Shifting: One-Inclusion Mistake Bounds and Sample Compression

Benjamin I. P. Rubinstein
Peter Bartlett
J. Hyam Rubinstein

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-86
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-86.html

June 25, 2007

Acknowledgement

We gratefully acknowledge the support of the NSF under award DMS-0434383.

Shifting: One-Inclusion Mistake Bounds and Sample Compression

Benjamin I. P. Rubinstein [a,*], Peter L. Bartlett [a,b] J. Hyam Rubinstein [c]

[a] *Computer Science Division, University of California, Berkeley, 387 Soda Hall #1776, Berkeley, CA 94720-1776, USA*

[b] *Department of Statistics, University of California, Berkeley, 367 Evans Hall #3860, Berkeley CA 94720-3860, USA*

[c] *Department of Mathematics & Statistics, the University of Melbourne, Parkville, VIC 3010, Australia*

Abstract

We present new expected risk bounds for binary and multiclass prediction, and resolve several recent conjectures on sample compressibility due to Kuzmin and Warmuth. By exploiting the combinatorial structure of concept class $\mathcal{F}$, Haussler *et al.* achieved a $\mathrm{VC}(\mathcal{F})/n$ bound for the natural one-inclusion prediction strategy. The key step in their proof is a $d = \mathrm{VC}(\mathcal{F})$ bound on the graph density of a subgraph of the hypercube—one-inclusion graph. The first main result of this report is a density bound of $n \binom{n-1}{\leq d-1} / \binom{n}{\leq d} < d$, which positively resolves a conjecture of Kuzmin and Warmuth relating to their unlabeled Peeling compression scheme and also leads to an improved one-inclusion mistake bound. The proof uses a new form of VC-invariant shifting and a group-theoretic symmetrization. Our second main result is an algebraic topological property of maximum classes of VC-dimension d as being d-contractible simplicial complexes, extending the well-known characterization that $d = 1$ maximum classes are trees. We negatively resolve a minimum degree conjecture of Kuzmin and Warmuth—the second part to a conjectured proof of correctness for Peeling—that every class has one-inclusion minimum degree at most its VC-dimension. Our final main result is a k-class analogue of the d/n mistake bound, replacing the VC-dimension by the Pollard pseudo-dimension and the one-inclusion strategy by its natural hypergraph generalization. This result improves on known PAC-based expected risk bounds by a factor of $O(\log n)$ and is shown to be optimal up to a $O(\log k)$ factor. The combinatorial technique of shifting takes a central role in understanding the one-inclusion (hyper)graph and is a running theme throughout.

Key words: one-inclusion mistake bounds, worst-case expected risk, multiclass prediction, sample compression, shifting

1. Introduction

In [13,12] Haussler, Littlestone and Warmuth proposed the one-inclusion prediction strategy as a natural approach to the prediction (or mistake-driven) model of learning, in which a prediction strategy maps a training sample and test point to a test prediction with hopefully guaranteed low probability of erring. The significance of their contribution was two-fold. On the one hand the derived $\mathrm{VC}(\mathcal{F})/n$ upper-bound on the

* Corresponding author.
Email address: benr@cs.berkeley.edu (Benjamin I. P. Rubinstein).

worst-case expected risk of the one-inclusion strategy learning from $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ improved on the previous-best bound for consistent learners by an order of $\log n$. This was achieved by taking the combinatorial structure of the underlying $\mathcal{F}$ into account—which had not been done in previous work—in order to break ties between hypotheses consistent with the training set but offering contradictory predictions on a given test point. At the same time Haussler [12] introduced the idea of *shifting* subsets of the n-cube down around the origin—an idea previously developed in combinatorics—as a powerful tool for learning-theoretic results. In particular, shifting admitted deeply insightful proofs of Sauer's Lemma and a VC-dimension bound on the density of the one-inclusion graph—the key result needed for the one-inclusion strategy's expected risk bound.

Recently shifting has impacted work towards the sample compressibility conjecture of [19], in [17]. In order to k-compress a concept class C, one must be able to compress any sample s consistent with C to a subsample of length at most k and then be able to map such a compressed-set to some s-consistent concept (not necessarily belonging to C). Given a k-compression scheme for bounded k, Littlestone and Warmuth demonstrated a proof for the learnability of C that is simpler than proofs based on finite VC-dimension. The necessity of having a bounded compression scheme for learnability motivated the compression conjecture, which states that every concept class C of VC-dimension d has a d (or order d)-compression scheme.

This paper continues the study of the one-inclusion graph—the natural graph structure induced by a subset of the n-cube—and its related prediction strategy under the lens of shifting. After the necessary background including the prediction model of learning, PAC-based expected risk bounds, the one-inclusion prediction strategy and sample compressibility summarized in Section 2, we develop the technique of shatter-invariant shifting in Section 3.1. While a subset's VC-dimension cannot be increased by Haussler's shifting, shatter-invariant shifting guarantees a finite sequence of shifts to a fixed-point under which the shattering of a chosen set remains *invariant*, thus preserving VC-dimension throughout.

In Section 3.2 we apply a group-theoretic symmetrization to tighten the mistake bound—the worst-case expected risk bound—of the deterministic one-inclusion strategy from d/n to $\lceil D_n^d \rceil / n$, where $D_n^d < d$ for all n, d; the bound for the randomized one-inclusion strategy is improved to D_n^d/n. The derived D_n^d density bound positively resolves a conjecture of Kuzmin and Warmuth which was suggested as a step towards a correctness proof of the Peeling unlabeled compression scheme [17]. In Section 5 we provide counter-examples to another conjecture of Kuzmin and Warmuth which is the second step of the conjectured correctness proof; Section 6 discusses the consequences of our combinatorial results for sample compression. Notably, a proof of correctness for Peeling would imply a result on the inembeddability of maximal classes into certain maximum classes.

Section 4 explores characterizations and properties of one-inclusion graphs and maximum/maximal concept classes. A colorability characterization of one-inclusion-isomorphic graphs, extending previous work on characterizing graphs embeddable in the n-cube [7,15,14], is provided. We extend the work on forbidden labels of Floyd [8] slightly to cubical characterizations of both maximum and non-maximum maximal classes on finite domains. Finally we extend the classic result of Dudley [5] that a maximum concept class of VC-dimension 1 is a tree. We show that maximum classes of VC-dimension d on finite domains are in fact d-contractible simplicial complexes, the natural generalization of trees in algebraic topology.

Finally we generalize the prediction model, the one-inclusion strategy and its bounds from binary to k-class learning in Section 7. To date, the best bound on expected risk in this case is $O(\alpha \log \alpha)$ for $\alpha = \Psi_{\mathrm{G}}\text{-dim}\,(\mathcal{F})\,/n$, where $\Psi_{\mathrm{G}}\text{-dim}\,(\mathcal{F})$ denotes the graph dimension of $\mathcal{F}$. We derive a bound of $\Psi_{\mathrm{P}}\text{-dim}\,(\mathcal{F})\,/n$, which improves the dependence on n by a log factor. Here, $\Psi_{\mathrm{P}}\text{-dim}\,(\mathcal{F})$ is the Pollard dimension of $\mathcal{F}$. We show that this bound is at most an $O(\log k)$ factor from optimal. Thus, as in the binary case, exploiting class structure enables significantly better bounds on expected risk for multiclass prediction.

A preliminary version of this report appeared as [21].

2. Definitions & background

We begin with some notation. Sets/random variables, scalars and vectors will be written in uppercase, lowercase and bolded typeface, respectively, as in $C, x, \boldsymbol{v}$. The set of natural numbers $\mathbb{N}$ is defined as the positive integers. We define $\binom{n}{\leq r} = \sum_{i=0}^{r} \binom{n}{i}$ to be the number of subsets of size at most r in a set of cardinality n. Let $[n] = \{1, \ldots, n\}$ and S_n be the set of permutations on $[n]$. We write the density of graph $G = (V, E)$ as $\operatorname{dens}(G) = |E|/|V|$ and graph minimum degree as $\delta(G)$. The bit-wise exclusive-OR of strings $\boldsymbol{u}, \boldsymbol{v} \in \{0,1\}^n$ will be written as $\boldsymbol{u} \operatorname{\mathsf{xor}} \boldsymbol{v}$. $\mathbf{1}[A]$ denotes the indicator function of A.

2.1. The prediction model of learning

We begin with the basic setup of [13]. The set $\mathcal{X}$ is the *domain* and $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ is a *concept class* on $\mathcal{X}$. For notational convenience we write $\operatorname{sam}(\boldsymbol{x}, f) = ((x_1, f(x_1)), \ldots, (x_n, f(x_n)))$ for $\boldsymbol{x} \in \mathcal{X}^n$, $f \in \mathcal{F}$. A *prediction strategy* is a mapping of the form $Q : \bigcup_{n>1} (\mathcal{X} \times \{0,1\})^{n-1} \times \mathcal{X} \to \{0,1\}$, taking a labeled sample and test point to a prediction of the point's label.

Definition 1 (Mistake bounds) *The* prediction model *of learning concerns the following scenario. Given full knowledge of strategy Q, an adversary picks a distribution P on $\mathcal{X}$ and concept $f \in \mathcal{F}$ so as to maximize the probability of $\{Q(\operatorname{sam}(X_1, \ldots, X_{n-1}, f), X_n) \neq f(X_n)\}$ where $X_i \overset{i.i.d.}{\sim} P$. Thus the measure of performance is the worst-case expected risk*

$$\hat{M}_{Q,\mathcal{F}}(n) = \sup_{f \in \mathcal{F}} \sup_{P} \mathbb{E}_{\boldsymbol{X} \sim P^n} \left[\mathbf{1}\left[Q\left(\operatorname{sam}\left((X_1, \ldots, X_{n-1}), f \right), X_n \right) \neq f(X_n) \right] \right] \ .$$

A mistake bound *for Q with respect to $\mathcal{F}$ is an upper-bound on $\hat{M}_{Q,\mathcal{F}}$.*

While Valiant's Probably Approximately Correct (PAC) model is concerned with showing that $\Pr\left(\mathbb{E}\left[\mathbf{1}\left[Q(\operatorname{sam}\left((X_1, \ldots, X_{n-1}), f\right), X_n) \neq f(X_n)\right] \mid X_1, \ldots, X_{n-1}\right] > \epsilon\right)$ is small (i.e. the risk is concentrated close to 0), the prediction model focuses on the size of the expectation $\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left[Q(\operatorname{sam}\left((X_1, \ldots, X_{n-1}), f\right), X_n) \neq f(X_n)\right] \mid X_1, \ldots, X_{n-1}\right]\right]$ (i.e. that the expected risk is close to 0). The following allows us to derive mistake-bounds by bounding a worst-case average [13, Corollary 2.1].

Lemma 2 (Permutation mistake bounds) *For any $n > 1$, concept class $\mathcal{F}$ and prediction strategy $\mathcal{Q}$,*

$$\begin{aligned} \hat{M}_{Q,\mathcal{F}}(n) &\leq \sup_{f \in \mathcal{F}} \sup_{\boldsymbol{x} \in \mathcal{X}^n} \frac{1}{n!} \sum_{g \in S_n} \mathbf{1}\left[Q\left(\operatorname{sam}\left(\left(x_{g(1)}, \ldots, x_{g(n-1)} \right), f \right), x_{g(n)} \right) \neq f\left(x_{g(n)} \right) \right] \\ &= \hat{\hat{M}}_{Q,\mathcal{F}}(n) \ . \end{aligned}$$

A permutation mistake bound *for Q with respect to $\mathcal{F}$ is an upper-bound on $\hat{\hat{M}}_{Q,\mathcal{F}}$.*

2.2. The capacity of function classes contained in $\{0, \ldots, k\}^{\mathcal{X}}$

For a finite set $\mathcal{Y}$, we denote by $\Pi_{\boldsymbol{x}}(\mathcal{F}) = \{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{F}\}$ the projection of $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ on $\boldsymbol{x} \in \mathcal{X}^n$—the equivalence classes of functions induced by labelings of $\boldsymbol{x}$.

Definition 3 *The* Vapnik-Chervonenkis dimension *of concept class $\mathcal{F}$ is defined as* $\operatorname{VC}(\mathcal{F}) = \sup\{n \mid \exists \boldsymbol{x} \in \mathcal{X}^n, \Pi_{\boldsymbol{x}}(\mathcal{F}) = \{0,1\}^n\}$. *Any $\boldsymbol{x}$ satisfying $\{0,1\}^{|\boldsymbol{x}|} = \Pi_{\boldsymbol{x}}(\mathcal{F})$ is said to be* shattered *by $\mathcal{F}$.*

Lemma 4 (Sauer's Lemma [22]) *For any $n \in \mathbb{N}$ and $V \subseteq \{0,1\}^n$, $|V| \leq \binom{n}{\leq \operatorname{VC}(V)}$. A subset V satisfying $\forall c \in \{0,1\}^n, \operatorname{VC}(V \cup \{c\}) > \operatorname{VC}(V)$ is known as* maximal*; if furthermore V meets Sauer's Lemma with equality then it is called* maximum*.*

It is well-known that the VC-dimension is an inappropriate measure of capacity when $|\mathcal{Y}| > 2$. The following unifying framework of class capacities for $|\mathcal{Y}| < \infty$ is due to [2].

Definition 5 (Translation framework for multiclass capacity) *Let $k \in \mathbb{N}$, $\mathcal{F} \subseteq \{0,\dots,k\}^{\mathcal{X}}$ and Ψ be a family of mappings $\psi : \{0,\dots,k\} \to \{0,1,\star\}$ called* translations. *For $\boldsymbol{x} \in \mathcal{X}^n$, $\boldsymbol{v} \in \Pi_{\boldsymbol{x}}(\mathcal{F}) \subseteq \{0,\dots,k\}^n$ and $\boldsymbol{\psi} \in \Psi^n$ we write $\boldsymbol{\psi}(\boldsymbol{v}) = (\psi_1(v_1),\dots,\psi_n(v_n))$ and $\boldsymbol{\psi}(\Pi_{\boldsymbol{x}}(\mathcal{F})) = \{\boldsymbol{\psi}(\boldsymbol{v}) : \boldsymbol{v} \in \Pi_{\boldsymbol{x}}(\mathcal{F})\}$. We say that $\boldsymbol{x} \in \mathcal{X}^n$ is* Ψ-shattered *by $\mathcal{F}$ if there exists a $\boldsymbol{\psi} \in \Psi^n$ such that $\{0,1\}^n \subseteq \boldsymbol{\psi}(\Pi_{\boldsymbol{x}}(\mathcal{F}))$. The* Ψ-dimension *of $\mathcal{F}$ is defined by*

$$\Psi\text{-dim}(\mathcal{F}) = \sup\{n \mid \exists \boldsymbol{x} \in \mathcal{X}^n, \boldsymbol{\psi} \in \Psi^n \ \ s.t.\ \ \{0,1\}^n \subseteq \boldsymbol{\psi}(\Pi_{\boldsymbol{x}}(\mathcal{F}))\} \ .$$

Example 6 *The following translation families and corresponding dimensions are used in this paper:*

(a) The Pollard pseudo-dimension $\Psi_{\mathrm{P}}\text{-dim}(V)$ *is induced by the family $\Psi_P = \{\psi_{P,i} : i \in [k]\}$ where $\psi_{P,i}(a) = \mathbf{1}[a < i]$.*

(b) The Graph dimension $\Psi_{\mathrm{G}}\text{-dim}(V)$ *is induced by the family $\Psi_G = \{\psi_{G,i} : i \in \{0,\dots,k\}\}$ where $\psi_{G,i}(a) = \mathbf{1}[a = i]$.*

(c) The Natarajan dimension $\Psi_{\mathrm{N}}\text{-dim}(V)$ *is induced by the family $\Psi_N = \{\psi_{N,i,j} : i,j \in \{0,\dots,k\}, i \neq j\}$ where*

$$\psi_{N,i,j}(a) = \begin{cases} 1 \ , & a = i \ , \\ 0 \ , & a = j \ , \\ \star \ , & a \notin \{i,j\} \ . \end{cases}$$

Finite Ψ-dimension, for certain 'distinguisher' translation families, characterizes multiclass learnability [2, Theorem 16]. The Ψs in Example 6 are all distinguishers.

2.3. Existing expected risk bounds for consistent multiclass learners

The following is a precise statement of the PAC-based multiclass expected risk bound referenced in Section 1. The statement and its proof both follow [13, Theorem 4.1] closely.

Theorem 7 *Let $\mathcal{F}, \mathcal{H} \subseteq \{0,\dots,k\}^{\mathcal{X}}$ be arbitrary with $d = \Psi_{\mathrm{G}}\text{-dim}(\mathcal{H}) \in \mathbb{N}$. Let Q be a prediction strategy such that for all $\boldsymbol{x} \in \bigcup_{n\geq 1} \mathcal{X}^n$ and all $f \in \mathcal{F}$, $Q(\mathrm{sam}(\boldsymbol{x}, f), \cdot) \in \mathcal{H}$ and $Q(\mathrm{sam}(\boldsymbol{x}, f), x_i) = f(x_i)$ for all $i \in [n]$. Equivalently Q is a learning algorithm that when given an $f \in \mathcal{F}$-labeled training set outputs a consistent*[1] *hypothesis from $\mathcal{H}$. Then $\hat{M}_{Q,\mathcal{F}}(n) \leq \frac{2(d+1)}{n} \log_2\left(\frac{4en}{d}\right)$ for all $n > d$.*

Proof. For $\boldsymbol{x} \in \bigcup_{n\in\mathbb{N}} \mathcal{X}^n$ define the risk functional $\mathrm{R}_{Q,f,P}(\boldsymbol{x}) = \mathbb{E}_P[Q(\mathrm{sam}(\boldsymbol{x}, f), X) \neq f(X)]$. Let $d = \Psi_{\mathrm{G}}\text{-dim}(\mathcal{F})$. By [2, Lemma 15] the VC-dimension of the 0-1 loss class induced by $\mathcal{F}$ equals d. Then by e.g. [4], for all $f \in \mathcal{F}$, distributions P on $\mathcal{X}$, $\epsilon > 0$ and $n > d$,

$$\Pr_{P^n}(\mathrm{R}_{Q,f,P}(\boldsymbol{X}) \geq \epsilon) \leq (2en/d)^d 2^{1-\epsilon n/2} \ .$$

By this inequality and the fact that the risk is most 1,

$$\begin{aligned} \mathbb{E}_{P^n}[\mathrm{R}_{Q,f,P}(\boldsymbol{X})] &= \mathbb{E}_{P^n}[\mathrm{R}_{Q,f,P}(\boldsymbol{X}) | \mathrm{R}_{Q,f,P}(\boldsymbol{X}) < \epsilon] \Pr_{P^n}(\mathrm{R}_{Q,f,P}(\boldsymbol{X}) < \epsilon) \\ &\quad + \ \mathbb{E}_{P^n}[\mathrm{R}_{Q,f,P}(\boldsymbol{X}) | \mathrm{R}_{Q,f,P}(\boldsymbol{X}) \geq \epsilon] \Pr_{P^n}(\mathrm{R}_{Q,f,P}(\boldsymbol{X}) \geq \epsilon) \\ &\leq \epsilon + (2en/d)^d 2^{1-\epsilon n/2} \ . \end{aligned}$$

Taking $\epsilon = 2t^{-1}\left(\log_2(nd^{-1}) + d\log_2(2end^{-1})\right)$ the result follows by Fubini's theorem. ∎

[1] In Section 1 we refer to the PAC-based bound as being in terms of $\Psi_{\mathrm{G}}\text{-dim}(\mathcal{F})$. Consistency of Q implies that $\Psi_{\mathrm{G}}\text{-dim}(\mathcal{F}) \leq \Psi_{\mathrm{G}}\text{-dim}(\mathcal{H})$ so we are being at worst generous to the PAC-based bound.

Algorithm 1 The deterministic multiclass one-inclusion prediction strategy $Q_{\mathcal{G},\mathcal{F}}$

Given: $\mathcal{F} \subseteq \{0,\ldots,k\}^{\mathcal{X}}$, $\text{sam}\left((x_1,\ldots,x_{n-1}),f\right) \in \left(\mathcal{X} \times \{0,\ldots,k\}\right)^{n-1}$ s.t. $f \in \mathcal{F}$, $x_n \in \mathcal{X}$
Returns: a prediction of $f(x_n)$ in $\{0,\ldots,k\}$

1. $V \longleftarrow \Pi_{\boldsymbol{x}}\left(\mathcal{F}\right)$;
2. $G \longleftarrow \mathcal{G}\left(V\right)$;
3. $\overrightarrow{G} \longleftarrow$ orient G to minimize the maximum outdegree ;
4. $V_{\text{space}} \longleftarrow \{\boldsymbol{v} \in V \mid v_1 = f(x_1),\ldots,v_{n-1} = f(x_{n-1})\}$;
5. **if** $V_{\text{space}} = \{\boldsymbol{v}\}$ **then return** v_n ;
6. **else return** the n^{th} component of the head of hyperedge V_{space} in $\overrightarrow{G}$;

Algorithm 2 The randomized multiclass one-inclusion prediction strategy $Q_{\mathcal{G}rand,\mathcal{F}}$

Given: $\mathcal{F} \subseteq \{0,\ldots,k\}^{\mathcal{X}}$, $\text{sam}\left((x_1,\ldots,x_{n-1}),f\right) \in \left(\mathcal{X} \times \{0,\ldots,k\}\right)^{n-1}$, $x_n \in \mathcal{X}$
Returns: a random prediction of $f(x_n)$

1. $V \longleftarrow \Pi_{\boldsymbol{x}}\left(\mathcal{F}\right)$;
2. $G = (V,E) \longleftarrow \mathcal{G}\left(V\right)$;
3. $P_e \longleftarrow$ distribution on $e \in E$ minimizing total probability incident to each vertex ;
4. $V_{\text{space}} \longleftarrow \{\boldsymbol{v} \in V \mid v_1 = f(x_1),\ldots,v_{n-1} = f(x_{n-1})\}$;
5. **if** $V_{\text{space}} = \{\boldsymbol{v}\}$ **then return** v_n ;
   else {
6. Select $\boldsymbol{V} \in V_{\text{space}}$ randomly according to distribution $P_{V_{\text{space}}}$;
7. **return** the n^{th} component of $\boldsymbol{V}$;
   }

2.4. The one-inclusion prediction strategy

A subset of the n-cube—the projection of some $\mathcal{F}$—induces the one-inclusion graph, which underlies a natural prediction strategy that is the focus of this section. The following definition generalizes the important data structure to subsets of $\{0,\ldots,k\}^n$.

Definition 8 (One-inclusion hypergraphs) *The* one-inclusion hypergraph $\mathcal{G}\left(V\right) = (V,E)$ *of* $V \subseteq \{0,\ldots,k\}^n$ *is the undirected graph with vertex-set* V *and hyperedge-set* E *of maximal (with respect to inclusion) sets of pairwise hamming-*1 *separated vertices. Under* $k = 1$*, the induced* E *is an edge-set and* $\mathcal{G}\left(V\right)$ *reduces to the* one-inclusion graph*.*

The one-inclusion graph's prediction strategy $Q_{\mathcal{G},\mathcal{F}}$ [13] immediately generalizes to the multiclass prediction strategy of Algorithm 1. For the remainder of this paper, barring Section 7, we will restrict our discussion to the $k = 1$ case, on which the following result focuses.

In words, the one-inclusion graph G of the projection of $\mathcal{F}$ on $\boldsymbol{x} \in \mathcal{X}^n$ is formed. G is then oriented to $\overrightarrow{G}$ so that maximum outdegree is minimized. Recall that an oriented hyperedge is a set with a single element identified as the head. The set V_{space} of vertices in G consistent with the labeled $n-1$-sample is formed. This set is either a singleton or an hyperedge in G. If V_{space} is a singleton $\boldsymbol{v}$, predict the label of $f(x_n)$ as the n^{th} component of $\boldsymbol{v}$, v_n. Otherwise predict the last component of the head of the directed hyperedge in $\overrightarrow{G}$.

Example 9 *Consider the subset* $V = \{(0,0,0),(1,0,0),(0,1,0),(1,1,0),(1,2,0),(2,2,0),(1,1,1),(2,1,1),(0,1,2),(1,1,2),(2,1,2),(0,2,2)\} \subset \{0,1,2\}^3$ *that is induced by points* $x_1,x_2,x_3 \in \mathcal{X}$ *and some class* $\mathcal{F} \subset \{0,1,2\}^{\mathcal{X}}$*. It is depicted in Figure 1 together with its induced hyperedge set. A possible orientation of the hypergraph, representing one of several possible prediction strategies for* $\mathcal{F}$ *on* $\{x_1,x_2,x_3\}$*, is shown in Figure 2; notice that the maximum outdegree is 2.*

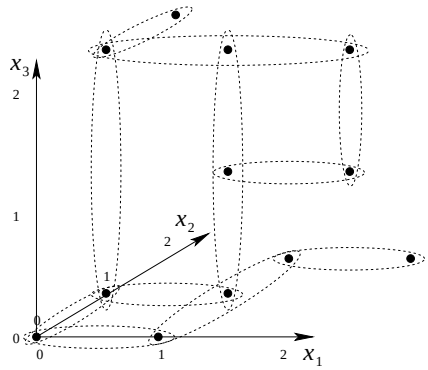

Fig. 1. The one-inclusion hypergraph of Example 9. Vertices are depicted as points, hyperedges as bounding ellipses.

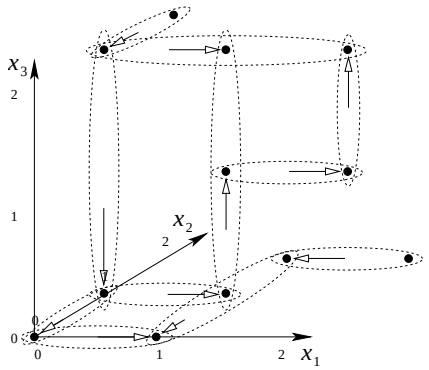

Fig. 2. The hypergraph of Figure 1 oriented with maximum outdegree 2. Predictions are made by following the head.

Replacing orientation with a distribution over each (hyper)edge induces a *randomized strategy* $Q_{\mathcal{G}rand,\mathcal{F}}$ in Algorithm 2. By exploiting the combinatorial structure of concept classes, Haussler *et al.* were able to improve on best-known bounds on worst-case expected risk by a factor of $\log n$ [13, Theorem 2.3].

Theorem 10 (The one-inclusion mistake bounds) $\hat{M}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}(n) \leq \frac{\mathrm{VC}(\mathcal{F})}{n}$ *for every concept class* $\mathcal{F}$ *and* $n > 1$. *The same holds for the randomized strategy.*

A lower bound in [18] showed that the one-inclusion strategy's performance is optimal within a factor of $1 + o(1)$. In the deterministic (randomized) one-inclusion prediction strategy, the orientation (respectively assignment of edge distributions) is achieved by a simple reduction to a network flow problem [13].

2.5. Sample compressibility

We recall the notions of labeled [19,8,9] and unlabeled [16,17] compression schemes. We begin with the former, essential definition of [19] on which all subsequent definitions are based. Informally, one k-compresses a concept class C by compressing any sample s of length at least k that is consistent with C, to a subsample of length at most k and then mapping such a compressed-set to some s-consistent concept (not necessarily belonging to C).

Definition 11 (Labeled compression schemes) *Let* $k \in \mathbb{N}$, *domain* $\mathcal{X}$ *and family* $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ *be arbitrary, and consider a pair of mappings of the following form*

$$\kappa_{\mathcal{F}} : \bigcup_{n=k}^{\infty} (\mathcal{X} \times \{0,1\})^n \longrightarrow \bigcup_{l=0}^{k} (\mathcal{X} \times \{0,1\})^l$$

$$\rho_{\mathcal{F}} : \left(\bigcup_{l=0}^{k} (\mathcal{X} \times \{0,1\})^l \right) \times \mathcal{X} \longrightarrow \{0,1\} \ .$$

If, for each $f \in \mathcal{F}$ *and* $\boldsymbol{x} \in \bigcup_{n=k}^{\infty} \mathcal{X}^n$, *the following conditions are satisfied,*

(i) [subsample condition]*: the* compression function $\kappa_{\mathcal{F}}$ *maps the sequence* $\mathrm{sam}(\boldsymbol{x}, f)$ *to a subsequence of length at most* k, *called the* representative *of* f*; and*

(ii) [consistency condition]*: the* reconstruction function $\rho_{\mathcal{F}}$ *labels* x_i *consistently with* $f(x_i)$ *for each* $i \in [n]$.

Then $(\kappa_{\mathcal{F}}, \rho_{\mathcal{F}})$ *is a* k-compression scheme. *A* k*-compression scheme is said to have* size k, *and if the representative size bound* k *is met with equality for some* $f \in \mathcal{F}$ *and* $\boldsymbol{x} \in \mathcal{X}^n$ *then we say it is of* size exactly k. *A compression scheme defines a hypothesis (not necessarily in* $\mathcal{F}$*) by the mapping* $\rho_{\mathcal{F}}(\kappa_{\mathcal{F}}(\mathrm{sam}(\boldsymbol{x}, f)), \cdot) : \mathcal{X} \to \{0,1\}$.

Littlestone and Warmuth [19] showed that k-compressibility in the above labeled sense, for any $k < \infty$, is *sufficient* for learnability. Furthermore their proof is considerably simpler than (the more traditional) learnability proofs based directly on finite VC-dimension. The authors asked the natural question of *necessity* [19,8,9,23,24,16,17], which corresponds to the following.

Problem 12 (Sample compression) *Does every concept class of VC-dimension $d < \infty$ have a labeled compression scheme of size $O(d)$?*

Floyd and Warmuth were the first to demonstrate a significant positive result on the problem, by showing that maximum classes of VC-dimension d can be d-compressed [8,9]. Over a decade later Kuzmin and Warmuth recently showed that d-maximum classes can in fact be compressed to unlabeled sets [16,17]; earlier, Ben-David and Litman demonstrated a special-case of this result in [3].

Definition 13 (Unlabeled compression schemes) *Let C be a maximum concept class of VC-dimension d on a finite domain $\mathcal{X}$. A* representation mapping *r of C satisfies:*

(i) r is a bijection between C and subsets of $\mathcal{X}$ of size at most d; and
*(ii) [*non-clashing*]: $c|\,(r(c) \cup r(c')) \neq c'|\,(r(c) \cup r(c'))$ for all $c, c' \in C$, $c \neq c'$.*

Given bijectivity, the non-clashing condition is equivalent to:

3. For each $\boldsymbol{x} \subseteq \mathcal{X}, c \in C$, there exists exactly one $c' \in C$ such that $\text{sam}\,(\boldsymbol{x}, c') = \text{sam}\,(\boldsymbol{x}, c)$ and $r(c) \subseteq \boldsymbol{x}$.

Such a representation mapping constitutes a d-unlabeled compression scheme for C.

This definition is sufficient for the unlabeled analogue of the labeled Definition 11, where a C-consistent sample is compressed to an unlabeled sample-subsequence of length at most d and which itself can be reconstructed to a concept consistent with the original sample. Trivially a k-unlabeled compression scheme can be transformed into a k-compression scheme; however it is still not clear whether including labels aids compression or not. An answer to Problem 12 must provide a general scheme that satisfies the two conditions laid out in Definition 11, and need go no further. Kuzmin and Warmuth were able to prove that a sophisticated Tail Matching algorithm successfully d-unlabeled compresses all d-maximum concept classes [16,17]. They also proposed a significantly simpler unlabeled Peeling algorithm but it is still not known whether this correctly compresses maximum classes. The algorithm assigns representatives to concepts by iteratively 'peeling' away a minimum degree vertex from the present one-inclusion graph of the class; the peeled vertex's representative is assigned to be the set of remaining edges adjoining that vertex, and then that vertex is removed. Two of the combinatorial results in this paper relate to Kuzmin and Warmuth's conjectured correctness proof for Peeling, as described in Section 6 below.

What is needed for the compressibility conjecture of Warmuth *et al.* is a general d-compression scheme of d-maximal classes. Any concept class C of VC-dimension d can be expanded by adding one concept at a time until it is d-maximal, and an un/labeled d-compression scheme for such a maximal class immediately induces a d-sized scheme for C. With this and the d-compressibility of maximum classes in mind, the following question naturally arises.

Problem 14 *As a function of $d < \infty$, what is an upper-bound on the supremum over each d-maximal class V, of the infimum of the VC-dimension of maximum classes containing V?*

An answer of $O(d)$ would immediately imply a positive solution to the compressibility conjecture. It is clear that maximum classes have a very special, recursive structure that is not shared by maximal classes. In particular consider the following products of projecting away an axis [25,10].

Definition 15 *For any $V \subseteq \{0,1\}^n$, define with respect to $i \in [n]$*

(i) The reduction*, $V^i = \Pi_{[n]\setminus\{i\}}\left(\left\{v \in V \mid i \in I_{\mathcal{G}(V)}(v)\right\}\right)$; and*
(ii) The tail*, $\text{tail}_i\,(V) = \left\{v \in V \mid i \notin I_{\mathcal{G}(V)}(v)\right\}$,*

where $I_{\mathcal{G}(V)}(v) \subseteq [n]$ denotes the set of labels of the edges incidental to vertex $v \in V$.

Welzl [25] proved that the reduction C^i and projection $\Pi_{[n]\setminus\{i\}}(C)$ of a d-maximum class C are $d-1$- and d-maximum concept classes respectively; and through the recursive decomposition of a given maximum class into these products, several sets of authors have shown that a maximum class can be *compressed* recursively [8,9,16,17]. In particular, to the best of our knowledge all maximum compression schemes appeal to this special structure that is not shared by non-maximum maximal classes. It is not yet clear how to compress maximal classes in general, or whether the specialized schemes developed for maximum classes can be brought to bear on this task.

3. Shifting and graph density

The key to proving the classic one-inclusion mistake bound of Theorem 10 is the following result on graph density [13, Lemma 2.4].

Lemma 16 (One-inclusion graph density bound) *For all $n \in \mathbb{N}, V \subseteq \{0,1\}^n$, $\operatorname{dens}(\mathcal{G}(V)) \leq \mathrm{VC}(V)$.*

An elegant proof of this deep result, due to Haussler [12], uses *shifting*. Shifting is the process of contracting a subset of the n-cube towards $\mathbf{0}$ along one direction $s \in [n]$ at a time – each point with a gap below in the s direction is translated down.

Definition 17 (Shifting operators) *For each $s \in [n]$ define the shift operators on vertex $\boldsymbol{v} \in V \subseteq \{0,1\}^n$ and vertex-set V, respectively, as*

$$S_s(\boldsymbol{v}; V) = \begin{cases} (v_1, \ldots, v_{s-1}, 1, v_{s+1}, \ldots, v_n) & \text{if } v_s = 1 \text{ and} \\ & \quad (v_1, \ldots, v_{s-1}, 0, v_{s+1}, \ldots, v_n) \in V \\ (v_1, \ldots, v_{s-1}, 0, v_{s+1}, \ldots, v_n) & \text{otherwise} \end{cases}$$
$$S_s(V) = \{S_s(\boldsymbol{v}; V) \mid \boldsymbol{v} \in V\} \quad .$$

One-inclusion graph $\mathcal{G}(V)$ is said to be shifted *to $\mathcal{G}(S_s(V))$ along s – that is, the 'shifted' edge-set is the edge-set induced by the shifted vertex-set.*

Closed-below sets are those subsets of the n-cube that are the fixed-points of shifting.

Definition 18 *Let $I \subseteq [n]$. We call a subset $V \subseteq \{0,1\}^n$ I-*closed-below *if $S_s(V) = V$ for all $s \in I$. If V is $[n]$-closed-below then we call it* closed-below*.*

The process of "shifting down to $\mathbf{0}$" can be generalized to *axis-parallel contractions to $\boldsymbol{v}^\star \in \{0,1\}^n$* (or equivalently shifting can be preceded by a relabeling of component-wise labels, and followed by the subsequent inverse re-labelings). For such cases the closed-below property simply generalizes to a fixed-point property. Indeed many of the following properties and their consequences for shifting also apply to these more general contractions.

A number of properties of shifting follow relatively easily [12]:

$$|S_s(V)| = |V| \ , \qquad \text{by the injectivity of } S_s(\,\cdot\,; V) \tag{1}$$
$$\mathrm{VC}(S_s(V)) \leq \mathrm{VC}(V) \ , \qquad \text{as } S_s(V) \text{ shatters } I \subseteq [n] \Rightarrow V \text{ shatters } I \tag{2}$$
$$|E| \leq |V| \cdot \mathrm{VC}(V) \ , \qquad \text{as } V \text{ closed-below} \Rightarrow \max_{\boldsymbol{v} \in V} \|\boldsymbol{v}\|_1 \leq \mathrm{VC}(V) \tag{3}$$
$$|S_s(E)| \geq |E| \ , \qquad \text{by cases} \tag{4}$$
$$\exists T \in \mathbb{N}, \boldsymbol{s} \in [n]^T \text{ s.t. } S_{s_T}(\ldots S_{s_1}(V)) \text{ is closed-below (a fixed-point)} \ . \tag{5}$$

Properties (1–2) and the justification of (3) together imply Sauer's lemma; Properties (1–5) lead to

$$\frac{|E|}{|V|} \;\leq\; \ldots \;\leq\; \frac{|S_{s_T}(\ldots S_{s_1}(E))|}{|S_{s_T}(\ldots S_{s_1}(V))|} \;\leq\; \mathrm{VC}(S_{s_T}(\ldots S_{s_1}(V))) \;\leq\; \ldots \;\leq\; \mathrm{VC}(V)$$

proving Lemma 16.

3.1. Shatter-invariant shifting

While Haussler shifts to bound density, the number of edges can increase *and* the VC-dimension can decrease—both contributing to the observed gap between graph density and capacity. Our first result demonstrates that shifting can in fact be controlled to preserve VC-dimension.

Lemma 19 *Consider arbitrary $n \in \mathbb{N}$, $I \subseteq [n]$ and $V \subseteq \{0,1\}^n$ that shatters I. There exists a finite sequence $s_1, \ldots, s_T$ in $[n]$ such that each $V_t = S_{s_t}(\ldots S_{s_1}(V))$ shatters I and V_T is closed-below. In particular* $\mathrm{VC}(V_T) = \mathrm{VC}(V_{T-1}) = \ldots = \mathrm{VC}(V)$.

Proof. $\Pi_I(\cdot)$ is invariant to shifting on $\overline{I} = [n] \backslash I$. So some finite number of shifts on $\overline{I}$ will produce a $\overline{I}$-closed-below family W that shatters I. Hence W must contain representatives for each element of $\{0,1\}^{|I|}$ on I with components equal to 0 outside I. Thus the shattering of I is invariant to the shifting of W on I, so that a finite number of shifts on I produces an I-closed-below W' that shatters I. Repeating the process a finite number of times until no non-trivial shifts are made produces a closed-below family that shatters I. The second claim now follows from (2). ■

In addition to the following interesting but inapplicable approach to bounding density, shatter-invariant shifting will be applied in Section 3.2 to prove that only maximum subsets can maximize density amongst all subsets with constant VC-dimension.

Remark 20 *Lemma 19 suggests that we study graph density by accounting for edges added during shifting—edges that must appear in the final closed-below graph W that are not present in the original V. If $d = \mathrm{VC}(V)$ then for each d-index-set I witnessing the VC-dimension of V, V can be shifted down to some fixed-point W_I while retaining the shattering of I. Such a W_I must contain an I-colored d-cube, and in particular each of that cube's $d2^{d-1}$ edges. We can thus maximize a lower-bound on the number of edges added to V when shifting to W_I, optimizing over the collection of index sets $\mathcal{S} = \{I \subseteq [n] : |I| = d, V \text{ shatters } I\}$ and witnessing subsets $V_I = \{U \subseteq V \mid U \text{ shatters } I\}$. This produces the density bound of*

$$\frac{|E|}{|V|} \leq d - \frac{d2^{d-1} - \min_{I \in \mathcal{S}} \max_{U \in V_I} |\mathrm{E}(\mathcal{G}(U))|}{|V|} \leq d \;.$$

Shifting can be further controlled to retain shattering of certain collections of sets, which can be applied to produce similar bounds.

3.2. Tightly bounding graph density by symmetrization

Kuzmin and Warmuth [17] introduced D_n^d as a potential bound on the graph density of maximum classes. We begin with properties of D_n^d, a technical lemma and then proceed to the main result which positively resolves the conjecture of Kuzmin and Warmuth. A discussion of the sample compressibility consequences of the result can be found in Section 6.

Definition 21 *Define $D_n^d = \frac{n\binom{n-1}{\leq d-1}}{\binom{n}{\leq d}}$ for all $n \in \mathbb{N}$ and $d \in [n]$. Denote by V_n^d the VC-dimension d closed-below subset of $\{0,1\}^n$ equal to the union of all $\binom{n}{d}$ closed-below embedded d-cubes.*

Lemma 22 *D_n^d*

(i) equals the graph density of V_n^d for each $n \in \mathbb{N}$ and $d \in [n]$;
(ii) is strictly upper-bounded by d, for all n;
(iii) equals $\frac{d}{2}$ for all $n = d \in \mathbb{N}$;
(iv) is strictly monotonic increasing in d (with n fixed);

(v) is strictly monotonic increasing in n (with d fixed); and
(vi) approaches d as $n \to \infty$.

Proof. By counting, for each $d \leq n < \infty$, the density of $\mathcal{G}\left(V_n^d\right)$ equals D_n^d:

$$\frac{\left|\mathrm{E}\left(\mathcal{G}\left(V_n^d\right)\right)\right|}{\left|V_n^d\right|} \;=\; \frac{\sum_{i=1}^{d} i\binom{n}{i}}{\sum_{i=0}^{d}\binom{n}{i}} \;=\; \frac{n\sum_{i=0}^{d-1}\frac{i+1}{n}\binom{n}{i+1}}{\binom{n}{\leq d}} \;=\; \frac{n\sum_{i=0}^{d-1}\binom{n-1}{i}}{\binom{n}{\leq d}} \;=\; \frac{n\binom{n-1}{\leq d-1}}{\binom{n}{\leq d}}$$

proving (i). Since for all $A, B, C, D > 0$, $\frac{A}{B} < \frac{A+C}{B+D}$ iff $\frac{A}{B} < \frac{C}{D}$, it is sufficient for (iv) to prove that $D_n^{d-1} < \frac{n\binom{n-1}{d-1}}{\binom{n}{d}}$. By (i) and Lemma 16 $D_n^d \leq d$, and so

$$D_n^{d-1} \leq d-1 < d = \frac{n\cdot(n-1)!}{n!}\frac{(n-d)!}{(n-d)!}\frac{d!}{(d-1)!} = \frac{n\frac{(n-1)!}{(n-d)!(d-1)!}}{\frac{n!}{(n-d)!d!}} = \frac{n\binom{n-1}{d-1}}{\binom{n}{d}} \;.$$

Monotonicity in d, (i) and Lemma 16 together prove (ii). Now for any $n \in \mathbb{N}$

$$\begin{aligned} D_n^n &= \frac{n\binom{n-1}{\leq n-1}}{\binom{n}{\leq n}} \\ &= \frac{n2^{n-1}}{2^n} \\ &= \frac{n}{2} \;, \end{aligned}$$

proving part (iii). Theorem 24 states that V_n^d uniquely maximizes density, at D_n^d, over all closed-below families of VC-dimension d in the n-cube. Thus $D_{n-1}^d = \mathrm{dens}\left(V_{n-1}^d\right) < D_n^d$ which is part (v). Part (vi) follows from the asymptotically matching lower-bound of [18]. ■

Lemma 23 *Consider arbitrary $U, V \subseteq \{0,1\}^n$ with $\mathrm{dens}\left(\mathcal{G}\left(V\right)\right) \geq \rho > 0$, $|U| \leq |V|$ and $\left|\mathrm{E}\left(\mathcal{G}\left(U\right)\right)\right| \geq \left|\mathrm{E}\left(\mathcal{G}\left(V\right)\right)\right|$. If $\mathrm{dens}\left(\mathcal{G}\left(U \cap V\right)\right) < \rho$ then $\mathrm{dens}\left(\mathcal{G}\left(U \cup V\right)\right) > \rho$.*

Proof. If $\mathcal{G}\left(U \cap V\right)$ has density less than ρ then

$$\begin{aligned} \frac{\left|\mathrm{E}\left(\mathcal{G}\left(U \cup V\right)\right)\right|}{|U \cup V|} &\geq \frac{\left|\mathrm{E}\left(\mathcal{G}\left(U\right)\right)\right| + \left|\mathrm{E}\left(\mathcal{G}\left(V\right)\right)\right| - \left|\mathrm{E}\left(\mathcal{G}\left(U \cap V\right)\right)\right|}{|U| + |V| - |U \cap V|} \\ &\geq \frac{2\left|\mathrm{E}\left(\mathcal{G}\left(V\right)\right)\right| - \left|\mathrm{E}\left(\mathcal{G}\left(U \cap V\right)\right)\right|}{2|V| - |U \cap V|} \\ &> \frac{2\rho|V| - \rho|U \cap V|}{2|V| - |U \cap V|} \;=\; \rho \end{aligned}$$

■

Theorem 24 (Symmetrization density bound) *Every family $V \subseteq \{0,1\}^n$ with $d = \mathrm{VC}(V)$ has $(V, E) = \mathcal{G}\left(V\right)$ with graph density*

$$\frac{|E|}{|V|} \;\leq\; D_n^d \;<\; d \;. \tag{6}$$

For $n \in \mathbb{N}$ and $d \in [n]$, V_n^d is the unique closed-below VC-dimension d subset of $\{0,1\}^n$ meeting (6) with equality. A VC-dimension d family $V \subseteq \{0,1\}^n$ meets (6) with equality only if V is maximum.

Proof. Allow a permutation $g \in S_n$ to act on vector $\boldsymbol{v} \in \{0,1\}^n$ and family $V \subseteq \{0,1\}^n$ by $g(\boldsymbol{v}) = \left(v_{g(1)}, \ldots, v_{g(n)}\right)$ and $g(V) = \{g(\boldsymbol{v}) \mid \boldsymbol{v} \in V\}$; and define $S_n(V) = \bigcup_{g \in S_n} g(V)$. Note that a closed-below VC-dimension d family $V \subseteq \{0,1\}^n$ satisfies $S_n(V) = V$ iff $V = V_n^d$, as $\mathrm{VC}(V) \geq d$ implies V contains

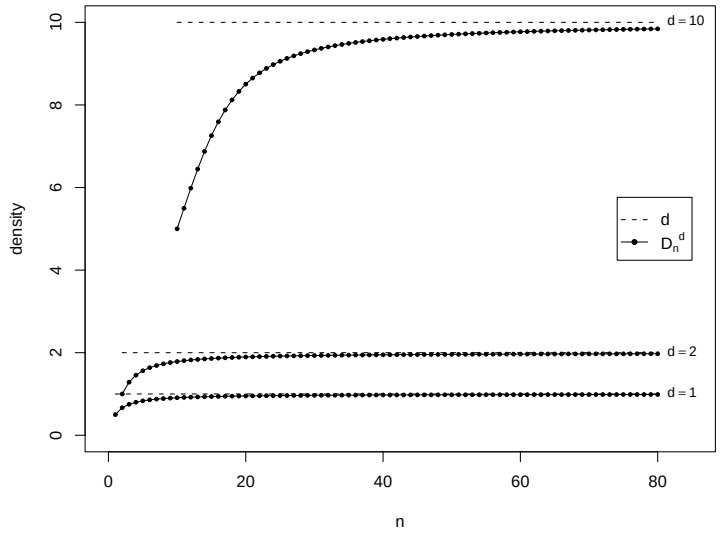

Fig. 3. The improved graph density bound of Theorem 24. The density bounding D_n^d is plotted (dotted solid) alongside the previous best d (dashed), for $d \in \{1, 2, 10\}$.

an embedded d-cube, invariance to S_n implies further that V contains all $\binom{n}{d}$ such cubes, and $\mathrm{VC}(V) \leq d$ implies that $V \subseteq V_n^d$. Consider now any

$$V_{n,d}^* \in \arg\min \left\{ |U| \;\middle|\; U \in \underset{\{U \subseteq \{0,1\}^n \mid \mathrm{VC}(U) \leq d, U \text{ closed-below}\}}{\arg\max} \mathrm{dens}\left(\mathcal{G}\left(U\right)\right) \right\} .$$

For the purposes of contradiction assume that $V_{n,d}^* \neq g(V_{n,d}^*)$ for some permutation $g \in S_n$. Then if $\mathrm{dens}\left(\mathcal{G}\left(V_{n,d}^* \cap g(V_{n,d}^*)\right)\right) \geq \mathrm{dens}\left(\mathcal{G}\left(V_{n,d}^*\right)\right)$ then $V_{n,d}^*$ would not have been selected above (i.e. a closed-below family at least as small and dense as $V_{n,d}^* \cap g(V_{n,d}^*)$ would have been chosen). Thus $\mathrm{dens}\left(\mathcal{G}\left(V_{n,d}^* \cup g(V_{n,d}^*)\right)\right) > \mathrm{dens}\left(\mathcal{G}\left(V_{n,d}^*\right)\right)$ by Lemma 23. But then again $V_{n,d}^*$ would not have been selected (i.e. a distinct family at least as dense as $V_{n,d}^* \cup g(V_{n,d}^*)$ would have been selected instead, since every vector in this union contains no more than d 1's). Hence $V_{n,d}^* = S_n(V_{n,d}^*)$ and so $V_{n,d}^* = V_n^{d'}$ and by Lemma 22.(i) $\mathrm{dens}\left(\mathcal{G}\left(V_{n,d}^*\right)\right) = D_n^{d'}$, for $d' = \mathrm{VC}(V_{n,d}^*) \leq d$. But by Lemma 22.(iv) this implies that $d = d'$ and (6) is true for all closed-below families; V_n^d uniquely maximizes density amongst all closed-below VC-dimension d families in the n-cube.

For an arbitrary $V \subseteq \{0,1\}^n$ with $d = \mathrm{VC}(V)$ consider any of its closed-below fixed-point (cf. (5)), $W \subseteq \{0,1\}^n$. Noting that $\mathrm{VC}(W) \leq d$ and $\mathrm{dens}\left(\mathcal{G}\left(V\right)\right) \leq \mathrm{dens}\left(\mathcal{G}\left(W\right)\right)$ by (2) and (1) & (4) respectively, the bound (6) follows directly for V. Furthermore if we shift to preserve VC-dimension then $\mathrm{VC}(W) = d$ while still $|V| = |W|$. And since $\mathrm{dens}\left(\mathcal{G}\left(W\right)\right) = D_n^d$ only if $W = V_n^d$, it follows that V maximizes density amongst all VC-dimension d families in the n-cube, with $\mathrm{dens}\left(\mathcal{G}\left(V\right)\right) = D_n^d$, only if it is maximum. ■

Theorem 24 improves on the VC-dimension density bound of Lemma 16 for low sample sizes (see Figure 3). This new result immediately implies the following one-inclusion mistake bounds (see Appendix A for the proof).

Theorem 25 (Symmetrization mistake bound) *Consider any $n \in \mathbb{N}$ and $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with $\mathrm{VC}(\mathcal{F}) = d < \infty$. Then $\hat{M}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}(n) \leq \left\lceil D_n^d \right\rceil / n$ and $\hat{M}_{Q_{\mathcal{G}rand,\mathcal{F}},\mathcal{F}}(n) \leq D_n^d / n$.*

For small d, $n^*(d) = \min\left\{n \geq d \mid d = \left\lceil D_n^d \right\rceil\right\}$—the first n for which the new and old deterministic one-inclusion mistake bounds coincide—appears to remain very close to $2.96d$ (see Fig. 4). The randomized strategy's mistake bound of Theorem 25 offers a strict improvement over that of [13].

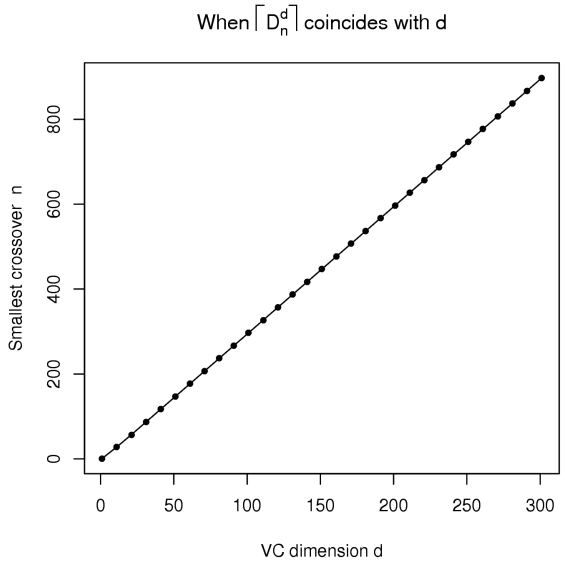

Fig. 4. Calculating the point at which the new mistake bound of Theorem 25 coincides with that of [13,12]. For each $1 \leq d \leq 300$, we see that the new bound provides a strict improvement iff n is no more than about $2.96d$.

Remark 26 *The symmetrization method of Theorem 24 can be extended over subgroups $G \subset S_n$ to gain even tighter estimates of density. Just as the S_n-invariant V_n^d is the maximizer of density among all closed-below $V \subseteq V_n^d$, there exist G-invariant families that maximize the density over all of* their *sub-families: to estimate a graph's density, find the smallest subgroup that admits an invariant family containing the given graph and count that invariant's density.*

4. Characterizations

We now consider several related characterizations and properties of the one-inclusion graph. Beginning with Section 4.1 we present an edge-colorability characterization of graphs isomorphic to a one-inclusion graph, extending results of [7,15,14]. The characterization fully justifies the use of 'color' in-place of 'dimension' when discussing edges, embedded cubes, etc. Section 4.2 introduces the complementary characterization of maximum and maximal subsets of the n-cube—an extension of the notion of forbidden labels [8]. Finally the well-known characterization of VC-dimension 1 maximum classes as trees composed of a single edge of each color [5], is extended in Section 4.3 to an algebraic topological property of maximum classes of arbitrary dimension.

4.1. Characterizing one-inclusion-isomorphic graphs

In this section we equate a set of n colors with the n dimensions of $\{0,1\}^n$, coloring each edge of a one-inclusion graph according to the axis to which it is parallel. Constraints on the structure of a one-inclusion graph can then be re-written in terms of conditions on such a coloring; and from this we can characterize arbitrary graphs isomorphic to one-inclusion graphs. In particular the colorability characterization facilitates the useful visualization of subgraphs of the n-cube for $n > 3$.

Definition 27 *Let $G = (V, E)$ be an arbitrary graph. Then an* edge-coloring *or simply* coloring *of G is a mapping $\mathrm{col} : E \to \mathcal{C}$ into some finite set of colors $\mathcal{C}$. The parity of a color $c \in \mathcal{C}$ in some subgraph (W, F) of G is defined as $\oplus_{(W,F)}(c) = \sum_{f \in F} \mathbf{1}[\mathrm{col}(f) = c] \mod 2$. If the parity is congruent to 0 (1) then we say that c has* even (odd) parity *in (W, F). If the subgraph is the whole graph G or is otherwise understood then we may drop the subgraph parameter as in $\oplus(c) = \oplus_G(c)$. The parity $\oplus_{(W,F)} = \left(\oplus_{(W,F)}(c_1), \ldots, \oplus_{(W,F)}(c_n)\right) \in \{0,1\}^n$ of an edge coloring $\mathrm{col}(\cdot)$ in subgraph (W, F) is the vector of parities taken over the colors $\{c_1, \ldots, c_n\} = \mathcal{C}$.*

We begin with necessary colorability conditions.

Proposition 28 *If $G = (V, E)$ is isomorphic to one-inclusion graph $\mathcal{G}(\phi(V))$ via the mapping $\phi : V \to \{0,1\}^n$, then there exists a coloring $\mathrm{col} : E \to \mathcal{C} = \{c_1, \ldots, c_n\}$ of G satisfying:*

- *(i) Each color has even parity in each cycle of G.*
- *(ii) There do not exist two walks in G with the same initial point and different end points, having the same color parities.*
- *(iii) If $x, y, z \in V$ are vertices such that x, y are connected to z by walks W_x, W_y with $\left|\oplus_{W_x} \mathsf{xor} \oplus_{W_y}\right| = 1$, then $\{x, y\} \in E$. Furthermore, if $\oplus_{\phi(W_x)} \mathsf{xor} \oplus_{\phi(W_y)} = \{i\} \subset [n]$ then the induced coloring in $\mathcal{G}(\phi(V))$ satisfies $\mathrm{col}(\{\phi(x), \phi(y)\}) = c_i$.*

In addition (ii) implies (iv) and is equivalent to (v):

- *(iv) At each $v \in V$ each color appears in the adjoining edges of v at most once.*
- *(v) Any walk in G with distinct start and end vertices must have some odd-parity color.*

In particular coloring each $\{x, y\} \in E$ by the index on which $\phi(x)$ and $\phi(y)$ differ, is one such coloring.

Proof. To prove (ii) $\Rightarrow$ (iv) consider distinct $u, v, w \in V$ such that $\{u, v\}, \{v, w\} \in E$. Then (ii) implies that the single-edge walks (v, u) and (v, w) must have different parities which implies that $\mathrm{col}(\{u, v\}) \neq \mathrm{col}(\{v, w\})$ leading to (iv). For the equivalence, suppose that (ii) were false, then take such a pair of falsifying walks W_1, W_2 both starting at some $s \in V$ and ending at $f_1 \neq f_2 \in V$ respectively; the walk $W = W_1^{-1} \circ W_2$ has all-even parities. But together with $f_1 \neq f_2$ this implies that W witnesses the falsification of (v). Now suppose that (ii) is true for G and consider any walk W starting and finishing at distinct $s, f \in V$ respectively. Pick any vertex m along W and consider the components W_1, W_2 along W starting (ending) at s (m) and m (f) respectively—at most one of these could be a walk with empty edge-set. It follows that walks W_1^{-1} and W_2 begin at m and end at $s \neq f$ so that W_1 and W_2 must have different parities. Hence the composition W must have at least one odd component-parity implying (v).

Now suppose that (V, E) is a isomorphic to the one-inclusion graph $\mathcal{G}(V) \subseteq \{0,1\}^n$, and we must construct an edge-coloring satisfying the given conditions. We color each edge $\{u, v\} \in E$ with $\mathrm{col}(\{u, v\}) := i \in [n]$ s.t. $\phi(u)_i \neq \phi(v)_i$. This is a well-defined function since $\{\phi(u), \phi(v)\}$ is an edge in $\mathcal{G}(\phi(V))$, and so exactly one such i exists. That is, we are coloring G and $\mathcal{G}(\phi(V))$ such that each edge's color is invariant under ϕ. (i) follows from the fact that a cycle in $\{0,1\}^n$, viewed as a walk with arbitrary start point along the cycle, must experience an even number of steps in any one direction since the end and start vertices must coincide. The end vertex of a $\{0,1\}^n$-walk with fixed starting vertex is invariant under permutations of the walk's step direction sequence, implying (ii) as a special case by reduction of color occurrence counts modulo 2. Take walks W_x, W_y as in condition (iii). Clearly $\oplus_{W_x} \mathsf{xor} \oplus_{W_y} = \phi(x) \mathsf{xor} \phi(y)$; thus $\phi(x)$ and $\phi(y)$ differ on exactly the single coordinate $i \in [n]$ and hence $\{\phi(x), \phi(y)\}$ is an edge of $\mathcal{G}(\phi(V))$ and $\mathrm{col}(\{x, y\}) = \mathrm{col}(\{\phi(x), \phi(y)\}) = i$. ■

Conditions (i) and (ii) together exactly characterize cycles by dictating that a walk is a cycle iff the walk has all even parities. Condition (iii) additionally says that if we can close a walk with a single one-inclusion edge e to make a valid cycle (with even parities) then e is indeed included in $\mathcal{G}(\phi(V))$.

Proposition 29 *Let $G = (V, E)$ be an arbitrary graph with $k \in \mathbb{N}$ connected components. If G can be edge-colored with $\mathcal{C} = \{c_1, \ldots, c_n\}$ such that conditions (i)–(iii) of Proposition 28 hold, then G is isomorphic to a*

one-inclusion graph in $\{0,1\}^{n+\lceil \log_2(k)+1\rceil}$ *for* $k>1$*, or to a one-inclusion graph in* $\{0,1\}^n$ *for connected* G*.*

Proof. Assume that G is connected and let $T=(V,E')$ be an arbitrary spanning tree for G, arbitrarily rooted at some $v_0 \in V$. For each $v \in V$ let P_v denote the unique path from v_0 to v in T, and define $\ell(v) = \oplus_{P_v}$. Trivially $\ell(v_0) = \oplus_{(\{v_0\},\emptyset)} = (0,\ldots,0)$. We claim that $\ell(v)$ represents a valid one-inclusion isomorphism: that $\ell(v)$ is an injection, that the image of V under $\ell(\cdot)$ induces a one-inclusion graph such that $\{u,v\} \in E$ iff $\mathrm{d}_{hamm}(\ell(u),\ell(v)) = 1$ and with $\{u,v\},\{u,v'\} \in E, v \neq v'$ implying that $\ell(u) \mathsf{xor} \ell(v) \neq \ell(u) \mathsf{xor} \ell(v')$.

By (ii) $\ell(\cdot)$ is an injection. Suppose that $\{u,v\} \in E$; if $\{u,v\} \in E'$ then $\mathrm{d}_{hamm}(\ell(u),\ell(v)) = 1$ by construction of ℓ, so suppose that $\{u,v\} \in E \backslash E'$. Then $W = P_u \circ \{u,v\} \circ P_v^{-1}$ is a cycle in G and by (i) must have all even parities. Thus the parities of P_u and P_v on colors in $\mathcal{C}\backslash\{\mathrm{col}(\{u,v\})\}$ must coincide and on $\mathrm{col}(\{u,v\})$ must differ. Thus again $\mathrm{d}_{hamm}(\ell(u),\ell(v)) = 1$. Suppose for distinct $u,v \in V$ that $\mathrm{d}_{hamm}(\ell(u),\ell(v)) = 1$ then by (iii) $\{u,v\} \in E$; furthermore $\mathrm{col}(\{u,v\})$ equals the coordinate on which $\ell(u)$ and $\ell(v)$ differ. Finally suppose that $\{u,v\},\{u,v'\} \in E, v \neq v'$. Then together with (iv) this second consequence of (iii) implies that $\ell(u) \mathsf{xor} \ell(v) = \{\mathrm{col}(\{u,v\})\} \neq \{\mathrm{col}(\{u,v'\})\} = \ell(u) \mathsf{xor} \ell(v')$.

If $k>1$ then map each component of G into a different copy of the n-cube. There may be common vectors within the different n-cubes, or vectors that are hamming-1 apart so that new unwanted edges would be necessary to maintain the one-inclusion property. Thus to maintain both the isomorphism and the one-inclusion property we embed each image into a different corner of an $\{0,1\}^{n+m}$-cube for sufficiently large m. It can be shown [2] that for any $m \in \mathbb{N}$ the $\{0,1\}^m$-cube contains a set of 2^{m-1} vectors that are pairwise no less than hamming-2 apart. Thus we can pack k points in an $\lceil \log_2(k)+1\rceil$-cube and so we embed the k disconnected one-inclusion graphs constructed as above in corners of the $n+\lceil \log_2(k)+1\rceil$-cube so that no new edges need to be added to maintain the one-inclusion property. ■

In [14], Havel and Morávek prove that a graph (V,E) with vertex-set $V \subseteq \{0,1\}^n$ admits a coloring satisfying conditions (i) and (v) iff $\{u,v\} \in E$ implies $\mathrm{d}_{hamm}(u,v) = 1$. This and other earlier work, such as [7,15], focus on identifying isomorphism with a subgraph of the n-cube, rather than isomorphism with a one-inclusion graph as considered here where the additional condition (iii) is required. These so-called *cubical* (as opposed to necessarily one-inclusion) graphs have applications in networks and parallel algorithms [20]; significant work has gone into enumerating classes of graphs that are cubical/non-cubical and also into the computational complexity of the corresponding decision problem.

4.2. The complementary view of one-inclusion graphs

Focusing on the complement of a subset of the n-cube turns out to provide a surprisingly useful view on the combinatorics of such subsets.

Definition 30 *The* complementary set *of a family* $V \subseteq \{0,1\}^n$ *is* $\overline{V} = \{0,1\}^n \backslash V$*. A collection of subcubes* $\mathcal{C}$ *contained/embedded in* $\overline{V}$ *is called* d-complete *if each subcube is of dimension* d *and for each choice of* $I \subset [n]$ *with* $|I| = d$ *there exists a* $C \in \mathcal{C}$ *shattering* I *(or equivalently* C *is* I*-colored). A* maximally overlapping d*-complete collection in the* n*-cube is a minimizer of* $\left|\bigcup_{C\in\mathcal{C}} C\right|$ *over all* d*-complete collections in the* n*-cube.*

The key to the usefulness of the complementary set is the following geometric characterization of a finite concept class VC-dimension.

Theorem 31 $V \subseteq \{0,1\}^n$ *has* $\mathrm{VC}(V) \leq d$ *iff* $\overline{V}$ *contains a* $(n-d-1)$*-complete collection of subcubes. In particular this implies that* $\mathrm{VC}(V) = d$ *iff* $\overline{V}$ *contains a* $(n-d-1)$*-complete collection of subcubes but no* $(n-d)$*-complete collection.*

Proof. For fixed $I \subseteq [n]$, $|I| = k+1$, $\Pi_I(V) \neq \{0,1\}^{k+1}$ iff there exists an $([n]\backslash I)$-colored $(n-k-1)$-subcube embedded in $\overline{V}$. Thus $\mathrm{VC}(V) \leq k$ iff $\overline{V}$ contains a $(n-k-1)$-complete collection of subcubes. Now apply

[2] Points in diagonally opposite corners—take all vectors with an even number of 1's.

this equivalence directly with $k = d$ and its inverse with $k = d-1$. This proves $\mathrm{VC}(V) \leq d$ iff $\overline{V}$ contains a $(n-d-1)$-complete collection and $\mathrm{VC}(V) \geq d$ iff $\overline{V}$ does not contain a $(n-d)$-complete collection. ■

From this result we gain the first natural characterization of maximal classes.

Lemma 32 (Complementary characterization of maximal sets) *$V \subseteq \{0,1\}^n$ of $\mathrm{VC}(V) = d$ is maximal iff $\overline{V}$ is a $(n-d-1)$-complete collection of subcubes and properly contains no $(n-d-1)$-complete collection.*

Proof. Consider any $V \subseteq \{0,1\}^n$ with $\overline{V}$ equal to a $(n-d-1)$-complete collection, and properly containing no other $(n-d-1)$-complete collection. Then Theorem 31 implies that $\mathrm{VC}(V) = d$. Adding any point $v \notin V$ to V corresponds to removing v from $\overline{V}$, thereby breaking at least one of the $(n-d-1)$-cubes in $\overline{V}$. Since $\overline{V \cup \{v\}}$ contains no $(n-d-1)$-complete collection, $\mathrm{VC}(V \cup \{v\}) \geq d+1$ which by definition implies that V is maximal.

Consider now any maximal $V \subseteq \{0,1\}^n$ of VC-dimension d. Then by Theorem 31, $\overline{V}$ contains a $(n-d-1)$-complete collection $\mathcal{C}$. By the maximality of V, $\overline{V} \setminus \bigcup_{C \in \mathcal{C}} C = \emptyset$ since any point $v \in \overline{V}$ not covered by $\mathcal{C}$ could be added to V so that $\overline{V} \setminus \{v\}$ would still contain $\bigcup_{C \in \mathcal{C}} C$ implying the contradictory $\mathrm{VC}(V \cup \{v\}) = d$. Thus $\overline{V}$ contains (but not properly contains) an $(n-d-1)$-complete collection. ■

We can also study the complement of special maximal classes—maximum classes.

Lemma 33 *$V_n^d = \{x \in \{0,1\}^n : \|x\|_1 \leq d\}$ is the only maximal closed-below family of VC-dimension d in the n-cube. Thus maximal and maximum coincide for closed-below families.*

Proof. Let $V \subseteq \{0,1\}^n$ be a maximal closed-below family of VC-dimension d. $\mathrm{VC}(V) = d$ implies that V contains at least one d-cube but no $(d+1)$-cube (where cubes are embedded in V and contain the origin). Maximality implies that, for every $v \in \overline{V}$, $\mathrm{VC}(V \cup \{v\}) > d$ and thus that v must have at least $d+1$ ones. Hence $V = V_n^d$. ■

Theorem 34 (Complementary characterization of maximum sets) *For any $n, d \in \mathbb{N}$ and set $V \subseteq \{0,1\}^n$, the following statements are equivalent:*

(i) V is maximum with $\mathrm{VC}(V) = d$;
(ii) $\overline{V}$ is the union of a maximally overlapping $(n-d-1)$-complete collection $\mathcal{C}$, in the sense that $\mathcal{C}$ covers a minimum number of distinct points in the n-cube ($|\bigcup_{C \in \mathcal{C}} C| = |\overline{V}|$ is minimum over all $(n-d-1)$-complete collections);
(iii) $\overline{V}$ is maximum with $\mathrm{VC}(\overline{V}) = n-d-1$.
(iv) V is the union of a maximally overlapping d-complete collection; and
(v) V and $\overline{V}$ contain a d-complete and a $(n-d-1)$-complete collection respectively.

Proof. Let $V \subseteq \{0,1\}^n$ be a maximum class with VC-dimension d. By Lemma 32 maximal V has complement $\overline{V}$ equal to the union of some $(n-d-1)$-complete collection $\mathcal{C}$. If $|\bigcup_{C \in \mathcal{C}} C|$ were not minimal over all $(n-d-1)$-collections then there would exist families of VC-dimension d in the n-cube of larger cardinality than V contradicting the choice of V as maximum. Thus (i) $\Rightarrow$ (ii). Conversely if $V \subseteq \{0,1\}^n$ is defined by $\overline{V} = \bigcup_{C \in \mathcal{C}} C$, for some maximally overlapping $(n-d-1)$-complete collection $\mathcal{C}$, then V can not properly contain an $(n-d-1)$-complete collection and so is maximal of VC-dimension d and furthermore has maximum cardinality over all VC-dimension d maximal subsets in the n-cube. So (ii) $\Rightarrow$ (i).

For (iii) and (iv), let $\overline{V}'$ denote $\overline{V}$ with all n components of each of its vertices flipped. Any sequence of shifts takes V down to a closed-below fixed-point iff the sequence takes $\overline{V}$ up to a closed-above fixed-point iff it takes $\overline{V}'$ down to the (correspondingly flipped) closed-below fixed-point. Since V is maximum, every sequence of shifts down to a fixed-point maps V to V_n^d as that is the unique closed-below family of cardinality $\binom{n}{\leq d}$ and VC-dimension at most d (see Lemma 33); such a sequence takes $\overline{V}$ up to $\overline{V_n^d}$ and $\overline{V}'$ down to V_n^{n-d-1}. Now consider a VC-invariant shifting of $\overline{V}'$ down to a closed-below family; this corresponds to a

shifting of V down to V_n^d. Hence the VC-invariant shifting of $\overline{V}'$ has fixed-point V_n^{n-d-1} and so $\overline{V}'$ is a maximum VC-dimension $n-d-1$ family. Since $\mathrm{VC}(\overline{V}) = \mathrm{VC}(\overline{V}')$ and $|\overline{V}| = |\overline{V}'|$, the results follows.

Consider now an arbitrary subset $V \subseteq \{0,1\}^n$ such that V and $\overline{V}$ contain a d and a $(n-d-1)$-complete collection of cubes respectively. Denote the unions of these collections U_d and U_{n-d-1} respectively. By Sauer's Lemma, (ii) and (iv), $|V| \geq |U_d| \geq \binom{n}{\leq d}$ and $|\overline{V}| \geq |U_{n-d-1}| \geq \binom{n}{\leq n-d-1}$. With $|V| = 2^{2^n} - |\overline{V}|$ this implies that $V = U_d$ and $\overline{V} = U_{n-d-1}$, that U_d and U_{n-d-1} are the unions of maximally overlapping collections and so that V is maximum of VC-dimension d. The converse is immediate, and so (i) $\Leftrightarrow$ (v). ■

Remark 35 *The equivalences of (i)–(iv) in Theorem 34, were first shown by Floyd in her thesis [8] under the guise of* forbidden labels*. Each complementary $n-d-1$-cube of a maximum class of VC-dimension d can be uniquely identified with the intersection over that cube's concepts' sets of support—i.e. the cube's concept with fewest 1's. Floyd referred to such a concept as a forbidden label since no concept in the class can be consistent with that complementary concept. In particular Floyd showed that a maximum class is characterized by its set of forbidden labels [8, Lemma 3.15] and that such a class has a maximum complement of the appropriate VC-dimension [8, Lemma 3.20]. She also considered maximum classes on infinite domains, which is beyond the scope of this paper. The relatively superficial change of viewpoint from forbidden labels to complementary simplicial complexes may provide a useful geometric characterization of maximum classes. The forbidden labels of maximal classes, as per Lemma 32, were not discussed in [8].*

With Theorem 34.(v), we can prove the following classic result (see e.g. [5,1]) characterizing VC-1 maximum classes.

Lemma 36 *$V \subseteq \{0,1\}^n$ is maximum of VC-dimension 1 iff $\mathcal{G}(V)$ is a tree with d uniquely colored edges.*

Proof. Consider maximum $V \subseteq \{0,1\}^n$ of VC-dimension 1. By Theorem 34, V equals a union of d uniquely colored edges and so is acyclic. By Sauer's Lemma $|V| = d+1$. Thus V is a tree with d uniquely colored edges. Conversely such a tree has VC-dimension 1 and has $d+1$ vertices, and thus is maximum. ■

4.3. An algebraic topological property: maximum classes are contractible

We now develop a natural extension of the tree characterization of maximum VC-1 classes of Lemma 36. The direction of extension replaces vertices and edges of a graph by higher dimensional cubes; in the language of algebraic topology we are interested in simplicial complexes (like graphs) that are contractible (like trees). We begin with some preliminaries needed only for this section, then state and prove the main theorem.

Definition 37 *A* homotopy *is a continuous map $F : X \times [0,1] \to Y$. The* initial map *is F restricted to $X \times \{0\}$ and the* final map *is F restricted to $X \times \{1\}$. We often say that the initial and final maps are* homotopic*; and for such maps we refer to the respective product domains as X with the short-hand understood by context. A* homotopy equivalence *between spaces X and Y is a pair of maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g$, is homotopic to the identity map on Y and $g \circ f$ is homotopic to the identity map on X.*

Definition 38 *A* cubical simplicial complex *is a union of solid cubes of the form $[a_1, b_1] \times \ldots \times [a_m, b_m]$ (for varying but bounded m) such that the intersection of any two cubes is either a cubical face of both cubes or the empty-set.*

Recall that a contractible complex X is one which has the same homotopy type as a point, that is, the identity map $I : X \to X$ is homotopic to the constant map $c : X \to p$ a point in X. (Note c is considered as a map from X to X with image p). Then in our situation of contractibility, the two maps are c considered as the map from X to $\{p\}$ and $i : \{p\} \to X$ which takes p to p but considered as a point of X. Then the composition $c \circ i$ is the identity on $\{p\}$ so the constant homotopy which is independent of the second variable in $[0,1]$ is the homotopy from $c \circ i$ to the identity on $\{p\}$ and $i \circ c$ is homotopic to the identity map I on X via the homotopy of the identity to the constant map.

Theorem 39 (Algebraic topological property of maximum classes) *Maximum classes of VC-dim. d in the n-cube form d-dimensional cubical complexes which are contractible subcomplexes of the cubical structure of the binary hypercube $[0,1]^n$.*

Proof. Consider the projection map f from the n-cube to the $(n-1)$-cube. We prove our result by induction on $n+d$. So by assumption, any Maximum(n',d') class with $d'+n' < d+n$ is contractible. Let X denote our Maximum(n,d) class, viewed as a d-dimensional cubical complex. Then we know that $f(X) = X'$ is a contractible d-dimensional cubical complex, since it is a Maximum$(n-1,d)$ class. Also f projects the reduction, which is of the form $Y \times [0,1]$, onto Y, where Y is a Maximum$(n-1,d-1)$ class, and hence by the inductive hypothesis is a contractible $(d-1)$-dimensional cubical complex.

Now we do some basic algebraic topology. Consider a pair of spaces such as $(X, Y \times [0,1])$. So the second space is a subspace of the first one. Then we can examine the effect of collapsing the subspace to a point. Write this as $X/Y \times [0,1]$ (a quotient space). Now by standard arguments, if the subspace is contractible, then the quotient space is homotopy equivalent to the original space. In other words, collapsing a contractible subspace to a point does not affect the homotopy properties of a space. Note here that X is a cubical complex and $Y \times [0,1]$ is a subcomplex, which is a sufficient condition to apply this collapsing result.

Next, consider the two quotient spaces, $X/Y \times [0,1]$ and $f(X)/Y$. It also follows by standard results that these are in fact homeomorphic. In fact, the map $f : X \mapsto f(X)$ is one-to-one on $X \backslash (Y \times [0,1])$ and projects $Y \times [0,1] \to Y$. So again, since $Y \times [0,1]$ is a subcomplex of X, it follows that the results of collapsing $Y \times [0,1]$ to a point in X and Y to a point in $f(X)$ are homeomorphic by the map induced by f. But now we can apply the result of the previous paragraph. Namely we know by induction that $f(X)$ is contractible and Y is contractible, so $f(X)/Y$ is contractible. But therefore it follows that $X/Y \times [0,1]$ is contractible. Finally we got this by collapsing a contractible subspace $Y \times [0,1]$ to a point (the extra factor $[0,1]$ makes no difference to contractibility as is easy to see). So X is homotopy equivalent to $X/Y \times [0,1]$ which we have just proved is contractible, hence X is contractible. (Anything homotopy equivalent to a contractible space is contractible). ■

Note that there are contractible cubical complexes, equal to the union of a complete collection of d-cubes, which are not maximum classes; and there are also such cubical complexes which are not contractible.

Example 40 *Consider the union of a complete collection of 2-cubes in $\{0,1\}^5$, shown in Figure 5. This class is contractible but not maximum: the subset's VC-dimension and cardinality are 3 and 17 respectively, whereas the cardinality of a 2-maximum class in the 5-cube is 16.*

Example 41 *Consider the union of a complete collection of 1-cubes in $\{0,1\}^4$, shown in Figure 6. This class is not contractible or maximum: the subset's VC-dimension is 2, its cardinality is 6, and it is not even connected.*

Remark 42 *Theorems 34 and 39 together lead to the interesting result that a d-complete cubical complex of fewest vertices is in fact contractible.*

5. One-inclusion minimum degree can exceed VC-dimension

Kuzmin and Warmuth conjectured [17] that every VC-dimension $d < \infty$ class has a one-inclusion graph with minimum degree $\delta \leq d$. This conjecture was motivated by their Peeling algorithm: if the conjecture were true then the Peeling algorithm would successfully compress all maximum classes. The following counter-example, motivated by the complementary view of the one-inclusion graph, resolves the conjecture as false. See Section 6 for an in-depth discussion of peeling as well as other consequences of this result.

Theorem 43 *There exists a family $V \subset \{0,1\}^{12}$ with VC-dimension 10 having vertices of graph degree in $\{11,12\}$.*

	x_1	x_2	x_3	x_4	x_5
c_1	0	0	0	0	0
c_2	1	0	0	0	0
c_3	0	1	0	0	0
c_4	0	0	1	0	0
c_5	1	1	0	0	0
c_6	0	1	1	0	0
c_7	1	0	1	0	0
c_8	0	0	0	1	0
c_9	1	0	0	1	0
c_{10}	0	1	0	1	0
c_{11}	0	1	1	1	0
c_{12}	0	0	0	1	1
c_{13}	1	0	0	1	1
c_{14}	1	0	0	0	1
c_{15}	0	0	1	0	1
c_{16}	0	1	1	0	1
c_{17}	0	1	0	0	1

Fig. 5. The non-maximum, contractible, 2-complete collection of Example 40.

	x_1	x_2	x_3	x_4
c_1	0	0	0	0
c_2	1	0	0	0
c_3	1	1	0	0
c_4	1	0	1	0
c_5	0	1	1	1
c_6	0	1	1	0

Fig. 6. The non-maximum, incontractible, 1-complete collection of Example 41.

Proof. We describe V by way of $\overline{V}$, which is composed of three vertex-disjoint 4-cubes that are pairwise hamming-4 separated:

$$\overline{V} = S_1 \cup S_2 \cup S_3 \ , \ \text{where} \ \begin{aligned} S_1 &= \{0,1\}^4 \times \{(0,0,1,1,0,0,1,1)\} \\ S_2 &= \{(0,0,1,1)\} \times \{0,1\}^4 \times \{(1,1,0,0)\} \\ S_3 &= \{(1,1,0,0,1,1,0,0)\} \times \{0,1\}^4 \ . \end{aligned}$$

We first establish that $\mathrm{VC}(V) = d = 10$. The three subcubes collectively contain edges along each direction in [12], thus $\overline{V}$ contains an $(n-d-1) = 1$-complete collection of cubes. The subcubes S_1, S_2, S_3 shatter $\{1,\ldots,4\}$, $\{5,\ldots,8\}$ and $\{9,\ldots,12\}$ respectively, and since they are pairwise-4 apart $\overline{V}$ cannot contain an $(n-d) = 2$-complete collection. Thus $\mathrm{VC}(V) = 10$ by Theorem 31.

Since S_1, S_2, S_3 are pairwise-4 separated, any vertex $v \in V$ hamming-1 from S_i must be hamming-1 from exactly one $w \in S_i$ and at least distance-2 from the other two complementary-subcubes; in particular every $v \in V$ can adjoin at most one $w \in \overline{V}$ and so at most one v's potential $\{0,1\}^{12}$-neighbors can be missing from V and so v has degree in $\{11, 12\}$. ■

5.1. The uniform degree-VC ratio

Although the proof of Theorem 43 is tied to the details of the counter-example—particularly that the δ-VC gap is 1—-the example does immediately extend to related examples of higher VC-dimension, embedded in higher-dimensional hypercubes.

Corollary 44 *For each $d \geq 10$ and $n \geq d+2$ there exists a family $V \subseteq \{0,1\}^n$ such that $\mathrm{VC}(V) = d$ and $\delta(\mathcal{G}(V)) = d+1$.*

Proof. For $d = 10$ Theorem 43 provides a graph with $V_{10} \subset \{0,1\}^{10+2}$ with VC-dimension 10 and minimum degree 11. For any $d \geq 10$ we can construct an appropriate $V_d \subset \{0,1\}^{d+2}$, as we did for V_{10}, with the following complementary set:

$$\overline{V}_d = S_{d,1} \cup S_{d,2} \cup S_{d,3}$$

where

$$S_{d,1} = \{0,1\}^4 \times \{(0,0,1,1,0,0,1,1)\} \times \{0\}^{d-10}$$
$$S_{d,2} = \{(0,0,1,1)\} \times \{0,1\}^4 \times \{(1,1,0,0)\} \times \{0\}^{d-10}$$
$$S_{d,3} = \{(1,1,0,0,1,1,0,0)\} \times \{0,1\}^{d-6}$$

The same arguments for V_{10} apply for general $d > 10$ to imply that $\mathrm{VC}(V_d) = d$ and $\delta(\mathcal{G}(V_d)) = d+1$. Now to get families in arbitrary n-cubes for $n \geq d+2$ (for $d \geq 10$) note that we can simply embed the appropriate V_d in the n-cube, i.e. as $V_d \times \{0\}^{n-d-2}$, which does not affect VC-dimension or minimum degree. ■

Thus there are 'many' counter-examples for which the δ-VC gap is one, but can larger gaps be achieved? A first step towards answering this question is provided by the following corollary.

Lemma 45 *For any $n \in \mathbb{N}$ and any $V \subseteq \{0,1\}^n$, $\mathrm{VC}(V \times V) = 2\mathrm{VC}(V)$ and $\delta(V \times V) = 2\delta(V)$.*

Proof. V shatters index-set $I \subseteq [n]$ iff $V \times V$ shatters $I \circ I$, where $\circ$ denotes concatenation. For fixed $u, v \in V$ consider the vertex $u \circ v \in V \times V$. $u \circ v$ is hamming-1 from some $x \circ y$, $x, y \in \{0,1\}^n$, iff either $u = x$ and $\mathrm{d}_{hamm}(v, y) = 1$ or $v = y$ and $\mathrm{d}_{hamm}(u, x) = 1$. Thus $\deg(u \circ v) = \deg(u) + \deg(v)$ and

$$\delta(\mathcal{G}(V \times V)) = \min_{v \in V \times V} \deg(v) = \min_{u,v \in V} \deg(u \circ v) = 2 \min_{v \in V} \deg(v) = 2\delta(V)$$

■

Corollary 46 *For each $i \in \mathbb{N}$ there exists a family $V \in \{0,1\}^{12i}$ with $\delta(\mathcal{G}(V)) - \mathrm{VC}(V) = i$.*

Proof. Consider the family $V \subset \{0,1\}^{12}$ with $\mathrm{VC}(V) = 10$ and $\delta(V) = 11$ constructed as the counter-example in Theorem 43. Then by induction on i Lemma 45 implies that for any $i \in \mathbb{N}$, the product family $V_i = \prod_{j=1}^{i} V \subset \{0,1\}^{12i}$ has $\mathrm{VC}(V_i) = 10i$ and $\delta(\mathcal{G}(V_i)) = 11i$. ■

Corollary 46 demonstrates arbitrary δ-VC gaps. We see that to achieve large gaps it is sufficient for both the minimum degree and VC-dimension to be large. Whether this is *necessary* motivates the next definition.

Definition 47 (The uniform degree-VC ratio) *The* uniform degree-VC ratio *is defined as*

$$\kappa = \sup_{n \in \mathbb{N}} \sup_{\substack{V \subseteq \{0,1\}^n \\ |V| \geq 1}} \frac{\delta(\mathcal{G}(V))}{\mathrm{VC}(V)} \ .$$

The classic density bound and the full n-cube establish basic upper- and lower-bounds on κ.

Lemma 48 $1 \leq \kappa < 2$.

Proof. The lower-bound is witnessed by the n-cube: for $n \in \mathbb{N}$, $\delta(\mathcal{G}(\{0,1\}^n)) = n$ and $\mathrm{VC}(\{0,1\}^n) = n$. The upper-bound follows from the density bound of Theorem 24: for any $n \in \mathbb{N}$, $V \subseteq \{0,1\}^n$ and $(V, E) = \mathcal{G}(V)$ we have that $\delta(\mathcal{G}(V)) \leq \frac{\sum_{v \in V} \deg(v)}{|V|} \leq \frac{2|E|}{|V|} \leq 2D_n^{\mathrm{VC}(V)} < 2\mathrm{VC}(V)$. ■

The Kuzmin-Warmuth degree conjecture and density bounds are naturally related.

Proposition 49 *The Kuzmin-Warmuth minimum degree conjecture [17] is true iff* $\kappa \leq 1$.

Corollary 50 $\kappa \geq 1.1$.

Proof. The example families V_i, $i \in \mathbb{N}$, of Theorem 43, Corollary 46 satisfy $\frac{\deg(\mathcal{G}(V_i))}{\mathrm{VC}(V_i)} = 1.1$. ■

The classic density bound shows that the VC-dimension is bounded below by graph density. Since the degree conjecture fails, this raises the natural question as to whether there is some intermediate bound for the VC-dimension that is less than the minimal degree but larger than the density.

6. Consequences for sample compression

Kuzmin and Warmuth [17] proposed the elegant Peeling algorithm (Algorithm 3) and conjectured that it is an unlabeled d-compression scheme for d-maximum classes. Given $V \subseteq \{0,1\}^n$ and $k \leq n$, one *k-peels* V by successively removing vertices of degree less than k from V, at each step removing a minimum-degree vertex. A successful peeling ultimately reaches $\emptyset$. At each stage the currently peeled vertex is assigned its present incident dimensions as its representative. Thus a k-peeled V admits a mapping r from concepts of V to representatives of size at most k.

Algorithm 3 The Min-Peeling Algorithm of [17]

Given: $C \subseteq \{0,1\}^{\mathcal{X}}$ with $|\mathcal{X}| < \infty$
Returns: a representation mapping r for C

$G \longleftarrow \mathcal{G}(\Pi_{\mathcal{X}}(C))$;
while $C \neq \emptyset$ **do**
  $(v, c) \longleftarrow$ a minimum-degree vertex in G and the concept of C in that vertex's version space ;
  $r(c) \longleftarrow I_G(v)$;
  $(G, C) \longleftarrow (G \backslash \{v\}, C \backslash \{c\})$;
end while
return r ;

Kuzmin and Warmuth's minimum degree conjecture [17] predicted that every VC-dimension d class has a one-inclusion graph with minimum degree $\delta \leq d$. If this were true then every d-dimensional class would have a d-peeling. As a refinement to this conjecture, Kuzmin and Warmuth also conjectured that D_n^d bounds the density of all one-inclusion graphs and that any graph G of VC-dimension d in the n-cube with $\mathrm{dens}(G) \leq D_n^d$ has $\delta(G) \leq D_n^d$. Although we have verified the D_n^d density bound with Theorem 24 our counter-examples in Section 5 negatively resolve both minimum degree conjectures. Note, however, that our examples are not maximum classes and so it is still possible that Peeling is a valid maximum unlabeled d-compression scheme.

An immediate consequence of a proof of the correctness of maximum peeling (together with our minimum degree counter-examples) would be an impossibility statement for embedding maximal classes in certain maximum classes, giving a lower bound on the quantity in Problem 14.

Proposition 51 (Peeling implies an increase of VC-dim when embedding) *If every maximum class of VC-dimension d in the n-cube can be d-peeled, then there exists a maximal class V which cannot be embedded in any maximum class of VC-dimension smaller than $\kappa \cdot \mathrm{VC}(V)$. In particular, for each $i \in \mathbb{N}$ there exist maximal classes of VC-dimension $10i$ that could not be embedding in any maximum class of VC-dimension equal to or smaller than $11i$.*

Proof. Suppose that d-maximum classes could be d-peeled and assume that it were possible to embed a maximal class L of minimum degree $\delta > d$ in a d-maximum class M. Then d-peeling the M would proceed by iteratively removing minimum degree vertices, each of degree at most d. Eventually a minimum degree vertex will come from the embedding of L; consider the first such vertex. It will have degree at least $\delta > d$ and so it follows that M could not be d-peeled. Thus any maximal class embeddable in a d-maximum class

must have minimum degree at most d. The particular $10i$-maximal classes can be found by adding concepts to the examples of Section 5. ■

7. Expected risk bounds for multiclass prediction

As in the $k = 1$ case, the key to developing the multiclass one-inclusion mistake bound is in bounding hypergraph density. We proceed by shifting a graph induced by the one-inclusion hypergraph.

Theorem 52 (One-inclusion hypergraph density bound) *For any $k, n \in \mathbb{N}$ and $V \subseteq \{0, \ldots, k\}^n$, the one-inclusion hypergraph $(V, E) = \mathcal{G}(V)$ satisfies $\frac{|E|}{|V|} \leq \Psi_{\mathrm{P}}\text{-dim}(V)$.*

Proof. We begin by replacing the hyperedge structure E with a related edge structure E'. Two vertices $\boldsymbol{u}, \boldsymbol{v} \in V$ are connected in the graph (V, E') iff there exists an $i \in [n]$ such that $\boldsymbol{u}, \boldsymbol{v}$ differ only at i and no $\boldsymbol{w} \in V$ exists such that $u_i < w_i < v_i$ and $w_j = u_j = v_j$ on $[n]\backslash\{i\}$. Trivially

$$\frac{|E|}{|V|} \leq \frac{|E'|}{|V|} \leq \frac{k|E|}{|V|} \ . \tag{7}$$

Consider now shifting vertex $\boldsymbol{v} \in V$ at shift label $t \in [k]$ along shift coordinate $s \in [n]$ by

$$\begin{aligned}
&S_{s,t}(\boldsymbol{v}; V) = \boldsymbol{v}^{s(v'_s)} \\
&\text{where} \\
&\qquad \boldsymbol{v}^{s(i)} = (v_1, \ldots, v_{s-1}, i, v_{s+1}, \ldots, v_n) \quad \text{for } i \in \{0, \ldots, k\} \\
&\qquad v'_s = \begin{cases} \min\left\{ x \in \{0, \ldots, v_s\} \;\middle|\; \boldsymbol{v}^{s(x)} \notin V \text{ or } x = v_s \right\} & \text{if } v_s = t \\ v_s & \text{otherwise} \end{cases}
\end{aligned}$$

We shift V on s at t as usual; we shift V on s alone by bubbling vertices down to fill gaps below:

$$\begin{aligned}
S_{s,t}(V) &= \{S_{s,t}(\boldsymbol{v}; V) \mid \boldsymbol{v} \in V\} \\
S_s(V) &= S_{s,k}(S_{s,k-1}(\ldots S_{s,1}(V))) \ .
\end{aligned}$$

Let $S_s(E')$ denote the *edge*-set induced by $S_s(V)$. The mapping S_s on a vertex-set is injective implying that

$$|S_s(V)| = |V| \ . \tag{8}$$

Consider any $\{\boldsymbol{u}, \boldsymbol{v}\} \in E'$ with $i \in [n]$ denoting the index on which $\boldsymbol{u}, \boldsymbol{v}$ differ. If $i = s$ then no other vertex $\boldsymbol{w} \in V$ can come between $\boldsymbol{u}$ and $\boldsymbol{v}$ during shifting by construction of E', so $\{S_s(\boldsymbol{u}; V), S_s(\boldsymbol{v}; V)\} \in S_s(E')$. Now suppose that $i \neq s$. If both vertices shift down by the same number of labels then they remain connected in $S_s(E')$. Otherwise assume WLOG that $S_s(\boldsymbol{u}; V)_s < S_s(\boldsymbol{v}; V)_s$ then the shifted vertices will lose their edge, however since v_s did not shift down to $S_s(\boldsymbol{u}; V)_s$ there must have been some $\boldsymbol{w} \in V$ different from $\boldsymbol{u}$ on $\{i, s\}$ such that $w_s < v_s$ with $S_s(\boldsymbol{w}; V)_s = S_s(\boldsymbol{u}; V)_s$. Thus, $S_s(\boldsymbol{w}; V)$ and $S_s(\boldsymbol{u}; V)$ differ only on $\{i\}$ and a new edge $\{S_s(\boldsymbol{w}; V), S_s(\boldsymbol{u}; V)\}$ is in $S_s(E')$ that was not in E' (otherwise $\boldsymbol{u}$ would not have shifted). Thus

$$|S_s(E')| \geq |E'| \ . \tag{9}$$

Suppose that $I \subseteq [n]$ is Ψ_P-shattered by $S_s(V)$. If $s \notin I$ then $\Pi_I(S_s(V)) = \Pi_I(V)$ and I is Ψ_P-shattered by V. If $s \in I$ then V Ψ_P-shatters I. Consider witnesses of $S_s(V)$'s Ψ_P-shattering of I equal to 1 at s, taking each value in $\{0, 1\}^{|I|-1}$ on $I\backslash\{s\}$. These were not shifted and so are witnesses for V. Since these vertices were not shifted they were blocked by vertices of V of equal values on $I\backslash\{s\}$ but equal to 0 at s. These are the remaining half of the witnesses of V's Ψ_P-shattering of I. Thus

$$S_s(V)\ \Psi_P\text{-shatters } I \subseteq [n] \Rightarrow V\ \Psi_P\text{-shatters } I\ . \tag{10}$$

In a finite number of shifts starting from (V, E'), a closed-below family W with induced edge-set F will be reached. If $I \subseteq [n]$ is Ψ_P-shattered by W and $|I| = d = \Psi_{\mathrm{P}}\text{-dim}(W)$, then since W is closed-below the translation vector $(\psi_{P,1}, \ldots, \psi_{P,1})(\cdot) = (\mathbf{1}[\cdot < 1], \ldots, \mathbf{1}[\cdot < 1])$ must witness this shattering. Hence each $\boldsymbol{w} \in W$ has at most d non-zero components. Counting edges in F by upper-adjoining vertices we have proved that

$$(V, E') \text{ finitely shifts to closed-below graph } (W, F) \text{ s.t. } |F| \leq |W| \cdot \Psi_{\mathrm{P}}\text{-dim}(W)\ . \tag{11}$$

Combining properties (7)–(11) we have that $\frac{|E|}{|V|} \leq \frac{|E'|}{|V|} \leq \frac{|F|}{|W|} \leq \Psi_{\mathrm{P}}\text{-dim}(W) \leq \Psi_{\mathrm{P}}\text{-dim}(V)$. ■

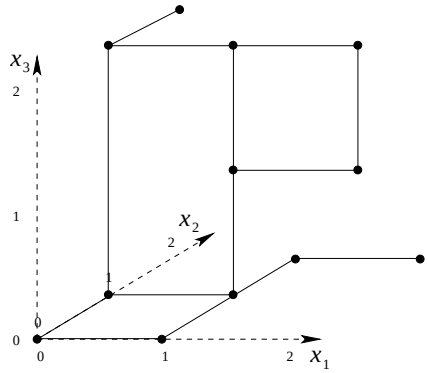

Fig. 7. The graph induced by the one-inclusion hypergraph of Figure 7, for the shifting process in Theorem 52. The graph's density increases to $\frac{14}{12}$ and pseudo-dimension remains fixed.

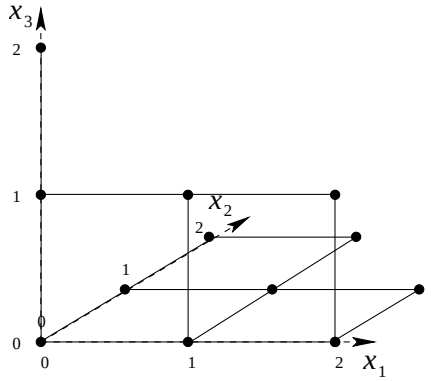

Fig. 8. The closed-below fixed-point reached by shifting the graph in Figure 7. The graph's density further increases to $\frac{16}{12}$ while the vertex-set's pseudo-dimension does not increase (it remains at 2).

Example 53 *Consider the class $V \subset \{0,1,2\}^3$ of Example 9, with one-inclusion hypergraph displayed in Figure 1. $\mathcal{G}(V)$ has density $\frac{11}{12}$ while $\Psi_{\mathrm{P}}\text{-dim}(V) = 2$. To illustrate the shifting process in the proof of Theorem 52, consider Figures 7 and 8. The former depicts the graph induced by the hypergraph $\mathcal{G}(V)$; it has density $\frac{14}{12} \geq \frac{11}{12}$ and dimension $\Psi_{\mathrm{P}}\text{-dim}(V)$. The latter figure depicts a closed-below fixed point reached by shifting on x_3 at 1, x_3 at 2, x_1 at 1, x_2 at 1 and finally on x_2 at 2. The fixed-point graph has density $\frac{16}{12} \geq \frac{14}{12}$ and dimension $2 \leq \Psi_{\mathrm{P}}\text{-dim}(V)$.*

The remaining arguments from the $k = 1$ case of [13,12] now imply the multiclass mistake bound.

Theorem 54 (One-inclusion multiclass mistake bounds) *Consider any $k, n \in \mathbb{N}$ and class $\mathcal{F} \subseteq \{0, \ldots, k\}^{\mathcal{X}}$ with $\Psi_{\mathrm{P}}\text{-dim}(\mathcal{F}) < \infty$. The multiclass one-inclusion prediction strategy satisfies $\hat{M}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}(n) \leq \Psi_{\mathrm{P}}\text{-dim}(\mathcal{F})/n$.*

7.1. Proof of the general multiclass mistake bound

We begin with the generalization of Lemma 2.2[13, Corollary 2.1].

Lemma 55 *For any $n > 1, k \in \mathbb{N}$, any $\mathcal{F} \subseteq \{0, \ldots, k\}^{\mathcal{X}}$ and any deterministic prediction strategy Q, $\hat{M}_{Q,\mathcal{F}}(n) \leq \hat{\hat{M}}_{Q,\mathcal{F}}(n)$.*

Proof. For initially fixed $f \in \mathcal{F}$, permutation $\sigma \in S_n$ and distribution P on $\mathcal{X}$, exchangeability of P^n and linearity of expectation imply

$$
\begin{aligned}
&\mathbb{E}_{P^n}\left[\mathbf{1}\left[Q\left(\operatorname{sam}\left(\left(X_1,\ldots,X_{n-1}\right),f\right),X_n\right)\neq f(X_n)\right]\right]\\
=&\,\mathbb{E}_{P^n}\left[\mathbf{1}\left[Q\left(\operatorname{sam}\left(\left(X_{\sigma(1)},\ldots,X_{\sigma(n-1)}\right),f\right),X_{\sigma(n)}\right)\neq f\left(X_{\sigma(n)}\right)\right]\right]\\
=&\,\mathbb{E}_{P^n}\left[\frac{1}{n!}\sum_{\sigma\in S_n}\mathbf{1}\left[Q\left(\operatorname{sam}\left(\left(X_{\sigma(1)},\ldots,X_{\sigma(n-1)}\right),f\right),X_{\sigma(n)}\right)\neq f\left(X_{\sigma(n)}\right)\right]\right]\\
\leq&\sup_{\boldsymbol{x}\in\mathcal{X}^n}\frac{1}{n!}\sum_{\sigma\in S_n}\mathbf{1}\left[Q\left(\operatorname{sam}\left(\left(x_{\sigma(1)},\ldots,x_{\sigma(n-1)}\right),f\right),x_{\sigma(n)}\right)\neq f\left(x_{\sigma(n)}\right)\right]\ .
\end{aligned}
$$

Taking the supremum over $\mathcal{F}$ of both sides of the inequality completes the proof. ■

We now generalize [13, Theorem 2.3] to derive multiclass permutation mistake bounds from directed one-inclusion hypergraph maximum outdegree.

Lemma 56 *Consider any $\mathcal{F}\subseteq\{0,\ldots,k\}^{\mathcal{X}}$. If $\Delta(V)$ upper-bounds the maximum outdegree of $\overrightarrow{\mathcal{G}(V)}$ for any $V\subseteq\{0,\ldots,k\}^n$ under some understood orientation strategy[3], then $\hat{\hat{M}}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}(n)\leq\sup_{\boldsymbol{x}\in\mathcal{X}^n}\frac{\Delta(\Pi_{\boldsymbol{x}}(\mathcal{F}))}{n}$ for all $n>1$.*

Proof. Observe for fixed $f\in\mathcal{F}$, $\boldsymbol{x}$ and sample-order permutation $\sigma\in S_n$, that given $\operatorname{sam}\left(\left(x_{\sigma(1)},\ldots,x_{\sigma(n-1)}\right),f\right)$ strategy $Q_{\mathcal{G},\mathcal{F}}$ makes a mistake on $x_{\sigma(n)}$ iff $\boldsymbol{v}=(f(x_1),\ldots,f(x_n))$ has an out-going edge in the $x_{\sigma(n)}{}^{\text{th}}$ direction. Secondly observe that x_i appears in n^{-1} of the $n!$ permutations of $\boldsymbol{x}$. Thus

$$
\begin{aligned}
&\frac{1}{n!}\sum_{\sigma\in S_n}\mathbf{1}\left[Q_{\mathcal{G},\mathcal{F}}\left(\operatorname{sam}\left(\left(x_{\sigma(1)},\ldots,x_{\sigma(n-1)}\right),f\right),x_{\sigma(n)}\right)\neq f\left(x_{\sigma(n)}\right)\right]\\
\leq&\frac{\operatorname{outdeg}\left(\left(f\left(x_1\right),\ldots,f\left(x_n\right)\right)\right)}{n}\ .
\end{aligned}
$$

And taking suprema of both sides leads to

$$
\hat{\hat{M}}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}\ \leq\ \sup_{\boldsymbol{x}\in\mathcal{X}^n}\sup_{f\in\mathcal{F}}\frac{\operatorname{outdeg}\left(\left(f\left(x_1\right),\ldots,f\left(x_n\right)\right)\right)}{n}\ =\ \sup_{\boldsymbol{x}\in\mathcal{X}^n}\frac{\Delta\left(\Pi_{\boldsymbol{x}}\left(\mathcal{F}\right)\right)}{n}\ .
$$

■

Next we follow [12] in a non-constructive orientation of $\mathcal{G}\left(\Pi_{\boldsymbol{x}}\left(\mathcal{F}\right)\right)$.

Lemma 57 *For any $V\subseteq\{0,\ldots,k\}^n$ the edges of one-inclusion hypergraph $\mathcal{G}(V)=(V,E)$ can be oriented to give directed one-inclusion hypergraph $\overrightarrow{\mathcal{G}(V)}$ with maximum outdegree at most $\lceil\operatorname{maxdens}\left(\mathcal{G}\left(V\right)\right)\rceil$, where $\operatorname{maxdens}(G)$ is the maximum density of all subgraphs of (hyper)graph G.*

Proof. The result follows from an application of Hall's Theorem [11] to subgraphs of the bipartite graph depicted in Figure 9. We construct the bipartite graph (V_b,E_b) by taking vertices $V_b=E\cup V^{(1)}\cup\ldots\cup V^{(d)}$, where $V^{(1)},\ldots,V^{(d)}$ are $d=\operatorname{maxdens}\left(\mathcal{G}\left(V\right)\right)$ copies of the hypergraph's vertex-set V. Then $(w^{(1)},w^{(2)})\in V_b\times V_b$ is in undirected edge-set E_b iff there exists $i,j\in\{1,2\}$ and $\boldsymbol{v}\in V$ s.t. $i\neq j$, $w^{(i)}$ is one of the d copies of $\boldsymbol{v}$, and $\boldsymbol{v}\in w^{(j)}\in E$. Denote the neighbors of a vertex $v\in V_b$ by $\Gamma_{\mathrm{b}}\left(v\right)\subseteq V_b$.

Consider now any subgraph (V',E') of $\mathcal{G}\left(V\right)$ induced by selecting $q=|E'|\leq|E|$ hyperedges from the one-inclusion hypergraph, so that all vertices of V' have positive degree in the subgraph—isolated vertices are removed. Then

[3] Notice that the way we orient is unimportant, just that Δ is a bound on outdegree that depends only on V.

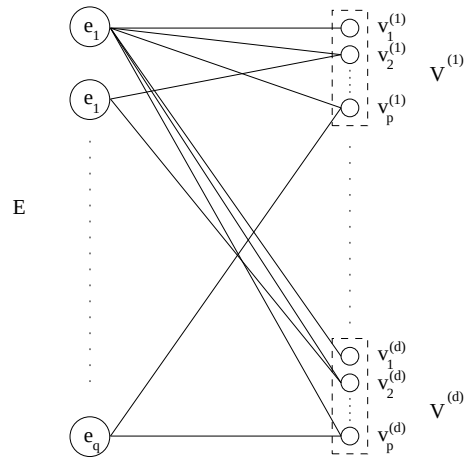

Fig. 9. The bipartite graph from the proof of Lemma 57. The vertex partitions are (on the left) the hyperedges of the one-inclusion hypergraph and (on the right) maxdensity-many copies of the vertices of the one-inclusion hypergraph. Each one-inclusion hyperedge is connected, in the bipartite graph, to the copies of its neighboring one-inclusion vertices.

$$\begin{aligned}\left|\bigcup_{e\in E'}\Gamma_{\mathrm{b}}\left(e\right)\right| &= d|V'| \\ &\geq |E'| \\ &= q \ .\end{aligned}$$

The first equality follows from the fact that the set of vertices adjoining E' in (V', E') is exactly V' and so in (V_b, E_b) the set of adjoining vertices are the d copies of V'. The inequality is the statement $\mathrm{dens}\left((V', E')\right) \leq d$ rearranged. Thus the family of $|E|$ neighbor sets $\mathcal{S}_E = \{\Gamma_{\mathrm{b}}\left(e\right) \mid e \in E\}$ satisfies the following: for all $1 \leq q \leq |E|$, the union of any q of the sets in $\mathcal{S}_E$ contains at least q distinct elements. Thus $\mathcal{S}_E$ satisfies the conditions of Hall's Theorem [11] so that each set of neighbors $\Gamma_{\mathrm{b}}\left(e\right)$ has a distinct representative $v_e^{(i)} \in \Gamma_{\mathrm{b}}\left(e\right)$ which is the i^{th} copy of some $(k+1)$-valued vector $\boldsymbol{v}^e \in V$ that adjoins e in $\mathcal{G}\left(V\right)$. Each such $\boldsymbol{v}^e$ provides an orientation for hyperedge e (arbitrarily) directed out from $\boldsymbol{v}$. As the neighbor set representatives $v_e^{(i)}$ are unique, when treating different copies of the same $\mathcal{G}\left(V\right)$ vertex as distinct, no one-inclusion hypergraph vertex $\boldsymbol{v}$ can be the representative of more than d hyperedges. Thus the outdegree for each $\boldsymbol{v} \in V$ in $\mathcal{G}\left(V\right)$ is at most d. ■

Finally note that Pollard dimension is non-decreasing with inclusion, so all subgraphs of a one-inclusion hypergraph $\mathcal{G}\left(V\right)$ have Pollard pseudo-dimension at most $\Psi_{\mathrm{P}}\text{-dim}\left(V\right)$.

Combining this observation with Lemmas 55–57 and Theorem 5.1 we see that

$$\begin{aligned}\hat{M}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}(n) &\leq \hat{\hat{M}}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}(n) \\ &\leq \sup_{\boldsymbol{x}\in\mathcal{X}^n} \frac{\lceil \mathrm{maxdens}\left(\Pi_{\boldsymbol{x}}\left(\mathcal{F}\right)\right)\rceil}{n} \\ &\leq \sup_{\boldsymbol{x}\in\mathcal{X}^n} \frac{\Psi_{\mathrm{P}}\text{-dim}\left(\Pi_{\boldsymbol{x}}\left(\mathcal{F}\right)\right)}{n} \\ &\leq \frac{\Psi_{\mathrm{P}}\text{-dim}\left(\mathcal{F}\right)}{n} \ .\end{aligned}$$

7.2. Towards a bound in terms of the Graph dimension

In addition to Theorem 54 the following analogous density bound is possible (implying the analogous mistake bound), but is in terms of the Ψ_G-dim instead of the Pollard pseudo-dimension. The result holds for the special case of all $k \in \mathbb{N}$ and $n = 2$. A general bound of this type would allow more direct comparison with the PAC-based result of Theorem 7.

Lemma 58 *For any $k \in \mathbb{N}$ and family $V \subseteq \{0, \ldots, k\}^2$, $\operatorname{dens}(\mathcal{G}(V)) \leq \Psi_{\mathrm{G}}\text{-dim}(V)$.*

Proof. Fix $n = 2$ and $k \in \mathbb{N}$. We will show that for each $V \subseteq \{0, \ldots, k\}^n$ there exists a translation vector $\boldsymbol{\psi} \in \Psi_G^n$ such that $\operatorname{dens}(\mathcal{G}(V)) \leq \operatorname{dens}(\mathcal{G}(\boldsymbol{\psi}(V)))$ which by Lemma 2.9 is in turn bounded above by $\mathrm{VC}(\boldsymbol{\psi}(V)) \leq \Psi_{\mathrm{G}}\text{-dim}(V)$.

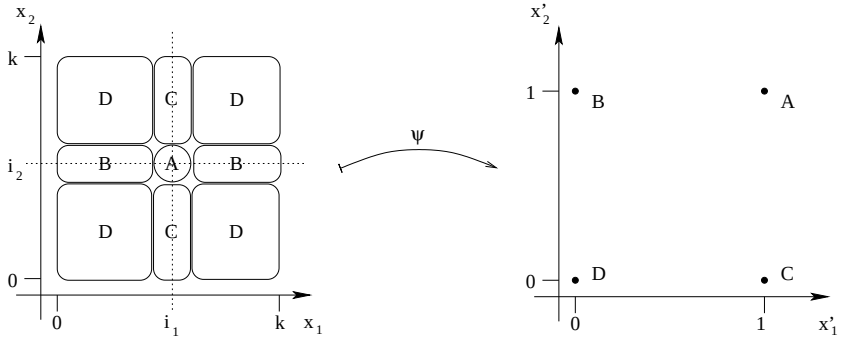

Fig. 10. The left-hand figure shows the pre-images for each of the possible elements of the image of $V \subseteq \{0, \ldots, k\}^2$ under some translation induced by a pair $(i_1, i_2) \in \{0, \ldots, k\}^2$.

We use translations $\boldsymbol{\psi} \in \Psi_G^n$ and the thresholding indices that induce them $(i_1, \ldots, i_n) \in \{0, \ldots, k\}^n$ interchangeably, as described in Example 2.6.

Let $\hat{\boldsymbol{\psi}} \in \Psi_G^2$, with its equivalent representation $(\hat{i_1}, \hat{i_2}) \in \{0, \ldots, k\}^2$, produce a maximally dense translation $\hat{\boldsymbol{\psi}} \in \arg\max_{\boldsymbol{\psi} \in \Psi_G^2} \operatorname{dens}(\mathcal{G}(\boldsymbol{\psi}(V)))$. At least one such translation must exist as $|\Psi_G^2| = (k+1)^2 < \infty$. We split on the density of the one-inclusion graph of the translated $T = \hat{\boldsymbol{\psi}}(V)$ (see Table 1), using the notation of Figure 10 for referring to the elements of T: $A = (1,1), B = (0,1), C = (1,0), D = (0,0)$.

Table 1
Enumeration of the possible densities of the translated one-inclusion graph.

$\operatorname{dens}(\mathcal{G}(T))$	T
0	$\phi, \{A\}, \{B\}, \{C\}, \{D\}, \{A, D\}, \{B, C\}$
$\frac{1}{2}$	$\{A, B\}, \{A, C\}, \{D, B\}, \{D, C\}$
$\frac{2}{3}$	$\{A, B, C\}, \{A, B, D\}, \{A, C, D\}, \{B, C, D\}$
1	$\{A, B, C, D\}$

Suppose that $\operatorname{dens}(\mathcal{G}(T)) = 0$. Assume that $\frac{|E|}{|V|} > 0$. Then $|E| \geq 1$, and we know that there is a row i (column j) hyperedge which adjoins at least two vertices along that row (column). This is a contradiction, as we could have positioned $(\hat{i_1}, \hat{i_2})$ over either of these vertices to get $\{A, B\} \subseteq T$ ($\{A, C\} \subseteq T$) and as a consequence $\operatorname{dens}(\mathcal{G}(T)) \geq 0.5$.

Suppose that $\operatorname{dens}(\mathcal{G}(T)) = \frac{1}{2}$ and assume that $\frac{|E|}{|V|} > \frac{1}{2}$. Note that for any non-empty hypergraph (V, E), $2\frac{|E|}{|V|} \leq \frac{1}{|V|} \sum_{v \in V} \deg(v)$. Thus at least one vertex in V must have degree 2 or more. This contradicts our assumption, as it implies that we could have positioned $(\hat{i_1}, \hat{i_2})$ over this vertex to have $\{A, B, C\} \subseteq T$ which would imply $\operatorname{dens}(\mathcal{G}(T)) \geq \frac{2}{3}$.

Suppose that $\mathrm{dens}(\mathcal{G}(T)) = \frac{2}{3}$ and assume that $\frac{|E|}{|V|} > \frac{2}{3}$. Again there must be at least one vertex in V of degree at least 2. Assume there was just one such vertex, then counting $|E| \leq \frac{|V|+1}{2} \leq \frac{2|V|}{3}$ provided $|V| \geq 3$ which is the case by the assumed density on V. This is a contradiction, so there must be at least two vertices of degree 2 or more, in V. But then we could have placed $(\hat{i_1}, \hat{i_2})$ over one of these, to get the full cube $\{A, B, C, D\} = T$.

Finally note that $\mathrm{dens}(\mathcal{G}(V)) \leq 1$ always holds, so combining cases we have proven that for $n = 2$ $\frac{|E|}{|V|} \leq \mathrm{dens}(\mathcal{G}(T))$. ■

7.3. A general lower bound

We now show that the general multiclass mistake bound of Theorem 54 is optimal to within a $O(\log k)$ factor, noting that Ψ_N is smaller than Ψ_P by at most such a factor [2, Theorem 10].

Definition 59 *We call a family $\mathcal{F} \subseteq \{0, \ldots, k\}^{\mathcal{X}}$* trivial *if either $|\mathcal{F}| = 1$ or there exist no $x_1, x_2 \in \mathcal{X}$ and $f_1, f_2 \in \mathcal{F}$ such that $f_1(x_1) \neq f_2(x_1)$ and $f_1(x_2) = f_2(x_2)$.*

Theorem 60 *Consider any deterministic or randomized prediction strategy Q and any $\mathcal{F} \subseteq \{0, \ldots, k\}^{\mathcal{X}}$ that has $2 \leq \Psi_{\mathrm{N}}\text{-dim}(\mathcal{F}) < \infty$ or is non-trivial with $\Psi_{\mathrm{N}}\text{-dim}(\mathcal{F}) < 2$. Then for all $n > \Psi_{\mathrm{N}}\text{-dim}(\mathcal{F})$, $\hat{M}_{Q,\mathcal{F}}(n) \geq \max\{1, \Psi_{\mathrm{N}}\text{-dim}(\mathcal{F}) - 1\}/(2en)$.*

Proof. Following [6], we use the probabilistic method to prove the existence of a target in $\mathcal{F}$ for which prediction under a distribution P supported by a Ψ_N-shattered subset is hard. Consider $d = \Psi_{\mathrm{N}}\text{-dim}(\mathcal{F}) \geq 2$ with $n > d$. Fix a $\mathcal{Z} = \{z_1, \ldots, z_d\}$ Ψ_N-shattered by $\mathcal{F}$ and then a subset $\mathcal{F}_{\mathcal{Z}} \subseteq \mathcal{F}$ of 2^d functions that Ψ_N-shatters $\mathcal{Z}$. Define a distribution P on $\mathcal{X}$ by $P(\{z_i\}) = n^{-1}$ for each $i \in [d-1]$, $P(\{z_d\}) = 1-(d-1)n^{-1}$ and $P(\{x\}) = 0$ for all $x \in \mathcal{X}\backslash\mathcal{Z}$. Observe that

$$\begin{aligned}
\Pr_{P^n}(\forall i \in [n-1], X_n \neq X_i) &\geq \Pr_{P^n}(X_n \neq z_d, \forall i \in [n-1], X_n \neq X_i) \\
&= \frac{d-1}{n}\left(1 - \frac{1}{n}\right)^{n-1} \\
&\geq \frac{d-1}{en} \ .
\end{aligned}$$

For any $f \in \mathcal{F}_{\mathcal{Z}}$ and $\boldsymbol{x} \in \mathcal{Z}^n$ with $x_n \neq x_i$ for all $i \in [n-1]$, exactly half of the functions in $\mathcal{F}_{\mathcal{Z}}$ consistent with $\mathrm{sam}((x_1, \ldots, x_{n-1}), f)$ output some $i \in \{0, \ldots, k\}$ on x_n and the remaining half output some $j \in \{0, \ldots, k\}\backslash\{i\}$. Thus $\mathbb{E}_{\mathrm{Unif}(\mathcal{F}_{\mathcal{Z}})}[\mathbf{1}[Q(\mathrm{sam}((x_1, \ldots, x_{n-1}, F), x_n) \neq F(x_n)]] = 0.5$ for such an $\boldsymbol{x}$ and so

$$\hat{M}_{Q,\mathcal{F}} \ \geq \ \hat{M}_{Q,\mathcal{F}_{\mathcal{Z}}} \ \geq \ \mathbb{E}_{\mathrm{Unif}(\mathcal{F}_{\mathcal{Z}}) \times P^n}[\mathbf{1}[Q(\mathrm{sam}((X_1, \ldots, X_{n-1}, F), X_n) \neq F(X_n)]] \ \geq \ \frac{d-1}{2en} \ .$$

The similar case of $d < 2$ is omitted here and shows that there is a distribution P on $\mathcal{X}$ and function $f \in \mathcal{F}$ such that $\mathbb{E}_{P^n}[\mathbf{1}[Q(\mathrm{sam}((X_1, \ldots, X_{n-1}), f), X_n) \neq f(X_n)]] \geq (2en)^{-1}$. ■

8. Conclusions and open problems

In this paper we have developed new shifting machinery and tightened the binary one-inclusion mistake bound from d/n to D_n^d/n ($\lceil D_n^d \rceil/n$ for the deterministic strategy). This was made possible through a symmetrization density bound, a result recently conjectured by Kuzmin and Warmuth [17]. We have described the k-class generalization of the prediction learning model and derived a mistake bound for the multiclass one-inclusion prediction strategy that improves on previous PAC-based expected risk bounds by $O(\log n)$ and that is within $O(\log k)$ of optimal. We also presented several characterizations and properties of one-inclusion graphs and their vertex-sets: a colorability characterization of one-inclusion isomorphic graphs,

the complementary characterizations of maximum (due to Floyd [8]) and maximal classes, and the algebraic topological property of maximum classes that d-maximum classes are d-contractible simplicial complexes. Finally we settled the minimum degree conjecture of Kuzmin and Warmuth [17] as being false, and introduced the uniform VC-degree ratio κ as a measure of how greatly a subset's dimension and minimum degree can differ.

Here shifting with invariance to the shattering of a single set was described, however we are aware of invariance to more complex shatterings. The symmetrization method of Theorem 24 can be extended over subgroups $G \subset S_n$ to gain tighter density bounds.

In addition to the general multiclass mistake bound of $\Psi_P\text{-dim}(\mathcal{F})/n$ (Theorem 54), Lemma 58 provides the analogous bound in terms of the Graph dimension for all $k \in \mathbb{N}$ but only the special case of $n = 2$. It is open as to whether this result generalizes to $n \in \mathbb{N}$. While a general Ψ_G-based bound would allow direct comparison with the PAC-based expected risk bound, it should also be noted that Ψ_P and Ψ_G are in fact incomparable—neither $\Psi_G \leq \Psi_P$ nor $\Psi_P \leq \Psi_G$ singly holds for all classes [2, Theorem 1].

While Theorem 24 resolves the conjectured density bound of Kuzmin and Warmuth [17], the remainder of the conjectured correctness proof for the Peeling compression scheme (and also the less refined minimum degree conjecture) is shown to be false. A consequence of a proof of correctness for d-peeling maximum classes of VC-dimension d would be an impossibility result for generally embedding maximal classes in maximum classes with only a constant additive increase in VC-dimension.

Acknowledgments

We gratefully acknowledge the support of the NSF under award DMS-0434383.

References

[1] N. Alon, D. Haussler, E. Welzl, Partitioning and geometric embedding of range spaces of finite Vapnik-Chervonenkis dimension. *Proceedings of the 3rd Annual ACM Symposium on Computational Geometry*, 1987, pp. 331–340.

[2] S. Ben-David, N. Cesa-Bianchi, D. Haussler, P.M. Long, Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, **50**(1) (1995) 74–86.

[3] S. Ben-David, A. Litman, Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, **86**(1) (1998) 3–25.

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association of Computing Machinery*, **36**(4) (1989) 929–965.

[5] R.M. Dudley, The structure of some Vapnik-Chervonenkis classes. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman*, Vol. II (Le Cam, L. M., Olshen, R. A., eds.), Wadsworth, New York, 1985, pp. 495–507.

[6] A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant, A general lower bound on the number of examples needed for learning. *Information and Computation*, **82**(3) (1989) 247–261.

[7] F. Firsov, On isometric embeddings of a graph into a boolean cube. *Cybernetics* **1** (1965) 112–113.

[8] S. Floyd, *On space-bounded learning and the Vapnik-Chervonenkis dimension*. PhD thesis, Technical report TR-89-061, International Computer Science Institute, Berkeley, CA, 1989.

[9] S. Floyd, M.K. Warmuth, Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, **21**(3) (1995) 269–304.

[10] B. Gärtner, E. Welzl, Vapnik-Chervonenkis Dimension and (Pseudo-) Hyperplane Arrangements. *Discrete & Computational Geometry*, **12** (1994) 399–432.

[11] P. Hall, On representatives of subsets. *Journal of the London Mathematical Society*, **10** (1935) 26–30.

[12] D. Haussler, Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory (A)* **69**(2) (1995) 217–232.

[13] D. Haussler, N. Littlestone, M.K. Warmuth, Predicting $\{0,1\}$ functions on randomly drawn points. *Information and Computation*, **115**(2) (1994) 284–293.

[14] I. Havel, J. Morávek, B-valuations of graphs. *Czech. Math. J.* **22** (1972) 338–351.

[15] J. Hlavička, Race-free assignment in asynchronous switching circuits. *Information Processing Machines* **13** (1967).

[16] D. Kuzmin, M.K. Warmuth, Unlabeled compression schemes for maximum classes. In Auer, P., Meir, R. (eds.) *Proceedings of the 18th Annual Conference on Learning Theory* **3559** of *Lecture Notes in Computer Science*, 2005, pp. 591–605.

[17] D. Kuzmin, M.K. Warmuth, Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 2006, to appear.
[18] Y. Li, P.M. Long, A. Srinivasan, The one-inclusion graph algorithm is near optimal for the prediction model of learning. *IEEE Transactions on Information Theory*, **47**(3) (2002) 1257–1261.
[19] N. Littlestone, M.K. Warmuth, Relating data compression and learnability. Unpublished manuscript, `http://www.cse.ucsc.edu/~manfred/pubs/lrnk-olivier.pdf`, 1986.
[20] M. Livingston, Q.F. Stout, Embeddings in Hypercubes. *Mathematical and Computational Modelling*, **11** (1988) 222–227.
[21] B.I.P. Rubinstein, P.L. Bartlett, J.H. Rubinstein, Shifting, One-Inclusion Mistake Bounds and Tight Multiclass Expected Risk Bounds, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19*, MIT Press, Cambridge, MA, 2007.
[22] N. Sauer, On the density of families of sets. *Journal of Combinatorial Theory (A)*, **13** (1972) 145–147.
[23] M. Warmuth, Sample Compression, Learnability, and the Vapnik-Chervonenkis Dimension. *Computational Learning Theory: Third European Conference, EuroCOLT '97*, in *Lecture Notes in Computer Science*, Springer, 1997.
[24] M.K. Warmuth, Compressing to VC Dimension Many Points. *Proceedings of the 16th Annual Conference on Learning Theory (COLT03)*, open problem paper, 2003.
[25] E. Welzl, Complete range spaces. Unpublished notes, 1987.

Appendix A. Proof of Theorem 25

The proof corresponds exactly to the proof of Theorem 10 [13, Theorem 2.3], using the symmetrization graph density bounded of Theorem 24 in place of the original density bound of Lemma 16 [13, Lemma 2.4]. We provide a high-level sketch of how the results are chained in [13]. The proof of Theorem 54 contains these results, generalized, in full detail.

For the deterministic strategy, a simple argument [13, Theorem 2.3] shows that this worst-case average over permutations is at most the supremum over $\boldsymbol{x} \in \mathcal{X}^n$ of the maximum outdegree of (the oriented) $\mathcal{G}\left(\Pi_{\boldsymbol{x}}\left(\mathcal{F}\right)\right)$, over n. The essential ingredients are that the strategy makes a mistake iff the correct vertex in the projected graph (e.g. the vertex corresponding to $(f(x_1), \ldots, f(x_n))$) has an out-going edge in the n^{th} direction—or that under permutation σ of the n-sample there is such an edge in the $\sigma(n)^{\text{th}}$ direction. Secondly x_i appears last in the sample in n^{-1} of the $n!$ permutations of the sample. Either the network flow construction of [13] or the application of Hall's Theorem [11] of [12] then show that $\mathcal{G}\left(V\right)$ can be oriented so that its maximum outdegree is at most $\lceil \text{maxdens}\left(\mathcal{G}\left(V\right)\right) \rceil$ where $\text{maxdens}\left(G\right)$ denotes the maximum density of all subgraphs of (hyper)graph G. Theorem 24 then bounds the density of all subgraphs of V by D_n^d, as each has VC-dimension at most $\text{VC}(V)$.

The randomized strategy follows roughly the same argument. In place of edge-orientation the goal is to assign a distribution on each edge—a probability on each of the two adjoining vertices. The same argument that upper-bounds $\hat{\hat{M}}(n)$ for the deterministic strategy, produces an upper-bound for the randomized strategy in terms of the sum of the out-going probabilities from a vertex, over all vertices. The network flow construction assigns probabilities so that each vertex's total probability is at most the maximum subgraph density. Again, Theorem 24 implies that this is in turn at most D_n^d.

Lemma 2 [13, Corollary 2.1] finally leads to the mistake bounds for both cases.