Two proofs for Shallow Packings

Kunal Dutta, Esther Ezra, Arijit Ghosh

Two proofs for Shallow Packings*

Kunal Dutta † Esther Ezra ‡ Arijit Ghosh *

August 8, 2016

Abstract

We refine the bound on the packing number, originally shown by Haussler, for shallow geometric set systems. Specifically, let $\mathcal{V}$ be a finite set system defined over an n-point set X; we view $\mathcal{V}$ as a set of indicator vectors over the n-dimensional unit cube. A δ-separated set of $\mathcal{V}$ is a subcollection $\mathcal{W}$, s.t. the Hamming distance between each pair $\mathbf{u}, \mathbf{v} \in \mathcal{W}$ is greater than δ, where $\delta > 0$ is an integer parameter. The δ-packing number is then defined as the cardinality of a largest δ-separated subcollection of $\mathcal{V}$. Haussler showed an asymptotically tight bound of $\Theta((n/\delta)^d)$ on the δ-packing number if $\mathcal{V}$ has VC-dimension (or *primal shatter dimension*) d. We refine this bound for the scenario where, for any subset, $X' \subseteq X$ of size $m \leq n$ and for any parameter $1 \leq k \leq m$, the number of vectors of length at most k in the restriction of $\mathcal{V}$ to X' is only $O(m^{d_1} k^{d-d_1})$, for a fixed integer $d > 0$ and a real parameter $1 \leq d_1 \leq d$ (this generalizes the standard notion of *bounded primal shatter dimension* when $d_1 = d$). In this case when $\mathcal{V}$ is "k-shallow" (all vector lengths are at most k), we show that its δ-packing number is $O(n^{d_1} k^{d-d_1}/\delta^d)$, matching Haussler's bound for the special cases where $d_1 = d$ or $k = n$. We present two proofs, the first is an extension of Haussler's approach, and the second extends the proof of Chazelle, originally presented as a simplification for Haussler's proof.

1 Introduction

Let $\mathcal{V}$ be a set system defined over an n-point set X. We follow the notation in [Hau95], and view $\mathcal{V}$ as a set of indicator vectors in $\mathbb{R}^n$, that is, $\mathcal{V} \subseteq \{0,1\}^n$. Given a subsequence of indices (coordinates) $I = (i_1, \ldots, i_k)$, $1 \leq i_j \leq n$, $k \leq n$, the *projection* $\mathcal{V}_{|_I}$ of $\mathcal{V}$ onto I (also referred to as the *restriction* of $\mathcal{V}$ to I) is defined as

$$\mathcal{V}_{|_I} = \{(\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_k}) \mid \mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \in \mathcal{V}\}.$$

With a slight abuse of notation we write $I \subseteq [n]$ to state the fact that I is a subsequence of indices as above. We now recall the definition of the primal shatter function of $\mathcal{V}$:

*Work on this paper by Kunal Dutta and Arijit Ghosh has been supported by the Indo-German Max-Planck Center for Computer Science (IMPECS). Work on this paper by Esther Ezra has been supported by NSF Grants CCF-11-17336, CCF-12-16689, and NSF CAREER CCF-15-53354. A preliminary version of this paper appeared in *Proc. Sympos. Computational Geometry*, 2015, pp. 96-110 [DEG15].

†D1: Algorithms & Complexity, Max-Planck-Institut für Informatik, 66123 Saarbrücken, Germany {kdutta,agosh}@mpi-inf.mpg.de

‡School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA eezra3@math.gatech.edu

Definition 1 (Primal Shatter Function [HW87, Mat99]) *The* primal shatter function *of* $\mathcal{V} \subseteq \{0,1\}^n$ *is a function, denoted by* $\pi_{\mathcal{V}}$*, whose value at* m *is defined by* $\pi_{\mathcal{V}}(m) = \max_{I \subseteq [n], |I|=m} |\mathcal{V}_{|_I}|$*. In other words,* $\pi_{\mathcal{V}}(m)$ *is the maximum possible number of distinct vectors of* $\mathcal{V}$ *when projected onto a subsequence of* m *indices.*

From now on we say that $\mathcal{V} \subseteq \{0,1\}^n$ has *primal shatter dimension* d if $\pi_{\mathcal{V}}(m) \leq Cm^d$, for all $m \leq n$, where $d > 1$ and $C > 0$ are constants. A notion closely related to the primal shatter dimension is that of the *VC-dimension*:

Definition 2 (VC-dimension [Hau95, VC71]) *An index sequence* $I = (i_1, \ldots, i_k)$ *is* shattered *by* $\mathcal{V}$ *if* $\mathcal{V}_{|_I} = \{0,1\}^k$*. The* VC-dimension *of* $\mathcal{V}$*, denoted by* d_0 *is the size of the longest sequence* I *shattered by* $\mathcal{V}$*. That is,* $d_0 = \max\{k \mid \exists I = (i_1, i_2, \ldots, i_k), 1 \leq i_j \leq n, \text{ with } \mathcal{V}_{|_I} = \{0,1\}^k\}$.

The notions of primal shatter dimension and VC-dimension are interrelated. By the Sauer-Shelah Lemma (see [Sau72, She72] and the discussion below) the VC-dimension of a set system $\mathcal{V}$ always bounds its primal shatter dimension, that is, $d \leq d_0$. On the other hand, when the primal shatter dimension is bounded by d, the VC-dimension d_0 does not exceed $O(d \log d)$ (which is straightforward by definition, see, e.g., [HP11]).

A typical family of set systems that arise in geometry with bounded primal shatter (resp., VC-) dimension consists of set systems defined over points in some low-dimensional space $\mathbb{R}^d$, where $\mathcal{V}$ represents a collection of certain simply-shaped regions, e.g., halfspaces, balls, or simplices in $\mathbb{R}^d$. In such cases, the primal shatter (and VC-) dimension is a function of d; see, e.g., [HP11] for more details. When we flip the roles of points and regions, we obtain the so-called *dual set systems* (where we refer to the former as *primal set systems*). In this case, the ground set is a collection $\mathcal{S}$ of algebraic surfaces in $\mathbb{R}^d$, and $\mathcal{V}$ corresponds to faces of all dimensions in the *arrangement* $\mathcal{A}(\mathcal{S})$ of $\mathcal{S}$, that is, this is the decomposition of $\mathbb{R}^d$ into connected open *cells* of dimensions $0, 1, \ldots, d$ induced by $\mathcal{S}$. Each cell is a maximal connected region that is contained in the intersection of a fixed number of the surfaces and avoids all other surfaces; in particular, the 0-dimensional cells of $\mathcal{A}(\mathcal{S})$ are called "vertices", and d-dimensional cells are simply referred to as "cells"; see [SA95] for more details. The distinction between primal and dual set systems in geometry is essential, and set systems of both kinds appear in numerous geometric applications, see, once again [HP11] and the references therein.

1.1 δ-packing

The *length* $\|\mathbf{v}\|$ of a vector $\mathbf{v} \in \mathcal{V}$ under the L^1 norm is defined as $\sum_{i=1}^n |\mathbf{v}_i|$, where $\mathbf{v}_i$ is the ith coordinate of $\mathbf{v}$, $i = 1, \ldots, n$. The *distance* $\rho(\mathbf{u}, \mathbf{v})$ between a pair of vectors $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ is defined as the L^1 norm of the difference $\mathbf{u} - \mathbf{v}$, that is, $\rho(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^n |\mathbf{u}_i - \mathbf{v}_i|$. In other words, it is the *symmetric difference distance* between the corresponding sets represented by $\mathbf{u}$, $\mathbf{v}$. [1]

Let $\delta > 0$ be an integer parameter. We say that a subset of vectors $\mathcal{W} \subseteq \{0,1\}^n$ is *δ-separated* if for each pair $\mathbf{u}, \mathbf{v} \in \mathcal{W}$, $\rho(\mathbf{u}, \mathbf{v}) > \delta$. The *$\delta$-packing number* for $\mathcal{V}$, denote it by $\mathcal{M}(\delta, \mathcal{V})$, is then defined as the cardinality of a largest δ-separated subset $\mathcal{W} \subseteq \mathcal{V}$. A key property, originally

[1] The symmetric difference distance between two sets A, B is the cardinality of their symmetric difference.

shown by Haussler [Hau95] (see also [Cha92, CW89, Dud78, Mat99, Wel92]), is that set systems of bounded primal shatter dimension admit small δ-packing numbers. That is:

Theorem 3 (Packing Lemma [Hau95, Mat99]) *Let $\mathcal{V} \subseteq \{0,1\}^n$ be a set of indicator vectors of primal shatter dimension d, and let $1 \leq \delta \leq n$ be an integer parameter. Then $\mathcal{M}(\delta, \mathcal{V}) = O((n/\delta)^d)$, where the constant of proportionality depends on d.*

We note that in the original formulation in [Hau95] the assumption is that the set system has a finite VC-dimension. However, its formulation in [Mat99], which is based on a simplification of the analysis of Haussler by Chazelle [Cha92], relies on the assumption that the primal shatter dimension is d, which is the actual bound that we state in Theorem 3 (we comment though that one component of the analysis uses the fact that the VC-dimension d_0 is finite, but this follows from the bound $O(d \log d)$ on d_0). We also comment that a closer inspection of the analysis in [Hau95] shows that this assumption can be replaced with that of having bounded primal shatter dimension (independent of the analysis in [Cha92]). We describe these considerations in Section 2.1.

1.2 Previous work

In his seminal work, Dudley [Dud78] presented the first application of *chaining*, a proof technique due to Kolmogorov, to empirical process theory, where he showed the bound $O((n/\delta)^{d_0} \log^{d_0}(n/\delta))$ on $\mathcal{M}(\delta, \mathcal{V})$, with a constant of proportionality depending on the VC-dimension d_0 (see also previous work by Haussler [Hau92] and Pollard [Pol84] for an alternative proof and a specification of the constant of proportionality). This bound was later improved by Haussler [Hau95], who showed $\mathcal{M}(\delta, \mathcal{V}) \leq e(d_0+1)\left(\frac{2en}{\delta}\right)^{d_0}$, where e is the base of the natural logarithm (see also Theorem 3), and presented a matching lower bound, which leaves only a constant factor gap, which depends exponentially in d_0. In fact, the aforementioned bounds are more general, and can also be applied to classes of real-valued functions of finite "pseudo-dimension" (the special case of set systems corresponds to Boolean functions), see, e.g., [Hau92], however, we do not discuss this generalization in this paper and focus merely on set systems $\mathcal{V}$ of finite primal shatter (resp., VC-) dimension.

The bound of Haussler [Hau95] (Theorem 3) is in fact a generalization of the so-called Sauer-Shelah Lemma [Sau72, She72], asserting that $|\mathcal{V}| \leq (en/d_0)^{d_0}$, and thus this bound is $O(n^{d_0})$. Indeed, when $\delta = 1$, the corresponding δ-separated set should include all vectors in $\mathcal{V}$, and then the bound of Haussler [Hau95] becomes $O(n^{d_0})$, matching the Sauer-Shelah bound up to a constant factor that depends on d_0.

There have been several studies extending Haussler's bound or improving it in some special scenarios. We name only a few of them. Gottlieb *et al.* [GKM12] presented a sharpening of this bound when δ is relatively large, i.e., δ is close to $n/2$, in which case the vectors are "nearly orthogonal". They also presented a tighter lower bound, which considerably simplifies the analysis of Bshouty *et al.* [BLL09], who achieved the same bound.

A major application of packing is in obtaining improved bounds on the *sample complexity* in machine learning. This was studied by Li *et al.* [LLS01] (see also [Hau92]), who presented an asymptotically tight bound on the sample complexity, in order to guarantee a small "relative error." This problem has been revisited by Har-Peled and Sharir [HPS11] in the context of geometric set sys-

Figure 1: A grid of $\frac{n}{\delta} \times \frac{n}{\delta}$ axis-parallel rectangles, each of which with multiplicity $\delta/2$. The multiplicity in the figure is depicted only for the leftmost vertical and the top horizontal rectangles. The small shaded rectangle in the figure is a δ-shallow cell of the arrangement.

tems, where they referred to a sample of the above kind as a "relative approximation", and showed how to integrate it into an *approximate range counting* machinery, which is a central application in computational geometry. The packing number has also been used by Welzl [Wel92] in order to construct spanning trees of low crossing number (see also [Mat99]) and by Matoušek [Mat95, Mat99] in order to obtain asymptotically tight bounds in geometric discrepancy.

1.3 Our Result

Shallow packing lemma. In the sequel, we refine the bound in the Packing Lemma (Theorem 3) so that it becomes sensitive to the length of the vectors $\mathbf{v} \in \mathcal{V}$, based on an appropriate refinement of the underlying primal shatter function. This refinement has several geometric applications. Our ultimate goal is to show that when the set system is "shallow" (that is, the underlying vectors are short), the packing number becomes much smaller than the bound in Theorem 3.

Nevertheless, we cannot always enforce such an improvement, as in some settings the worst-case asymptotic bound on the packing number is $\Omega((n/\delta)^d)$ even when the set system is shallow. See Figure 1 and the paragraph at the end of this section where we give a more detailed description of this construction to the non-expert reader.

Therefore, in order to obtain an improvement on the packing number of shallow set systems, we may need further assumptions on the primal shatter function. Such assumptions stem from the random sampling technique of Clarkson and Shor [CS89], which we define as follows. Let $\mathcal{V}$ be our set system. We assume that for any sequence I of $m \leq n$ indices, and for any parameter $1 \leq k \leq m$, the number of vectors in $\mathcal{V}_{|_I}$ of length at most k is only $O(m^{d_1} k^{d-d_1})$, where d is the primal shatter dimension and $1 \leq d_1 \leq d$ is a real parameter.[2] When $k = m$ we obtain $O(m^d)$ vectors in total, in accordance with the assumption that the primal shatter dimension is d, but the above bound is also sensitive to the length of the vectors as long as $d_1 < d$. From now on, we say that a primal shatter function of this kind has the (d, d_1) *Clarkson-Shor property.*

Let us now denote by $\mathcal{M}(\delta, k, \mathcal{V})$ the δ-packing number of $\mathcal{V}$, where the vector length of each element in $\mathcal{V}$ is at most k, for some integer parameter $1 \leq k \leq n$. With this notation we can assume, without loss of generality, that $k \geq \delta/2$, as otherwise the distance between any two elements in $\mathcal{V}$ must be strictly less than δ, in which case the packing contains at most one element. In Sections 2– 3 we present two proofs for our main result, stated below:

Theorem 4 (Shallow Packing Lemma) *Let $\mathcal{V} \subseteq \{0,1\}^n$ be a set of indicator vectors, whose primal shatter function has a (d, d_1) Clarkson-Shor property (where d is the primal shatter dimension and $1 \leq d_1 \leq d$ is a real parameter). Let $\delta \geq 1$ be an integer parameter, and k an integer*

[2] We ignore the cases where $d_1 < 1$, as it does not seem to appear in natural set systems—see below.

parameter between 1 *and* n*, and suppose that* $k \geq \delta/2$*. Then:*

$$\mathcal{M}(\delta, k, \mathcal{V}) = O\left(\frac{n^{d_1} k^{d-d_1}}{\delta^d}\right),$$

where the constant of proportionality depends on d*.*

The actual dependence of the constant of proportionality on d is analyzed at end of Section 2 for the first proof, and end of Section 3 for the second proof.

This problem has initially been addressed by the second author in [Ezr16] as a major tool to obtain size-sensitive discrepancy bounds in set systems of this kind, where it has been shown $\mathcal{M}(\delta, k, \mathcal{V}) = O\left(\frac{n^{d_1} k^{d-d_1} \log^d(n/\delta)}{\delta^d}\right)$. The analysis in [Ezr16] is a refinement over the technique of Dudley [Dud78] combined with the existence of small-size *relative approximations* (see [Ezr16] for more details). In the current analysis we completely remove the extra $\log^d(n/\delta)$ factor appearing in the previous bound. In particular, when $d_1 = d$ (where we just have the original assumption on the primal shatter function) or $k = n$ (in which case each vector in $\mathcal{V}$ has an arbitrary length), our bound matches the tight bound of Haussler, and thus appears as a generalization of the Packing Lemma (when replacing VC-dimension by primal shatter dimension). We present two proofs for Theorem 4, the first is an extension of Haussler's approach (Section 2), and the second is an extension of Chazelle's proof [Cha92] to the Packing Lemma (Section 3).

We note that after the submission of this paper we found a third proof by Mustafa [Mus16], based on the simplification of Chazelle's proof [Cha92] and combining Markov's inequality. While we find this proof simpler than ours, we emphasize the fact that the problem of shallow packings has initially been posed by the second author in [Ezr16], and the preliminary version of this paper [DEG15] is the first to resolve it (in the sense of Theorem 4). We thus believe the results in [Mus16] were somewhat inspired by the preliminary version in [DEG15] .

Applications. Theorem 4 implies smaller packing numbers for several natural geometric set systems under the shallowness assumption. These set systems are described in detail in Section 4.1. In Section 4.2 we show an application in geometric discrepancy, where the goal is to obtain discrepancy bounds that are sensitive to the length of the vectors in $\mathcal{V}$. Due to the bound in Theorem 4 we obtain an improvement over the one presented in [Ezr16].

An example of a shallow set system with large packing numbers Consider a ground set that is a collection of axis-parallel rectangles, and the vectors $\mathcal{V}$ represent subsets of rectangles that cover a point in the plane. For simplicity of exposition, we define these vectors to represent all two-dimensional cells in the arrangement of the given rectangles. In this case, the (d, d_1) Clarkson-Shor property holds only for $d = d_1 = 2$. Indeed, even when the arrangement is very shallow (say, each point in the plane is covered by at most two rectangles), one can still have quadratically many cells by just taking a grid of long and skinny rectangles, and thus in general we cannot obtain size-sensitive bounds on $|\mathcal{V}|$.

It is well known that the primal shatter function of $\mathcal{V}$ is quadratic (see, e.g., [HP11]), and therefore, by Theorem 3, the packing number is $O((n/\delta)^2)$. Nevertheless, we claim that even when

the arrangement is shallow, the asymptotic bound on the packing number is not any better than $(n/\delta)^2$. Indeed, fix a positive even parameter $\delta > 0$, and suppose, without loss of generality, that n/δ is an integer number. Consider now an $\frac{n}{\delta} \times \frac{n}{\delta}$ grid of long and skinny rectangles, where each rectangle in the grid is duplicated $\delta/2$ times (with a possibly infinitesimal perturbation), as illustrated in Figure 1. Clearly, each (two-dimensional) cell in the arrangement is covered by at most δ rectangles, and thus the set system is δ-shallow. Consider now only the cells at "depth" δ (that is, they are covered by precisely δ rectangles), and let $\mathcal{F} \subset \mathcal{V}$ be the set of their representative vectors. It is easy to verify that, for each pair $v, v' \in \mathcal{F}$, the distance $\rho(v, v')$ is at least δ (see once again Figure 1), and thus $\mathcal{F}$ is both δ-separated and δ-shallow. However, by construction, we have $|\mathcal{F}| = \Omega((n/\delta)^2)$, and thus the δ-shallowness assumption does not yield an improvement over the general case.

2 First Proof: Refining Haussler's Approach

2.1 Preliminaries

Relative approximations and ε-nets. We mentioned in the introduction the notion of *relative* (ε, η)*-approximations*. We now define them formally: Following the definition from [HPS11], given a set system $\mathcal{V} \subseteq \{0,1\}^n$ and two parameters, $0 < \varepsilon < 1$ and $0 < \eta < 1$, we say that a subsequence I of indices is a *relative* (ε, η)*-approximation* if it satisfies, for each vector $\mathbf{v} \in \mathcal{V}$,

$$\left| \frac{\|\mathbf{v}_{|_I}\|}{|I|} - \frac{\|\mathbf{v}\|}{n} \right| \leq \eta \frac{\|\mathbf{v}\|}{n}, \quad \text{if } \frac{\|\mathbf{v}\|}{n} \geq \varepsilon, \quad \text{and}$$

$$\left| \frac{\|\mathbf{v}_{|_I}\|}{|I|} - \frac{\|\mathbf{v}\|}{n} \right| \leq \eta\varepsilon, \quad \text{otherwise,}$$

where $\mathbf{v}_{|_I}$ is the projection of $\mathbf{v} \in \mathcal{V}$ onto I.

As observed by Har-Peled and Sharir [HPS11], the analysis of Li *et al.* [LLS01] implies that if $\mathcal{V}$ has primal shatter dimension d, then a random sample of $\frac{cd\log(1/\varepsilon)}{\varepsilon\eta^2}$ indices (each of which is drawn independently) is a relative (ε, η)-approximation for $\mathcal{V}$ with constant probability, where $c > 0$ is an absolute constant. More specifically, success with probability at least $1 - q$ is guaranteed if one samples $\frac{c(d\log(1/\varepsilon)+\log(1/q))}{\varepsilon\eta^2}$ indices.[3]

It was also observed in [HPS11] that ε*-nets* arise as a special case of relative (ε, η)-approximations. Specifically, an ε-net is a subsequence of indices I with the property that any vector $\mathbf{v} \in \mathcal{V}$ with $\|\mathbf{v}\| \geq n\varepsilon$ satisfies $\|\mathbf{v}_{|_I}\| \geq 1$. In other words, N is a hitting set for all the "long" vectors. In this case, if we set η to be some constant fraction, say, $1/4$, then a relative $(\varepsilon, 1/4)$-approximation becomes an ε-net. Moreover, a random sample of $O\left(\frac{d\log(1/\varepsilon)+\log(1/q)}{\varepsilon}\right)$ indices (with an appropriate choice of the constant of proportionality) is an ε-net for $\mathcal{V}$, with probability at least $1 - q$; see [HPS11] for further details.

[3] We note that although in the original analysis for this bound d is the VC-dimension, this assumption can be replaced by having just a primal shatter dimension d; see, e.g., [HP11] for the details of the analysis.

Overview of Haussler's approach. For the sake of completeness, we repeat some of the details in the analysis of Haussler [Hau95] and use similar notation for ease of presentation.

Let $\mathcal{V} \subseteq \{0,1\}^n$ be a collection of indicator vectors of bounded primal shatter dimension d, and denote its VC-dimension by d_0. By the discussion above, $d_0 = O(d \log d)$. From now on we assume that $\mathcal{V}$ is δ-separated, and thus a bound on $|\mathcal{V}|$ is also a bound on the packing number of $\mathcal{V}$. The analysis in [Hau95] exploits the method of "conditional variance" in order to conclude

$$|\mathcal{V}| \leq (d_0+1)\,\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right] = O\left(d \log d\,\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right]\right), \tag{1}$$

where $\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right]$ is the expected size of $\mathcal{V}$ when projected onto a subset $I = \{i_1, \ldots, i_{m-1}\}$ of $m-1$ indices chosen uniformly at random without replacements from $[n]$, and

$$m := \left\lceil \frac{(2d_0+2)(n+1)}{\delta + 2d_0 + 2} \right\rceil = O\left(\frac{d_0 n}{\delta}\right) = O\left(\frac{nd \log d}{\delta}\right). \tag{2}$$

We justify this choice in Appendix A, as well as the facts that $m \leq n$ and I consists of precisely $m-1$ indices.

Moreover, we refine Haussler's analysis to include two natural extensions (see Appendix A for details): *(i) Obtain a refined bound on* $\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right]$*:* In the analysis of Haussler $\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right]$ is replaced by its upper bound $O(m^d)$, resulting from the fact that the primal shatter dimension of $\mathcal{V}$ (and thus of $\mathcal{V}_{|_I}$) is d, from which we obtain that *for any* choice of I, $|\mathcal{V}_{|_I}| = O((m-1)^d) = O(m^d)$, with a constant of proportionality that depends on d, and thus the packing number is $O((n/\delta)^d)$, as asserted in Theorem 3.[4] However, in our analysis we would like to have a more subtle bound on the actual expected value of $|\mathcal{V}_{|_I}|$. In fact, the scenario imposed by our assumptions on the set system eventually yields a much smaller bound on the expectation of $|\mathcal{V}_{|_I}|$, and thus on $|\mathcal{V}|$ due to Inequality (1). We review this in more detail below. *(ii) Relaxing the bound on* m*.* We show that Inequality (1) is still applicable when the sample I is slightly larger than the bound in (2), as a stand alone relation, this may result in a suboptimal bound on $|\mathcal{V}|$, however, this property will assist us to obtain local improvements over the bound on $|\mathcal{V}|$, eventually yielding the bound in Theorem 4. Specifically, in our analysis we proceed in iterations, where at the first iteration we obtain a preliminary bound on $|\mathcal{V}|$ (Corollary 6), and then, at each subsequent iteration $j > 1$, we draw a sample I_j of $m_j - 1$ indices where

$$m_j := m \log^{(j)}(n/\delta) = O\left(\frac{d_0 n \log^{(j)}(n/\delta)}{\delta}\right), \tag{3}$$

m is our choice in (2), and $\log^{(j)}(\cdot)$ is the jth iterated logarithm function. Then, by a straightforward generalization of Haussler's analysis (described in Appendix A), we obtain, for each $j = 2, \ldots, \log^*(n/\delta)$:

$$|\mathcal{V}| \leq (d_0+1)\,\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right]. \tag{4}$$

We note that since the bounds (1)–(4) involve a dependency on the VC-dimension d_0, we will sometimes need to explicitly refer to this parameter in addition to the primal shatter dimension d. Nevertheless, throughout the analysis we exploit the relation $d \leq d_0 = O(d \log d)$, mentioned in Section 1.

[4] We note, however, that the original analysis of Haussler [Hau95] does not rely on the primal shatter dimension, and the bound on $\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right]$ is just $O(m^{d_0})$ due to the Sauer-Shelah Lemma.

2.2 Overview of our Approach

We next present the proof of Theorem 4. In what follows, we assume that $\mathcal{V}$ is δ-separated, and denote by d its primal shatter dimension and by d_0 its VC-dimension. We first recall the assumption that the primal shatter function of $\mathcal{V}$ has a (d, d_1) Clarkson-Shor property, and that the length of each vector $\mathbf{v} \in \mathcal{V}$ under the L^1 norm is at most k. This implies that $\mathcal{V}$ consists of at most $O(n^{d_1}k^{d-d_1})$ vectors.

Since the Clarkson-Shor property is hereditary, then this also applies to any projection of $\mathcal{V}$ onto a subset of indices, implying that the bound on $|\mathcal{V}_{|_I}|$ is at most $O(m^{d_1}k^{d-d_1})$, where I is a subset of $m-1$ indices as above. However, due to our sampling scheme we expect that the length of each vector in $\mathcal{V}_{|_I}$ should be much smaller than k, (e.g., in expectation this value should not exceed $k(m-1)/n$), from which we may conclude that the actual bound on $|\mathcal{V}_{|_I}|$ is smaller than the trivial bound of $O(m^{d_1}k^{d-d_1})$. Ideally, we would like to show that this bound is $O(m^{d_1}(km/n)^{d-d_1}) = O(n^{d_1}k^{d-d_1}/\delta^d)$, which matches our asymptotic bound in Theorem 4 (recall that $m = O(n/\delta)$). However, this is likely to happen only in case where the length of each vector in $\mathcal{V}_{|_I}$ does not exceed its expected value, or that there are only a few vectors whose length deviates from its expected value by far, whereas, in the worst case there might be many leftover "long" vectors in $\mathcal{V}_{|_I}$. Nevertheless, our goal is to show that, with some care one can proceed in iterations, where initially I is a slightly larger sample, and then at each iteration we reduce its size, until eventually it becomes $O(m)$ and we remain with only a few long vectors. At each such iteration $\mathcal{V}_{|_I}$ is a random structure that depends on the choice of I and may thus contain long vectors, however, in expectation they will be scarce!

Specifically, we proceed over at most $\log^*(n/\delta)$ iterations, where we perform local improvements over the bound on $|\mathcal{V}|$, as follows. Let $|\mathcal{V}|^{(j)}$ be the bound on $|\mathcal{V}|$ after the jth iteration is completed, $1 \leq j \leq \log^*(n/\delta)$. We first show in Corollary 6 that for the first iteration, $|\mathcal{V}| \leq |\mathcal{V}|^{(1)} = O\left(\frac{n^{d_1}k^{d-d_1}\log^d(n/\delta)}{\delta^d}\right)$, with a constant of proportionality that depends on d. Then, at each further iteration $j \geq 2$, we select a set I_j of $m_j - 1 = O(n\log^{(j)}(n/\delta)/\delta)$ indices uniformly at random without replacements from $[n]$ (see (3) for the bound on m_j). Our goal is to bound $\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right]$ using the bound $|\mathcal{V}|^{(j-1)}$, obtained at the previous iteration, which, we assume by induction to be $O\left(\frac{n^{d_1}k^{d-d_1}(\log^{(j-1)}(n/\delta))^d}{\delta^d}\right)$ (see Lemma 8 for the recursive relation that we derive, as well as our observation that the coefficient of the recursive term is at most 1), where the base case $j = 2$ is shown in Corollary 6.

A key property in the analysis is to show that the probability that the length of a vector $\mathbf{v} \in \mathcal{V}_{|_{I_j}}$ (after the projection of $\mathcal{V}$ onto I_j) deviates from its expectation decays exponentially (Lemma 7). Note that in our case this expectation is at most $k(m_j-1)/n$. This, in particular, enables us to claim that *in expectation* the overall majority of the vectors in $\mathcal{V}_{|_{I_j}}$ have length at most $O(k(m_j-1)/n)$, whereas the remaining longer vectors are scarce. This is the key idea to derive a recursive inequality for $|\mathcal{V}|^{(j)}$ using the bound on $\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right]$ (Lemma 8). Roughly speaking, since the Clarkson-Shor property is hereditary, we apply it to $\mathcal{V}_{|_{I_j}}$ and conclude that the number of its vectors of length at most $O(k(m_j-1)/n)$ is only $O\left(\frac{n^{d_1}k^{d-d_1}(\log^{(j)}(n/\delta))^d}{\delta^d}\right)$, with a constant of proportionality that depends on d. On the other hand, due to Lemma 7 and our inductive hypothesis, the number

of longer vectors does not exceed $O\left(\frac{n^{d_1}k^{d-d_1}}{\delta^d}\right)$, which is dominated by the first bound. We thus conclude $\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right] = O\left(\frac{n^{d_1}k^{d-d_1}(\log^{(j)}(n/\delta))^d}{\delta^d}\right)$. Then we apply Inequality (4) in order to complete the inductive step, whence we obtain the bound on $|\mathcal{V}|^{(j)}$, and thus on $|\mathcal{V}|$. Note that the full analysis is somewhat more subtle, as we need to show that the actual coefficient of the recursive term $|\mathcal{V}|^{(j)}$ is at most 1 (Lemma 8). We emphasize the fact that the sample I_j is always chosen from the *original* ground set $[n]$, and thus, at each iteration we construct a new sample *from scratch*, and then exploit our observation in Inequality (4).

In what follows, we also assume that $\delta < n/2^{(d_0+1)}$ (where d_0 is the VC-dim), as otherwise the bound on the packing number is a constant that depends on d and d_0 by the Packing Lemma (Theorem 3). This assumption is crucial for the recursive analysis presented in this section—see below.

2.3 The First Iteration

In order to show our bound on $|\mathcal{V}^{(1)}|$, we form a subset $I_1 = (i_1, \ldots, i_{m_1})$ of $m_1 = |I_1| = O\left(\frac{dn\log(n/\delta)}{\delta}\right)$ indices[5] with the following two properties: (i) each vector in $\mathcal{V}$ is mapped to a distinct vector in $\mathcal{V}_{|_{I_1}}$, and (ii) the length of each vector in $\mathcal{V}_{|_{I_1}}$ does not exceed $O(k \cdot m_1/n)$.

A set I_1 as above exists by the considerations in [Ezr16]. Specifically:

Lemma 5 *A sample I_1 as above satisfies properties (i)–(ii), with probability at least 1/2.*

Proof : We rely on the notions of "relative approximations" and "ε-nets", defined in Section 2.1.

In order to show (i), we first form the set system corresponding to all symmetric difference pairs induced by $\mathcal{V}$. That is, we form the vector set $\mathcal{D}$, where $\mathcal{D} = \{(|\mathbf{u}_1 - \mathbf{v}_1|, \ldots, |\mathbf{u}_n - \mathbf{v}_n|) \mid \mathbf{u}, \mathbf{v} \in \mathcal{V}\}$. Since we assume that $\mathcal{V}$ is δ-separated, we have $\|\mathbf{w}\| \geq \delta$, for each $\mathbf{w} \in \mathcal{D}$. As observed in [Dud78] (see also [HP11]), $\mathcal{D}$ has a finite VC-dimension.

We now construct an *ε-net* for $\mathcal{D}$, with $\varepsilon = \delta/n$. By our discussion above a sample I_1 of $O(d(n/\delta)\log(n/\delta))$ indices has this property with probability greater than, say, 3/4 (for a sufficiently large constant of proportionality).

Thus, by definition, any vector $\mathbf{w} \in \mathcal{D}$ (recall that its length is at least δ) must satisfy $|\mathbf{w}_{|_{I_1}}| \geq 1$, where $\mathbf{w}_{|_{I_1}}$ denotes the projection of $\mathbf{w}$ onto I_1. But this implies that we must have $\mathbf{u}_{|_{I_1}} \neq \mathbf{v}_{|_{I_1}}$, for each pair $\mathbf{u}, \mathbf{v} \in \mathcal{V}$, and thus $\mathbf{u}$, $\mathbf{v}$ must be mapped to distinct vectors in the projection of $\mathcal{V}$ onto I_1, from which property (i) follows.

In order to have property (ii) we observe that the same sample I_1 is a *relative* $(\delta/n, 1/4)$-*approximation* for $\mathcal{V}$ with probability at least 3/4 (for an appropriate choice of the constant of

[5] In this particular step we use a different machinery than that of Haussler [Hau95]; see the proof of Lemma 5 and our remark after Corollary 6. Therefore, $|I_1| = m_1$, rather than $m_1 - 1$. Furthermore, the constant of proportionality in the bound on m_1 depends just on the primal shatter dimension d instead of the VC-dimension d_0 as in (3).

proportionality). Given this property of I_1, this implies that any vector $\mathbf{v} \in \mathcal{V}$ satisfies

$$\left| \frac{\|\mathbf{v}_{|_{I_1}}\|}{m_1} - \frac{\|\mathbf{v}\|}{n} \right| \leq \frac{1}{4} \cdot \frac{\|\mathbf{v}\|}{n},$$

if $\|\mathbf{v}\|/n \geq \delta/n$, and

$$\left| \frac{\|\mathbf{v}_{|_{I_1}}\|}{m_1} - \frac{\|\mathbf{v}\|}{n} \right| \leq \frac{1}{4} \cdot \frac{\delta}{n},$$

otherwise. Since $\|\mathbf{v}\| \leq k$, and $k \geq \delta/2$ by assumption, it is easy to verify that we always have $\|\mathbf{v}_{|_{I_1}}\| \leq 3/2 \cdot km_1/n$. In other words, $\|\mathbf{v}_{|_{I_1}}\| = O(k \cdot m_1/n)$, as asserted.

Combining the two roles of I_1 (each with probability 3/4), it follows that it is both a (δ/n)-net for $\mathcal{D}$ and a relative $(\delta/n, 1/4)$-approximation for $\mathcal{V}$, with probability at least 1/2, and thus it satisfies properties (i)–(ii) with this probability. This completes the proof of the lemma. □

We next apply Lemma 5 in order to bound $|\mathcal{V}_{|_{I_1}}|$. We first recall that the (d, d_1) Clarkson-Shor property of the primal shatter function of $\mathcal{V}$ is hereditary. Incorporating the bound on m_1 and property (ii), we conclude that

$$|\mathcal{V}_{|_{I_1}}| = O\left(m_1^{d_1} \left(\frac{km_1}{n} \right)^{d-d_1} \right) = O\left(\frac{n^{d_1} k^{d-d_1} \log^d (n/\delta)}{\delta^d} \right),$$

with a constant of proportionality that depends on d. Now, due to property (i), $|\mathcal{V}| \leq |\mathcal{V}_{|_{I_1}}|$, we thus conclude:

Corollary 6 *After the first iteration we have:* $|\mathcal{V}| \leq |\mathcal{V}|^{(1)} = O\left(\frac{n^{d_1} k^{d-d_1} \log^d (n/\delta)}{\delta^d} \right)$*, with a constant of proportionality that depends on* d*.*

Remark: We note that the preliminary bound given in Corollary 6 is crucial for the analysis, as it constitutes the base for the iterative process described in Section 2.4. In fact, this step of the analysis alone bypasses our refinement to Haussler's approach, and instead exploits the approach of Dudley [Dud78].

2.4 The Subsequent Iterations: Applying the Inductive Step

Let us now fix an iteration $j \geq 2$. As noted above, we assume by induction on j that the bound $|\mathcal{V}|^{(j-1)}$ on $|\mathcal{V}|$ after the $(j-1)$th iteration is $O\left(\frac{n^{d_1} k^{d-d_1} (\log^{(j-1)} (n/\delta))^d}{\delta^d} \right)$. Let I_j be a subset of $m_j - 1$ indices, chosen uniformly at random without replacements from $[n]$, with m_j given by (3). Let $\mathbf{v} \in \mathcal{V}$, and denote by $\mathbf{v}_{|_{I_j}}$ its projection onto I_j. The expected length $\mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|]$ of $\mathbf{v}_{|_{I_j}}$ is at most $k(m_j - 1)/n = O(d_0 k \log^{(j)} (n/\delta)/\delta)$. We next show:

Lemma 7 (Exponential Decay Lemma)

$$\mathbf{Prob}\left[\|\mathbf{v}_{|_{I_j}}\| \geq t \cdot \frac{k(m_j - 1)}{n} \right] < 2^{-tk(m_j-1)/n},$$

where $t \geq 2e$ *is a real parameter and* e *is the base of the natural logarithm.*

Proof : We first observe that the length of $\mathbf{v}_{|I_j}$ is a random variable with a hypergeometric distribution. Indeed, this is precisely the question of uniformly choosing $m_j - 1$ elements at random (into our set I_j) from a given set of n elements without replacements, and then, for a given $\|\mathbf{v}\|$-element subset of the full set (recall that the length of $\mathbf{v}$ corresponds to the cardinality of an appropriate subset in the set system), we consider how many of its elements have been chosen into I_j. Specifically, we have:

$$\mathbf{Prob}\left[\|\mathbf{v}_{|I_j}\| = s\right] = \frac{\binom{\|\mathbf{v}\|}{s}\binom{n-\|\mathbf{v}\|}{m_j-1-s}}{\binom{n}{m_j-1}},$$

for each non-negative integer $s \leq \min\{\|\mathbf{v}\|, m_j - 1\}$.

Our goal is to show a Chernoff-type bound over the probability that $\|\mathbf{v}_{|I_j}\|$ deviates from its expectation. However, we face the difficulty that the corresponding indicator variables are not independent, and thus we cannot apply a Chernoff bound directly (see, e.g. [AS08]). Nevertheless, in our scenario a Chernoff bound is still applicable, this can be viewed by various approaches, see, e.g., [AD11, Mulz, PS97]. For the sake of completeness we describe the proof in detail, and rely on the analysis of Panconesi and Srinivasan [PS97], which implies that when the underlying indicator variables are "negatively correlated", one can still apply a Chernoff bound (see also [AD11]).

We enumerate all non-zero coordinates of $\mathbf{v}$ in an arbitrary order, let $L = \{l_1, \ldots l_{\|\mathbf{v}\|}\}$ be this set of indices (in this notation we ignore all the zero-coordinates), and attach an indicator variable X_i to each index $l_i \in L$, which is defined to be one if and only if $l_i \in I_j$ (in other words, the corresponding element in the underlying set induced by $\mathbf{v}$ has been chosen to be included into the sample of the $m_j - 1$ elements). According to this notation, $\|\mathbf{v}_{|I_j}\|$ is represented by the sum $X = \sum_{i=1}^{\|\mathbf{v}\|} X_i$. It is now easy to verify that $\mathbf{Prob}[X_i = 1] = (m_j - 1)/n$, and by linearity of expectation $\mathbf{Exp}[\|\mathbf{v}_{|I_j}\|] = \|\mathbf{v}\| \cdot (m_j - 1)/n$.

However, the variables X_i are not independent due to our probabilistic model (that is, I_j is chosen without replacements), nevertheless, they are *negatively correlated*. This implies that for each subset $K \subseteq \{1, \ldots, \|\mathbf{v}\|\}$

$$\mathbf{Prob}[\bigwedge_{i \in K} X_i = 0] \leq \prod_{i \in K} \mathbf{Prob}[X_i = 0],$$

and

$$\mathbf{Prob}[\bigwedge_{i \in K} X_i = 1] \leq \prod_{i \in K} \mathbf{Prob}[X_i = 1].$$

Indeed, following the considerations in [AD11], let us show first the latter inequality. Put $L_K = \bigcup_{i \in K}\{l_i\}$. Then $\mathbf{Prob}[\bigwedge_{i \in K} X_i = 1] = \mathbf{Prob}[L_K \subseteq I_j]$, and since I_j is uniformly chosen, in order to bound the latter we need to take the proportion between the number of subsets of size $m_j - 1$ that contain L_K and the entire number of subsets of size $m_j - 1$ that can be chosen from an n-element set. Hence

$$\mathbf{Prob}[L_K \subseteq I_j] = \frac{\binom{n-|K|}{m_j-1-|K|}}{\binom{n}{m_j-1}} = \frac{(m_j-1)(m_j-2)\cdots(m_j-|K|)}{n(n-1)\cdots(n-|K|+1)},$$

and the latter is smaller than $\left(\frac{m_j-1}{n}\right)^{|K|}$, as is easily verified. Using similar arguments for the first

correlation inequality, we obtain

$$\mathbf{Prob}[\bigwedge_{i\in K} X_i = 0] = \frac{\binom{n-|K|}{m_j-1}}{\binom{n}{m_j-1}} < \left(1 - \frac{m_j - 1}{n}\right)^{|K|}.$$

We are now ready to apply [PS97, Theorem 3.4] stating that if the indicator variables X_i are negatively correlated then (recall that $X = \sum_{i=1}^{\|\mathbf{v}\|} X_i$)[6]:

$$\mathbf{Prob}[X > \rho\,\mathbf{Exp}[X]] < \left(\frac{e^{\rho-1}}{\rho^\rho}\right)^{\mathbf{Exp}[X]},$$

for any $\rho > 1$. In particular, when $\rho \geq 2e$ (where e is the base of the natural logarithm), the latter term is bounded by $2^{-\rho\,\mathbf{Exp}[X]}$. Recall that we assumed $\|\mathbf{v}\| \leq k$, and thus $\mathbf{Exp}[X] = \mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|] \leq k \cdot (m_j - 1)/n$. Thus, for any $t \geq 2e$, we obtain:

$$\mathbf{Prob}\left[\|\mathbf{v}_{|_{I_j}}\| > tk(m_j-1)/n\right] = \mathbf{Prob}\left[\|\mathbf{v}_{|_{I_j}}\| > \frac{tk(m_j-1)/n}{\mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|]} \cdot \mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|]\right],$$

observe that in this case $\rho := \frac{tk(m_j-1)/n}{\mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|]} \geq 2e$, due to our assumption on t and the fact that $\frac{k(m_j-1)/n}{\mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|]} \geq 1$, and thus the latter term is bounded by:

$$2^{-\frac{tk(m_j-1)/n}{\mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|]}\,\mathbf{Exp}[\|\mathbf{v}_{|_{I_j}}\|]} = 2^{-tk(m_j-1)/n},$$

as asserted. □

We now proceed as follows. Recall that we assume $k \geq \delta/2$, and by (3) we have $m_j = O\left(\frac{d_0 n \log^{(j)}(n/\delta)}{\delta}\right)$. Thus it follows from Lemma 7 that

$$\mathbf{Prob}\left[\|\mathbf{v}_{|_{I_j}}\| \geq C \cdot \frac{k(m_j-1)}{n}\right] < \frac{1}{(\log^{(j-1)}(n/\delta))^D}, \tag{5}$$

where $C \geq 2e$ is a sufficiently large constant, and $D > d_0$ is another constant whose choice depends on C and d_0, and can be made arbitrarily large. Since $d_0 \geq d$ we obviously have $D > d$. We next show:

Lemma 8 *Under the assumption that $k \geq \delta/2$, we have, at any iteration $j \geq 2$:*

$$|\mathcal{V}|^{(j)} \leq A(d_0+1) \cdot \frac{n^{d_1} k^{d-d_1} (\log^{(j)}(n/\delta))^d}{\delta^d} + (d_0+1) \cdot \frac{|\mathcal{V}|^{(j-1)}}{(\log^{(j-1)}(n/\delta))^D}, \tag{6}$$

where $|\mathcal{V}|^{(l)}$ is the bound on $|\mathcal{V}|$ after the lth iteration, and $A > 0$ is a constant that depends on d (and d_0) and the constant of proportionality determined by the Clarkson-Shor property of $\mathcal{V}$.

[6]We note that in the original formulation in [PS97], one needs to have a set of *independent* random variables $\hat{X}_i$, $i \in \{1, \ldots, \|\mathbf{v}\|\}$ with $\hat{X} = \sum_{i=1}^{\|\mathbf{v}\|} \hat{X}_i$, such that $\mathbf{Exp}[X] \leq \mathbf{Exp}[\hat{X}]$. In the scenario of our problem $\hat{X}_i$ is taken to be a Bernoulli indicator random variable, which takes value one with probability $(m_j - 1)/n$, in which case $\mathbf{Exp}[X] = \mathbf{Exp}[\hat{X}] = \|\mathbf{v}\| \cdot (m_j - 1)/n$.

Proof : We in fact show:

$$\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right] \leq A \cdot \frac{n^{d_1} k^{d-d_1} (\log^{(j)}(n/\delta))^d}{\delta^d} + \frac{|\mathcal{V}|^{(j-1)}}{(\log^{(j-1)}(n/\delta))^D},$$

and then exploit the relation $|\mathcal{V}| \leq (d_0+1)\,\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right]$ (Inequality (4)), in order to prove Inequality (6).

In order to obtain the first term in the bound on $\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right]$, we consider all vectors of length at most $C \cdot \frac{k(m_j-1)}{n}$ (where $C \geq 2e$ is a sufficiently large constant as above) in the projection of $\mathcal{V}$ onto a subset I_j of $m_j - 1$ indices (in this part of the analysis I_j can be arbitrary). Since the primal shatter function of $\mathcal{V}$ has a (d, d_1) Clarkson-Shor property, which is hereditary, we obtain at most

$$O(m_j{}^{d_1}(k(m_j-1)/n)^{d-d_1}) = O\left(\frac{n^{d_1} k^{d-d_1} (\log^{(j)}(n/\delta))^d}{\delta^d}\right)$$

vectors in $\mathcal{V}_{|_{I_j}}$ of length smaller than $C \cdot \frac{k(m_j-1)}{n} = O(\frac{k \log^{(j)}(n/\delta)}{\delta})$. It is easy to verify that the constant of proportionality A in the bound just obtained depends on d, d_0, and the constant of proportionality determined by the (d, d_1) Clarkson-Shor property of $\mathcal{V}$ (observe that the latter one is at most $(2(d_0+1))^d$, since m_j depends linearly in d_0).

Next, in order to obtain the second term, we consider the vectors $\mathbf{v} \in \mathcal{V}$ that are mapped to vectors $\mathbf{v}_{|_{I_j}} \in \mathcal{V}_{|_{I_j}}$ with $\|\mathbf{v}_{|_{I_j}}\| > C \cdot \frac{k(m_j-1)}{n}$. By Inequality (5):

$$\mathbf{Exp}\left[\left|\left\{\mathbf{v} \in \mathcal{V} \;\mid\; \|\mathbf{v}_{|_{I_j}}\| > C \cdot \frac{k(m_j-1)}{n}\right\}\right|\right] < \frac{|\mathcal{V}|}{(\log^{(j-1)}(n/\delta))^D},$$

and recall that $|\mathcal{V}|^{(j-1)}$ is the bound on $|\mathcal{V}|$ after the previous iteration $j-1$. This completes the proof of the lemma.

□

Remark: We note that the bound on $\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right]$ consists of the *worst-case* bound on the number of short vectors of length at most $C \cdot k(m_j-1)/n$, obtained by the (d, d_1) Clarkson-Shor property, plus the *expected* number of long vectors.

Wrapping up. We now complete the analysis and solve Inequality (6). Our initial assumption that $\delta < n/2^{(d_0+1)}$, and the fact that $D > d$ is sufficiently large, imply that the coefficient of the recursive term is smaller than 1, for any $2 \leq j \leq \log^*(n/\delta) - \log^*(d_0+1)$.[7] Then, using induction on j, one can verify that the solution is

$$|\mathcal{V}|^{(j)} \leq 2A(d_0+1)\frac{n^{d_1} k^{d-d_1} (\log^{(j)}(n/\delta))^d}{\delta^d}, \tag{7}$$

[7]We observe that $2 \leq \log^*(n/\delta) - \log^*(d_0+1) \leq \log^*(n/\delta)$, due to our assumption that $\delta < n/2^{(d_0+1)}$, and the fact that $d_0 \geq 1$.

for any $2 \leq j \leq \log^*(n/\delta) - \log^*(d_0+1)$.

We thus conclude $|\mathcal{V}|^{(j)} = O\left(\frac{n^{d_1}k^{d-d_1}(\log^{(j)}(n/\delta))^d}{\delta^d}\right)$. In particular, at the termination of the last iteration $j^* = \log^*(n/\delta) - \log^*(d_0+1)$, we obtain:

$$|\mathcal{V}| \leq |\mathcal{V}|^{(j^*)} = O\left(\frac{n^{d_1}k^{d-d_1}}{\delta^d}\right),$$

with a constant of proportionality that depends exponentially on d (and d_0). Specifically, it is easy to verify that due to our choice of j^* and the fact that A is at most $(2(d_0+1))^d$, the resulting constant of proportionality in the bound on $|\mathcal{V}|$ is at most $(d_0+1)^{O(d)}$. This at last completes the proof of Theorem 4.

3 Second Proof: Refining Chazelle's Approach

In this section, we shall prove a size-sensitive version of Haussler's upper bound for δ-separated set of vectors, having bounded primal shatter dimension, building on Chazelle's presentation of Haussler's proof as explained in [Mat99]. By Haussler's result [Hau95], we know that

$$\mathcal{M}(\delta, k, \mathcal{V}) = g^d \cdot \left(\frac{n}{\delta}\right)^{d_1} \left(\frac{k}{\delta}\right)^{d-d_1},$$

where $g := g(n,k,\delta)$ is such that $g^d = O\left(\left(\frac{n}{k}\right)^{d-d_1}\right)$. We shall show that the optimal (up to constants) upper bound for g, is in fact c^*, where c^* is independent of n, k and δ.

3.1 Overview of the second approach

Consider a preliminary attempt to extend Chazelle's proof to shallow packings. Like Chazelle, one chooses a random subsequence of indices I of size roughly n/δ, and estimates the number of projections on I caused by δ-separated vectors of size bounded by k. Suppose that the projection of every vector in the δ-separated system $\mathcal{V}$ onto I, were nearly equal to its expected value k/δ. By the definition of the (d, d_1) Clarkson-Shor property, the number of projections on the subsequence I would then be bounded by $c\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1} = c\left(\frac{n^{d_1}k^{d-d_1}}{\delta^d}\right)$, for some constant c, and Chazelle's proof would go through in a straightforward manner. However, for a given vector, its projection on I can be much larger than expected. To handle this, we shall choose I in a way that the *number* of such "bad" vectors, is at most a constant times their expected number. Also, we treat the unknown "extra" factor g (at most the Haussler packing bound, divided by $\frac{n^{d_1}k^{d-d_1}}{\delta^d}$) as a function of n, k, δ, whose value we show to be bounded by a constant independent of n, k, and δ (see Inequality (12) and succeeding paragraph). This allows us to get the final bound in a single iteration.

3.2 Details of the Proof

Before we give the details of the second proof, we will need the definition of the *unit distance graph* of a family of vectors which will play a central role in the proof of the theorem.

Definition 9 (Unit Distance Graph) *For a family $\mathcal{V} \subseteq \{0,1\}^n$ of indicator vectors, the* unit distance graph $\mathcal{UD}(\mathcal{V}) = (\mathcal{V}, E)$ *is a graph with vertex set $\mathcal{V}$ and $\{\mathbf{v}_1, \mathbf{v}_2\}$ is an edge, i.e. $\{\mathbf{v}_1, \mathbf{v}_2\} \in E$, if $\rho(\mathbf{v}_1, \mathbf{v}_2) = 1$.*

Proof : Consider a random subsequence of indices $I = (i_1, \ldots, i_s)$ obtained by selecting each $i \in [n]$ with probability $p = \frac{11 d_0 K}{\delta}$, and then taking a random permutation of the selected indices. Here $K = K(n, k, \delta) \geq 4$ is a parameter to be fixed later. Define $\mathcal{V}_1 := \mathcal{V}_{|_I}$. Consider the unit distance graph $\mathcal{UD}(\mathcal{V}_1)$. For each vector $\mathbf{v}_1 \in \mathcal{V}_1$, define the weight of $\mathbf{v}_1$ as:

$$w(\mathbf{v}_1) := \left|\left\{\mathbf{v} \in \mathcal{V} : \mathbf{v}_{|_I} = \mathbf{v}_1\right\}\right|.$$

Observe that

$$\sum_{\mathbf{v}_1 \in \mathcal{V}_1} w(\mathbf{v}_1) = \mathcal{M}(\delta, k, \mathcal{V}).$$

Now define the weight of an edge $\{\mathbf{v}_1, \mathbf{v}_1'\} \in E$ as $w\left(\{\mathbf{v}_1, \mathbf{v}_1'\}\right) := \min\{w(\mathbf{v}_1), w(\mathbf{v}_1')\}$. Let $W := \sum_{\{\mathbf{v}_1, \mathbf{v}_1'\} \in E} w\left(\{\mathbf{v}_1, \mathbf{v}_1'\}\right)$. We now need the following lemma, whose proof we include from Matoušek's book [Mat99], in order to make the exposition self-contained.

Lemma 10 $W \leq 2d_0 \sum_{\mathbf{v}_1 \in \mathcal{V}_1} w(\mathbf{v}_1) = 2d_0 \mathcal{M}(\delta, k, \mathcal{V})$.

Proof : The proof is based on the following lemma, proved by Haussler [Hau95] for set systems with bounded VC-dimension. The following version appears in [Mat99]:

Lemma 11 ([Hau95]) *Let $\mathcal{V} \subseteq \{0,1\}^n$ be a set of indicator vectors with VC-dimension d_0. Then the unit-distance graph $\mathcal{UD}(\mathcal{V})$ has at most $d_0|\mathcal{V}|$ edges.*

Since the VC-dimension of $\mathcal{V}_1$ is bounded by d_0 from the hereditary property of VC-dimension, the lemma implies that there exists a vertex $\mathbf{v}_1 \in \mathcal{V}_1$, whose degree is at most $2d_0$. Removing $\mathbf{v}_1$, the total vertex weight will be reduce by $w(\mathbf{v}_1)$, and the total edge weight will be reduce by at most $2d_0 w(\mathbf{v}_1)$. By continuing the process and the argument until all vertices are removed, we get the lemma. □

Next, we shall prove a lower bound on the expectation $\mathbf{Exp}[W]$. Set $I' := (i_1, i_2, \ldots, i_{s-1})$, and let $\mathcal{V}_2 := \mathcal{V}_{|_{I'}}$. Let $E_1 \subseteq E$ be those edges $\{\mathbf{v}_1, \mathbf{v}_1'\}$ of E where vectors $\mathbf{v}_1$ and $\mathbf{v}_1'$ differ in the coordinate i_s, and let

$$W_1 := \sum_{\{\mathbf{v}_1, \mathbf{v}_1'\} \in E_1} w\left(\{\mathbf{v}_1, \mathbf{v}_1'\}\right).$$

We need to lower bound $\mathbf{Exp}[W_1]$. Given I', let

$$Y = Y(I') := \left|\left\{\mathbf{v} \in \mathcal{V} : \|\mathbf{v}_{|_{I'}}\| > ekp\right\}\right|,$$

i.e., the number of vectors in $\mathcal{V}$, each of whose length after projecting onto I' is more than ekp, (where e is the base of the natural logarithm function).

Let N_Y and N_S denote the events $(Y \leq 8\,\mathbf{Exp}[Y])$ and $\left(\frac{np}{2} \leq s \leq \frac{3np}{2}\right)$ respectively. Also, let $Nice$ denote the event $N_Y \cap N_S$. Conditioning W on $Nice$, we get:

$$\begin{aligned}\mathbf{Exp}[W] &= \mathbf{Prob}\left[Nice\right]\mathbf{Exp}[W|Nice] + \mathbf{Prob}\left[\overline{Nice}\right]\mathbf{Exp}[W|\overline{Nice}] \\ &\geq \mathbf{Prob}\left[Nice\right]\mathbf{Exp}[W|Nice].\end{aligned}$$

By Markov's Inequality, we have

$$\mathbf{Prob}\left[\overline{N_Y}\right] = \mathbf{Prob}\left[Y \geq 8\,\mathbf{Exp}[Y]\right] \leq 1/8,$$

and using Chernoff Bounds, (see [AS08, App. A]), with the fact that $n/\delta \geq 1$, we get

$$\mathbf{Prob}\left[\overline{N_S}\right] = \mathbf{Prob}\left[|s - np| > \frac{np}{2}\right] \leq 2\exp\left(\frac{-11d_0Kn}{3\cdot 2^2\delta}\right) \leq 2\exp\left(-11K/12\right) \ll 1/8,$$

since $d_0 \geq 1$ and $K \geq 4$. This implies that

$$\begin{aligned}\mathbf{Prob}\left[Nice\right] \;=\; \mathbf{Prob}\left[N_Y \cap N_S\right] &= \mathbf{Prob}\left[\overline{\overline{N_Y} \cup \overline{N_S}}\right] \\ &= 1 - \mathbf{Prob}\left[\overline{N_Y} \cup \overline{N_S}\right] \\ &\geq 1 - \mathbf{Prob}\left[\overline{N_S}\right] - \mathbf{Prob}\left[\overline{N_Y}\right] \\ &\geq 1 - \frac{1}{8} - \frac{1}{8} \;=\; \frac{3}{4},\end{aligned}$$

where the last inequality follows from the bounds on $\mathbf{Prob}\left[\overline{N_S}\right]$ and $\mathbf{Prob}\left[\overline{N_Y}\right]$. Hence,

$$\mathbf{Exp}[W] \geq \frac{3}{4}\,\mathbf{Exp}[W|Nice] = \frac{3}{4}s\cdot\mathbf{Exp}[W_1|Nice] \geq \left(\frac{3np}{8}\right)\mathbf{Exp}[W_1|Nice], \tag{8}$$

where the equality follows by symmetry of the choice of i_s from the indices in I, and the last inequality comes from the lower bound on s under the condition $Nice$.

So to lower bound $\mathbf{Exp}[W]$, it suffices to lower bound $\mathbf{Exp}[W_1|Nice]$. Let W_2 denote $W_1|Nice$. We shall lower bound $\mathbf{Exp}[W_2|I']$, that is, the expected weight contributed by i_s, under the event $Nice$, for a fixed choice of I'.

By definition, $W_1 = \sum_{\{\mathbf{v}_1,\mathbf{v}_1'\}\in E_1} w\left(\{\mathbf{v}_1,\mathbf{v}_1'\}\right)$. Partition $\mathcal{V}$ into equivalence classes $\mathcal{V}_1', \ldots, \mathcal{V}_r'$ according to their projection onto I':

$$\mathcal{V} = \mathcal{V}_1' \sqcup \ldots \sqcup \mathcal{V}_r',$$

where $r \leq c(3np/2)^d = O((n/\delta)^d)$. Here, c is the constant independent of n, δ, occuring in the definition of primal shatter dimension. Define $B \subset [r]$ as follows:

$$B := \{j \in [r] : \mathbf{v} \in \mathcal{V}_j' \Leftrightarrow \|\mathbf{v}_{|_{I'}}\| > ekp\}.$$

Further, let G be $[r] \setminus B$. Since $Nice$ holds, we have:

$$\sum_{j\in B} |\mathcal{V}_j'| \leq 8\,\mathbf{Exp}[Y]. \tag{9}$$

We will estimate the contribution of the classes in G to the expected weight $\mathbf{Exp}[W_2|I']$, that is, the contribution of the edges of $\mathcal{UD}(\mathcal{V}_1)$ which have both of their endpoints in G. Consider a class $\mathcal{V}'_i$ such that $i \in G$. Let $\mathcal{V}''_1 \subset \mathcal{V}'_i$ be those vectors in $\mathcal{V}'_i$ with 1 in the i_s-th coordinate, and let $\mathcal{V}''_2 = \mathcal{V}'_i \setminus \mathcal{V}''_1$. Let $b = |\mathcal{V}'_i|$, $b_1 = |\mathcal{V}''_1|$ and $b_2 = |\mathcal{V}''_2|$. Then the edge $\{\mathbf{v}, \mathbf{v}'\} \in E_1$ formed by the projection of $\mathcal{V}'_i$ onto I, has weight

$$w\left(\{\mathbf{v}, \mathbf{v}'\}\right) = \min(b_1, b_2) \geq \frac{b_1 b_2}{b}. \tag{10}$$

Observe that in Inequality (10), b depends only on the subset I'. Also, since I' is fixed, the product $b_1 \cdot b_2$ depends on the choice of i_s. The product $b_1 b_2$ is the number of ordered pairs of vectors $(\mathbf{v}, \mathbf{v}')$, with $\mathbf{v}$ and $\mathbf{v}'$ in $\mathcal{V}'_i$, such that $\mathbf{v}$ and $\mathbf{v}'$ differs only in the i_s-th coordinate. Observe that under the event $Nice$, for a fixed I', the probability that any given element $i \in [n]$ is the chosen element i_s, is $\frac{1}{n-|I'|} = \frac{1}{n-s+1} \geq \frac{1}{n}$ [8]. Since for a given ordered pair $(\mathbf{v}, \mathbf{v}')$ of distinct vectors $\mathbf{v}, \mathbf{v}' \in \mathcal{V}'_i$, $\mathbf{v}$ and $\mathbf{v}'$ differ in $\rho(\mathbf{v}, \mathbf{v}') > \delta$ coordinates, the probability that $(\mathbf{v}, \mathbf{v}')$ differ in the chosen i_s-th coordinate is

$$\mathbf{Prob}\left[|\mathbf{v}_{i_s} - \mathbf{v}'_{i_s}| = 1\right] = \frac{\rho(\mathbf{v}, \mathbf{v}')}{n - s + 1} \geq \frac{\delta}{n}.$$

Therefore, the expected contribution of $(\mathbf{v}, \mathbf{v}')$ to $b_1 b_2$ is at least $\frac{\delta}{n}$ and this implies

$$\begin{aligned} \mathbf{Exp}[w\left(\{\mathbf{v}, \mathbf{v}'\}\right)|Nice] &\geq \mathbf{Exp}\left[(b_1 b_2/b)|Nice\right] \\ &\geq \frac{b(b-1)\delta}{bn} = \frac{\delta(b-1)}{n} = \frac{\delta(|\mathcal{V}'_i| - 1)}{n}, \end{aligned}$$

where the inequality in the second line comes from Inequality (10).

Hence, the expected contribution of G to the weight W_2 is:

$$\mathbf{Exp}[W_2|I'] \geq \sum_{\{\mathbf{v},\mathbf{v}'\} \in G} \mathbf{Exp}[w\left(\{\mathbf{v}, \mathbf{v}'\}\right)|Nice] \geq \sum_{i \in G} (|\mathcal{V}'_i| - 1)\frac{\delta}{n}. \tag{11}$$

Observe that by the (d, d_1) Clarkson-Shor property and the fact that $|I'| < \frac{3np}{2}$, we have

$$|G| \leq C|I'|^{d_1} (ekp)^{d-d_1} < C\left(3np/2\right)^{d_1} (ekp)^{d-d_1},$$

where C is a constant. Substituting the bound of $|G|$ in Inequality (11) and the fact that $\sum_{j \in B} |\mathcal{V}'_j| \leq$

[8] For any $p \in (0, 1)$, the binomial distribution with parameter p on a finite set X, under the conditions that only one element is selected and that it does not belong to a fixed subset $I' \subset X$, is just the uniform distribution on $X \setminus I'$.

$8\,\mathbf{Exp}[Y]$ (see Inequality (9)) we get:

$$\begin{aligned}
\mathbf{Exp}[W_2|I'] &\geq \left(\left(\sum_{i\in G}|\mathcal{V}'_i|\right) - |G|\right)\frac{\delta}{n} \\
&\geq \left(\left(\sum_{i\in G}|\mathcal{V}'_i|\right) - C(3np/2)^{d_1}(ekp)^{d-d_1}\right)\frac{\delta}{n} \\
&= \left(|\mathcal{V}| - \left(\sum_{i\in B}|\mathcal{V}'_i|\right) - C(3np/2)^{d_1}(ekp)^{d-d_1}\right)\frac{\delta}{n} \\
&\geq \left(|\mathcal{V}| - 8\,\mathbf{Exp}[Y] - C(11d_0K)^d.(3/2)^{d_1}e^{d-d_1}\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}\right)\frac{\delta}{n} \\
&\geq \left(\mathcal{M}(\delta,k,\mathcal{V}) - 8\,\mathbf{Exp}[Y] - C_1K^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}\right)\frac{\delta}{n},
\end{aligned}$$

where $C_1 = C\cdot(11d_0)^d(3/2)^{d_1}e^{d-d_1}$.

Since the above holds for each I' which satisfies $Nice$, we get that

$$\mathbf{Exp}[W_2] \geq \left(\mathcal{M}(\delta,k,\mathcal{V}) - 8\,\mathbf{Exp}[Y] - C_1K^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}\right)\frac{\delta}{n},$$

Using Inequality (8), and comparing with the upper bound on W,

$$(3np/8)\,\mathbf{Exp}[W_2] \leq \mathbf{Exp}[W] \leq 2d_0\mathcal{M}(\delta,k,\mathcal{V}),$$

and substituting the lower bound on $\mathbf{Exp}[W_2]$ and solving for $\mathcal{M}(\delta,k,\mathcal{V})$, we get

$$\mathcal{M}(\delta,k,\mathcal{V}) \leq \frac{33K\left(8\,\mathbf{Exp}[Y] + C_1K^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}\right)}{33K-16}.$$

Next, the following lemma connects the parameters K and g:

Lemma 12 *For* $K = \max\left\{4, \frac{2d\log g}{11d_0}\right\}$, $\mathbf{Exp}[Y] \leq \frac{n^{d_1}k^{d-d_1}}{\delta^d}$.

Substituting the choice of K and the value of $\mathbf{Exp}[Y]$ from Lemma 12, we get that

$$\begin{aligned}
g^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1} &= \mathcal{M}(\delta,k,\mathcal{V}) \leq \frac{C_1K^d\left(\frac{n}{\delta}\right)^{d_1}\cdot\left(\frac{k}{\delta}\right)^{d-d_1} + 8\left(\frac{n}{\delta}\right)^{d_1}\cdot\left(\frac{k}{\delta}\right)^{d-d_1}}{1-16/33K} \\
&\leq C_2K^d\left(\frac{n}{\delta}\right)^{d_1}\cdot\left(\frac{k}{\delta}\right)^{d-d_1} \\
&\leq C_2\left(\max\left\{4,\frac{2d\log g}{11d_0}\right\}\right)^d\cdot\left(\frac{n}{\delta}\right)^{d_1}\cdot\left(\frac{k}{\delta}\right)^{d-d_1},
\end{aligned}$$

(where in the second step we used that $1-16/33K \geq 1/2$, and hence C_2 is a positive constant.)

The last inequality above implies that $g^d \leq C_2 \left(\max\{4, \frac{2d\log g}{11d_0}\}\right)^d$, or with $C_3 = C_2^{1/d}$:

$$g \leq C_3 \max\left\{4, \frac{2d\log g}{11d_0}\right\}. \tag{12}$$

We claim that the above implies that g, and hence K, must be bounded from above by a constant independent of n, k, δ. If g is not increasing with n, k, and δ, then g is bounded, and we are done. Otherwise, for any $g = g(n,k,\delta)$ growing with n, k, or δ, we have $g \gg C_4 \log g$ (where $C_4 = \frac{2dC_3}{11d_0}$) for sufficiently large n, k or δ. This inequality can only be satisfied when g is bounded by a constant function of n, k and δ, i.e., $g(n,k,\delta) \leq c^*$, where c^* is independent of n, k and δ. Thus in either case, $g = g(n,k,\delta)$ is a function bounded by an absolute constant. Therefore, it suffices to take $K \geq \frac{2d\log c^*}{11d_0}$, where c^* is the greatest positive value satisfying Inequality (12).

The constants C_1, C_2 are of the order of d_0^d. Therefore C_3 is $O(d_0)$, and C_4 is $O(d)$. Hence, $c^* = O(d\log d)$, and therefore $g^d = O((d\log d)^d)$. Thus the constant of proportionality in the bound, in the proof based on Chazelle's approach, can be upper bounded by $O((d\log d)^d) = O(2^{d\log d})$. □

It only remains to prove Claim 12:

Proof of Lemma 12: The proof follows easily from Chernoff Bounds. Fix $\mathbf{v} \in \mathcal{V}$, and let $Z = \|\mathbf{v}_{|_{I'}}\|$ and $Z' = \|\mathbf{v}_{|_I}\|$. Observe that $Z' \geq Z$. Since $\|\mathbf{v}\| \leq k$, we have $\mathbf{Exp}[Z'] \leq kp$. The probability that $Z \geq ekp$ is upper bounded using Chernoff bounds, see [AS08, Theorem A.1.12], as follows:

$$\begin{aligned}
\mathbf{Prob}\left[Z > ekp\right] &\leq \mathbf{Prob}\left[Z' - \mathbf{Exp}[Z'] > \left(\beta - 1\right)\mathbf{Exp}[Z']\right], \quad \text{where } \beta = \frac{ekp}{\mathbf{Exp}[Z']}, \\
&\leq \exp\left(-\mathbf{Exp}[Z'](\beta\log\beta - \beta + 1)\right) \\
&\leq \exp\left(-ekp(\log\beta - 1 + \frac{1}{\beta})\right) \\
&\leq \exp(-kp) \\
&\leq \exp\left(-\frac{11d_0Kk}{\delta}\right),
\end{aligned}$$

since $\beta = \frac{ekp}{\mathbf{Exp}[Z']} \geq e$, and as can be verified by elementary calculus, for any $\beta \geq e$, $\log\beta - 1 + 1/\beta \geq 1/e$. Hence the expected number $\mathbf{Exp}[Y]$ of vectors whose projections on I' have norm at least $(11ed_0Kk/\delta)$, is at most:

$$\mathbf{Exp}[Y] \leq \mathcal{M}(\delta,k,\mathcal{V})\exp\left(-\frac{11d_0Kk}{\delta}\right) \leq \mathcal{M}(\delta,k,\mathcal{V})\exp\left(-11d_0K/2\right),$$

since $k \geq \delta/2$. Substituting the value of $\mathcal{M}(\delta,k,\mathcal{V})$ in terms of g, we have

$$\begin{aligned}
\mathbf{Exp}[Y] &\leq g^d\left(\frac{n}{\delta}\right)^{d_1} \cdot \left(\frac{k}{\delta}\right)^{d-d_1} \cdot \exp(-11d_0K/2) \\
&\leq \left(\frac{n}{\delta}\right)^{d_1} \cdot \left(\frac{k}{\delta}\right)^{d-d_1} \cdot \exp\left(d\log g - \frac{11d_0K}{2}\right) \\
&\leq \left(\frac{n}{\delta}\right)^{d_1} \cdot \left(\frac{k}{\delta}\right)^{d-d_1},
\end{aligned}$$

for $K \geq \frac{2d \log g}{11 d_0}$. $\square$

4 Applications

4.1 Realization to Geometric Set Systems

We now incorporate the random sampling technique of Clarkson and Shor [CS89] with Theorem 4 in order to conclude that small shallow packings exist in several useful scenarios. This includes the case where $\mathcal{V}$ represents: (i) A collection of halfspaces defined over an n-point set in d-space (that is, each set in the set system is the subset of points contained in a halfspace). In this case, for any integer parameter $0 \leq k \leq n$, the number of halfspaces that contain at most k points is $O(n^{\lfloor d/2 \rfloor} k^{\lceil d/2 \rceil})$, and thus the primal shatter function has a $(d, \lfloor d/2 \rfloor)$ Clarkson-Shor property. (ii) A collection of balls defined over an n-point set in d-space. Here, the number of balls that contain at most k points is $O(n^{\lfloor (d+1)/2 \rfloor} k^{\lceil (d+1)/2 \rceil})$, and therefore the primal shatter function has a $(d+1, \lfloor (d+1)/2 \rfloor)$ Clarkson-Shor property. (iii) A collection of *parallel slabs* (that is, each of these regions is enclosed between two parallel hyperplanes and has an arbitrary width), defined over an n-point set in d-space. The number of slabs, which contains at most k points is $O(n^d k)$. (iv) A dual set system of points in d-space and a collection F of n $(d-1)$-variate (not necessarily continuous or totally defined) functions of *constant description complexity*. Specifically, the graph of each function is a semi-algebraic set in $\mathbb{R}^d$ defined by a constant number of polynomial equalities and inequalities of constant maximum degree (see [SA95, Chapter 7] for a detailed description of these properties, which we omit here).[9] In this case, $\mathcal{V}$ is represented by the cells (of all dimensions) in the *arrangement* of the graphs of the functions in F (see Section 1 for the definition) that lie below at most k function graphs. This portion of the arrangement is also referred to as the *at most k-level*, and its combinatorial complexity is $O(n^{d-1+\varepsilon} k^{1-\varepsilon})$, for any $\varepsilon > 0$, where the constant of proportionality depends on d and ε. Thus the primal shatter function has a $(d, d-1+\varepsilon)$ Clarkson-Shor property.

All bounds presented in (i)–(iv) are well known in the field of computational geometry; we refer the reader to [CS89, Mat02, SA95] for further details. We thus conclude:

Corollary 13 *Let $\mathcal{V} \subseteq \{0,1\}^n$ be a set of indicator vectors representing a set system of halfspaces defined over an n-point set in d-space, and let δ, k be two integer parameters as in Theorem 4. Then:*

$$\mathcal{M}(\delta, k, \mathcal{V}) = O\left(\frac{n^{\lfloor d/2 \rfloor} k^{\lceil d/2 \rceil}}{\delta^d}\right),$$

where the constant of proportionality depends on d.

Corollary 14 *Let $\mathcal{V} \subseteq \{0,1\}^n$ be a set of indicator vectors representing a set system of balls defined over an n-point set in d-space, and let δ, k be two integer parameters as in Theorem 4. Then:*

$$\mathcal{M}(\delta, k, \mathcal{V}) = O\left(\frac{n^{\lfloor (d+1)/2 \rfloor} k^{\lceil (d+1)/2 \rceil}}{\delta^{d+1}}\right),$$

[9] In [SA95] it is also required that the projection of each function onto the plane $x_d = 0$ has a constant description complexity.

where the constant of proportionality depends on d.

Corollary 15 *Let $\mathcal{V} \subseteq \{0,1\}^n$ be a set of indicator vectors representing a set system of parallel slabs defined over an n-point set in d-space, and let δ, k be two integer parameters as in Theorem 4. Then:*

$$\mathcal{M}(\delta, k, \mathcal{V}) = O\left(\frac{n^d k}{\delta^{d+1}}\right),$$

where the constant of proportionality depends on d.

Corollary 16 *Let $\mathcal{V} \subseteq \{0,1\}^n$ be a set of indicator vectors representing a dual set system of n $(d-1)$-variate (not necessarily continuous or totally defined) functions of constant description complexity and points in d-space. Let δ, k be two integer parameters as in Theorem 4. Then:*

$$\mathcal{M}(\delta, k, \mathcal{V}) = O\left(\frac{n^{d-1+\varepsilon} k^{1-\varepsilon}}{\delta^d}\right),$$

where the constant of proportionality depends on d and on ε.

4.2 Size-Sensitive Discrepancy Bounds

Suppose we are given a set system Σ defined over an n-point set X (for simplicity of exposition, we refer to Σ as a collection of subsets of X rather than a collection of vectors defined over the unit cube). We now wish to color the points of X by two colors, such that in each set of Σ the deviation from an even split is as small as possible.

Formally, a *two-coloring* of X is a mapping $\chi : X \to \{-1, +1\}$. For a set $S \in \Sigma$ we define $\chi(S) := \sum_{x \in S} \chi(x)$. The *discrepancy* of Σ is then defined as $\mathrm{disc}(\Sigma) := \min_\chi \max_{S \in \Sigma} |\chi(S)|$.

In a previous work [Ezr16], the second author presented size-sensitive discrepancy bounds for set systems of halfspaces defined over n points in d-space. These bounds were achieved by combining the *entropy method* [LM12] with δ-packings, and, as observed in [Ezr16], they are optimal up to a poly-logarithmic factor. Incorporating our bound in Theorem 4 into the analysis in [Ezr16], the bounds on $\chi(S)$ improve by a logarithmic factor. Specifically, we obtain (see Appendix B for the proof):

Theorem 17 *Let Σ be a set system, defined over an n-point set, whose primal shatter function has a (d, d_1) Clarkson-Shor property. Then, there is a two-coloring χ, such that for each $S \in \Sigma$:*

$$\chi(S) = \begin{cases} O\left(|S|^{1/2 - d_1/(2d)} n^{(d_1-1)/(2d)} \log^{1/(2d)} n\right), & \text{if } d_1 > 1 \\ O\left(|S|^{1/2 - 1/(2d)} \log^{1+1/2d} n\right), & \text{if } d_1 = 1 \end{cases}$$

Such a two-coloring χ can be computed in expected polynomial time.

Concluding Remarks. The analogy between shallow packings for dual set systems and shallow cuttings [Mat92] may cause one to interpret shallow packings as the "primal version" of shallow cuttings. For an example of getting shallow packing, though now explicitly stated as such, for dual set systems with small union complexity via shallow cuttings refer to Mustafa and Ray's paper on Mnets [MR14]. In this paper we presented discrepancy applications for shallow packings. We hope to find additional applications in geometry and beyond.

Acknowledgments. We authors would like to thank two anonymous referees for their useful comments. The second author wishes to thank Boris Aronov, Sariel Har-Peled, Aryeh Kontorovich, and Wolfgang Mulzer for useful discussions and suggestions. Last but not least, the second author thanks Ramon Van Handel, for various discussions and for spotting an error in an earlier version of this paper.

References

[AS08] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley, New York, 3rd edition, 2008.

[AD11] A. Auger and B. Doerr. *Theory of Randomized Search Heuristics:Foundations and Recent Developments*, World Scientific Publishing, 2011.

[BLL09] N. H. Bshouty, Y. Li, and P. M. Long. Using the Doubling Dimension to Analyze the Generalization of Learning Algorithms. *J. Comput. Syst. Sci.*, 75(6):323–335, 2009.

[Cha92] B Chazelle. A note on Haussler's packing lemma. 1992. Unpublished manuscript, Princeton.

[CS89] K. L. Clarkson and P. W. Shor. Applications of random sampling in computational geometry, ii. *Discrete Comput. Geom.*, 4(5):387–421, 1989.

[CW89] B. Chazelle and E. Welzl. Quasi-optimal Range Searching in Spaces of Finite VC-dimension. *Discrete Comput. Geom.*, 4(5):467–489, 1989.

[Dud78] R. M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–1049, 1978.

[DEG15] K. Dutta, E. Ezra, and A. Ghosh. Two proofs for shallow packings. In *Proc. 31st Int. Sympos. Comput. Geom*, pp.96–110, 2015.

[Ezr16] E. Ezra. A Size-sensitive Discrepancy Bound for Set Systems of Bounded Primal Shatter Dimension. *SIAM J. Comput.*, 45(1): 84–101, 2016. A preliminary version appeard in *Proc. 25th Ann. ACM-SIAM Symposium on Discrete Algorithms*, pp. 1378–1388. SIAM, 2014.

[GKM12] L.-A. Gottlieb, A. Kontorovich, and E. Mossel. VC Bounds on the Cardinality of Nearly Orthogonal Function Classes. *Discrete Math.*, 312(10):1766–1775, 2012.

[Hau92] D. Haussler. Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Inf. Comput.*, 100(1):78–150, 1992.

[Hau95] D. Haussler. Sphere Packing Numbers for Subsets of the Boolean N-cube with Bounded Vapnik-Chervonenkis Dimension. *J. Comb. Theory Ser. A*, 69(2):217–232, 1995.

[HLW94] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting 0, 1-functions on Randomly Drawn Points. *Inf. Comput.*, 115(2):248–292, 1994.

[HP11] S. Har-Peled. *Geometric Approximation Algorithms.* American Mathematical Society, Boston, MA, USA, 2011.

[HPS11] S. Har-Peled and M. Sharir. Relative (ε, ρ)-Approximations in Geometry. *Discrete Comput. Geom.*, 45(3):462–496, 2011.

[HW87] D. Haussler and E. Welzl. Epsilon-Nets and Simplex Range Queries. *Discrete & Computational Geometry*, 2:127–151, 1987.

[LLS01] Y. Li, P. M. Long, and A. Srinivasan. Improved Bounds on the Sample Complexity of Learning. *J. Comput. Syst. Sci.*, 62(3):516–527, 2001.

[LM12] S. Lovett and R. Meka. Constructive Discrepancy Minimization by Walking on the Edges. In *Proc. IEEE 53rd Ann. Symposium on Foundations of Computer Science*, FOCS '12, pp. 61–67, Washington, DC, USA, 2012. IEEE Computer Society.

[Mat92] J. Matoušek. Reporting Points in Halfspaces. *Comput. Geom. Theory Appl.*, 2(3):169–186, 1992.

[Mat95] J. Matousek. Tight Upper Bounds for the Discrepancy of Half-Spaces. *Discrete & Computational Geometry*, 13:593–601, 1995.

[Mat99] J. Matousek. *Geometric Discrepancy: An Ilustrated Guide (Algorithms and Combinatorics).* Springer, 1999.

[Mat02] J. Matousek. *Lectures on Discrete Geometry.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2002.

[Mulz] W. Mulzer. Chernoff Bounds, *Personal note.* `http://page.mi.fu-berlin.de/mulzer/notes/misc/chernoff.pdf`.

[Mus16] N. H. Mustafa. A simple proof of the shallow packing lemma. *Discrete & Computational Geometry*, 55(3):739–743, 2016.

[MR14] N. H. Mustafa and S. Ray. Near-Optimal Generalisations of a Theorem of Macbeath. In *Proc. 31st International Symposium on Theoretical Aspects of Computer Science, STACS '14, pp. 578–589, Lyon, France, 2014.*

[PS97] *A. Panconesi and A. Srinivasan. Randomized distributed edge coloring via an extension of the Chernoff-Hoeffding bounds.* SIAM J. Comput., *26:350–368 (1997).*

[Pol84] *D. Pollard.* Convergence of Stochastic Processes. *Springer-Verlag, 1984.*

[SA95] *M. Sharir and P. K. Agarwal.* Davenport-Schinzel Sequences and Their Geometric Applications. *Cambridge University Press, New York, NY, USA, 1995.*

[Sau72] *N. Sauer. On the density of families of sets.* J. Combin. Theory, Ser A*, 13(1):145–147, 1972.*

[She72] *S. Shelah. A combinatorial problem, stability and order for models and theories in infinitary languages.* Pacific J. Math.*, 41:247–261, 1972.*

[VC71] *V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities.* Theory Prob. Appl.*, 16(2):264–280, 1971.*

[Wel92] *E. Welzl. On Spanning Trees with Low Crossing Numbers. In* Data Structures and Efficient Algorithms, Final Report on the DFG Special Joint Initiative*, pp. 233–249, London, UK, UK, 1992. Springer-Verlag.*

A Overview of Haussler's Approach

Let $\mathcal{V} \subseteq \{0,1\}^n$ be a collection of indicator vectors of primal shatter dimension d. We denote its VC-dimension by d_0; as discussed in Section 1 $d_0 = O(d \log d)$.

We first form a probability distribution P over $\mathcal{V}$, implying that $\mathcal{V}$ can be viewed[10] as an n-dimensional random variable taking values in $\{0,1\}^n$. Thus its components $\mathcal{V}_i$, $i = 1, \ldots, n$, represent n correlated indicator random variables (Bernoulli random variables), and each of their values is determined by randomly selecting a vector $\mathbf{v} \in \mathcal{V}$, and letting $\mathcal{V}_i$ be the ith component of $\mathbf{v}$. The variance of a Bernoulli random variable B is known to be $\mathbf{Prob}[B = 1]\,\mathbf{Prob}[B = 0]$, and then, for a sequence $B_1, \ldots, B_m$ of Bernoulli random variables, the *conditional variance* of B_m given $B_1, \ldots, B_{m-1}$ is defined as

$$\mathbf{Var}(B_m | B_1, \ldots, B_{m-1}) = \sum_{\mathbf{v} \in \{0,1\}^{m-1}} \mathbf{Prob}(\mathbf{v})\,\mathbf{Prob}(B_m = 1|\mathbf{v})\,(1 - \mathbf{Prob}(B_m = 1|\mathbf{v}))\,,$$

where $\mathbf{Prob}(\mathbf{v}) = \mathbf{Prob}(B_1 = \mathbf{v}_1, B_2 = \mathbf{v}_2, \ldots, B_{m-1} = \mathbf{v}_{m-1})$, and $\mathbf{Prob}(B_m = 1|\mathbf{v}) = \mathbf{Prob}(B_m = 1 | B_1 = \mathbf{v}_1, B_2 = \mathbf{v}_2, \ldots, B_{m-1} = \mathbf{v}_{m-1})$.

A key property in the analysis of Haussler [Hau95] lies in the density of a *unit distance graph* $G = (\mathcal{V}, E)$, defined over $\mathcal{V}$, whose edges correspond to all pairs $\mathbf{u}, \mathbf{v} \in \mathcal{V}$, whose symmetric difference distance is (precisely) 1. In other words, $\mathbf{u}$, $\mathbf{v}$ appear as neighbors on the unit cube $\{0,1\}^n$. It has been shown by Haussler *et al.* [HLW94] that the density of G is bounded by the VC-dimension of $\mathcal{V}$, that is, $|E|/|\mathcal{V}| \leq d_0$; see also [Hau95] for an alternative proof using the technique of "shifting".[11] Then, this low density property is exploited in order to show that once we have chosen $(n-1)$

[10] For the time being, P is an arbitrary distribution, but later on (Lemma 19) it is taken to be the uniform distribution in the obvious way, where each vector in $\mathcal{V}$ is equally likely to be chosen. This distribution, however, may not remain uniform after the projection of $\mathcal{V}$ onto a proper subsequence $I' = (i_1, \ldots, i_m)$ of $m < n$ indices, as several vectors in $\mathcal{V}$ may be projected onto the same vector in $\mathcal{V}_{|_{I'}}$.

[11] We cannot guarantee such a relation when the VC-dimension d_0 is replaced by the primal shatter dimension d, and therefore we proceed with the analysis using this ratio.

coordinates of the random variable $\mathcal{V}$, the variance in the choice of the remaining coordinate is relatively small. That is:

Lemma 18 ([Hau95]) *For any distribution P on $\mathcal{V}$,*

$$\sum_{i=1}^{n} \mathbf{Var}(\mathcal{V}_i | \mathcal{V}_1, \ldots, \mathcal{V}_{i-1}, \mathcal{V}_{i+1}, \ldots, \mathcal{V}_n) \leq d_0.$$

As observed in [Hau95], Lemma 18 continues to hold on any restriction of $\mathcal{V}$ to a sequence $I' = \{i_i, \ldots, i_m\}$ of $m \leq n$ indices. Indeed, when projecting $\mathcal{V}$ onto I' the VC-dimension in the resulting set system remains d_0. Furthermore, the conditional variance is now defined w.r.t. the induced probability distribution on $\mathcal{V}_{|_{I'}}$ in the obvious way, where the probability to obtain a sequence of m values corresponds to an appropriate marginal distribution, that is, $\mathbf{Prob}_{|_{I'}}(u_1, \ldots, u_m) = \mathbf{Prob}(\mathbf{v} \in \mathcal{V} \mid v_{i_j} = u_j, 1 \leq j \leq m)$. With this observation, we can rewrite the inequality stated in Lemma 18 as

$$\sum_{i=1}^{m} \mathbf{Var}(\mathcal{V}_{i_j} | \mathcal{V}_{i_1}, \ldots, \mathcal{V}_{i_{j-1}}, \mathcal{V}_{i_{j+1}}, \ldots, \mathcal{V}_{i_m}) \leq d_0.$$

If I' is a sequence chosen uniformly at random (over all such m-tuples), then when averaging over all choices of I' we clearly obtain:

$$\mathbf{Exp}\left[\sum_{i=1}^{m} \mathbf{Var}(\mathcal{V}_{i_j} | \mathcal{V}_{i_1}, \ldots, \mathcal{V}_{i_{j-1}}, \mathcal{V}_{i_{j+1}}, \ldots, \mathcal{V}_{i_m})\right] \leq d_0,$$

or

$$\sum_{i=1}^{m} \mathbf{Exp}\left[\mathbf{Var}(\mathcal{V}_{i_j} | \mathcal{V}_{i_1}, \ldots, \mathcal{V}_{i_{j-1}}, \mathcal{V}_{i_{j+1}}, \ldots, \mathcal{V}_{i_m})\right] \leq d_0,$$

by linearity of expectation. In fact, by symmetry of the random variables $\mathcal{V}_{i_j}$ (recall that I' is a random m-tuple) each of the summands in the above inequality has an equal contribution, and thus, in particular (recall once again that the expectation is taken over all choices of $I' = \{i_i, \ldots, i_m\}$):

$$\mathbf{Exp}_{I'}\left[\mathbf{Var}(\mathcal{V}_{i_m} | \mathcal{V}_{i_1}, \ldots, \mathcal{V}_{i_{m-1}})\right] \leq \frac{d_0}{m}, \tag{13}$$

where we write $\mathbf{Exp}_{I'}[\cdot]$ to emphasize the fact that the expectation is taken over all choices of I'. The above bound is now integrated with the next key property:

Lemma 19 ([Hau95]) *Let $\mathcal{V}$ be δ-separated subset of $\{0,1\}^n$, for some $1 \leq \delta \leq n$ integer, and form a uniform distribution P on $\mathcal{V}$. Let $I = (i_1, \ldots, i_{m-1})$ be a sequence of $m-1$ distinct indices between 1 and n, where m is any integer between 1 and n. Suppose now that another index i_m is drawn uniformly at random from the remaining $n - m + 1$ indices. Then*

$$\mathbf{Exp}\left[\mathbf{Var}\left(\mathcal{V}_{i_m} | \mathcal{V}_{i_1} \ldots, \mathcal{V}_{i_{m-1}}\right)\right] \geq \frac{\delta}{2(n-m+1)}\left(1 - \frac{|\mathcal{V}_{|_I}|}{|\mathcal{V}|}\right),$$

where the conditional variance is taken w.r.t. the distribution P, and the expectation is taken w.r.t. the random choice of i_m.

We now observe that when the entire sequence $I' = (i_1, \ldots, i_m)$ is chosen uniformly at random, then the bound in Lemma 19 continues to hold when averaging on the entire sequence I' (rather than just on i_m), that is, we have:

$$\mathbf{Exp}_{I'}\left[\mathbf{Var}\left(\mathcal{V}_{i_m}|\mathcal{V}_{i_1}\ldots,\mathcal{V}_{i_{m-1}}\right)\right] \geq \mathbf{Exp}_{I'}\left[\frac{\delta}{2(n-m+1)}\left(1-\frac{|\mathcal{V}_{|_I}|}{|\mathcal{V}|}\right)\right] \tag{14}$$

$$= \frac{\delta}{2(n-m+1)}\left(1-\frac{\mathbf{Exp}_I[|\mathcal{V}_{|_I}|]}{|\mathcal{V}|}\right).$$

Note that under this formulation, $|\mathcal{V}_{|_I}|$ (the number of sets in the projection of $\mathcal{V}$ onto I) is a random variable that depends on the choice of $I = (i_1, \ldots, i_{m-1})$, in particular, since it does not depend on the choice of i_m, we have $\mathbf{Exp}_{I'}[|\mathcal{V}_{|_I}|] = \mathbf{Exp}_I[|\mathcal{V}_{|_I}|]$.

The analysis of Haussler [Hau95] then proceeds as follows. We assume that $\mathcal{V}$ is δ-separated as in Lemma 19 (then a bound on $|\mathcal{V}|$ is the actual bound on the packing number), and then choose

$$m := \left\lceil \frac{(2d_0+2)(n+1)}{\delta+2d_0+2} \right\rceil,$$

indices $i_1, \ldots, i_m$ uniformly at random without replacements from $[n]$ (without loss of generality, we can assume that $\delta \geq 3$ as, otherwise, we set the bound on the packing to be $O(n^{d_1}k^{d-d_1})$, as asserted by the (d, d_1) Clarkson-Shor property of $\mathcal{V}$. Moreover, we can assume, without loss of generality, $n \geq d_0, \delta$, and thus we have $m \leq n$). Put $I' = \{i_1, \ldots, i_m\}$, $I = I' \setminus \{i_m\}$. Then the analysis in [Hau95] combines the two bounds in Inequalities (13) and (14) in order to derive an upper bound on $|\mathcal{V}|$, from which the bound in the Packing Lemma (Theorem 3) is obtained. Specifically, using simple algebraic manipulations, we obtain:

$$|\mathcal{V}| \leq \frac{\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right]}{1-\frac{2d_0(n-m+1)}{m\delta}} \tag{15}$$

It has been shown in [Hau95] that due to our choice of m, we have:

$$\frac{2d_0(n-m+1)}{m\delta} \leq \frac{d_0}{d_0+1}, \tag{16}$$

from which we obtain Inequality (1), as asserted. This completes the description of our first extension of Haussler's analysis, as described in Section 2.1.

In order to apply our second extension, we observe that one only needs to assume $m \leq n$ when obtaining Inequalities (13) and (14). In addition, the choice of m in Inequality (15) can be made slightly larger (but still smaller than n), since the term $\frac{2d_0(n-m+1)}{m\delta}$ in Inequality (16) is a decreasing function of m, as can easily be verified. Recall that in our analysis we replace m by $m_j := m\log^{(j)}(n/\delta)$, where $2 \leq j \leq \log^*(n/\delta)$, in which case we still obtain $|\mathcal{V}| \leq (d_0+1)\,\mathbf{Exp}_{I_j}\left[\left|\mathcal{V}_{|_{I_j}}\right|\right]$, where I_j is a set of $m_j - 1$ indices chosen uniformly at random without replacements from $[n]$.

B Proof of Theorem 17.

We use the framework of Ezra [Ezr16], together with the Shallow Packing Lemma (Theorem 4), to show Theorem 17. To this end, we use the machinery and notation in [Ezr16] and only revise the proof of [Ezr16, Theorem 3.5], where we now plug into the analysis the new bound in (Theorem 4). In particular, we use the same notation and definitions for $\mathcal{F}_j^i$ (representing an appropriate collection of "canonical sets") and Δ_j^i (representing a desired discrepancy bound on the canonical sets), and then bound an appropriate entropy function in order to apply the mechanism of Lovett and Meka [LM12]. We do not repeat these details here, and instead refer the reader for the notation and machinery presented in [Ezr16]. In the sequel we only show the derivation of the bound for Δ_j^i, then the derivation of the discrepancy bound for the original sets $S \in \Sigma$ is straightforward by the analysis in [Ezr16], in which S is represented by the disjoint union of the symmetric difference of pairs of canonical sets.

Assume without loss of generality that $\log n$ is an integer, and let $k := \log n$. By Theorem 4 and the analysis in [Ezr16], we have that for each $i = 1, \ldots, k$ and $j = i-1, \ldots, k$,

$$|\mathcal{F}_j^i| \leq C \cdot \frac{2^{jd}}{2^{(i-1)(d-d_1)}},$$

where $C > 0$ is an appropriate constant as stated in Theorem 4. By the construction in [Ezr16], each set $F_j^i \in \mathcal{F}_j^i$ satisfies $|F_j^i| = O(n/2^{i-1})$, for a fixed index i, and any $j = i-1, \ldots, k$.

Our discrepancy parameter Δ_j^i is chosen as follows:

$$\Delta_j^i := A \cdot \frac{1}{(1+|j-j_0|)^2} \left(\frac{n^{1/2-1/(2d)}}{2^{(i-1)\cdot(1/2-d_1/(2d))}} \right) \log^{1/2d} n, \tag{17}$$

where

$$j_0 := (1/d)\log n + (1-d_1/d)(i-1) - (1/d)\log\log n - B,$$

for an appropriate constant $B > 5 + \log C$, and for a sufficiently large constant of proportionality $A > 0$, whose choice depends on B, and will be determined shortly (note that all the three constants A, B, and C depend on d).

In order to apply the constructive discrepancy minimization technique of Lovett and Meka [LM12], we need to show:

Lemma 20 *Put $s_j := n/2^{j-1}$. The choice in (17), for $A > 0$ sufficiently large (whose choice depends on C and thus on d), satisfies*

$$\sum_{i=1}^{k} \sum_{j=i-1}^{k} C \cdot \frac{2^{jd}}{2^{(d-d_1)(i-1)}} \exp\left(-\frac{\left(\Delta_j^i\right)^2}{16 s_j} \right) \leq \frac{n}{16}. \tag{18}$$

We next proceed almost verbatim as in [Ezr16].

We first note that at $j = j_0$ the above exponent becomes a constant, whereas the bound $C \cdot \frac{2^{jd}}{2^{(d-d_1)(i-1)}}$ (representing an appropriate packing number) becomes roughly $n/\log n$ (for a fixed

index i). Indeed, applying our choice in (17), we have

$$\exp\left(-\frac{\left(\Delta_j^i\right)^2}{16s_j}\right) = \exp\left(-\frac{A^2\cdot 2^{j-1}\log^{1/d} n}{16(1+|j-j_0|)^4 n^{1/d} 2^{(1-d_1/d)\cdot(i-1)}}\right),$$

which is $\exp\left(-\frac{A^2}{16\cdot 2^{B+1}}\right)$ at $j=j_0=(1/d)\log n+(1-d_1/d)(i-1)-(1/d)\log\log n-B$. Concerning the bound $C\cdot\frac{2^{jd}}{2^{(d-d_1)(i-1)}}$, at $j=j_0$ we obtain:

$$C\cdot\frac{n2^{(d-d_1)(i-1)}}{2^{(d-d_1)(i-1)}2^{dB}\log n} = C\cdot\frac{n}{2^{dB}\log n},$$

as asserted.

We now fix an index i, split the summation into the two parts $j\geq j_0$ and $i-1\leq j<j_0$, and then bound each part in turn. In the first part, the exponent will "take over" the summation in the sense that it decreases superexponentially, making the other factors (with $j>j_0$) insignificant, and in the second part, the packing size will decrease geometrically. Thus the "peak" of this summation is obtained at $j=j_0$, and is decreasing as we go beyond or below j.

For the first part, put $j:=j_0+l$, for an integer $l\geq 0$, and then

$$\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=j_0}^{k} C\cdot\frac{2^{jd}}{2^{(d-d_1)(i-1)}}\exp\left(-\frac{\left(\Delta_j^i\right)^2}{16s_j}\right)\\
&\leq \sum_{i=1}^{k}\sum_{l=0}^{k-j_0} C\cdot\frac{n2^{ld}}{2^{dB}\log n}\exp\left(-\frac{A^2}{16}\cdot\frac{2^{l-(B+1)}}{(1+l)^4}\right)\\
&\leq C\cdot n2^{-dB}\sum_{l=0}^{k-j_0} 2^{ld}\exp\left(-\frac{A^2}{16}\cdot\frac{2^{l-(B+1)}}{(1+l)^4}\right),
\end{aligned}$$

where the logarithmic factor is now eliminated due to the summation over i. The exponents in the above sum decrease superexponentially. Choosing A sufficiently large (say, $A>2^{6+(B+1)+\log d}$) and having $B>5+\log C$ as above, we can guarantee that the latter sum is strictly smaller than $n/32$.

When $j<j_0$, put $j:=j_0-l$, $l>0$ as above. We now obtain, by just bounding the exponent from above by 1, and using similar considerations as above:

$$\sum_{i=1}^{k}\sum_{j=i-1}^{j_0-1} C\cdot\frac{2^{jd}}{2^{(d-d_1)(i-1)}}\exp\left(-\frac{{\Delta_j^i}^2}{16s_j}\right)\leq\sum_{i=1}^{k}\sum_{j=i-1}^{j_0-1} C\cdot\frac{n}{\log n 2^{d(l+B)}}\leq\sum_{l=1}^{j_0-(i-1)} C\cdot\frac{n}{2^{d(l+B)}}.$$

Once again, our choice for B guarantees that the above (geometrically decreasing) sum is strictly smaller than $n/32$. Thus the entire summation is bounded by $n/16$, as asserted.

This completes the proof of Theorem 17.