{1/2} w_j^{-1} B_j$. As $\{w_j\} < 1$,

$$\|\widetilde{A}_i\|_2^2 \leq w_j^{-2} B_T^T B_j = w_j^{-2} A_j^T (A^T W^{-1} A)^{-1} A_j = \frac{1}{d} \leq \frac{2d}{m'}.$$

Note that the last inequality follows from the following calculation:

$$m' = \sum_{i=1}^{m} \lceil w_i \rceil \leq 2 \cdot \sum_{i=1}^{m} w_i = 2D \sum_{i=1}^{m} \sum_{i=1}^{m} \overline{w_i} \leq 2D \cdot d.$$

This concludes the proof that $\widetilde{K}$ is normalized. We now show that $\frac{4}{5}\widetilde{K} \subseteq T(K) \subseteq \widetilde{K}$, which immediately implies $\frac{4}{5}\mathrm{vb}(\widetilde{K}, \widetilde{K}) \leq \mathrm{vb}(K, K) \leq \frac{5}{4}\mathrm{vb}(\widetilde{K}, \widetilde{K})$. We fix an arbitrary $y \in \frac{4}{5}\widetilde{K}$, which by definition can be written as

$$y = \frac{4}{5}\sqrt{\tfrac{d}{m'}} \sum_{i=1}^{m} \Big(\sum_{p=1}^{\lfloor w_i \rfloor} x_{i,p} w_i^{-1} B_i + x_{i,\lceil w_i \rceil} \{w_i\}^{1/2} w_i^{-1} B_i \Big) \in \frac{4}{5}\sqrt{\tfrac{d}{m'}} \widetilde{A}^T B_\infty^{m'} = \frac{4}{5}\widetilde{K},$$

and rewrite it as

$$y = \frac{4}{5}\sqrt{\tfrac{d}{m'}} \sum_{i=1}^{m} \Bigg(\underbrace{w_i^{-1}\Big(\sum_{p=1}^{\lfloor w_i \rfloor} x_{i,p} + x_{i,\lceil w_i \rceil}\{w_i\} \Big)}_{\in [-1,1]} + \underbrace{x_{i,\lceil w_i \rceil} \tfrac{\{w_i\}^{1/2} - \{w_i\}}{w_i}}_{\in [-1/4, 1/4]} \Bigg) B_i \in \sqrt{\tfrac{d}{m'}} B^T B_\infty^m = T(K).$$

For the second inclusion, select an arbitrary $y := \sqrt{\frac{d}{m'}} \sum_{i=1}^{m} x_i B_i \in B^T B_\infty^m = T(K)$, note that it can be rewritten as

$$y = \sqrt{\tfrac{d}{m'}} \sum_{i=1}^{m} \Bigg(\sum_{p=1}^{\lfloor w_i \rfloor} x_i w_i^{-1} B_i + x_i \{w_i\} w_i^{-1} B_i \Bigg) \in \sqrt{\tfrac{d}{m'}} \widetilde{A}^T B_\infty^{m'} = \widetilde{K}.$$

□

To conclude this section, we show the following decomposition property of approximately regular matrices, which will be essential in our proof. In particular, it will allow us to approximately break down our original zonotope into a sum of smaller, simpler zonotopes in the proof of Theorem 4.1.3.

Lemma 4.2.3. *There is a universal constant $C > 0$ so that the following holds: let $A \in \mathbb{R}^{m \times d}$ be an approximately regular matrix. Then there are disjoint subsets $J_1 \dot\cup \cdots \dot\cup J_k \subseteq [m]$ with $k \geq \frac{m}{Cd}$ and $|J_\ell| \leq Cd$, and $\sum_{i \in J_\ell} A_i A_i^T \succeq \frac{1}{Ck} I_d$ for all $\ell \in [k]$.*

The key ingredient of the proof of Lemma 4.2.3 will be the Kadison-Singer theorem [48] from operator theory.

Theorem 4.2.2. *Let $v_1, ..., v_m \in \mathbb{R}^n$ with $\sum_{i=1}^m v_i v_i^T = I_d$, and let $\varepsilon > 0$ so that $\|v_i\|_2^2 \leq \varepsilon$ for all $i \in [m]$. Then there is a partition $[m] = S_1 \dot\cup S_2$ so that for both $j \in \{1, 2\}$, one has*

$$\Big\| \sum_{i \in S_j} v_i v_i^T - \frac{1}{2} I_d \Big\|_{op} \leq 3\sqrt{\varepsilon}.$$

We formulate the following slight variant of Theorem 4.2.2, which is more suitable for our purposes.

Lemma 4.2.4. *Let $v_1, ..., v_m \in \mathbb{R}^d$ be vectors with $\sum_{i=1}^m v_i v_i^T \succeq L \cdot I_d$ for some $L > 0$, and let $\varepsilon := \max_{i=1,...,m} \|v_i\|_2^2$. Then there exists a partition $[m] = S_1 \dot\cup S_2$ so that*

$$\sum_{i \in S_j} v_i v_i^T \succeq \Big(\tfrac{L}{2} - 3\sqrt{L\varepsilon} \Big) I_d \quad \forall j \in \{1, 2\}.$$

Proof of Lemma 4.2.4. We define $M := \sum_{i=1}^m v_i v_i^T$, a PSD matrix satisfying $M \succeq L \cdot I_d$. We define a new collection of vectors by $v_i' := M^{-1/2} v_i$, so that

$$\sum_{i=1}^m v_i' (v_i')^T = M^{-1/2} \Big(\sum_{i=1}^m v_i v_i^T \Big) M^{-1/2} = I_d.$$

Define $\varepsilon' : -\varepsilon / L$ and note that for each $i \in [m]$,

$$\|v_i'\|_2^2 = v_i^T M^{-1} v_i \leq v_i^T (\tfrac{1}{L} I_d) v_i = \tfrac{\|v_i\|_2^2}{L} \leq \varepsilon'.$$

Thus we can apply Theorem 4.2.2 to $v_1', ..., v_m'$ to obtain a partition $[m] = S_1 \dot\cup S_2$ so that for $j \in 1, 2$, one has

$$M^{-1/2} \Big(\sum_{i \in S_j} v_i v_i^T \Big) M^{-1/2} = \sum_{i \in S_j} v_i' (v_i')^T \overset{\text{Thm 4.2.2}}{\succeq} \Big(\frac{1}{2} - 3\sqrt{\varepsilon / L} \Big) I_d. \tag{4.2}$$

Note that if $A \succeq B$, then $M^{1/2} A M^{1/2} \succeq M^{1/2} B M^{1/2}$; thus combining this fact with (4.2), we conclude that

$$\sum_{i \in S_j} v_i v_i^T \succeq \Big(\frac{1}{2} - 3\sqrt{\varepsilon / L} \Big) M^{1/2} I_d M^{1/2} \succeq \Big(\frac{L}{2} - 3\sqrt{L\varepsilon} \Big) I_d.$$

□

The proof of Lemma 4.2.3 is based on iteratively applying Lemma 4.2.4.

Proof of Lemma 4.2.3. In the case that $\frac{m}{d} \leq C$, we can take $k = 1$ and $J_1 = [m]$ to prove the statement, thus we can assume that $m \geq Cd$. We set $\varepsilon := 4\frac{d}{m}$ so that by approximate regularity of A, $\|A_i\|_2^2 \leq \varepsilon$ for all $i \in [m]$. Set $t \in \mathbb{N}$ to be a parameter to be fixed later. For $s \in \{0, ..., t\}$, we obtain partitions $\mathcal{P}_s$ of the row indices, beginning with $\mathcal{P}_0 := \{[m]\}$, so that $\mathcal{P}_{s+1}$ is a refinement of $\mathcal{P}_s$, and $|\mathcal{P}_s| = 2^s$. In more detail: for each iteration $s \in \{0, ..., t-1\}$, and for each $S \in \mathcal{P}_s$, we apply Lemma 4.2.4 to the collection of vectors $\{A_i\}_{i \in S}$. Denoting the obtained partition as $S = S_1 \dot{\cup} S_2$, we add $\{S_1, S_2\}$ to $\mathcal{P}_{s+1}$. Defining $L_s := 2^{-s} - 15\sqrt{2^{-s}\varepsilon}$, we first analyze how well we can approximate the identity at each step of our partition process. Specifically, we show that if $2^t \leq \frac{m}{Cd}$ for a sufficiently large constant $C > 0$, then for all $s \in \{0, ..., t\}$ one has $\sum_{i \in S} A_i A_i^T \succeq L_s I_d$ for all $S \in \mathcal{P}_s$. It is immediate that $L_s \leq 2^{-s}$ for all $s \geq 0$. We prove the desired bound by induction on s. For $s = 0$, $\mathcal{P}_0 = \{[m]\}$ and the claim is true, as $L_0 \leq 1$. We next consider an iteration $s \in \{0, ..., t-1\}$, and suppose $S \in \mathcal{P}_s$ is partitioned as $S = S_1 \dot{\cup} S_2$. Then by Lemma 4.2.4,

$$\sum_{i \in S_j} A_i A_i^T \succeq \Big(\frac{L_s}{2} - 3\sqrt{L_s \varepsilon}\Big) I_d \quad \forall j \in \{1, 2\}. \tag{4.3}$$

To see that this is at least L_{s+1},

$$\begin{aligned}
\frac{L_s}{2} - 3\sqrt{L_s\varepsilon} &\geq \frac{L_s}{2} - 3\sqrt{2^{-s}\varepsilon} && L_s \leq 2^{-s} \\
&\geq 2^{-(s+1)} - \frac{15}{2}\sqrt{2^{-s}\varepsilon} - 3\sqrt{2^{-s}\varepsilon} \\
&\geq 2^{-(s+1)} - 15\sqrt{2^{-(s+1)}\varepsilon} && 15/2 + 3 \leq 15\sqrt{2^{-1}}.
\end{aligned}$$

To complete the proof, fix a large enough constant C and $t \in \mathbb{N}$ so that $\frac{m}{2Cd} \leq 2^t \leq \frac{m}{Cd}$. As long as C is sufficiently large,

$$L_t \geq \frac{Cd}{m} - 15\sqrt{\frac{2Cd}{m} \cdot 4\frac{d}{m}} = \frac{d}{m} \cdot (C - 15\sqrt{8C}) \geq \frac{C}{2} \cdot \frac{d}{m}.$$

In addition, $\mathbb{E}_{S \sim \mathcal{P}_t}[|S|] = \frac{m}{2^t} \leq 2Cd$. Thus by applying Markov's inequality, at least half of the sets $S \in \mathcal{P}_t$ have at most $4Cd$ indices, and those sets satisfy the lemma. □

4.3 Technical Details: the Gaussian Measure of Sections of Zonotopes

In Section 4.2, we showed that one can always approximately decompose a normalized zonotope into a sum of "simpler" zonotopes. We will prove a Gaussian measure lower bound on sections of normalized zonotopes by first proving a bound on sections of such "simpler" zonotopes, and then applying *log-concavity* of γ_d to infer such a result for all normalized zonotopes, based on our decomposition lemma in Section 4.2. Recall from Section 1.2 that there is a well-established connection between the Gaussian measure of sections of convex bodies and the existence of good partial colorings for such convex bodies; see [30] for a nice explanation and summary of such results.

To begin this section, we introduce the Gaussian measure γ_d on $\mathbb{R}^d$ with the standard Gaussian density $\frac{1}{(2\pi)^{d/2}} e^{-\|x\|_2^2/2}$ and distribution denoted by $N(0, I_d)$. Given any subspace $F \subseteq \mathbb{R}^d$, denote by $I_F \in \mathbb{R}^{d\times d}$ the identity on the subspace. In particular, $I_F = \sum_{i=1}^{\dim(F)} u_i u_i^T$, where $u_1, ..., u_{\dim(F)}$ is any orthonormal basis of F. It is well-known that the Gaussian distribution is log-concave, which means by definition that for all compact subsets $S, T \subseteq \mathbb{R}^d$ and $0 \leq \lambda \leq 1$ one has

$$\gamma_d(\lambda S + (1-\lambda)T) \geq \gamma_d(S)^\lambda \cdot \gamma_d(T)^{1-\lambda}.$$

By induction one can extend this claim to any collection of compact subsets $S_1, ..., S_k \subseteq \mathbb{R}^d$ and $\lambda_1, ..., \lambda_k \geq 0$ with $\sum_{i=1}^k \lambda_i = 1$, we have $\gamma_d(\lambda_1 S_1 + \cdots + \lambda_k S_k) \geq \prod_{\ell=1}^k \gamma_d(S_\ell)^{\lambda_\ell}$.

The Sidak-Khatri theorem, stated below, will be useful for our proof. A *strip* is a symmetric convex body of the form $P = \{x \in \mathbb{R}^d : |\langle a, x\rangle| \leq 1\}$, for some $a \in \mathbb{R}^d$.

Theorem 4.3.1. *For any two symmetric convex bodies $P, Q \subseteq \mathbb{R}^d$, where at least one is a strip, one has*

$$\gamma_d(P \cap Q) \geq \gamma_d(P) \cdot \gamma_d(Q).$$

More recently, it has been proved that this theorem holds for any pair of symmetric convex bodies [59], but the weaker result suffices for our purposes.

The following estimate gives a lower bound on the Gaussian measure of sections of symmetric convex bodies.

Lemma 4.3.1. *For any symmetric convex body K and any subspace $H \subseteq \mathbb{R}^d$, one has*

$$\gamma_H(K \cap H) \geq \gamma_d(K).$$

The next lemma gives a lower estimate on the Gaussian measure of a strip.

Lemma 4.3.2. *For any $a \in \mathbb{R}^d$ with $\|a\|_2 \leq 1$ and $t \geq 1$,*

$$\Pr_{y \sim N(0,I_d)}[|\langle a, y\rangle| \leq t] \geq \exp(-e^{-t^2/2} \cdot \|a\|_2^2).$$

Proof. We apply the following tail inequality, due to Szarek and Werner [71], for $t > -1$:

$$\Pr_{g \sim N(0,1)}[g > t] < \frac{1}{\sqrt{2\pi}} \frac{4e^{-t^2/2}}{3t + (t^2+8)^{1/2}}.$$

From this bound we see that for $t \geq 1$,

$$\Pr g \sim N(0,1)[g > t] < \frac{1}{\sqrt{2\pi}} \frac{4e^{-t^2/2}}{6},$$

hence

$$\Pr_{g \sim N(0,1)}[|g| \leq t] \geq 1 - \frac{4}{3\sqrt{2\pi}} e^{-t^2/2}.$$

By convexity of the function $f(z) = e^{-2z/3}$, we have $1 - \frac{4}{3\sqrt{2\pi}} z \geq e^{-2z/3}$ for any $z \in [0, e^{-1/2}]$. Thus for $t \geq 1$, we have shown that

$$\Pr[|g| \leq t] \geq \exp(-\tfrac{2}{3} e^{-t^2/2}).$$

Thus for any $a \in \mathbb{R}^d$ with $\|a\|_2 \leq 1$ and $t \geq 1$,

$$\Pr_{y \sim N(0,I_d)}[|\langle a, y\rangle| \leq t] = \Pr_{g \sim N(0,1)}\left[|g| \leq \tfrac{t}{\|a\|_2}\right] \geq \exp(-\tfrac{2}{3} e^{-t^2/(2\|a\|_2^2)}) \geq \exp(-e^{-t^2/2} \cdot \|a\|_2^2).$$

The last inequality follows from the following calculation:

$$\frac{2}{3} \exp\left(\frac{t^2}{2} - \frac{t^2}{2\|a\|_2^2}\right) \leq \frac{2}{3} \exp\left(\frac{1}{2} - \frac{1}{2\|a\|_2^2}\right) \leq \frac{2}{3} e^{1/2} \cdot \frac{2}{e} \cdot \|a\|_2^2 \leq \|a\|_2^2.$$

The second to last inequality follows from observing that $e^z \geq ez$ for $z := 1/(2\|a\|_2^2)$. □

The next lemma is a useful comparison inequality (see for example [42]).

Lemma 4.3.3. *Let K be a symmetric convex body and let $0 \preceq A \preceq B$. Then*

$$\Pr_{y\sim N(0,A)}[y \in K] \geq \Pr_{y\sim N(0,B)}[y \in K]$$

Proof. Sample a random variable $z \sim N(0, B - A)$. Applying log-concavity,

$$\Pr_{y\sim N(0,A)}[y \in K] \geq \Pr_{y\sim N(0,A)}\Big[\Pr_{z\sim N(0,B-A)}[y + z \in K]\Big] = \Pr_{y\sim N(0,B)}[y \in K].$$

□

Finally, the following lemma allows us to dismiss constant scaling factors [74].

Lemma 4.3.4. *Let $K \subset \mathbb{R}^d$ be a measurable set and B be a Euclidean ball centered at the origin such that $\gamma_d(K) = \gamma_d(B)$. Then $\gamma_d(tK) \geq \gamma_d(tB)$ for all $t \in [0, 1]$. In particular, if $\gamma_d(C_1 \cdot K) \geq e^{-C_1 d}$ for some constant $C_1 \geq 1$, then also $\gamma_d(K) \geq e^{-C_2 d}$ for some $C_2 := C_2(C_1) > 0$.*

We are now ready to prove Theorem 4.1.3, which we will do in three steps. First, we will show a simple bound for any zonotope defined by a matrix with orthonormal columns; next, we will loosen the assumption that the columns are orthonormal to match the conditions of the matrices obtained via Lemma 4.2.3; finally, we combine Lemma 4.2.3 with log-concavity of γ_d to prove Theorem 4.1.3.

Lemma 4.3.5. *Let $K := A^T B_\infty^m \subseteq \mathbb{R}^d$ be a zonotope, where $A \in \mathbb{R}^{m\times d}$ has orthonormal columns. Then for any subspace $H \subseteq \mathbb{R}^d$ with $n := \dim(H)$, and any $t \geq 1$,*

$$\gamma_H(t \cdot K \cap H) \geq \exp(-e^{-t^2/2} \cdot n).$$

Proof. Let $U \in \mathbb{R}^{d\times n}$ be a matrix with orthonormal columns $U^1, ..., U^n$ spanning H. Sampling $y \sim N(0, I_n)$, Uy is a standard Gaussian random variable in H. As we assume that $\sum_{i=1}^m A_i A_i^T = I_d$, we can always write

$$Uy = \sum_{j=1}^n y_j I_d U^j = \sum_{i=1}^m A_i \sum_{j=1}^n y_j \langle A_i, U^j\rangle = \sum_{i=1}^m A_i \langle y, U^T A_i\rangle. \tag{4.4}$$

We interpret $U^T A_i \in \mathbb{R}^n$ as the coordinates of $\Pi_H(A_i)$ (the orthogonal projection of A_i onto H) in terms of the basis U of H. Note that equation (4.4) shows the following: for all $y \in \mathbb{R}^n$ and $s > 0$ in order to show that $Uy \in sK$, it is sufficient to show that $|\langle y, U^T A_i \rangle| \leq s$ for all $i \in [m]$. We combine this observation with the Sidak-Khatri inequality (see Theorem 4.3.1 above) to lower bound the Gaussian measure:

$$\begin{aligned}
\gamma_H(t \cdot K \cap H) &= \Pr_{y \sim N(0,I_n)}[Uy \in t \cdot K] \\
&\geq \Pr_{y \in N(0,I_n)}[|\langle U^T A_i, y \rangle| \leq t \quad \forall i \in [m]] \\
&\geq \prod_{i=1}^{m} \Pr_{y \sim N(0,I_n)}[|\langle U^T A_i, y \rangle| \leq t] && \text{Theorem 4.3.1} \\
&\geq \prod_{i=1}^{m} \exp(-e^{-t^2/2} \|U^T A_i\|_2^2) && \text{Lemma 4.3.2} \\
&= \exp(-e^{-t^2/2} \sum_{i=1}^{m} \|U^T A_i\|_2^2) \\
&= \exp(-e^{-t^2/2} n),
\end{aligned}$$

where we have used that by orthonormality of the columns of A, $\|U^T A_i\|_2 \leq \|A_i\|_2 \leq 1$. $\square$

We now proceed to the second step: showing that the same bound holds if the matrix has nearly orthonormal columns by adding a rescaling argument.

Lemma 4.3.6. *Let $K = A^T B_\infty^m \subseteq \mathbb{R}^d$ be a zonotope where $A \in \mathbb{R}^{m \times d}$ is a matrix so that $\sum_{i=1}^m A_i A_i^T \succeq \alpha I_d$ for some $\alpha > 0$. Then for any n-dimensional subspace $H \subseteq \mathbb{R}^d$ and any $t \geq 1$, one has*

$$\gamma_H(\tfrac{t}{\sqrt{\alpha}} \cdot K \cap H) \geq \exp(-e^{-t^2/2} \cdot n).$$

Proof. As scaling K by a factor of $1/\sqrt{\alpha}$ is equivalent to scaling $\sum_{i=1}^m A_i A_i^T$ by $1/\alpha$, we can assume that $\alpha = 1$. Define the symmetric positive definite matrix $M := \sum_{i=1}^m A_i A_i^T \succeq I_d$, and define $\widetilde{A} \in \mathbb{R}^{m \times d}$ with rows rescaled as $\widetilde{A}_i := M^{-1/2} A_i$, so that $\sum_{i=1}^m \widetilde{A_i} \widetilde{A_i}^T = I_d$. We use $\widetilde{A}$ to define a rescaled zonotope $\widetilde{K} := \widetilde{A}^T B_\infty^m = M^{-1/2}(K)$ and rescaled subspace $\widetilde{H} := M^{-1/2}(H)$. For $U^1, ..., U^n$ any orthonormal basis of H, $\widetilde{U} := M^{-1/2} U$ with columns $\widetilde{U}^1, ..., \widetilde{U}^n$

will be the basis of $\widetilde{H}$, though it may not be orthonormal. Thus we use the comparison inequality from Lemma 4.3.3 and the observation that for $y \sim N(0, I_n)$, $\mathrm{Cov}(\widetilde{U}y) = \widetilde{U}\widetilde{U}^T = M^{-1/2}UU^TM^{-1/2} \preceq I_{\widetilde{H}}$, which allows us to return to the setting of Lemma 4.3.5:

$$\Pr_{y\sim N(0,I_d)}[Uy \in tK] = \Pr_{y\sim N(0,I_d)}[\widetilde{U}y \in t\widetilde{K}] \overset{Lem4.3.3}{\geq} \Pr_{y\sim N(0,I_{\widetilde{H}})}[y \in t\widetilde{K}] \overset{Lem4.3.5}{\geq} \exp(-e^{-t^2/2}\cdot n)$$

□

We are now ready for the final step: the proof of Theorem 4.1.3. In what follows, for a matrix $A \in \mathbb{R}^{m\times d}$ and indices $J \subseteq [m]$, $A_j \in \mathbb{R}^{|J|\times d}$ is the submatrix of A with rows indexed by J.

Proof of Theorem 4.1.3. Fix a normalized zonotope $K \subseteq \mathbb{R}^d$ and a subspace $H \subseteq \mathbb{R}^d$ with dimension n. As K is normalized, $K = \sqrt{\frac{d}{m}}A^TB_\infty^m$ for $A \in \mathbb{R}^{m\times d}$ an approximately regular matrix. We apply Lemma 4.2.3 to obtain disjoint subsets $J_1\dot\cup\cdots\dot\cup J_k \subseteq [m]$ with $k \geq \frac{m}{Cd}$ satisfying $\sum_{i\in J_\ell} A_iA_i^T \succeq \frac{d}{Cm}I_d$ for $C > 0$ constant. We now consider the zonotopes generated by these matrices: $K_\ell := \sqrt{\frac{d}{m}}A_{J_\ell}^TB_\infty^{|J_\ell|}$. As $J_1, ...J_k$ partition a subset of the rows of A, we also have

$$K_1 + \cdots + K_k \subseteq K \quad \text{and} \quad (K_1 \cap H) + \cdots + (K_k \cap H) \subseteq K \cap H.$$

Becuse we know that $k \geq m/Cd$, we also have $kK_\ell \subseteq \sqrt{\frac{k}{C}}A_{J_\ell}^TB_\infty^{|J_\ell|}$, so in particular we obtain

$$\sum_{i\in J_\ell}\left(\sqrt{\tfrac{k}{C}}A_i\right)\left(\sqrt{\tfrac{k}{C}}A_i\right)^T \succeq \frac{k}{C}\cdot\frac{d}{Cm}I_d \succeq \frac{1}{C^3}I_d.$$

Thus we can apply Lemma 4.3.6 to kK_ℓ with $\alpha := \frac{1}{C^3}$ to obtain that

$$\gamma_H(tC^{3/2}kK_\ell \cap H) \geq \exp(-e^{t^2/2}\cdot n) \quad \forall \ell \in K,\ t \geq 1.$$

The last step is to apply log-concavity of the Gaussian measure with scaling factors $\lambda_1 =$

$\cdots = \lambda_k = 1/k$:

$$\begin{aligned}\gamma_H(tC^{3/2}K \cap H) &\geq \gamma_H\Big(\tfrac{1}{k}(tC^{3/2}kK_1 \cap H) + \cdots + \tfrac{1}{k}(tC^{3/2}kK_k \cap H)\Big) \\ &\geq \prod_{\ell=1}^{k} \gamma_H(tC^{3/2} \cdot kK_\ell \cap H)^{1/k} && \text{Log-concavity of } \gamma_d \\ &\geq \exp(-e^{-t^2/2} \cdot n) && \text{Lemma 4.3.6.}\end{aligned}$$

□

4.4 Proof of Main Results

In this section we will use Theorem 4.1.3 to prove upper bounds on the vector balancing constant of a zonotope with itself and of two distinct zonotopes. We begin with the former, for which we will use the *partial coloring method* introduced in Section 1.2. To begin, we show the following argument for obtaining a constant discrepancy partial coloring when selecting vectors from B_2^d.

Lemma 4.4.1. *Let $v_1, ..., v_n \in B_2^d$ and let $K \subseteq \mathbb{R}^d$ be a symmetric convex body with $\gamma_H(K \cap H) \geq e^{-\alpha n}$ for some $\alpha > 0$, where $H = \operatorname{span}\{v_1, ..., v_n\}$. Then there is a randomized polynomial time algorithm that given a shift $y \in (-1, 1)^n$ finds a good partial coloring $x + y \in [-1, 1]^n$ with $\sum_{j=1}^n x_j v_j \in cK$ where $c := c(\alpha)$ is a constant.*

Proof. Define the random variable $Z \sim \sum_{j=1}^n z_j v_j$ where $z_i \sim N(0, 1)$ are i.i.d. Gaussian random variables so that $\mathbb{E}[ZZ^T] = \sum_{j=1}^n v_j v_j^T$ has trace $\operatorname{Tr}[\mathbb{E}[ZZ^T]] = \sum_{j=1}^n \|v_j\|_2^2 \leq n$. Take $u_1, ..., u_r$ any orthonormal basis of H; then $r \leq n$, and we can write $\sum_{j=1}^n v_j v_j^T = \sum_{j=1}^r \sigma_j u_j u_j^T$. As $\sum_{j=1}^n \succeq 0$, we know that $\sigma_j \geq 0$ for all j, hence without loss of generality we may reindex so that $0 \leq \sigma_1 \leq \cdots \leq \sigma_r$. As $\sum_{j=1}^n \|v_j\|_2^2 \leq n$, Markov's inequality implies that $\sigma_{2n/3} \leq 3/2$ (where we denote $\sigma_j = 0$ for $j > r$). We now define the subspaces

$$F := \operatorname{span}\{u_1, ..., u_{2n/3}\}, \qquad V := \Big\{g \in \mathbb{R}^n : \sum_{j=1}^{2n/3} g_j v_j \in F\Big\}.$$

Thus

$$\begin{aligned}
\Pr_{g\sim N(0,I_V)}\Big[\sum_{j=1}^{n} g_j v_j \in 3/2\cdot K\Big] &= \Pr_{g\sim N(0,I_{2n/3})}\Big[\sum_{j=1}^{2n/3} g_j\cdot\sigma_j u_j u_j^T \in 3/2\cdot K\Big] \\
&\geq \Pr_{g\sim N(0,I_{2n/3})}\Big[\sum_{j=1}^{2n/3} g_j\cdot 3/2\cdot u_j u_j^T \in 3/2\cdot K\Big] \quad \text{Lemma 4.3.3} \\
&= \gamma_F(K\cap F) \\
&\geq \gamma_H(K\cap H) \quad \text{Lemma 4.3.1} \\
&\geq e^{-\alpha n}.
\end{aligned}$$

Applying Theorem 1.2.4 to the symmetric convex body

$$Q := \Big\{x\in\mathbb{R}^n : \sum_{j=1}^{n} x_j v_j \in K\Big\}$$

contains a good partial coloring in $Q\cap F$. □

The main purpose of Lemma 4.4.1 for our proof is to show the existence of a partial coloring with optimal bounds as long as n is of the order of d.

Corollary 4.4.1. *Let $K\subseteq\mathbb{R}^d$ be a normalized zonotope and let $v_1,...,v_n\in K$. Then there is a randomized polynomial time algorithm to find a good partial coloring $x\in[-1,1]^n$ so that $\|\sum_{j=1}^n x_j v_j\|_K = O(\sqrt{d})$.*

Proof. Denoting $H := \operatorname{span}\{v_1,...,v_n\}$, Theorem 4.1.3 implies that $\gamma_H(C\cdot K\cap H)\geq e^{-n}$. By Lemma 4.3.4, there exists a constant $\alpha>0$ such that $\gamma_H(K\cap H)\geq e^{-\alpha n}$. As we showed in Lemma 4.2.1, our choice of scaling for normalized zonotopes implies that $v_i\in\sqrt{d}B_2^d$ for each $i\in[n]$, and the statement follows from Lemma 4.4.1. □

Equipped with the guaranteed existence of such partial colorings, we can prove an $O(\sqrt{d}\log\log\log d)$ bound on $\operatorname{vb}(K,K)$ for any zonotope K.

Proof of Theorem 4.1.2. We may assume that K is generated by $m=O(d\log d)$ segments by Theorem 4.1.1 and that $K=\sqrt{\frac{d}{m}}A^TB_\infty^m$ for an approximately regular $A\in\mathbb{R}^{m\times d}$ by

Lemma 4.2.2. By Theorem 1.2.1 we may assume that $n = d$, although for clarify we only apply this result in the final bound. As in the previous proof, we define the partial coloring body $Q := \{x \in \mathbb{R}^n : \sum_{j=1}^n x_j v_j \in K\}$. We iteratively apply Lemma 4.4.1 t times to obtain a partial coloring $x' \in Q \cap [-1,1]^n$ such that the set $I := \{i : |x'_i| < 1\}$ of partially colored indices satisfies $|I| \leq n/2^t$, hence by the triangle inequality,

$$\Big\| \sum_{j=1}^n x'_j v_j \Big\|_K = O(\sqrt{d} \cdot t).$$

For each $j \in I$, we can write $v_j = \sqrt{\frac{d}{m}} A^T u_i$ for some $u_i \in B_\infty^m$. By applying Spencer's theorem to $\{u_i\}_{i \in I}$, we obtain signs $\widetilde{x} \in \mathbb{R}^n$ (with $\widetilde{x}_i := 0$ for $i \notin I$) so that $x := \widetilde{x} + x' \in \{-1,1\}^n$, and

$$\sum_{i \in I} \widetilde{x}_i u_i \in \sqrt{|I| \cdot \log(2m/|I|)} \cdot c \cdot B_\infty^m.$$

Setting $t := \log\log(2m/n)$,

$$\begin{aligned}
\Big\| \sum_{j=1}^n x_j v_j \Big\|_K &\leq \Big\| \sum_{j=1}^n x'_j v_j \Big\|_K + \Big\| \sum_{j \in I} \widetilde{x}_j v_j \Big\|_K \\
&\lesssim \sqrt{d} \cdot t + \sqrt{\frac{n}{2^t} \cdot \log\Big(\frac{2m}{n/2^t}\Big)} \\
&= \sqrt{d} \log\log\Big(\frac{2m}{n}\Big) + \sqrt{\frac{n}{\log(2m/n)} \cdot \log\Big(\frac{2m}{n}\log(2m/n)\Big)} \\
&\lesssim \sqrt{d} \log\log(2m/n) \\
&\lesssim \sqrt{d} \log\log\Big(\frac{d\log d}{n}\Big).
\end{aligned}$$

Finally, applying Theorem 1.2.1

$$\mathrm{vb}(K,K) \leq 2\mathrm{vb}_d(K,K) = O(\sqrt{d} \log\log\log d).$$

□

To conclude the section, we prove that for any zonotopes $K, Q \subseteq \mathbb{R}^d$, $\mathrm{vb}(K,Q) = O(\sqrt{d \log d})$. First we note that we cannot apply the same proof as for Theorem 4.1.2, as we

relied on Spencer's theorem, which implies that $\mathrm{vb}_n(K,K) = O(\sqrt{n\log(2m/n)})$, in particular that giving a bound that improves as n decreases; such a bound does not hold in the setting where $K \neq Q$. As such an example, take $H \in \{-1,1\}^{d\times d}$ any Hadamard matrix (i.e. all rows and columns are orthogonal). Then one can check that $K := \frac{1}{\sqrt{d}}H^T B_\infty^d$ is a normalized zonotope (in fact, a rotated cube). Fix any $n \leq d$ and the points $v_i = \frac{1}{\sqrt{d}}H^T H^i = \sqrt{d}\cdot e_i$ for $i \in [n]$, $v_1, ..., v_n \in K$. Taking $Q := B_\infty^d$ as the second normalized zonotope, then we note that any good partial coloring $x \in [-1,1]^n$ must have a coordinate i with $|x_i| \geq 1/2$, thus $\|\sum_{j=1}^n x_j v_j\|_Q \geq \sqrt{d}|x_i| \geq \sqrt{d}/2$. Thus iterating Corollary 4.4.1 only implies a bound of $\sqrt{d}\log d$, but we can do better by applying Banaszczyk's theorem (see Theorem 1.2.3 and A.0.1).

Proof of Theorem 4.1.4. Let $K, Q \subseteq \mathbb{R}^d$ be normalized zonotopes and $v_1, ..., v_n \in K$ the vectors to be balanced. Let $H := \mathrm{span}\{v_1, ..., v_n\}$ with $r := \dim(H) \leq \min\{d, n\}$. By applying Theorem 4.1.3 to Q with subspace H and $t := \sqrt{2\log 2r}$, we conclude

$$\gamma_H\Big(\sqrt{2\log 2r}C'Q \cap H\Big) \geq e^{-1/2} > 1/2.$$

As Lemma 4.2.1 shows, $v_i \in \sqrt{d}B_2^d$ for each $i \in [n]$, thus by Theorem A.0.1, one can compute in polynomial time signs $x \in \{-1,1\}^n$ so that

$$\sum_{j=1}^n x_j v_j \in \sqrt{d}C''\Big(\sqrt{2\log 2r}C'Q \cap H\Big) \subseteq C\sqrt{d\log\min\{d,n\}Q},$$

as desired. In particular, $\mathrm{vb}(K,Q) = O(\sqrt{d\log d})$. □

Chapter 5

CONCLUSIONS AND FUTURE WORK

We conclude this dissertation by briefly summarizing the three collections of results presented and discussing potential future research directions.

In Chapter 2 we introduced the kernel discrepancy problem, motivated by the coreset problem for kernel density estimation from machine learning, and gave improved bounds for many kernels of interest, in addition to developing a new discrepancy based technique that allows for more data driven approaches in future work. Our approach opens several interesting directions for future research. The main remaining direction of interest for this problem would be to show a $\Theta(\sqrt{d})$ bound on the kernel discrepancy for positive definite kernels, which amounts to showing a matching upper bound to the lower bound proved in [55]. Towards this goal I propose multiple strategies. First, our chaining argument operates by transforming the problem into a discrepancy-style problem in a Hilbert space by applying the theory of reproducing kernel Hilbert spaces. We conjecture that the image of the domain of interest in the Hilbert space may have bounded Hilbert norm, which would show the tight upper bound. Another related direction will be to explore a stronger formulation of the chaining technique, which relies on bounds specific to the Hilbert space geometry [79]. Second, by removing the need to discretize the domain, we are able to apply additional geometric, analytic, and other conditions on the data points sampled from the distribution ρ to obtain stronger results. For example, if we make stronger assumptions on ρ (i.e. bounded variance, unimodality, etc), what assumptions can we consequently make about the sampled data points (with high probability), and–to combine with the previous approach–what impact do such assumptions have on the geometric properties of the point sets in the associated Hilbert space?

In Chapter 3 we describe a new application of the discrepancy method to the problem of quantizing the weights of neural networks. In particular, we show a computationally efficient algorithm for quantizing the weights of a neural network and prove theoretical guarantees on its expected performance on unseen samples. Potential extensions of this work could include making stronger spectral assumptions on the matrix. Our result assumes that the eigenvalues decay at some fixed rate, an assumption that is informed by the data; however, there are further assumptions we could make. For example, experimental evidence shows that the worst-case matrices for our solution, where all eigenvectors are concentrated in one direction, do not occur in practice. A similar combined experimental and theoretical approach as in our work could be taken to determine what assumptions on the matrix are both reasonable and lead to stronger bounds. As a first step towards solving this problem, we could consider the case that we have access to the full covariance matrix, rather than sampling and approximating the matrix as we do in our current results. Though practically infeasible, this approach introduces interesting connections to integer programming as well, opening other potential solution avenues.

Finally, in Chapter 4 we show an improved bound for the vector balancing constant of a zonotope with itself, as well as for two distinct zonotopes. In order to prove this result, we also prove a Gaussian measure lower bound on sections of normalized (in a precise sense) zonotopes, which is of independent interest. The main remaining open question in this area is to prove the tight $O(\sqrt{d})$ bound conjectured by Schechtman [60]; as mentioned in Section 4.1, this would follow from showing that any d-dimensional zonotope can be approximated up to a constant factor by a zonotope with $O(d)$ segments.

Using polar convex bodies, and noting that the polar body of a zonotope $A^\top B_\infty^m \subseteq \mathbb{R}^d$ is the preimage $A^{-1}(B_1^m) := \{x \in \mathbb{R}^d : \|Ax\|_1 \leq 1\}$, we can equivalently restate this sparsification question as follows:

Question 5.0.1. *Does there exist a universal constant $C > 0$ such that given any matrix $A \in \mathbb{R}^{m \times d}$ with $m \geq d$ and $0 < \varepsilon \leq \frac{1}{2}$, one can always find another matrix $\tilde{A} \in \mathbb{R}^{Cd/\varepsilon^2 \times d}$*

with $\|\tilde{A}x\|_1 \leq \|Ax\|_1 \leq (1+\varepsilon)\|\tilde{A}x\|_1$ *for all* $x \in \mathbb{R}^d$*?*

If one replaces the ℓ_1 norm by the ℓ_2 norm, a similar statement directly follows from the existence of linear-size spectral sparsifiers [12]. In that setting, each row of $\tilde{A}$ is a scalar multiple of a row of A, and it is reasonable to hope that a different rescaling of the rows may work for the ℓ_1 norm. Finally, the following question relates to the existence of good partial colorings in the ℓ_1 norm.

Question 5.0.2. *Given any matrix* $A \in \mathbb{R}^{m\times d}$*, does the set*

$$K := \left\{ x \in \mathbb{R}^m : \left| \sum_{i=1}^m x_i |\langle A_i, z\rangle| \right| \leq \sqrt{\tfrac{d}{m}} \|Az\|_1 \ \forall z \in \mathbb{R}^d \right\}$$

have large Gaussian measure $\gamma_m(K) \geq e^{-Cm}$ *where* $C > 0$ *is a universal constant?*

BIBLIOGRAPHY

[1] Noga Alon and Joel H. Spencer. *The Probabilistic Method*. Wiley Publishing, 4th edition, 2016.

[2] Gergely Ambrus and Rainie Bozzai. Colourful vector balancing. *Mathematika*, 70(4), August 2024.

[3] Gergely Ambrus and Bernardo González Merino. Large signed subset sums. *Mathematika*, 67(3):579–595, 2021.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1355–1362, Madison, WI, USA, 2012. Omnipress.

[6] Arturs Backurs, Jerry Chee, Sivakanth Gopi, Laurel Heck, Janardhan Kulkarni, and Thomas Rothvoss. Discquant: A quantization method for neural networks inspired by discrepancy theory. *Submitted*, 2024.

[7] W. Banaszczyk. Balancing vectors and Gaussian measures of n-dimensional convex bodies. *Random Struct. Algorithms*, 12(4):351–360, 1998.

[8] Wojciech Banaszczyk. Balancing vectors and convex bodies. *Studia Mathematica*, 106(1):93–100, 1993.

[9] N. Bansal, D. Dadush, S. Garg, and S. Lovett. The Gram-Schmidt walk: a cure for the Banaszczyk blues. In *STOC*, pages 587–597. ACM, 2018.

[10] Nikhil Bansal. Constructive algorithms for discrepancy minimization. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 3–10, 2010.

[11] 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*, STOC 2018, page 587–597, New York, NY, USA, 2018. Association for Computing Machinery.

[12] J. Batson, D. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 255–262, 2009.

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

[14] József Beck and Tibor Fiala. "Integer-making" theorems. *Discrete Applied Mathematics*, 3(1):1–8, 1981.

[15] Jon Louis Bentley and James B Saxe. Decomposable searching problems i. static-to-dynamic transformation. *Journal of Algorithms*, 1(4):301–358, 1980.

[16] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Mathematica*, 162(none):73 – 141, 1989.

[17] Rainie Bozzai and Thomas Rothvoss. Stronger coreset bounds for kernel density estimators via chaining. *Submitted*, 2023.

[18] B. Bukh. An improvement of the Beck-Fiala theorem. *Combinatorics, Probability and Computing*, 25(3):380–398, 2016.

[19] Imre Bárány. On the power of linear dependencies. In *Building bridges: Between Mathematics and Computer Science*, volume 19 of *Bolyai Society Mathematical Studies*, pages 31–45. Bolyai Society and Springer, 2010.

[20] Bernard Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, 2000.

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

[22] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 109–116, Arlington, Virginia, USA, 2010. AUAI Press.

[23] M. B. Cohen and R. Peng. ℓ_p row sampling by Lewis weights. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 183–192, New York, NY, USA, 2015. Association for Computing Machinery.

[24] Efren Cruz and Clayton Scott. Sparse approximation of a kernel mean. *IEEE Transactions on Signal Processing*, PP, 03 2015.

[25] D. Dadush, H. Jiang, and V. Reis. A new framework for matrix discrepancy: partial coloring bounds via mirror descent. In *STOC*, pages 649–658. ACM, 2022.

[26] Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *Forty-First International Conference on Machine Learning*, 2024.

[27] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301 – 2339, 2014.

[28] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[29] Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, 2022.

[30] Apostolos A Giannopoulos. On some vector balancing problems. *Studia Mathematica*, 122:225–234, 1997.

[31] Victor Grinberg and Sergey Sevast'yanov. Value of the steinitz constant. *Functional Analysis and its Applications*, 14:125–126, 1980.

[32] Nick Harvey and Samira Samadi. Near-optimal herding. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1165–1182, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[33] Babak Hassibi, Daivd G Stork, and Gregory J Wolff. optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, 1993.

[34] Laurel Heck, Victor Reis, and Thomas Rothvoss. The vector balancing constant for zonotopes. *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1292–1300, 2022.

[35] R. Hoberg and T. Rothvoss. A logarithmic additive integrality gap for bin packing. In *SODA*, pages 2616–2625. SIAM, 2017.

[36] S. B. Hopkins, P. Raghavendra, and A. Shetty. Matrix discrepancy from quantum communication. STOC 2022, pages 637–648, New York, NY, USA, 2022. Association for Computing Machinery.

[37] 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*, SoCG '11, page 47–56, New York, NY, USA, 2011. Association for Computing Machinery.

[38] 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*, SoCG '11, page 47–56, New York, NY, USA, 2011. Association for Computing Machinery.

[39] Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1975–1993. PMLR, 25–28 Jun 2019.

[40] Janardhan Kulkarni, Victor Reis, and Thomas Rothvoss. Optimal online discrepancy minimization. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, page 1832–1840, New York, NY, USA, 2024. Association for Computing Machinery.

[41] Simon Lacoste-Julien, Fredrik Lindsten, and Francis R. Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. *ArXiv*, abs/1501.02056, 2015.

[42] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[43] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1452–1461. JMLR.org, 2015.

[44] Shachar Lovett and Raghu Meka. Constructive discrepancy minimization by walking on the edges. In *FOCS*, pages 61–67. IEEE Computer Society, 2012.

[45] L. Lovász, J. Spencer, and K. Vesztergombi. Discrepancy of set-systems and matrices. *European Journal of Combinatorics*, 7(2):151–160, 1986.

[46] Ben Lund and Alexander Magazinov. The sign-sequence constant of the plane. *Acta Mathematica Hungarica*, 151:117–123, 2017.

[47] Eric Lybrand and Rayan Saab. A greedy algorithm for quantizing neural networks. *Journal of Machine Learning Research*, 22(156):1–38, 2021.

[48] A. W. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families ii: Mixed characteristic polynomials and the Kadison-Singer problem. *Annals of Mathematics*, 182(1):327–350, 2015.

[49] Jiri Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 1999.

[50] R. Meka. Discrepancy and beating the union bound (blog post), 2014.

[51] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. 2017.

[52] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7197–7206. PMLR, 13–18 Jul 2020.

[53] A. Nikolov, K. Talwar, and L. Zhang. The geometry of differential privacy: the sparse and approximate cases. In *STOC*, pages 351–360. ACM, 2013.

[54] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[55] Jeff Phillips and Wai Tai. Near-optimal coresets of kernel density estimates. *Discrete and Computational Geometry*, 63, 06 2020.

[56] Jeff M. Phillips. Algorithms for ε-approximations of terrains. In *Proceedings of the 35th International Colloquium on Automata, Languages and Programming - Volume Part I*, ICALP '08, page 447–458, Berlin, Heidelberg, 2008. Springer-Verlag.

[57] Jeff M. Phillips. ε-samples for Kernels. In *Proceedings of the Twenty-Fourth Annual Symposium on Discrete Algorithms*, pages 1622–1632. SIAM, 2013.

[58] Victor Reis and Thomas Rothvoss. Vector balancing in lebesgue spaces. *Random Structures & Algorithms*, 62:667 – 688, 2020.

[59] T. Royen. A simple proof of the gaussian correlation conjecture extended to multivariate gamma distributions. *arXiv: Probability*, 2014.

[60] Gideon Schectman. Fourier analytic methods in convex geometry. Workshop by the American Institute of Mathematics, 2007.

[61] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. A Wiley-interscience publication. Wiley, 1992.

[62] Sergey V. Sevast'yanov. On the approximate solution of the problem of calendar planning. *Upravlyaemye Systemy*, 20:49–63, 1980. (in Russian).

[63] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[64] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[65] Le Song, Xinhua Zhang, Alex Smola, Arthur Gretton, and Bernhard Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 992–999, New York, NY, USA, 2008. Association for Computing Machinery.

[66] J. Spencer. Six standard deviations suffice. *Trans. Amer. Math. Soc.*, 289(2):679–706, 1985.

[67] Joel Spencer. Balancing games. *Journal of Combinatorial Theory, Series B*, 23(1):68–74, 1977.

[68] Joel Spencer. Balancing unit vectors. *Journal of Combinatorial Theory*, 30:349–350, 1981.

[69] Joel Spencer. Balancing vectors in the max norm. *Combinatorica*, 6:55–65, 1986.

[70] Konrad Swanepoel. Balancing unit vectors. *Journal of Combinatorial Theory*, 89:105–112, 2000.

[71] S. J. Szarek and E. Werner. A nonsymmetric correlation inequality for gaussian measure. *Journal of Multivariate Analysis*, 68(2):193–211, 1999.

[72] Wai Ming Tai. Optimal Coreset for Gaussian Kernel Density Estimation. In Xavier Goaoc and Michael Kerber, editors, *38th International Symposium on Computational Geometry (SoCG 2022)*, volume 224 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 63:1–63:15, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[73] Michel Talagrand. Embedding subspaces of l_1 into ℓ_1^n. *Proceedings of the American Mathematical Society*, 108(2):363–369, 1990.

[74] T. Tkocz. *High-dimensional Phenomena: Dilations, Tensor Products and Geometry of L_1*. University of Warwick, 2015.

[75] Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better llm quantization with hadamard incoherence and lattice codebooks. In *Forty-First International Conference on Machine Learning*, 2024.

[76] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008.

[77] J. D. Vaaler. A geometric inequality with applications to linear forms. *Pacific Journal of Mathematics*, 83(2):543 – 553, 1979.

[78] Mart van Baalen, Andrey Kuzmin, Markus Nagel, Peter Couperus, Cedric Bastoul, Eric Mahurin, Tijmen Blankevoort, and Paul Whatmough. Gptvq: The blessing of dimensionality in llm quantization. *arXiv preprint arXiv:2402.15319*, 2024.

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

[80] A. Zouzias. A matrix hyperbolic cosine algorithm and applications. In Artur Czumaj, Kurt Mehlhorn, Andrew Pitts, and Roger Wattenhofer, editors, *Automata, Languages, and Programming*, pages 846–858, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

Appendix A

SUBGAUSSIAN RANDOM VARIABLES AND CHAINING

We begin by giving one of several equivalent definitions of subgaussian random variables, as well as the definition of the *subgaussian norm*, which we will need in Chapter 2 (see [79] for more details).

Definition A.0.1. *A random variable X is* K-subgaussian *if the tails of X satisfy*

$$\mathbb{P}[|X| \geq t] \leq 2\exp(-t^2/K^2) \quad \forall t \geq 0.$$

The subgaussian norm of X is then

$$\|X\|_{\psi_2} := \inf\left\{s > 0 : \ \mathbb{P}[|X| \geq t] \leq 2\exp(-t^2/s^2) \quad \forall t \geq 0\right\}.$$

Using the subgaussian norm, we can formulate the following version of the algorithmic version of Theorem 1.2.3). Due to its reliance on Gram-Schmidt normalization, the algorithm is called the *Gram-Schmidt Walk*.

Theorem A.0.1 (Gram-Schmidt Walk [11])**.** *There is a polynomial-time randomized algorithm that takes as input vectors $v_1, \ldots, v_n \in \mathbb{R}^m$ of ℓ_2 norm at most 1 and outputs random signs $\beta \in \{\pm 1\}^n$ such that the (mean zero) random variable $\sum_{i=1}^n \beta_i v_i$ is $O(1)$-subgaussian.*

This is the formulation that Dadush et. al. used to prove the algorithmic version of Banaszczyk's theorem, and it is also the version that will be more useful for our proofs.

The subgaussian norm will also be essential to the *chaining*-based approach that we develop in Chapter 2. The term chaining refers to a technique for creating a multi-layer ε-net, often allowing one to obtain better bounds than would be obtained by taking an ε-net over the entire space. For more details see [79].

To develop the key results related to this technique, we will need the following definitions.

Definition A.0.2. *Given a (pseudo)metric space (T, d) and $r > 0$:*

- $B_d(s, r) = \{t \in T : \ d(s, t) \leq r\}$;
- $\operatorname{diam}(d) := \sup_{t,s \in T} d(t, s)$;
- $\mathcal{N}(T, d, r)$ *is the size of a minimal r-cover of T w.r.t. d, i.e.*

$$\mathcal{N}(T, d, r) = \min \Big\{ \big|S\big| : \ S \subseteq T, \ T \subseteq \bigcup_{s \in S} B_d(s, r) \Big\}.$$

The key results regarding chaining are captured in the following collection of inequalities. The multi-step ε-net construuction can be seen by the fact that we are in essence integrating over the size of the ε-net, and thus able to appropriately weight the contributions accrued for different mesh sizes.

Theorem A.0.2 (Dudley's Integral Inequality)**.** *Let $(X_t)_{t \in T}$ be a random process on a pseudometric space (T, d) satisfying*

$$\|X_t - X_s\|_{\psi_2} \leq d(t, s) \quad \forall t, s \in T.$$

Then for any $t_0 \in T$,

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| \lesssim \int_0^{\operatorname{diam}(d)} \sqrt{\log \mathcal{N}(T, d, r)} \ \mathrm{d}r.$$

For our application we need to control the absolute value, which can be done as follows:

Theorem A.0.3 (Dudley's Integral Inequality II)**.** *Let $(X_t)_{t \in T}$ be a random process on a pseudometric space (T, d) satisfying*

$$\|X_t - X_s\|_{\psi_2} \leq d(t, s) \quad \forall t, s \in T.$$

Then for any $t_0 \in T$,

$$\mathbb{E} \sup_{t \in T} |X_t| \lesssim \int_0^{\operatorname{diam}(d)} \sqrt{\log \mathcal{N}(T, d, r)} \ \mathrm{d}r + \|X_{t_0}\|_{\psi_2}.$$

Proof. By the triangle inequality $\mathbb{E}\sup_{t\in T}|X_t| \leq \mathbb{E}\sup_{t\in T}|X_t - X_{t_0}| + \mathbb{E}[|X_{t_0}|]$. The claim then follows from Theorem A.0.2 and the fact that $\mathbb{E}[|X_{t_0}|] \lesssim \|X_{t_0}\|_{\psi_2}$. □

There is also a concentration version of this inequality.

Theorem A.0.4 (Dudley's Concentration Inequality)**.** *Let $(X_t)_{t\in T}$ be a random process on a pseudometric space (T, d) satisfying*

$$\|X_t - X_s\|_{\psi_2} \leq d(t, s) \ \forall t, s \in T.$$

Then, for every $u \geq 0$, the event

$$\sup_{t,s\in T}|X_t - X_s| \leq C\left[\int_0^{\operatorname{diam}(T)} \sqrt{\log \mathcal{N}(T, d, r)}\ \mathrm{d}r \ + \ u \cdot \operatorname{diam}(T)\right]$$

holds with probability at least $1 - 2\exp(-u^2)$, where $C > 0$ is a universal constant.