VC-dimensions of random function classes

Bernard Ycart, Joel Ratsaby

Discrete Mathematics and Theoretical Computer Science DMTCS vol. **10:1**, 2008, 113–128

VC-dimensions of random function classes

Bernard Ycart[1] and Joel Ratsaby[2]

[1]*LJK, CNRS UMR 5224, BP 53, 38041 Grenoble cedex 9, France,* ycart@imag.fr
[2]*Department of Electrical and Electronic Engineering Ariel University Center, Ariel 40700, Israel,* ratsaby@ariel.ac.il

received February 20, 2006, revised December 11, 2007, accepted January 1, 2008.

For any class of binary functions on $[n] = \{1, \dots, n\}$ a classical result by Sauer states a sufficient condition for its VC-dimension to be at least d: its cardinality should be at least $O(n^{d-1})$. A necessary condition is that its cardinality be at least 2^d (which is $O(1)$ with respect to n). How does the size of a 'typical' class of VC-dimension d compare to these two extreme thresholds ? To answer this, we consider classes generated randomly by two methods, repeated biased coin flips on the n-dimensional hypercube or uniform sampling over the space of all possible classes of cardinality k on $[n]$. As it turns out, the typical behavior of such classes is much more similar to the necessary condition; the cardinality k need only be larger than a threshold of 2^d for its VC-dimension to be at least d with high probability. If its expected size is greater than a threshold of $O(\log n)$ (which is still significantly smaller than the sufficient size of $O(n^{d-1})$) then it shatters every set of size d with high probability. The behavior in the neighborhood of these thresholds is described by the asymptotic probability distribution of the VC-dimension and of the largest d such that all sets of size d are shattered.

Keywords: Random binary functions, Vapnik-Chervonenkis dimension, Poisson approximation

1 Introduction

Let n and d be two integers such that $1 \leqslant d \leqslant n$ and denote by $[n]$ the set $\{1, \dots, n\}$. A class of binary functions is a subset of $\{0,1\}^{[n]}$. As only binary functions are considered we refer to them simply as functions. They will also be viewed as binary vectors: $f = (f(1), \dots, f(n))$. Let $\mathbf{x} = \{x_1, \dots, x_d\}$ be a subset of $[n]$ and $F \subset \{0,1\}^{[n]}$ be a class of functions. For any function $f \in F$, denote by $f_{|\mathbf{x}}$ its restriction to $\mathbf{x}$, *i.e.*, $f_{|\mathbf{x}} = (f(x_1), \dots, f(x_d))$. The class F is said to *shatter* $\mathbf{x}$ if:

$$\left|\{f_{|\mathbf{x}} : f \in F\}\right| = 2^d,$$

where $|\cdot|$ denotes the cardinality of a finite set. The *Vapnik-Chervonenkis dimension* of F, denoted as $\mathrm{VC}(F)$, is defined as the size of the largest set $\mathbf{x}$ shattered by F. The following Sauer-Shelah lemma (Sauer (1972); Shelah (1972); Vapnik and Chervonenkis (1971)) is a fundamental result relating the VC-dimension of a class of functions to its cardinality.

Lemma 1 *Let F be a class of functions on $[n]$ with*

$$|F| > \sum_{i=0}^{d-1} \binom{n}{i}. \tag{1}$$

Then F shatters at least one set $\mathbf{x} \subset [n]$ of cardinality $|\mathbf{x}| = d$.

An interesting extension (see Theorem 1 in Frankl (1983)) states that such a threshold (which holds for any class F) arises due to the simple fact that any *ideal* class F_0 of this size must shatter some set of size d. More generally, the lemma holds for classes on infinite domains X where instead of $|F|$ one has $\max_{Y \subset X} |F_{|Y}|$ with Y running over all finite subsets such that $|Y| \geqslant d$. Aside of being an interesting combinatorial result in set theory (Chapter 17 in Bollobás (1986)), Lemma 1 has been extended in various directions notably by Frankl (1983); Haussler and Long (1995); Alon et al. (1997); Anstee et al. (2005) and found applications in numerous fields such as combinatorial geometry (Pach and Agarwal (1995); Matoušek (1998)), graph theory (Haussler and Welzl (1987); Anthony et al. (1995a)), empirical processes (Pollard (1984)) and statistical learning theory (Haussler (1992); Vapnik (1998)).

The VC-dimension has numerous extensions, for instance, the pseudo-dimension for real-valued function classes, the scale-sensitive (or fat-shattering) dimension which characterizes the so-called Glivenko-Cantelli classes (Alon et al. (1997)), and the *testing dimension* (Romanik and Smith (1994); Anthony et al. (1995b)) of F, denoted as $\mathrm{TD}(F)$, which is defined as the maximal integer d such that *all* sets of size d are shattered by F. For other related dimensions see Haussler and Long (1995) and Anthony and Bartlett (1999).

Observe that as n tends to infinity and d remains fixed, the right hand side of (1) is of order $O(n^{d-1})$. Thus $O(n^{d-1})$ is a threshold point that dictates a sufficient cardinality for F to shatter at least one set of size d. As we shall show, it is not a necessary condition since it typically takes a class only of size $O(\log n)$ to shatter *all* sets of size d. In order to show this, our primary aim in this paper is to investigate the size of sets that are shattered by a *random* class F of functions (a random element of the power set $\mathcal{P}(\{0,1\}^{[n]})$).

We consider two natural approaches: fixing the size to be k and drawing a class F with equal probability from all classes of size k (uniform model, Definition 2), or drawing each individual function from $\mathcal{P}(\{0,1\}^{[n]})$ with an equal probability p in 2^n random trials (binomial model, Definition 1). We state several results on the asymptotic behavior of the size of shattered sets with an explicit dependence on k or p as n tends to infinity.

As a preview of our results, let us sketch the evolution of the size of sets shattered by a random class F under the uniform model with increasing $k = k_n$. Initially, when k is fixed, sets of size d are shattered only if $k \geqslant 2^d$. It turns out that at least one such set is shattered with high probability (*w.h.p.*, *i.e.* tending to 1 as n tends to infinity). As soon as k_n starts to increase to infinity there are shattered sets of *any* size and, moreover, any fixed set of size d (in particular the set $[d]$) is shattered *w.h.p.* If the speed at which k_n grows is sufficiently slow, $k_n \ll \log n$, then regardless of the value of d, there exists at least one set of size d which is *not* shattered (here $a_n \ll b_n$ denotes $\lim a_n/b_n = 0$). For $k_n = \alpha \log n + O(1)$, there exists a finite d such that *all* sets of size d are shattered and at least one set of size $d+1$ which is not shattered. Finally, for $k_n \gg \log n$ and any fixed d, *all* sets of size d are shattered. A similar behavior holds for the binomial model, where k_n is now replaced by $2^n p_n$, *i.e.*, the expected size of the random class F.

The results are stated in details in Sections 2 – 5 where three kinds of events are studied: (i) shattering at least one set of size d, (ii) shattering a given set $\mathbf{x}$ of size d, (iii) shattering all sets of size d. Clearly, the

VC-dimension and the testing-dimension of a random class F are related to (i) and (iii) respectively. In order to have a more complete comparison we introduce between them an intermediate dimension which is related to (ii) and is defined as follows.

Let F be a class of functions from $[n]$ to $\{0, 1\}$. The *initial-dimension* of F, denoted as $\mathrm{ID}(F)$, is the maximal integer d such that the set $[d]$ is shattered by F.

Clearly, the three dimensions are related by

$$\mathrm{VC}(F) \geqslant \mathrm{ID}(F) \geqslant \mathrm{TD}(F). \tag{2}$$

In subsequent sections we obtain the asymptotic distribution of $\mathrm{VC}(F)$ and $\mathrm{ID}(F)$ under the binomial and uniform models where the (expected) number of functions in the class is fixed (Propositions 1 and 2). The asymptotics of $\mathrm{TD}(F)$ turns out to be much sharper as for $k_n = c \log n$ its distribution is concentrated on one or two values *w.h.p.* (Proposition 3). This has striking similarities with the well known result on the concentration of the clique number for random graphs (see Chap. 11 of Bollobás (2001)).

In our analysis we use standard techniques from discrete probability and the theory of random graphs (available in Janson et al. (2000)). In Ycart and Ratsaby (2007), we applied these to study the VC-dimension of a random class of functions with a fixed number of ones, *i.e.* random hypergraphs.

The remainder of the paper is organized as follows: Section 2 defines the shattering events, the two probability models and describes their interdependence. Sections 3, 4 and 5 study the asymptotics of the VC-dimension, the initial-dimension and the testing-dimension respectively.

2 Probability models and events

We start by defining the events under study. They are associated with the shattering of sets by a random class of functions, *i.e.* a random variable with values in the power set $\mathcal{P}(\{0,1\}^{[n]})$. For clarity, we defer the precise definition of probability distributions for random classes of functions to Section 2.2.

2.1 Events of interest

Let $n \geqslant 1$, $1 \leqslant d \leqslant n$ and $\mathbf{x} = \{x_1, \dots, x_d\} \subseteq [n]$. Let $\eta \in \{0,1\}^{\mathbf{x}}$ denote a function from $\mathbf{x}$ to $\{0, 1\}$. The events that we are interested in are the subsets of $\mathcal{P}(\{0,1\}^n)$ defined as follows:

$$\begin{aligned}
C_{\mathbf{x},\eta} &:= \{\, F \text{ containing a function } f \text{ such that } f_{|\mathbf{x}} = \eta \,\} \\
S_{\mathbf{x}} &:= \{\, F \text{ shattering } \mathbf{x} \,\} \\
E_d &:= \{\, F \text{ such that } \mathrm{VC}(F) \geqslant d \,\} \\
A_d &:= \{\, F \text{ such that } \mathrm{TD}(F) \geqslant d \,\} \,.
\end{aligned}$$

The event $S_{[d]}$ is the set of classes that shatter $[d]$. It is thus equal to the event which is defined as follows:

$$S_d \;:=\; \{\, F \text{ such that } \mathrm{ID}(F) \geqslant d \,\} \,.$$

The main goal of the paper is to evaluate the probabilities of E_d, S_d, and A_d. The events E_d and A_d can be expressed in terms of the $S_{\mathbf{x}}$'s, for $\mathbf{x} \subset [n]$ and $|\mathbf{x}| = d$. These may in turn be expressed in terms of the $C_{\mathbf{x},\eta}$'s, for $\eta \in \{0,1\}^{\mathbf{x}}$. By definition, the class $\mathcal{F}$ shatters $\mathbf{x}$ if and only if for every $\eta \in \{0,1\}^{\mathbf{x}}$ there exists a function f whose restriction to $\mathbf{x}$ is η. Thus $S_{\mathbf{x}}$ is the intersection of the events $C_{\mathbf{x},\eta}$, $\eta \in \{0,1\}^{\mathbf{x}}$, *i.e.*,

$$S_{\mathbf{x}} = \bigcap_{\eta \in \{0,1\}^{\mathbf{x}}} C_{\mathbf{x},\eta} \ . \tag{3}$$

The VC-dimension of $\mathcal{F}$ is at least d if and only if there exists a set $\mathbf{x}$ with cardinality d which is shattered by $\mathcal{F}$. Therefore E_d is the union of the events $S_{\mathbf{x}}$ over all $\mathbf{x} \subset [n]$ such that $|\mathbf{x}| = d$, *i.e.*,

$$E_d = \bigcup_{\mathbf{x} \subset [n], |\mathbf{x}|=d} S_{\mathbf{x}} \ . \tag{4}$$

Finally, the testing-dimension of $\mathcal{F}$ is at least d if and only if all sets $\mathbf{x}$ of cardinality d are shattered by $\mathcal{F}$. Therefore A_d is the intersection of all the events $S_{\mathbf{x}}$, $\mathbf{x} \subset [n]$ with $|\mathbf{x}| = d$, *i.e.*,

$$A_d = \bigcap_{\mathbf{x} \subset [n], |\mathbf{x}|=d} S_{\mathbf{x}} \ . \tag{5}$$

2.2 Probability models

In this subsection we describe the underlying probability models with which a random class is generated. In addition to mathematical definitions, we also discuss how such generation may be carried out in practice. When stating probability quantities we will either refer to the probability law of a random variable, *e.g.*, $P_{n,p}$, $P^*_{n,k}$ or simply write $\mathbb{P}$ for the underlying probability distribution of the space on which the random variables are defined.

The two probability models by which a random class $\mathcal{F}$ of binary functions on $[n]$ will be created are denoted by 'binomial' and 'uniform'. In the binomial model, a probability parameter $0 \leqslant p \leqslant 1$ is defined and the random class $\mathcal{F}$ is constructed through 2^n independent coin tossings, one for each function in $\{0,1\}^{[n]}$, with a probability of success (*i.e.* selecting a function into $\mathcal{F}$), equal to p.

Definition 1 (Binomial model) *Let n be a positive integer. Let p be a real such that $0 \leqslant p \leqslant 1$. We call* binomial class *with parameters n and p a random class of functions containing any given function with probability p, independently of the others. We shall denote by $\mathcal{F}_{n,p}$ a binomial class with parameters n and p and by $P_{n,p}$ its probability distribution. Thus if F is any element of $\mathcal{P}(\{0,1\}^{[n]})$,*

$$\mathbb{P}(\mathcal{F}_{n,p} = F) = P_{n,p}(F) = p^{|F|}(1-p)^{2^n - |F|} \ . \tag{6}$$

An alternate way to construct a random class $\mathcal{F}$ is to first choose its cardinality k, $0 \leqslant k \leqslant 2^n$, and then select a class by a uniform random drawing from the family $\mathcal{F}^{(k)} \subset \mathcal{P}(\{0,1\}^{[n]})$ of all subsets of $\{0,1\}^{[n]}$ having k elements.

Definition 2 (Uniform model) *Let n be a positive integer. Let k be an integer such that $1 \leqslant k \leqslant 2^n$. We call* uniform class *with parameters n and k a random class with uniform distribution over all classes of k functions. We shall denote by $\mathcal{F}^*_{n,k}$ a uniform class with parameters n and k and by $P^*_{n,k}$ its probability distribution. Thus if F is any element of $\mathcal{P}(\{0,1\}^{[n]})$,*

$$\mathbb{P}(\mathcal{F}^*_{n,k} = F) = P^*_{n,k}(F) = \begin{cases} 1/\binom{2^n}{k} & \textit{if} \quad |F| = k \\ 0 & \textit{if} \quad |F| \neq k \ . \end{cases} \tag{7}$$

These models are the most natural ways to define probability distributions on the set $\mathcal{P}(\{0,1\}^{[n]})$ of all classes of functions on $[n]$. They match the two basic models of random graphs (see Chap. 1 of Janson et al. (2000)).

The denomination 'binomial model' comes from the fact that the total number of functions in a binomial class follows a binomial distribution as does the total number of edges in a binomial graph. More precisely, if we denote by K the total number of functions in a binomial class, it is immediate from (6) that

$$\forall k = 0, \ldots, 2^n, \quad P_{n,p}(K = k) = \binom{2^n}{k} p^k (1-p)^{2^n - k} . \tag{8}$$

The two models are obviously related. Indeed, the conditional distribution of a binomial class, conditioned on having a cardinality k, is that of a uniform class, *i.e.*, for any $B \subseteq \mathcal{P}(\{0,1\}^{[n]})$,

$$P_{n,p}(B \,|\, K = k) = P^*_{n,k}(B) . \tag{9}$$

Conversely, knowing the values of $P^*_{n,k}(B)$ for all k, one can compute $P_{n,p}(B)$ using the formula of total probabilities:

$$P_{n,p}(B) = \sum_{k=0}^{2^n} P^*_{n,k}(B)\, P_{n,p}(K = k) . \tag{10}$$

In practice, one can construct a binomial class by first selecting its cardinality K according to the binomial distribution (8) and then choose a uniform class of this size. The expected number of functions in a binomial class is clearly $\mathbb{E}(K) = p2^n$. When n is large, we expect both models to have the same behavior provided that $k \sim p2^n$. This intuition is partially justified by the theoretical results presented in Section 2.1 of Bollobás (2001) or Section 1.4 of Janson et al. (2000). Some minor discrepancies between both models will be pointed out in Sections 3 and 5.

One practical approach to construct a uniform class of cardinality k is to construct a random $n \times k$ binary matrix with the nk entries taking values 0 or 1 independently with probability $1/2$. Denoting by $Q^*_{n,k}$ the corresponding probability measure, then for any $M \in \mathcal{M}_{n\times k}(\{0,1\})$,

$$Q^*_{n,k}(M) = \frac{1}{2^{nk}}. \tag{11}$$

Clearly, the columns of a binary matrix M are vectors of length n which are binary functions on $[n]$. Hence the set of columns of M represents a class of binary functions. It contains k elements if and only if all columns are distinct, or less than k elements if two columns are the same. A binary matrix whose columns are distinct is called *simple* (see *e.g.* Anstee et al. (2005)), and we shall denote by S their set. We claim that the conditional distribution of the set of columns of a random binary matrix, knowing that it belongs to S, is the uniform distribution $P^*_{n,k}$. To see this, observe that the probability for a random binary matrix to be simple is

$$Q^*_{n,k}(S) = \frac{2^n(2^n-1)\cdots(2^n-k+1)}{2^{nk}} . \tag{12}$$

For any fixed class F of k binary functions there are $k!$ corresponding simple matrices in $\mathcal{M}_{n\times k}(\{0,1\})$. Therefore the conditional probability for the set of columns of a random matrix to be F is

$$Q^*_{n,k}(F \,|\, S) = \frac{k!}{2^{nk}} \frac{2^{nk}}{2^n(2^n-1)\cdots(2^n-k+1)} = \frac{1}{\binom{2^n}{k}} = P^*_{n,k}(F) . \tag{13}$$

Thus the process of independently drawing random binary matrices until a simple one is obtained, yields a uniform class of functions. In practice, if k is reasonably small compared to 2^n, then the probability of the conditioning event S is close to 1 so, typically, after one or two random matrix generations a uniform class F of cardinality k is obtained.

In Sections 3 and 5, we use this alternate representation of $P^*_{n,k}$ in order to compute the asymptotic probability of several types of shattering events (described in Section 2.1) as n tends to infinity, and for values of k_n such that $Q^*_{n,k_n}(S)$ tends to 1. From (12), it is easy to deduce that this is true provided that $k_n \ll 2^{n/2}$.

The next auxiliary lemma is a technical result which will permit the interchange of Q^*_{n,k_n} and P^*_{n,k_n} and hence simplify some of the analysis in subsequent sections.

Lemma 2 *Assume $k_n \ll 2^{n/2}$ and let $B \subset \mathcal{P}(\{0,1\}^{[n]})$. Then either $P^*_{n,k_n}(B)$ and $Q^*_{n,k_n}(B)$ both converge to the same limit or diverge.*

Proof: As seen above,

$$P^*_{n,k_n}(B) = Q^*_{n,k_n}(B \mid S) = \frac{Q^*_{n,k_n}(B \cap S)}{Q^*_{n,k_n}(S)}.$$

We have

$$Q^*_{n,k_n}(B) + Q^*_{n,k_n}(S) - 1 \leqslant Q^*_{n,k_n}(B \cap S) \leqslant Q^*_{n,k_n}(B).$$

Since $k \ll 2^{n/2}$ then $Q^*_{n,k_n}(S)$ tends to 1. Hence the limits of $Q^*_{n,k_n}(B)$, $Q^*_{n,k_n}(B \cap S)$ and $Q^*_{n,k_n}(B|S)$ are the same. □

In the next sections we present for each of the events E_d, S_d and A_d two types of results: the first type describes the values of the parameter (k_n or p_n) for which the probability of the event tends to 0 or 1 (these are presented as Lemmas, Corollaries and Remarks). The second type gives the behavior of the probability around the critical value of the parameter where the transition occurs (these are presented as Propositions).

3 Asymptotics for the VC-dimension

By definition of the VC-dimension, a class of fewer than 2^d functions cannot shatter a set of size d. As we show in this section, as soon as the number of functions is at least 2^d then at least one set of size d is shattered by a random class $\mathcal{F}$ *w.h.p.*

Lemma 3 *For any integer $d > 0$ let k be an integer satisfying $k \geqslant 2^d$. Then*

$$\lim_{n\to\infty} P^*_{n,k}(E_d) = 1.$$

Proof: It suffices to prove the result for $k = 2^d$ since $P^*_{n,k}(E_d) \geqslant P^*_{n,2^d}(E_d)$. Using Lemma 2 it suffices to show that $Q^*_{n,2^d}(E_d)$ tends to 1. Denote by M_d the 'complete' matrix with d rows and 2^d columns formed by all 2^d binary vectors of length d, ranked for instance in alphabetical order. The event E^*_d occurs if there exists an $\mathbf{x} = \{x_1, \ldots, x_d\} \subset [n]$ such that the submatrix whose rows and columns are indexed by $\mathbf{x}$ and $[2^d]$ respectively, is equal to M_d. Let $m = \lfloor n/d \rfloor$. Let

$$\mathbf{x}_i = \{di+1, di+2, \ldots, d(i+1)\}\,,\; 0 \leqslant i \leqslant m-1, \tag{14}$$

Consider the event defined as the set of those matrices having a submatrix whose rows are indexed by $\mathbf{x}_i$ and equal to M_d. For $i = 0, \ldots, m-1$, these m events are clearly independent and have the same probability 2^{-d2^d}. Hence the probability that at least one of them is fulfilled is

$$1 - (1 - 2^{-d2^d})^{\lfloor n/d \rfloor},$$

which tends to 1 as n increases. □

Remark 1 *When $k_n < 2^d$ no set of size d is shattered and hence $P^*_{n,k_n}(E_d) = 0$. For $k_n > 2^d$, $P^*_{n,k_n}(E_d)$ tends to 1. Hence for a uniform class, the critical value of the cardinality k_n for the event E_d is 2^d.*

Remark 2 *For any fixed $k > 0$, it follows that w.h.p. a uniform class $\mathcal{F}$ of cardinality k has a VC-dimension of at least $\lfloor \log_2 k \rfloor$, where $\lfloor \cdot \rfloor$ denotes the integer part and $\log_2(a)$ the logarithm in base 2.*

Since any class of cardinality k cannot shatter a subset of $[n]$ of size greater than $\lfloor \log_2 k \rfloor$ then this is also an upper bound on the VC-dimension of $\mathcal{F}$. Hence, under the uniform model $P^*_{n,k}$, the VC-dimension of $\mathcal{F}$ converges in probability to $\lfloor \log_2 k \rfloor$.

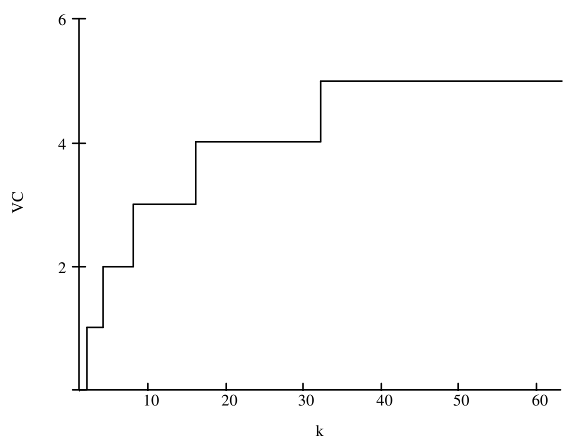

Fig. 1: VC-dimension of a random class of cardinality k ($n \to \infty$)

We proceed now to obtain the asymptotic distribution of VC($\mathcal{F}$) where $\mathcal{F}$ is a random class under the binomial model P_{n,p_n}. Since the number of functions in $\{0,1\}^{[n]}$ increases exponentially fast with n, in order to keep the expected cardinality of the random class a constant we choose a rate of decrease for p_n as $p_n = c2^{-n}$ for some $c > 0$. From Section 2, the number K of functions in a random class $\mathcal{F}_{n,p}$ follows the binomial law (8) with parameters 2^n and p_n. It therefore converges to the Poisson distribution with parameter c, *i.e.*,

$$\lim_{n\to\infty} P_{n,p_n}(K = k) = e^{-c}\frac{c^k}{k!} .$$

Conditioned on having a cardinality of $K = k$, by Remark 2, a random class $\mathcal{F}$ will have VC($\mathcal{F}$) $= \lfloor \log_2 k \rfloor$ *w.h.p.* Hence the event that the VC-dimension of the binomial class $\mathcal{F}_{n,p_n}$ is at least d is asymptotically distributed as $\lfloor \log_2 K \rfloor$, where K follows the Poisson distribution with parameter c. This is stated formally in the next result.

Proposition 1 *Fix any constant $c > 0$ and $d \geqslant 1$. Assume that p_n obeys $\lim_{n\to\infty} p_n 2^n = c$. Then*

$$\lim_{n\to\infty} P_{n,p_n}(E_d) = \sum_{k=2^d}^{\infty} e^{-c}\frac{c^k}{k!} \;.$$

Proof: From (10), we have

$$P_{n,p_n}(E_d) = \sum_{k=0}^{2^n} P^*_{n,k}(E_d)\binom{2^n}{k} p_n^k (1-p_n)^{2^n-k} \;.$$

Clearly, by definition of the VC-dimension, the factor $P^*_{n,k}(E_d) = 0$ equals zero for all $k < 2^d$. For $k \geqslant 2^d$, $P^*_{n,k}(E_d) \geqslant P^*_{n,2^d}(E_d)$ hence

$$P^*_{n,2^d}(E_d) \sum_{k=2^d}^{2^n} \binom{2^n}{k} p_n^k (1-p_n)^{2^n-k} \leqslant P_{n,p_n}(E_d) \leqslant \sum_{k=2^d}^{2^n} \binom{2^n}{k} p_n^k (1-p_n)^{2^n-k} \;.$$

The result follows since $P^*_{n,2^d}(E_d)$ tends to 1 (by Lemma 3) and the binomial distribution with parameters 2^n and p_n converges to the Poisson distribution with parameter c. □

While Lemma 3 established that a random class of functions of cardinality at least 2^d shatters one or more sets of size d *w.h.p.*, one can expect at least $O(n)$ sets of size d to be shattered. We now proceed to show that any *particular* set of size d has a positive probability of being shattered.

4 Asymptotics for the initial-dimension

Consider a random class $\mathcal{F}$ of k binary functions on $[n]$. Can a *fixed* set of size d in $[n]$, for instance $[d]$, be shattered by $\mathcal{F}$? The answer is affirmative, provided k is large enough. As it turns out, the initial-dimension of $\mathcal{F}$ has an asymptotic behavior which is similar to the distribution of its VC-dimension. We start with computing the probability of the event $S_{\mathbf{x}}$ (see (3)) under the binomial model $P_{n,p}$.

Lemma 4 *Let $n \geqslant 1$, $1 \leqslant d \leqslant n$ and $\mathbf{x} = \{x_1, \ldots, x_d\} \subset [n]$. Then under the binomial model with parameters n and $p \in [0,1]$ we have*

$$P_{n,p}(S_{\mathbf{x}}) = \left(1 - (1-p)^{2^{n-d}}\right)^{2^d} \;. \tag{15}$$

Proof: From (3), $S_{\mathbf{x}}$ is the intersection of the events $C_{\mathbf{x},\eta}$, $\eta \in \{0,1\}^{\mathbf{x}}$. We will prove that these events are independent under $P_{n,p}$ and that each has probability

$$P_{n,p}(C_{\mathbf{x},\eta}) = 1 - (1-p)^{2^{n-d}}. \tag{16}$$

We start with the latter. For an event B let us denote by $\overline{B}$ its complement

$$\overline{B} = \mathcal{P}(\{0,1\}^{[n]}) \setminus B.$$

We will use calligraphic capital letters to denote classes of functions and regular capital letters to denote events. Let $\mathcal{C}$ be a fixed class and $\mathcal{F}$ be a random class. Consider the event $C :=$ '$\mathcal{F}$ contains at least one element of $\mathcal{C}$'. Its complement $\overline{C}$ is: 'no function of $\mathcal{C}$ is contained in $\mathcal{F}$'. By Definition 1, its probability is

$$P_{n,p}(\overline{C}) = (1-p)^{|\mathcal{C}|}. \tag{17}$$

Let $\mathcal{C}_{\mathbf{x},\eta}$ be the class of functions that coincide with η on $\mathbf{x}$, *i.e.*, $f_{|\mathbf{x}} = \eta$ (there are 2^{n-d} such functions). Applying (17) to the event $\overline{C}_{\mathbf{x},\eta}$ yields (16).

It remains to show that the events $C_{\mathbf{x},\eta}$, $\eta \in \{0,1\}^{\mathbf{x}}$ are independent. It is a basic fact (see for instance Feller (1968), p. 115) that the independence of $(B_i)_{i\in I}$ is equivalent to the independence of their complements $(\overline{B_i})_{i\in I}$. For $1 \leqslant h \leqslant 2^d$, let $\eta_1, \ldots, \eta_h$ be distinct elements of $\{0,1\}^{\mathbf{x}}$. Consider the class $\mathcal{C} = \bigcup_{i=1,\ldots,h} \mathcal{C}_{\mathbf{x},\eta_i}$ whose cardinality is $|\mathcal{C}| = h2^{n-d}$. The event '$\mathcal{F} \cap \mathcal{C} = \emptyset$' means that for all functions f in $\mathcal{F}$, $f_{|\mathbf{x}} \neq \eta_i$, $1 \leqslant i \leqslant h$, or equivalently, the event $\bigcap_{i=1,\ldots,h} \overline{C}_{\mathbf{x},\eta_i}$ occurs. Resorting once more to (17) we have

$$P_{n,p}\left(\bigcap_{i=1,\ldots,h} \overline{C}_{\mathbf{x},\eta_i}\right) = (1-p)^{h2^{n-d}} = \prod_{i=1}^{h} P_{n,p}(\overline{C}_{\mathbf{x},\eta_i}).$$

From this it follows that the events $(\overline{C}_{\mathbf{x},\eta})_{\eta\in\{0,1\}^{\mathbf{x}}}$ are mutually independent, and hence, so are the events $(C_{\mathbf{x},\eta})_{\eta\in\{0,1\}^{\mathbf{x}}}$. □

From (15), and using the standard expansion

$$(1-t)^N = \exp\left(-N\sum_{i\geqslant 1}\frac{t^i}{i}\right) \tag{18}$$

which holds for $-1 < t < 1$ (see Feller (1968), p. 49), it follows that $P_{n,p_n}(S_{\mathbf{x}})$ tends to zero if $p_n \ll 2^{-n}$ and to 1 if $p_n \gg 2^{-n}$. As it was previously done for the event E_d (Lemma 3 and Remark 1), we now state the critical value of the expected cardinality for the event S_d under the binomial model.

Corollary 1 *The probability $P_{n,p_n}(S_d)$ that the initial-dimension is at least d tends to 0 or 1 according to whether the expected cardinality $p_n 2^n$ of the class tends to 0 or ∞, respectively. In particular, for* any *fixed $d > 0$, if the expected cardinality tends to infinity then the initial dimension is at least d w.h.p.*

The proof follows from Lemma 4 and the above arguments. When $p_n = c2^{-n}$ the initial-dimension converges in distribution according to the following result.

Proposition 2 *Fix any $c > 0$ and $d \geqslant 1$ and assume p_n satisfies $\lim_{n\to\infty} p_n 2^n = c$. Then*

$$\lim_{n\to\infty} P_{n,p_n}(S_d) = \left(1 - \exp(-c2^{-d})\right)^{2^d}.$$

Proof: From (18) we have $\lim_{N\to\infty}(1 - c/N)^N = e^{-c}$ hence

$$\lim_{n\to\infty}(1-p_n)^{2^{n-d}} = \exp(-c2^{-d}).$$

The result directly follows from Lemma 4. □

For $p_n = c2^{-n}$, it is interesting to compare the asymptotic probability distributions of the VC-dimension and the initial dimension, deduced from Propositions 1 and 2.

$$\mathbb{P}(\mathrm{VC}(\mathcal{F}) = d) \equiv P_{n,p_n}(E_d) - P_{n,p_n}(E_{d+1}) \quad \text{and} \quad \mathbb{P}(\mathrm{ID}(\mathcal{F}) = d) \equiv P_{n,p_n}(S_d) - P_{n,p_n}(S_{d+1}) \ .$$

They turn out to be remarkably close. Table 1 gives their significant values, for $c = 10$. The two distributions differ by approximately one unit.

d	1	2	3	4
$\mathbb{P}(\mathrm{VC}(\mathcal{F}) = d)$	.0098	.2099	.7310	.0487
$\mathbb{P}(\mathrm{ID}(\mathcal{F}) = d)$	.2766	.6428	.0672	.0000

Tab. 1: Asymptotic distributions of the VC and the initial-dimensions under the binomial model P_{n,p_n} for $p_n = c2^{-n}$ and $c = 10$.

Remark 3 *From Propositions 1 and 2, the asymptotic probability distributions of the events E_d and S_d have a similar functional form with respect to the expected cardinality c, but the former has a significantly earlier transition from zero to one. On Figure 2, the asymptotic probability of E_4 and S_4 are plotted against the expected cardinality of the class.*

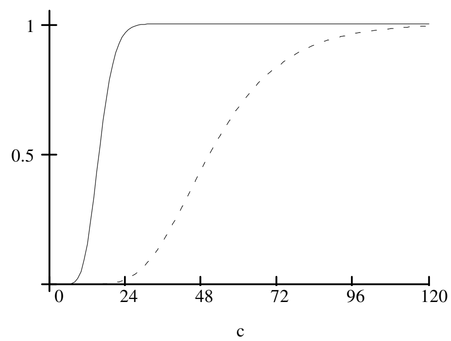

Fig. 2: Limiting probability distribution of the events E_d (solid) and S_d (dots) for $d = 4$, with respect to the expected cardinality c of the random class. Inflection points at $c = 15$ and $c = 44$, respectively.

Suppose that p_n is chosen so that the binomial class $\mathcal{F}$ has expected size $p_n 2^n = c = 2^d$. Using the above probability distributions for the VC-dimension and the initial-dimension, the corresponding expected values $\mathbb{E}(\mathrm{VC}(\mathcal{F}))$ and $\mathbb{E}(\mathrm{ID}(\mathcal{F}))$ can be computed and analyzed in terms of the expected class size c. Table 2 displays these values for $c = 2^d$, $d = 1, \ldots, 10$. As seen, the expected value of the VC-dimension of $\mathcal{F}$ is just slightly smaller than the size d of a set that could in theory be shattered by some class of the same cardinality as $\mathcal{F}$.

5 Asymptotics for the testing-dimension

From Lemma 3 (Corollary 1) it follows that as the cardinality k (or expected cardinality $p_n 2^n$) of a random class tends to infinity, the VC-dimension and the initial-dimension both tend to infinity. It is still possible

d	1	2	3	4	5	6	7	8	9	10
$\mathbb{E}(\mathrm{VC}(\mathcal{F}))$	0.74	1.52	2.51	3.52	4.52	5.52	6.51	7.51	8.51	9.50
$\mathbb{E}(\mathrm{ID}(\mathcal{F}))$	0.42	0.91	1.55	2.24	2.96	3.75	4.55	5.30	6.07	6.97

Tab. 2: Expectations of the asymptotic probability distributions of the VC-dimension and initial-dimension under the binomial model P_{n,p_n} for $p_n 2^n = c = 2^d$, $d = 1, \ldots, 10$.

however that as this occurs, the event A_d of the random class shattering *all* sets of size d (see Section 2.1) does *not* occur, even for $d = 1$. As we now show, the expected value of the number of unshattered sets of size d may tend to infinity even when the cardinality of the class tends to infinity.

In order to show this, we use the standard first-moment method (see Janson et al. (2000), p. 54). Let X be the random variable that counts the number of sets of size d which are *not* shattered by a random class $\mathcal{F}$. We may express X as a sum of indicators $\mathbb{I}_{\overline{S}_{\mathbf{x}}}(\mathcal{F})$ of the complement of the events $S_{\mathbf{x}}$ (see Section 2.1), over all sets $\mathbf{x}$ with $|\mathbf{x}| = d$. Hence the events '$X = 0$' and A_d are identical. Using Lemma 4, we may express the expected value of X as

$$\begin{aligned}
\mathbb{E}_{n,p}(X) &= \sum_{\mathbf{x}:|\mathbf{x}|=d} \mathbb{E}_{n,p}(\mathbb{I}_{\overline{S}_{\mathbf{x}}}(\mathcal{F})) \\
&= \sum_{\mathbf{x}:|\mathbf{x}|=d} \mathbb{P}_{n,p}(\overline{S}_{\mathbf{x}}) \\
&= \sum_{\mathbf{x}:|\mathbf{x}|=d} \left(1 - \left(1 - (1-p)^{2^{n-d}}\right)^{2^d}\right) \\
&= \binom{n}{d} \left(1 - \left(1 - (1-p)^{2^{n-d}}\right)^{2^d}\right). \qquad (19)
\end{aligned}$$

Assume $p = p_n = c\, 2^{-n} \log n$, for some positive constant c. As n tends to infinity then

$$\mathbb{E}_{n,p_n}(X) = \frac{2^d}{d!} n^{d - c2^{-d}} (1 + o(1)). \qquad (20)$$

Thus $\mathbb{E}_{n,p_n}(X)$ tends to 0 if $c < d2^d$, to $+\infty$ if $c > d2^d$, so $p_n = d2^{d-n} \log n$ appears as a threshold for the expected number of unshattered sets. As we have previously done for the events E_d and S_d (Lemma 3 and Corollary 1) we now state the behavior of the probability of A_d at the extreme values of the expected cardinality of the class. It is reasonable to expect that the same threshold as for the expected value of X, also holds for the probability of A_d. Lemma 5 below states that the probability of A_d tends to 1 if $p_n \gg d2^{d-n} \log n$. We believe that it tends to 0 if $p_n \ll d2^{d-n} \log n$. However, the first moment method only gives the first claim. For the second claim, the second-moment method (see for instance Janson et al. (2000) p. 54 or Spencer (1991) Theorem 3.1) may be used. This requires estimating the correlation between pairs of events $\overline{S}_{\mathbf{x}}$, $\overline{S}_{\mathbf{y}}$ which we were not able to do under the binomial model.

Lemma 5 *If $p_n \gg d2^{d-n} \log n$, then the probability $P_{n,p_n}(A_d)$ that the testing-dimension is at least d tends to 1.*

Proof: Assume that $p_n = c2^{-n} \log n$. Then from (20), $\mathbb{E}_{n,p_n}(X)$ tends to zero if $c > d2^d$. Hence

$$\lim_{n\to\infty} P_{n,p_n}(\overline{A}_d) = \lim_{n\to\infty} P_{n,p_n}(X > 0) \leqslant \lim_{n\to\infty} \mathbb{E}_{n,p_n}(X) = 0$$

which follows from an application of Markov's inequality for a non-negative integer-valued random variable. □

To obtain more precise estimates around the threshold, we shall use the uniform model. Proposition 3 below gives the asymptotic probability of the event A_d at the threshold, as was previously done for the events E_d and S_d in Propositions 1 and 2.

Proposition 3 *Let $d \geqslant 1$ and $k = k_n$ be positive integers such that*

$$\lim_{n\to\infty} k_n - \alpha_d \log n = c,$$

where

$$\alpha_d = -\frac{d}{\log(1-2^{-d})}$$

and c is a real constant. Then

$$\lim_{n\to\infty} P^*_{n,k_n}(A_d) = \exp\left(-\frac{2^d}{d!}(1-2^{-d})^c\right) .$$

Remark 4 *Denote $c_n = k_n - \alpha_d \log n$. If c_n tends to $+\infty$, then $P^*_{n,k_n}(A_d)$ tends to 1. If c_n tends to $-\infty$, then $P^*_{n,k_n}(A_d)$ tends to 0. Hence the critical value of the class-cardinality k_n is $\alpha_d \log n$: between the binomial model and the uniform one, the threshold for the class cardinality shifts from $2^d \log n$ to $\alpha_d \log n$.*

Consider a random class of cardinality $k_n = \alpha \log n$. If $\alpha \neq \alpha_d$ for all d, then *w.h.p.* the testing dimension of the class equals the largest integer d such that $\alpha > \alpha_d$. If $\alpha = \alpha_d$ for some positive integer d, then *w.h.p.* all sets of size $d-1$ are shattered (since $\alpha > \alpha_{d-1}$), at least one set of size $d+1$ is not (since $\alpha < \alpha_{d+1}$), and there may be a positive probability that all sets of size d are shattered. Therefore the testing dimension of the class in this case ($\alpha = \alpha_d$) is either $d-1$ or d *w.h.p.* Hence for any value of α the testing-dimension of a random class with cardinality $\alpha \log n$ concentrates on one or two values. Thus Proposition 3 expresses a concentration result for the testing-dimension which is similar to the well-known concentration theorem by Matula for the clique number of random graphs (see Theorem 7.1 in Janson et al. (2000) or Theorem 11.4 p. 228 of Bollobás (2001)).

We now proceed with the proof.

Proof: As in Section 3, we use Lemma 2 and replace $P^*_{n,k}$ by $Q^*_{n,k}$. In terms of matrices, the event $S_{\mathbf{x}}$ is identical to the event that the submatrix M with row indices $i \in \mathbf{x} \subset [n]$ and column indices $1 \leqslant j_1, \ldots, j_{2^d} \leqslant k$, is equal to the matrix M_d up to a permutation of columns (see the proof of Lemma 3).

As before, let X count the subsets $\mathbf{x}$ of size d that are not shattered by the random class $\mathcal{F}$, *i.e.*,

$$X = \sum_{\mathbf{x}:|\mathbf{x}|=d} \mathbb{I}_{\overline{S}_{\mathbf{x}}}(\mathcal{F}) .$$

Consider any $\mathbf{x} \subset [n]$ then recall from (3) that $S_{\mathbf{x}} = \bigcap_\eta C_{\mathbf{x},\eta}$ hence $\overline{S}_{\mathbf{x}} = \bigcup_\eta \overline{C}_{\mathbf{x},\eta}$ where η runs over the set $\{0,1\}^{\mathbf{x}}$.

Consider any h distinct functions $\eta_1, \ldots, \eta_h \in \{0,1\}^{\mathbf{x}}$. A random matrix has every one of its k columns different from every η_i, $1 \leqslant i \leqslant h$, with probability

$$Q^*_{n,k}(\overline{C}_{\mathbf{x},\eta_1} \cap \ldots \cap \overline{C}_{\mathbf{x},\eta_h}) = (1 - h2^{-d})^k \;.$$

Using the expression for the probability of a union of events (see Feller (1968) (1.5), p. 89) one obtains

$$Q^*_{n,k}(\overline{S_{\mathbf{x}}}) = \sum_{h=1}^{2^d} (-1)^{h-1} \binom{2^d}{h} (1 - h2^{-d})^k \;.$$

If $k = k_n$ tends to infinity, this sum tends to zero and the first term dominates. Hence:

$$Q^*_{n,k_n}(\overline{S_{\mathbf{x}}}) = 2^d (1 - 2^{-d})^{k_n} (1 + o(1)) \;.$$

For $k_n = \alpha_d \log n + c + o(1)$,

$$Q^*_{n,k_n}(\overline{S_{\mathbf{x}}}) = 2^d (1 - 2^{-d})^c n^{-d} (1 + o(1)) \;. \tag{21}$$

Denoting by $\mathbb{E}_{n,k}(X)$ the expectation of X with respect to $Q^*_{n,k}$, we have

$$\mathbb{E}_{n,k_n}(X) = \binom{n}{d} Q^*_{n,k_n}(\overline{S}_{\mathbf{x}}) \;, \tag{22}$$

which tends to $\frac{2^d}{d!}(1 - 2^{-d})^c$ with increasing n. It remains to show that

$$\lim_{n\to\infty} Q^*_{n,k_n}(X = 0) = \lim_{n\to\infty} \exp(-\mathbb{E}_{n,k_n}(X)) \;,$$

i.e. that a Poisson approximation holds for X (Barbour et al., 1992). The technique of proof, based on the Stein-Chen method, is quite standard: we shall use the results stated in Janson (1994).

For $d = 1$, the subsets $\mathbf{x}$ are singletons and the events $\overline{S}_{\mathbf{x}}$ are independent. Their common probability is $Q^*_{n,k_n}(\overline{S}_{\mathbf{x}}) = n^{-1}(1 + o(1))$. Hence

$$Q^*_{n,k_n}(X = 0) = \Big(1 - n^{-1}(1 + o(1))\Big)^n = \exp(-1)(1 + o(1)) \;.$$

For $d \geqslant 2$, the family of indicators $(\mathbb{I}_{\overline{S}_{\mathbf{x}}})$ is *dissociated* in the sense of Janson (1994) p. 10: the two sets of random variables $\{\mathbb{I}_{\overline{S}_{\mathbf{x}}}\,,\ \mathbf{x} \in J\}$ and $\{\mathbb{I}_{\overline{S}_{\mathbf{y}}}\,,\ \mathbf{y} \in K\}$ are independent whenever every $\mathbf{x} \in J$ is disjoint from every $\mathbf{y} \in K$. Denote by Γ the set of all $\mathbf{x} \subset [n]$ with $|\mathbf{x}| = d$. For $\mathbf{x} \in \Gamma$, denote by $\Gamma_{\mathbf{x}}$ the set of all $\mathbf{y}$ such that $\mathbf{x} \cap \mathbf{y} \neq \emptyset$. By Theorem 4 p. 10 of Janson (1994), the total variation distance between the distribution of X and the Poisson distribution with parameter $\mathbb{E}_{n,k_n}(X)$ is bounded above by

$$(1 \wedge \mathbb{E}(X)^{-1}) \left(\sum_{\mathbf{x}\in\Gamma} \sum_{\mathbf{y}\in\Gamma_{\mathbf{x}}} Q^*_{n,k_n}(\overline{S_{\mathbf{x}}}) Q^*_{n,k_n}(\overline{S_{\mathbf{y}}}) + \sum_{\mathbf{x}\in\Gamma} \sum_{\mathbf{y}\in\Gamma_{\mathbf{x}}\setminus\{\mathbf{x}\}} Q^*_{n,k_n}(\overline{S_{\mathbf{x}}} \cap \overline{S_{\mathbf{y}}}) \right) \tag{23}$$

The result will follow by proving that each of the two sums in (23) converges to zero. The first sum has $O(n^{2d-1})$ terms, each of order $O(n^{-2d})$, by (21). We decompose the second sum according to the number of elements in $\mathbf{x}\cap\mathbf{y}$ as follows:

$$\sum_{\mathbf{x}\in\Gamma}\sum_{\mathbf{y}\in\Gamma_{\mathbf{x}}\setminus\{\mathbf{x}\}} Q^*_{n,k_n}(\overline{S_{\mathbf{x}}}\cap\overline{S_{\mathbf{y}}}) = \sum_{h=1}^{d-1}\Delta_h\;,$$

where

$$\Delta_h = \sum_{|\mathbf{x}\cap\mathbf{y}|=h} Q^*_{n,k_n}(\overline{S}_{\mathbf{x}}\cap\overline{S}_{\mathbf{y}}).$$

Clearly, there are $O(n^{2d-h})$ terms in Δ_h. From (3) we have

$$\overline{S}_{\mathbf{x}}\cap\overline{S}_{\mathbf{y}} = \left(\bigcup_{\eta\in\{0,1\}^{\mathbf{x}}}\overline{C}_{\mathbf{x},\eta}\right)\bigcap\left(\bigcup_{\zeta\in\{0,1\}^{\mathbf{y}}}\overline{C}_{\mathbf{y},\zeta}\right) = \bigcup_{\eta,\zeta}\left(\overline{C}_{\mathbf{x},\eta}\cap\overline{C}_{\mathbf{y},\zeta}\right).$$

A randomly drawn function f on $[n]$ generated by n coin flips with probability $1/2$ has a probability of $(1/2)^d$ for each of the individual events '$f_{|\mathbf{x}}=\eta$' and '$f_{|\mathbf{y}}=\zeta$'. The probability that both events '$f_{|\mathbf{x}}=\eta$' and '$f_{|\mathbf{y}}=\zeta$' occur simultaneously is $(1/2)^{2h-d}$ provided that $\eta_{|\mathbf{x}\cap\mathbf{y}}=\zeta_{|\mathbf{x}\cap\mathbf{y}}$ and 0 otherwise. It follows that

$$Q^*_{n,k}(\overline{C}_{\mathbf{x},\eta}\cap\overline{C}_{\mathbf{y},\zeta}) = \begin{cases}(1-2^{-d+1}+2^{-2d+h})^k & \text{if } \eta\equiv\zeta \text{ on } \mathbf{x}\cap\mathbf{y}\\[1ex] (1-2^{-d+1})^k & \text{otherwise.}\end{cases}$$

Hence

$$\Delta_h \leqslant a\,n^{2d-h}(1-2^{-d+1}+2^{-2d+h})^k$$

for some positive a, not depending on n and k. For $k=k_n=\alpha_d\log n+c+o(1)$, there exists a positive constant b such that

$$\Delta_h \leqslant b\,n^{2d-h+\alpha_d\log(1-2^{-d+1}+2^{-2d+h})}\;.$$

That Δ_h tends to zero follows from having a negative exponent, *i.e.*,

$$2d-h+\alpha_d\log(1-2^{-d+1}+2^{-2d+h})<0 \tag{24}$$

for $d\geqslant 2$ and $h=1,\dots,d-1$. Indeed, the left hand side of (24) vanishes both for $h=0$ and $h=d$. As a function of h, its second derivative is positive on $[0,d]$ hence it is strictly convex. Therefore it is strictly negative for all $h=1,\dots,d-1$. Hence Δ_h tends to zero with increasing n. $\square$

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):616–631, 1997.

R. Anstee, B. Fleming, Z. Füredi, and A. Sali. Color critical hypergraphs and forbidden configurations. In *Proc. EuroComb2005*, pages 117–122. DMTCS, 2005.

M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

M. Anthony, G. Brightwell, and C. Cooper. The Vapnik-Chervonenkis dimension of a random graph. *Discrete Math.*, 138(1-3):43–56, 1995a.

M. Anthony, G. Brightwell, and J. Shawe-Taylor. On specifying Boolean functions by labelled examples. *Discrete Appl. Math.*, 61:1–25, 1995b.

A. Barbour, L. Holst, and S. Janson. *Poisson approximation*. Clarendon Press, Oxford, 1992.

B. Bollobás. *Random Graphs*. Cambridge University Press, 2001.

B. Bollobás. *Combinatorics: Set Systems, Hypergraphs, Families of vectors, and combinatorial probability*. Cambridge University Press, 1986.

W. Feller. *An introduction to probability theory and its applications*, volume I. Wiley, London, 3rd edition, 1968.

P. Frankl. On the trace of finite sets. *J. Combinatorial Theory*, 34:41–45, 1983.

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

D. Haussler and P. Long. A generalization of Sauer's lemma. *J. Combinatorial Theory*, 71(2):219–240, 1995.

D. Haussler and E. Welzl. Epsilon-nets and simplex range queries. *Discrete Computational Geometry*, 2: 127–151, 1987.

S. Janson. Coupling and Poisson approximation. *Acta Appl. Math.*, 34:7–15, 1994.

S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley, New York, 2000.

J. Matoušek. Geometric set systems. In *European Congress of Mathematics, Vol. II (Budapest 1996)*, volume 169 of *Progr. Math.*, pages 1–27. Birkhäuser, Basel, 1998.

J. Pach and P. Agarwal. *Combinatorial Geometry*. Wiley, New York, 1995.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

K. Romanik and C. Smith. Testing geometric objects. *Comput. Geometry: Theory and Appl.*, 4:157–176, 1994.

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

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

J. Spencer. Nine lectures on Random Graphs. In P. Bernard, editor, *Ecole d'été de probabilités de Saint-Flour XXI*, volume 1541 of *L.N. in Mathematics*, pages 293–343. Springer-Verlag, New York, 1991.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

B. Ycart and J. Ratsaby. The VC dimension of k-uniform random hypergraphs. *Random Struct. Algorithms*, 30(4):564–572, 2007.