Two Proofs for Shallow Packings

Kunal Dutta[1], Esther Ezra[2], and Arijit Ghosh[1]

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

2 **Department of Computer Science and Engineering**
Polytechnic Institute of NYU, Brooklyn, NY 11201-3840, USA; and
School of Mathematics
Georgia Institute of Technology, Atlanta, Georgia 30332, USA
esther@courant.nyu.edu

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 the 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.

1998 ACM Subject Classification F.2.2 [Nonnumerical Algorithms and Problems] Computations on discrete structures, Geometrical problems and computations, F.1.2 [Modes of Computation] Probabilistic computation

Keywords and phrases Set systems of bounded primal shatter dimension, δ-packing and Haussler's approach, relative approximations, Clarkson-Shor random sampling approach

Digital Object Identifier 10.4230/LIPIcs.SOCG.2015.96

1 Introduction

Let $\mathcal{V}$ be a set system defined over an n-point set X. We follow the notation in [19], 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 $\mathcal{V}$:


© Kunal Dutta, Esther Ezra, and Arijit Ghosh;
licensed under Creative Commons License CC-BY
31st International Symposium on Computational Geometry (SoCG'15).
Editors: Lars Arge and János Pach; pp. 96–110

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

▶ **Definition 1** (Primal Shatter Function [21, 27])**.** 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 [19, 32])**.** 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 [29, 31] 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., [16]).

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., [16] 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 $\mathcal{A}(S)$ are called "vertices", and d-dimensional cells are simply referred to as "cells"; see [30] 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 [16] and the references therein.

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 $\mathbf{u}$, $\mathbf{v}$.

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 the largest δ-separated subset $\mathcal{W} \subseteq \mathcal{V}$. A key property, originally shown by Haussler [19] (see also [8, 9, 11, 27, 33]), is that set systems of bounded primal shatter dimension admit small δ-packing numbers. That is:

▶ **Theorem 3** (Packing Lemma [19, 27])**.** *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 [19] the assumption is that the set system has a finite VC-dimension. However, its formulation in [27], which is based on a simplification of

the analysis of Haussler by Chazelle [8], relies on the assumption that the primal shatter dimension is d, which is the actual bound that we state in Theorem 3. We also comment that a closer inspection of the analysis in [19] shows that this assumption can be replaced with that of having bounded primal shatter dimension (independent of the analysis in [8]). We describe these considerations in Section 2.1.

Previous work. In his seminal work, Dudley [11] 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 [18] and Pollard [28] for an alternative proof and a specification of the constant of proportionality). This bound was later improved by Haussler [19], who showed $\mathcal{M}(\delta, \mathcal{V}) \leq e(d_0+1)\left(\frac{2en}{\delta}\right)^{d_0}$ (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., [18], 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 [19] (Theorem 3) is in fact a generalization of the so-called Sauer-Shelah Lemma [29, 31], asserting that $|\mathcal{V}| \leq (en/d_0)^{d_0}$, where e is the base of the natural logarithm, 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 [19] 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.* [15] 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.* [6], who achieved the same tightening.

A major application of packing is in obtaining improved bounds on the *sample complexity* in machine learning. This was studied by Li *et al.* [22] (see also [18]), 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 [17] in the context of geometric set systems, 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 [33] in order to construct spanning trees of low crossing number (see also [27]) and by Matoušek [26, 27] in order to obtain asymptotically tight bounds in geometric discrepancy.

Our result

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 realizations. 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 [14] for an example.

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 [10], 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.[1] 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 $1 \leq k \leq n$. By these assumptions, 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 is empty. 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, and whose VC-dim is d_0. Let $\delta \geq 1$ be an integer parameter, and k an integer 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 (and d_0).

This problem has initially been addressed by the second author in [13] 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 [13] is a refinement over the technique of Dudley [11] combined with the existence of small-size *relative approximations* (see [13] 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 [8] to the Packing Lemma (Section 3).

2 First Proof: Refining Haussler's Approach

2.1 Preliminaries

Overview of Haussler's Approach

For the sake of completeness, we repeat some of the details in the analysis of Haussler [19] 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

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

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 [19] 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

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

See a preliminary version of this paper for details, as well as the facts that $m \leq n$ and I consists of precisely $m-1$ indices [14, Appendix B].

Moreover, we refine Haussler's analysis to include two natural extensions (see [14, Appendix B] for details): *(i) Obtain a refined bound on* $\mathbf{Exp}_I\left[|\mathcal{V}_{|_I}|\right]$*:* This extension is a direct consequence of Inequality (1). 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.[2] 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 $|\mathcal{V}|$. 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 [14, Appendix B]), 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.

---

[2] We note, however, that the original analysis of Haussler [19] 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 the approach.

We next present the proof of Theorem 4. In what follows, we assume that $\mathcal{V}$ is δ-separated. 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 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 $O(m^{d_1}k^{d-d_1})$. 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 carefulness 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 $O\left(\frac{n^{d_1}k^{d-d_1}(\log^{(j-1)}(n/\delta))^d}{\delta^d}\right)$ (note that the actual constant of proportionality in our recursive scheme is 1, see Lemma 8), where the base case $j = 2$ is shown in Corollary 6.

A key property in the analysis is then 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. Specifically, 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 $|\mathcal{V}|$. These properties are described more rigorously in Lemma 8, where derive a recursive inequality for $|\mathcal{V}|^{(j)}$ using the bound on $\mathbf{Exp}_{I_j}\left[|\mathcal{V}_{|_{I_j}}|\right]$. 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 (4).

In what follows, we also assume that $\delta \leq n/2^{(d_0+1)}$ (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[3] 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 $O(k \cdot m_1/n)$.

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

A set I_1 as above exists by the considerations in [13]. See also a preliminray version of this paper for further details [14, Appendix C].

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 [11].

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 (see a preliminary version of this paper [14, Appendix D] for the proof):

[3] In this particular step we use a different machinery than that of Haussler [19]; 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).

▶ **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.

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

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 (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 Clarkson-Shor property of $\mathcal{V}$.

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):

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

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 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 $\delta \leq n/2^{(d_0+1)}$, and the fact that $D > d$ is sufficiently large, imply that the coefficient of the recursive term is smaller than 1, for any $2 \leq j \leq 1 + \log^*(n/\delta) - \log^*(d_0 + 1)$.[4] Then, using induction on j, one can verify that the solution is

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

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

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 $j^* = 1 + \log^*(n/\delta) - \log^*(d_0 + 1)$, 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 on d (and d_0). 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 systems in set-systems of bounded primal shatter dimension building on Chazelle's presentation of Haussler's proof, (which has been described by Matoušek as "a magician's trick") as explained in [27]. By Haussler's result [19], we know that $M = O(n/\delta)^d = (n/\delta)^{d_1}(l/\delta)^{d_2}.g(n,l,\delta)^d$, where $g(n,l,d) = O((n/l)^{d_2})$. We would like to show the optimum upper bound for g is independent of n, l. We shall show that the optimal bound (up to constants) is in fact, $g = c^*$, where c^* is the fixed point of $f(x) = c' \log x$, with $c' > 1$ independent of n, l, δ.

Intuition

We provide some intuition for our extension of the Haussler/Chazelle proof below (at least to the reader familiar with it). A naïve attempt to extend Chazelle's proof to shallow packings, fails, because (as in the previous proof), one chooses a random subsequence I, and estimates the number of projections on I, caused by δ-packed vectors of bounded size. For a given vector, its projection on I can be much larger than expected. However, we shall choose A' in a way that the *number* of such "bad" vectors, is at most a constant times their expected *number*. This allows us to get the final bound in a single iteration.

Details

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

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

▶ **Definition 9** (Unit distance graph)**.** For a set system $\mathcal{V}$, unit distance graph $\mathcal{UD}(\mathcal{V})$ is a graph with vertex set $\mathcal{V}$ and a pair $\{\mathbf{v}_1, \mathbf{v}_2\}$ is an edge if $\rho(\mathbf{v}_1, \mathbf{v}_2) = 1$.

Consider a random subsequence of indices $I = (i_1, \ldots, i_s)$ where each $i \in [n]$ is selected with probability $p = \frac{36 d_0 K}{\delta}$, where $K \geq 1$ 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 set $\mathbf{v}_1 \in \mathcal{V}_1$, define the weight of $\mathbf{v}_1$ as: $w(\mathbf{v}_1) := \#\{\mathbf{v} \in \mathcal{V} : \mathbf{v}_{|_I} = \mathbf{v}_1\}$. Observe that

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

Let E be the edge set of $\mathcal{UD}(\mathcal{V}_1)$. Now define the weight of an *edge* $e = \{\mathbf{v}_1, \mathbf{v}_1'\} \in E$ as $w(e) := \min(w(\mathbf{v}_1), w(\mathbf{v}_1'))$. Let $W := \sum_{e \in E} w(e)$. We claim that

▶ **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 [19] for set systems with bounded VC-dimension. The following version appears in Matoušek's book [27]:

▶ **Lemma 11** ([19])**.** *Let $\mathcal{V}$ be a set-system 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 drops by $w(\mathbf{v}_1)$, and the total edge weight drops by at most $2d_0 w(\mathbf{v}_1)$. Continuing the argument until all vertices are removed, we get the claim. ◀

Next, we shall prove a lower bound on the expectation $\mathbf{Exp}[W]$. Choose a random element $i_j \in \{i_1, \ldots, i_s\}$. Let $\mathcal{V}_2 := \mathcal{V}_{|_{I'}}$ where $I' = (i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_s)$, i.e., by abuse of notation $I' = I \setminus \{i_j\}$. Note that I' is a random subsequence where each $i \in [n]$ was chosen with probability $p' = p - 1/n$. Crucially, one can consider the above process equivalent to first choosing I' by selecting each element of $[n]$ with probability p', and then selecting a uniformly random element $i_j \in [n] \setminus I'$ with probability $1/n$.

Let $E_1 \subset E$ 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_j, and let

$$W_1 := \sum_{e \in E_1} w(e).$$

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

$$Y = Y(I') := \#\{\mathbf{v} \in \mathcal{V} : \|\mathbf{v}_{|_{I'}}\| > c(k/\delta)\},$$

i.e., the number of vectors in $\mathcal{V}$, each of whose norm after projecting onto I' is more than $c(l/\delta)$, (where c shall be chosen appropriately). Let $Nice$ denote the event

$$(Y \leq 8\,\mathbf{Exp}[Y]) \cap \left(\frac{np}{2} \leq s \leq \frac{3np}{2}\right) = N_Y \cap N_S.$$

Conditioning W on $Nice$, we get:

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

By Markov's Inequality, see [4, App. A], 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 [4, App. A], with the fact that $n/\delta \geq 1$, we get

$$\mathbf{Prob}\left[\overline{N_S}\right] = \mathbf{Prob}\left[|s - np| > np/2\right] \leq 2e^{(-36d_0Kn/3.2^2\delta)} << 1/4.$$

This implies

$$\mathbf{Prob}\left[Nice = N_Y \cap N_S\right] \quad \geq \quad 1 - \mathbf{Prob}\left[\overline{N_S}\right] - \mathbf{Prob}\left[\overline{N_Y}\right] \geq 7/8 - e^{(-4d_0K)} \geq 3/4,$$

where the last inequality follows from the fact that $d_0K \geq 1$.

Hence,

$$\mathbf{Exp}[W] \geq (3/4)\,\mathbf{Exp}[W|Nice] \geq \frac{3(np/2)}{4}\,\mathbf{Exp}[W_1|Nice], \tag{8}$$

where the last inequality follows by symmetry of the choice of i_j from I, and the lower bound on s when the event $Nice$ holds.

Hence, $\mathbf{Exp}[W] \geq \left(\frac{3np}{8}\right)\mathbf{Exp}[W_1|Nice]$. So to lower bound $\mathbf{Exp}[W]$ up to constants, it suffices just to lower bound $\mathbf{Exp}[W_1|Nice]$. Let W_2 denote $W_1|Nice$. Consider now $\mathbf{Exp}[W_2|A']$. That is, consider a fixed subsequence I' whose length is between $np/2$ and $3np/2$, and which is such that the number of vectors $\mathbf{v} \in \mathcal{V}$ whose norm after projection onto I' in more than ck/δ, is at most $8\,\mathbf{Exp}[Y]$. We shall lower bound $\mathbf{Exp}[W_2|I']$ for this choice of I'.

By definition, $W_1 = \sum_{e \in E_1} w(e)$. Consider the equivalence classes of $\mathcal{V}$ formed by their projection onto I':

$$\mathcal{V} = \mathcal{V}'_1 \cup \ldots \cup \mathcal{V}'_r.$$

Define $Bad \subset [r]$ to be those indices j for which $\mathcal{V}'_j$ is such that

$$\forall \mathbf{v} \in \mathcal{V}'_j : \|\mathbf{v}_{|_{I'}}\| > 8c(k/\delta).$$

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

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

We first estimate the contribution of the classes in $Good$, to the total weight. Consider a class $\mathcal{V}'_i$ such that $i \in Good$. Let $\mathcal{V}''_1 \subset \mathcal{V}'_i$ be those vectors in $\mathcal{V}'_i$ which contains 1 in the i_j-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 $e \in E_1$ formed by the projection of $\mathcal{V}'_i$ onto I, has weight

$$w(e) = \min(b_1, b_2) \geq \frac{b_1 b_2}{b}. \tag{9}$$

Observe that in Inequality (9), b is a constant as the subsequence I' is fixed and the product b_1b_2 is the random variable that depends on the choice of i_j. The product b_1b_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_j-th coordinate. For a given ordered pair $(\mathbf{v}, \mathbf{v}')$ of distinct vectors $\mathbf{v}, \mathbf{v}' \in \mathcal{V}'_i$, the probability $\mathbf{v}$ and $\mathbf{v}'$ differ in the i_j-th coordinate is $\frac{\delta}{n-s+1}$, which is at least $\frac{\delta}{n}$. Therefore, the expected contribution of $(\mathbf{v}, \mathbf{v}')$ to b_1b_2 is at least $\frac{\delta}{n}$ and this implies

$$\mathbf{Exp}[b_1b_2] \geq \frac{b(b-1)\delta}{n}.$$

And this further implies the together with Inequality (9) that the weight of e (conditioned on $Nice$ and I') is at least:

$$\mathbf{Exp}[w(e)|Nice \cap I'] \geq \frac{1}{b}\,\mathbf{Exp}[b_1b_2] \geq \frac{b(b-1)}{b}.\frac{\delta}{n} = (b-1)\frac{\delta}{n} = (|\mathcal{V}'_i| - 1)\frac{\delta}{n}.$$

Hence, the expected weight of $\mathbf{Exp}[W_2|I']$ is:

$$\mathbf{Exp}[W_2|I'] \geq \sum_{e\in E_1} \mathbf{Exp}[w(e)|Nice \cap I'] \geq \sum_{i\in Good} (|\mathcal{V}'_i| - 1)\frac{\delta}{n}$$

But by (d, d_1) Clarkson-Shor property, we have that

$$\forall i \in Good,\ \ |(\mathcal{V}'_i)_{|_{I'}}| \leq Cs^{d_1}(ckp)^{d-d_1}.$$

Substituting in the lower bound for $\mathbf{Exp}[W_2]$, we get:

$$\begin{aligned}
\mathbf{Exp}[W_2|A'] &\geq \left(\left(\sum_{i\in Good} |\mathcal{V}'_i|\right) - C(1.5np)^{d_1}(ckp)^{d-d_1}\right)\frac{\delta}{n} \\
&\geq \left(|\mathcal{V}| - 8\,\mathbf{Exp}[Y] - C(6dK)^d.(1.5)^{d_1}c^{d-d_1}\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}\right)\frac{\delta}{n} \\
&\geq \left(\mathcal{M}(\delta,k,\mathcal{V}) - 8\,\mathbf{Exp}[Y] - C_1K^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}\right)\frac{\delta}{n}
\end{aligned}$$

where in the first inequality, we used the fact that the event $N_S \subset Nice$ holds, and in the last line, $C_1 = C.(6d)^d 2^{d_1} c^{d-d_1}$. Since the above holds for each I' which satisfies $Nice$, we get that

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

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

$$(3np/8)\,\mathbf{Exp}[W_1|Nice] \leq \mathbf{Exp}[W] \leq 2d_0\mathcal{M}(\delta,k,\mathcal{V}),$$

and substituting the lower bound $\mathbf{Exp}[W_1|Nice]$, and solving for $\mathcal{M}(\delta,k,\mathcal{V})$, we get

$$\mathcal{M}(\delta,k,\mathcal{V}) \leq \frac{(27K/4)\left(8\,\mathbf{Exp}[Y] + C_1K^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}\right)}{(27K/4-1)}.$$

The following lemma therefore, completes the proof:

▶ **Lemma 12.** *For* $K = \max\{1, (\ln g)/36\}$, $\mathbf{Exp}[Y] \leq C_2\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d_2}$.

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

$$\begin{aligned}
g^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1} &= \mathcal{M}(\delta,k,\mathcal{V}) \\
&\leq \frac{C_1K^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1} + 8C_2\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}}{1-4/27K} \\
&\leq C_3K^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1} \leq C_4(\max\{1,\log g\})^d\left(\frac{n}{\delta}\right)^{d_1}\left(\frac{k}{\delta}\right)^{d-d_1}
\end{aligned}$$

where the shorthand $g = g(n,l,\delta)$. This implies that $g^d \leq C_4(\max\{1,\log g\})^d$, or $g \leq C_5\max\{1,\log g\}$. Since for any g growing with n, l, or δ, we would have $g >> C_5\log g$ for sufficiently large n, k or δ, this inequality is only satisfiable when g is a constant function of n, l, δ. i.e. $g \leq c^*$, where c^* is independent of n, k, δ.

It only remains to prove Claim 12:

Proof of Lemma 12. The proof follows easily from Chernoff Bounds. Fix $\mathbf{v} \in \mathcal{V}$. Let $Z = \|\mathbf{v}_{|_{I'}}\|$. Then $\mathbf{Exp}[Z] = \|\mathbf{v}\|p' = kp'$. Since I' is a random subsequence chosen with probability $p' = p - 1/n$, the probability that

$$Z \geq ckp' = \frac{36cdKk}{\delta} - \frac{ck}{n}$$

is upper bounded using Chernoff bounds, see [4, App. A], as:

$$\mathbf{Prob}\left[Z - \mathbf{Exp}[Z] > (c-1)\,\mathbf{Exp}[Z]\right] \leq e^{(-\mathbf{Exp}[Z])} \leq e^{(-36dKk/\delta)},$$

for $c = 1.01e$ and $n \geq 100$, say. Hence the expected number $\mathbf{Exp}[Y]$ of vectors, each of whose norm when projected onto I' in more than $36cdKk/\delta$ elements, is at most:

$$\mathbf{Exp}[Y] \leq \mathcal{M}(\delta, k, \mathcal{V})e^{(-36dKk/\delta)} \leq \mathcal{M}(\delta, k, \mathcal{V})e^{(-18dK)},$$

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

$$\mathbf{Exp}[Y] \leq g^d \left(\frac{n}{\delta}\right)^{d_1} \left(\frac{k}{\delta}\right)^{d-d_1} e^{(-18dK)} \leq \left(\frac{n}{\delta}\right)^{d_1} \left(\frac{k}{\delta}\right)^{d-d_1} e^{d(\ln g - 18K)} \leq \left(\frac{n}{\delta}\right)^{d_1} \left(\frac{k}{\delta}\right)^{d-d_1}$$

for $K \geq (\ln g)/18$. ◀

This completes the proof of Theorem 4.

4 Concluding Remarks and Further Research

We briefly mention a few applications of Theorem 4:

(i) Smaller packing numbers for several natural geometric set systems under the shallowness assumption. Letting $d > 1$ be an integer parameter, this includes set systems of points and halfspaces in d-dimensions, balls in d-dimensions, parallel slabs of arbitrary width in d-dimensions, as well as dual set systems defined over $(d-1)$-variate (not necessarily continuous or totally defined) functions F of *constant description complexity*. These results are described in detail in a preliminary version of this paper [14, Appendix B].

(ii) Spanning trees with low total conflict number. This is based on the machinery of Welzl [33] to construct spanning trees of low crossing number (see also [27]). Here the tree spans $\mathcal{V}$ (representing, say, a set of regions defined over n points in d-space), and the "conflict number" of an edge (u, v) is the symmetric difference distance between u and v. See [14, Appendix B] for further details.

(iii) Geometric discrepancy. Following the previous work of the second author [13], the new bound in Theorem 4 leads to an improved discrepancy bound that is sensitive to the size of the sets in various geometric set systems, including point and halfspaces in d-dimensions, this is mentioned in [14] and described in detail in the preliminary work of the first and the third author [12]. As a consequence, it is shown in [12] how to derive an improved bound on relative (ε, δ)-approximations by adapting the approach in [13]. Last, but not least, it is shown in [12] that the bound in Theorem 4 leads to better bounds on the discrepancy of geometric set systems of low degree, as long as $d_1 = 1$.

References

1 P. K. Agarwal, A. Efrat, and M. Sharir. Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications. *SIAM J. Comput.*, 29(2000):912–953.

2 P. K. Agarwal and J. Erickson. Geometric range searching and its relatives. *Discrete Comput. Geom.* (1997).

3 P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measures of Ppoints. *J. ACM*, 51(4):606–635 (2004).

4 N. Alon and J. H. Spencer. *The Probabilistic Method.* 2*nd* Edition, Wiley-Interscience, New York, USA, 2000.

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

6 N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *J. Comput. System Sci.*, 75(6):323–335 (2009).

7 T. M. Chan. Dynamic coresets. *Discrete Comput. Geom.*, 42: 469–488 (2009).

8 B. Chazelle. A note on Haussler's packing lemma. Unpublished manuscript, Princeton (1992).

9 B. Chazelle and E. Welzl. Quasi-optimal range searching in spaces of finite VC-dimension. *Discrete Comput. Geom.*, 4:467–489 (1989).

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

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

12 K. Dutta and A. Ghosh. Size sensitive packing number for Hamming cube and its consequences. CoRR abs/1412.3922 (2014).

13 E. Ezra. A size-sensitive discrepancy bound for set systems of bounded primal shatter dimension. In *Proc. Twenty-Fifth Annu. ACM-SIAM Sympos. Discrete Algorithms*, pages 1378–1388 (2014).

14 E. Ezra. Shallow Packings in Geometry. CoRR abs/1412.5215 (2014).

15 L. Gottlieb, A. Kontorovich, and E. Mossel. VC bounds on the cardinality of nearly orthogonal function classes. Discrete Math., 312(10):1766–1775 (2012).

16 S. Har-Peled. *Geometric Approximation Algorithms*, Mathematical Surveys and Monographs, Vol. 173 (2011).

17 S. Har-Peled and M. Sharir, Relative (p, ε)-approximations in geometry, *Discrete Comput. Geom.*, 45(3):462–496 (2011).

18 D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. In *Information and Computation*, 100(1):78–150 (1992).

19 D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combinatorial Theory Ser. A*, 69:217–232 (1995).

20 D. Haussler, N. Littlestone, M. K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2), 248–292 (1994).

21 D. Haussler and E. Welzl. ε-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151 (1987).

22 Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. Sys. Sci.*, 62(3):516–527 (2001).

23 S. Lovett and R. Meka. Constructive discrepancy minimization by walking on the edges. In *Proc. 53th Annu. IEEE Symp. Found. Comput. Sci.*, 61–67, (2012).

24 J. Matoušek, *Lectures on Discrete Geometry*, Springer-Verlag New York (2002).

25 J. Matoušek. Reporting points in halfspaces. *Comput. Geom. Theory Appl.*, 2:169–186 (1992).

26 J. Matoušek. Tight upper bounds for the discrepancy of halfspaces. *Discrete Comput. Geom.*, 13:593–601 (1995).

27 J. Matoušek. *Geometric Discrepancy*, Algorithms and Combinatorics, Vol. 18, Springer Verlag, Heidelberg (1999).

28 D. Pollard. *Convergence of Stochastic Processes*, Springer-Verlag (1984).

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

30 M. Sharir and P. K. Agarwal. *Davenport-Schinzel Sequences and Their Geometric Applications*. Cambridge University Press, New York (1995).

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

32 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).

33 E. Welzl. On spanning trees with low crossing numbers. In *Data Structures and Efficient Algorithms, Final Report on the DFG Special Joint Initiative*, volume 594 of *Lect. Notes in Comp. Sci.*, Springer-Verlag, Heidelberg, pp. 233–249 (1992).