Available online at www.sciencedirect.com

ScienceDirect

Discrete Applied Mathematics 156 (2008) 2753–2767

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

Constrained versions of Sauer's lemma

Joel Ratsaby

Department of Electrical and Electronic Engineering, Ariel University Center of Samaria, Ariel 40700, Israel

Received 8 July 2006; received in revised form 10 November 2007; accepted 18 November 2007
Available online 3 January 2008

Abstract

Let $[n] = \{1, \ldots, n\}$. For a function $h : [n] \to \{0, 1\}$, $x \in [n]$ and $y \in \{0, 1\}$ define by the *width* $\omega_h(x, y)$ of h at x the largest nonnegative integer a such that $h(z) = y$ on $x - a \le z \le x + a$. We consider finite VC-dimension classes of functions h constrained to have a width $\omega_h(x_i, y_i)$ which is larger than N for all points in a sample $\zeta = \{(x_i, y_i)\}_1^\ell$ or a width no larger than N over the whole domain $[n]$. Extending Sauer's lemma, a tight upper bound with closed-form estimates is obtained on the cardinality of several such classes.

Keywords: Sauer's lemma; Integer partitions; Binary functions

1. Introduction

Let $[n] = \{1, \ldots, n\}$ and denote by $2^{[n]}$ the class of all 2^n functions $h : [n] \to \{0, 1\}$. Let $\mathcal{H}$ be a class of functions and for a set $A = \{x_1, \ldots, x_k\} \subseteq [n]$ denote by $h_{|A} = [h(x_1), \ldots, h(x_k)]$ the *restriction* of h on A. A class $\mathcal{H}$ is said to *shatter* A if $\left|\{h_{|A} : h \in \mathcal{H}\}\right| = 2^k$. The Vapnik–Chervonenkis dimension of $\mathcal{H}$, denoted as $VC(\mathcal{H})$, is defined as the cardinality of the largest set shattered by $\mathcal{H}$. The following well-known result obtained by [19,21,24] states a tight upper bound on the cardinality of classes $\mathcal{H}$ of VC-dimension d.

Lemma 1 (*Sauer's Lemma*). *For any* $1 \le d < n$ *let*

$$\mathbb{S}(n, d) = \sum_{k=0}^{d} \binom{n}{k}.$$

Then

$$\max_{\mathcal{H} \subset 2^{[n]} : VC(\mathcal{H}) = d} |\mathcal{H}| = \mathbb{S}(n, d).$$

More generally, the lemma holds for classes of finite VC-dimension on infinite domains.

E-mail address: ratsaby@ariel.ac.il.
URL: http://www.ariel.ac.il/ee/pf/ratsaby.

0166-218X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2007.11.017

y					$y_1=1$									$y_2=0$							
h_1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
h_2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	0
[n]	1	2	.	.	x_1	.	.	.	.	.	.	.	.	x_2	.	.	.	.	.	.	n

Fig. 1. $\omega_{h_1}(\zeta) = \omega_{h_2}(\zeta) = 3$.

Aside of being an interesting combinatorial result (see Chapter 17 in [9]), Lemma 1 has been instrumental in statistical learning theory [23], combinatorial geometry [17], graph theory [4,16] and in the theory of empirical processes [18]. In such areas, the complexity of analysis of algorithms on discrete structures, for instance, learning an unknown target binary function, typically involves a simpler structure constrained by some ‘smoothness’ property which is induced by the underlying algorithmics. In learning, the constraint is induced by a finite sample.

Consider a binary function $h : [n] \to \{0, 1\}$, $x \in [n]$ and $y \in \{0, 1\}$ and define by $\omega_h(x, y)$ the largest a, $0 \le a \le \min\{x, n - x\}$ such that $h(z) = y$ for all $x - a \le z \le x + a$; if no such a exists then let $\omega_h(x, y) = -1$. We call this the *width* of h at x with respect to y. Denote by $\Xi = [n] \times \{0, 1\}$. By a *sample* $\zeta = \{(x_i, y_i)\}_{i=1}^{\ell} \in \Xi^{\ell}$, we mean a set of ℓ pairs with different x-components. Define by $\omega_h(\zeta) = \min_{1\le i\le \ell} \omega_h(x_i, y_i)$ the width of h with respect to ζ. For instance, Fig. 1 displays a sample $\zeta = \{(x_1, y_1), (x_2, y_2)\}$ and two functions h_1, h_2 which have a width of 3 with respect to ζ. In [3], the complexity of learning binary functions by aiming to maximize this sample width has been investigated.

The main question posed in this paper is as follows: starting from a class $\mathcal{H}$ as above with $\mathrm{VC}(\mathcal{H}) = d$ consider a subset of $\mathcal{H}$ of functions which are ‘smooth’, i.e., constrained to have large sample widths and therefore consecutive runs of 1’s or runs of 0’s of a certain minimal length. Does Sauer’s lemma hold for such a subset? How does its cardinality increase with respect to n and how is it affected by the size of the allowed sample width?

The area of research on Poisson approximations (see for instance [5–7]) includes many results on the number of binary sequences of length n that have ‘long’ repetitive runs (with various definitions of a long run). Our question above differs in that we add the condition of having a known VC-dimension. To our knowledge, this is the first instance of a study which considers estimating the complexity of a class constrained structurally by both an extremal set property (having a finite VC-dimension) and a repetitive-run type property.

Let $N \ge 0$ be a width parameter. We study the complexity of classes of the form

$$\mathcal{H}_N(\zeta) = \{h \in \mathcal{H} : \omega_h(\zeta) > N\}, \quad \mathrm{VC}(\mathcal{H}) = d \tag{1}$$

where $\zeta = \{(x_i, y_i)\}_{i=1}^{\ell} \in \Xi^{\ell}$ is a given sample.

We obtain tight bounds in the form of Sauer’s Lemma 1 on the cardinality of such classes. It turns out that the bounds have subtle nonlinear dependence on n and N. This is investigated in detail in subsequent sections.

For a function $h : [n] \to \{0, 1\}$ let the *difference* function be defined as

$$\delta_h(x) = \begin{cases} 1 & \text{if } h(x-1) = h(x) \\ 0 & \text{otherwise} \end{cases}$$

where we assume that any h satisfies $h(0) = 0$ (see Fig. 2). Define

$$\mathcal{D}_{\mathcal{H}} \equiv \{\delta_h : h \in \mathcal{H}\}, \tag{2}$$

or $\mathcal{D}$ for brevity. It is easy to see that the class $\mathcal{D}$ is in one-to-one correspondence with $\mathcal{H}$. For $N \ge 0$ and any sample ζ, if $\omega_h(x, y) \le N$ for $(x, y) \in \zeta$ then the corresponding δ_h has $\omega_{\delta_h}(x, 1) \le N$. In order to estimate the cardinality of classes $\mathcal{H}_N(\zeta)$ we will estimate the cardinality of the corresponding difference classes $\mathcal{D}_N(\zeta_+)$ which are defined based on $\zeta_+ = \{(x_i, 1) : (x_i, y_i) \in \zeta, 1 \le i \le l\}$. We denote by

$$\mathrm{VC}_{\Delta}(\mathcal{H}) \equiv \mathrm{VC}(\mathcal{D})$$

the VC-dimension of the difference class $\mathcal{D} = \{\delta_h : h \in \mathcal{H}\}$ and use it to characterize the complexity of $\mathcal{H}$ (it is easy to show that $\mathrm{VC}(\mathcal{D}) \le c\mathrm{VC}(\mathcal{H})$ for some small constant $c > 1$). We henceforth denote by $d \equiv \mathrm{VC}_{\Delta}(\mathcal{H})$.

The rest of the paper is organized as follows: in Section 2 we state the main results, Section 3 contains the lemmas used for proving these results.

h	1	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	1	0	1	0	0	0
δ_h	0	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	0	0	0	0	1	1
[n]	1	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	n

Fig. 2. h and the corresponding δ_h.

2. Main results

The first main result concerns a class of functions which is constrained by an upper bound on the width. Let $N \geq 0$ be a width parameter and for any class $\mathcal{H}$ of binary functions on $[n]$ define

$$\mathcal{H}_N = \{h \in \mathcal{H} : \omega_h(x, h(x)) \leq N, x \in [n]\}. \tag{3}$$

Denote by

$$w_{m,\nu}(n) = \sum_{k=0}^{m} (-1)^k \binom{m}{k} \binom{n+m-1-k(\nu+1)}{m-1},$$

$$\begin{aligned} c(k, n-k; m, N) = \binom{n-k}{m-1} & (w_{m,2N}(k-m+1) + w_{m-1,2N}(k-m-2N) \\ & + w_{m-1,2N}(k-m-2N-1)) \end{aligned} \tag{4}$$

and

$$\beta_r^{(N)}(n) \equiv \sum_{k=0}^{r} \sum_{m=1}^{n} c(k, n-k; m, N). \tag{5}$$

Theorem 1. *Let* $1 \leq d \leq n$ *and* $N \geq 0$*. Then*

$$\max_{\mathcal{H} \subset 2^{[n]} : VC_{\triangle}(\mathcal{H}) = d} |\mathcal{H}_N| = \beta_d^{(N)}(n). \tag{6}$$

The proof follows directly from Lemma 4 in Section 3.1 which combines the theory of integer partitions with the classic shifting method in extremal set theory. Our second main result is Theorem 2 which states an estimate on $w_{m,\nu}(n)$ which, as is later shown, is the number of constrained ordered integer partitions. With this estimate in place we then state a simpler closed-form approximation for $\beta_d^{(N)}(n)$. Let c_i, $i = 1, 2, \ldots$ denote constants between 0 and 1. For an integer $i > 0$, denote by $k^{\underline{i}} \equiv k(k-1)\cdots(k-i+1)$. Let

$$\mu_N \equiv \frac{N}{2}, \quad \sigma_N^2 \equiv \frac{2}{N+1} \left(\frac{(\mu_N+1)^{\underline{3}}}{3} + \frac{(\mu_N+1)^{\underline{2}}}{2} \right). \tag{7}$$

Denote by

$$A(n, N) \equiv (1-c_1) \left(\ln(N+1) - (\mu_N+1)^2 \frac{(1-c_2)}{2\sigma_N^2} \right) - c_1(1 + c_3 \ln n)$$

$$p(n, N) \equiv \frac{\mathrm{e}^{A(n,N)}}{1 + \mathrm{e}^{A(n,N)}}$$

and let $b(n, p, r)$ denote the probability that a binomial random variable with parameters n and p takes a value which is no larger than r where $0 \leq r \leq n$. Then as we show in Section 3.2, the function $\beta_d^{(N)}(n)$ may be approximated by

$$\hat{\beta}_d^{(N)}(n) = \frac{(2N+1)c_4}{\sigma_{2N}} \left(1 + \mathrm{e}^{A(n,2N)}\right)^n b(n, p(n, 2N), d).$$

The next two results concern classes of functions with a lower-bound on the width as defined in (1). They are simple and direct applications of Sauer's lemma (Lemma 1) and of Theorem 1 and hence are stated as propositions.

Proposition 1. *Let $1 \le d, \ell \le n$ and $N \ge 0$. Then*

$$\max_{\mathcal{H}\subset 2^{[n]},\zeta\in\Xi^{\ell}:VC_{\Delta}(\mathcal{H})=d} |\mathcal{H}_N(\zeta)| = \mathbb{S}(n-\ell-2N-1,d).$$

The proof is in Section 3.3. Next, consider an extremal case where the width of h is larger than N *only* on elements of ζ, for all $h \in \mathcal{H}_N(\zeta)$. In this case the class is defined as

$$\mathcal{H}^*_N(\zeta) = \{h \in \mathcal{H} : \omega_h(x,h(x)) > N \text{ iff } (x,h(x)) \in \zeta\}, \quad N \ge 0.$$

This type of class arises in certain applications where given a sample ζ an algorithm obtains a solution, i.e., a binary function, which maximizes the width on ζ. We are interested in the number of functions that have such maximal width since it represents the richness of the class of possible hypotheses. This is stated in the next result.

Proposition 2. *Let $1 \le d, \ell \le n$ and $N \ge 0$. Then*

$$\max_{\mathcal{H}\subset 2^{[n]},\zeta\in\Xi^{\ell}:VC_{\Delta}(\mathcal{H})=d} |\mathcal{H}^*_N(\zeta)| = \beta_d^{(N)}(n-\ell-2N-1).$$

The proof is in Section 3.4. We proceed to the technical work used to prove the above results.

3. Technical work

We start with several lemmas used in proving Theorem 1.

3.1. Lemmas for Theorem 1

Let $\binom{n}{k}$ denote the following function

$$\binom{n}{k} = \begin{cases} n!/(k!(n-k)!) & \text{if } 0 \le k \le n \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbb{I}(E)$ denote the indicator function which equals 1 if the expression E is true and 0 otherwise.

Lemma 2. *For any nonnegative integers n, $\nu \ge 0$ and $m \le n$ define by $w_{m,\nu}(n)$ the number of standard (one-dimensional) ordered partitions of n into m parts each no larger than ν. Then*

$$w_{m,\nu}(n) = \begin{cases} 0 & \text{if } n < 0 \\ \mathbb{I}(n=0) & \text{if } m = 0 \text{ or } \nu = 0 \\ \displaystyle\sum_{i=0,\nu+1,2(\nu+1),\ldots}^{n} (-1)^{i/(\nu+1)} \binom{m}{i/(\nu+1)} \binom{n-i+m-1}{n-i} & \text{if } m \ge 1. \end{cases} \tag{8}$$

Proof. By definition of $w_{m,\nu}(n)$, its generating function equals $\sum_{n\ge 0} w_{m,\nu}(n)x^n = (1+x+x^2+\cdots+x^{\nu})^m$ since the coefficient of x^n in this expression is the number of monomials $x^{i_1+i_2+\cdots+i_m}$ such that $i_1+i_2+\cdots+i_m = n$ and $0 \le i_j \le \nu$, $1 \le j \le m$. But this expression equals

$$\left(\frac{1-x^{\nu+1}}{1-x}\right)^m \equiv W(x).$$

When $m = 0$ or $\nu = 0$ the only nonzero coefficient is of x^0 and it equals 1 so $w_{m,\nu}(n) = \mathbb{I}(n=0)$. Let $T(x) = (1-x^{\nu+1})^m$ and $S(x) = \left(\frac{1}{1-x}\right)^m$. Then

$$T(x) = \sum_{i=0}^{m} (-1)^i \binom{m}{i} x^{i(\nu+1)}$$

which generates the sequence $t_\nu(n) = \binom{m}{n/(\nu+1)}(-1)^{n/(\nu+1)}\mathbb{I}(n \bmod (\nu+1) = 0)$. Similarly, for $m \geq 1$, it is easy to show that $S(x)$ generates $s(n) = \binom{n+m-1}{n}$. The product $W(x) = T(x)S(x)$ generates their convolution $t_\nu(n) \star s(n)$, namely,

$$w_{m,\nu}(n) = \sum_{i=0,\nu+1,2(\nu+1),\ldots,}^{n} (-1)^{i/(\nu+1)} \binom{m}{i/(\nu+1)} \binom{n-i+m-1}{n-i}. \quad \square$$

Remark 1. While our interest is in $[n] = \{1, \ldots, n\}$, we allow $w_{m,\nu}(n)$ to be defined on $n \leq 0$ for use by Lemma 3.

Remark 2. This expression may alternatively be expressed as

$$w_{m,\nu}(n) = \sum_{k=0}^{m} (-1)^k \binom{m}{k} \binom{n+m-1-k(\nu+1)}{m-1},$$

over $m \geq 1$.

We need two additional lemmas for proving (6) of Theorem 1.

Lemma 3. *Let the integer $N \geq 0$ and consider the class F of all binary-valued functions f on* $[n]$, *or equivalently, sequences $f = f(1), \ldots, f(n)$, satisfying:* (a) *f has no more than r* 1' s (b) *every run of consecutive* 1*'s in f is no longer than* $2N+1$, *except for a run that starts at $f(1)$ which may be of length* $2(N+1)$. *Then*

$$|F| = \beta_r^{(N)}(n).$$

Remark 3. When $r \leq 2N+1$, $\beta_r^{(N)}(n) = \mathbb{S}(n, r)$.

Proof. Consider the integer pair $[k, n-k]$, where $n \geq 1$ and $0 \leq k \leq n$. A two-dimensional ordered m-partition of $[k, n-k]$ is an ordered partition into m two-dimensional parts, $[a_j, b_j]$ where $0 \leq a_j, b_j \leq n$ but not both are zero and where $\sum_{j=1}^{m}[a_j, b_j] = [k, n-k]$. For instance, $[2, 1] = [0, 1] + [2, 0] = [1, 1] + [1, 0] = [2, 0] + [0, 1]$ are three partitions of $[2, 1]$ into two parts (for more examples see [1]).

Suppose we add the constraint that only a_1 or b_m may be zero while all remaining

$$a_j, b_k \geq 1, \quad 2 \leq j \leq m, 1 \leq k \leq m-1. \tag{9}$$

Denote any partition that satisfies this as *valid*. For instance, let $k = 2$, $m = 3$ then the valid m-partitions of $[k, n-k]$ are: $\{[0, 1][1, 1][1, n-4]\}$, $\{[0, 1][1, 2][1, n-5]\}, \ldots, \{[0, 1][1, n-3][1, 0]\}$, $\{[0, 2][1, 1][1, n-5]\}$, $\{[0, 2][1, 2][1, n-6]\}, \ldots, \{[0, 2][1, n-4][1, 0]\}, \ldots, \{[0, n-3][1, 1][1, 0]\}$. For $[k, n-k]$, let $\mathcal{P}_{n,k}$ be the collection of all valid partitions of $[k, n-k]$.

Let F_k denote all binary functions on $[n]$ which take the value 1 over exactly k elements of $[n]$. Define the mapping $\Pi : F_k \to \mathcal{P}_{n,k}$ where for any $f \in F_k$ the partition $\Pi(f)$ is defined by the following procedure: Start from the first element of $[n]$, i.e., 1. If f takes the value 1 on it then let a_1 be the length of the constant 1-segment, i.e., the set of all elements starting from 1 on which f takes the constant value 1. Otherwise if f takes the value 0 let $a_1 = 0$. Then let b_1 be the length of the subsequent 0-segment on which f takes the value 0. Let $[a_1, b_1]$ be the first part of $\Pi(f)$. Next, repeat the following: if there is at least one more element of $[n]$ which has not been included in the preceding segment, then let a_j be the length of the next 1-segment and b_j the length of the subsequent 0-segment. Let $[a_j, b_j]$, $j = 1, \ldots, m$, be the resulting sequence of parts where m is the total number of parts. Only the last part may have a zero valued b_m since the function may take the value 1 on the last element n of $[n]$ while all other parts, $[a_j, b_j]$, $2 \leq j \leq m-1$, must have $a_j, b_j \geq 1$. The result is a valid partition of $[k, n-k]$ into m parts.

Clearly, every $f \in F_k$ has a unique partition. Therefore Π is a bijection. Moreover, we may divide $\mathcal{P}_{n,k}$ into mutually exclusive subsets V_m consisting of all valid partitions of $[k, n-k]$ having exactly m parts, where $1 \leq m \leq n$. Thus

$$|F_k| = \sum_{m=1}^{n} |V_m|.$$

Consider the following constraint on components of parts:

$$a_i \le \begin{cases} 2N+1 & \text{if } 2 \le i \le m \\ 2(N+1) & \text{if } i = 1. \end{cases} \tag{10}$$

Denote by $V_{m,N} \subset \mathcal{P}_{n,k}$ the collection of valid partitions of $[k, n-k]$ into m parts each of which satisfies this constraint.

Let $F_{k,N} = F \cap F_k$ consist of all functions satisfying the run-constraint in the statement of the lemma and having exactly k ones. If f has no run of consecutive 1's starting at $f(i)$ of length larger than $2N+1$ then there does not exist a segment a_i of length larger than $2N+1$, $i \ge 2$ (and similarly with a run of size $2(N+1)$ starting at $f(1)$). Hence the parts of $\Pi(f)$ satisfy (10) and for any $f \in F_{k,N}$, its unique valid partition $\Pi(f)$ must be in $V_{m,N}$. We therefore have

$$|F_{k,N}| = \sum_{m=1}^{n} |V_{m,N}|. \tag{11}$$

By definition of F all its elements f have no more than r 1's hence it follows that

$$|F| = \sum_{k=0}^{r} |F_{k,N}|. \tag{12}$$

Let us denote by

$$c(k, n-k; m, N) \equiv |V_{m,N}| \tag{13}$$

the number of valid partitions of $[k, n-k]$ into exactly m parts whose components satisfy (10). In order to determine $|F|$ it therefore suffices to determine $c(k, n-k; m, N)$.

We next construct the generating function

$$G(t_1, t_2) = \sum_{\alpha_1 \ge 0} \sum_{\alpha_2 \ge 0} c(\alpha_1, \alpha_2; m, N) t_1^{\alpha_1} t_2^{\alpha_2}. \tag{14}$$

For any real number a let $(a)_+$ denote the value a if $a \ge 0$ and 0 otherwise. For $m \ge 1$,

$$\begin{aligned} G(t_1, t_2) = {} & (t_1^0 + t_1^1 + \cdots + t_1^{2N+2})(t_2^1 + t_2^2 + \cdots)^{\mathbb{I}(m \ge 2)} \\ & \times \left((t_1^1 + \cdots + t_1^{2N+1})(t_2^1 + t_2^2 + \cdots) \right)^{(m-2)_+} \\ & \times (t_1^1 + \cdots + t_1^{2N+1})^{\mathbb{I}(m \ge 2)} (t_2^0 + t_2^1 + \cdots) \end{aligned} \tag{15}$$

where the values of the exponents of all terms in the first and second factors represent the possible values for a_1 and b_1, respectively. The values of the exponents in the middle $m-2$ factors are for the values of a_j, b_j, $2 \le j \le m-1$ and those in the factor before last and last are for a_m and b_m, respectively. Equating this to (14) implies the coefficient of $t_1^{\alpha_1} t_2^{\alpha_2}$ equals $c(\alpha_1, \alpha_2; m, N)$ which we seek.

The right-hand side of (15) equals

$$t_1^{m-1} t_2^{m-1} \left(\frac{1}{1-t_2} \right)^m \left(\left(\frac{1-t_1^{2N+1}}{1-t_1} \right)^m + t_1^{2N+1}(1+t_1) \left(\frac{1-t_1^{2N+1}}{1-t_1} \right)^{m-1} \right). \tag{16}$$

Let $W(x) = \left(\frac{1-x^{2N+1}}{1-x} \right)^{m-1}$ generate $w_{m-1,2N}(n)$ which is defined in Lemma 2. Similarly, also from this lemma we recall that $s(n) = \binom{n+m-1}{n}$ corresponds to $(1/(1-x))^m$ and thus $\left(\frac{1}{1-t_2} \right)^m$ in (16) generates the sequence $s(\alpha_2)$. So (16) becomes

$$\sum_{\alpha_1, \alpha_2 \ge 0} s(\alpha_2) t_2^{\alpha_2+m-1} \left(w_{m,2N}(\alpha_1) t_1^{\alpha_1+m-1} + w_{m-1,2N}(\alpha_1) t_1^{\alpha_1+m+2N}(1+t_1) \right). \tag{17}$$

Equating the coefficients of $t_1^{\alpha'_1} t_2^{\alpha'_2}$ in (14) and (17) yields

$$c(\alpha'_1, \alpha'_2; m, N) = s(\alpha'_2 - m + 1)\left(w_{m,2N}(\alpha'_1 - m + 1) \right. \\ \left. + w_{m-1,2N}(\alpha'_1 - m - 2N) + w_{m-1,2N}(\alpha'_1 - m - 2N - 1)\right).$$

Replacing $s(\alpha'_2 - m + 1)$ by $\binom{\alpha'_2}{m-1}$, substituting k for α'_1, $n-k$ for α'_2 and combining (11)–(13) yields the result. □

The next lemma extends the result of Lemma 3 to the class $\mathcal{H}_N$ defined in (3).

Lemma 4. *Let $1 \le d \le n$ and $N \ge 0$. For any class $\mathcal{H}$ with $VC_\Delta(\mathcal{H}) = d$, the cardinality of the corresponding class $\mathcal{H}_N$ defined in (3) is no larger than $\beta_d^{(N)}(n)$. This bound is tight.*

Proof. Denote by $\mathcal{D}_N = \{\delta_h : h \in \mathcal{H}_N\}$. Clearly, $|\mathcal{D}_N| = |\mathcal{H}_N|$. Consider any $h \in \mathcal{H}_N$. Since for all $x \in [n]$, $\omega_h(x, h(x)) \le N$ then the corresponding δ_h in $\mathcal{D}_N$ satisfies the following: every run of consecutive 1's is of length no larger than $2N + 1$, except for a run which starts at $x = 1$ whose length may be as large as $2(N + 1)$. Let $\mathcal{F}_N$ be the set system corresponding to the class $\mathcal{D}_N$ which is defined as follows:

$$\mathcal{F}_N = \{A_\delta : \delta \in \mathcal{D}_N\}, \ A_\delta = \{x \in [n] : \delta(x) = 1\}.$$

Clearly, $|\mathcal{F}_N| = |\mathcal{D}_N|$. Note that the above constraint on δ translates to A_δ possessing the property P_N defined as having every subset $E \subseteq A_\delta$ which consists of consecutive elements $E = \{i, i + 1, \ldots, j - 1, j\}$ be of cardinality $|E| \le 2N + 1$, except for such an E that contains the element $\{1\}$ which may have cardinality as large as $2(N + 1)$. Hence for every element $A \in \mathcal{F}_N$, A satisfies P_N. This is denoted by $A \models P_N$. Define $G_\mathcal{F}(k) = \max\{|\{A \cap E : A \in \mathcal{F}_N\}| : E \subseteq [n], |E| = k\}$. The corresponding notion of VC-dimension for a class $\mathcal{F}_N$ of sets is the so-called *trace number* ([9], p. 131) which is defined as $tr(\mathcal{F}_N) = \max\{m : G_{\mathcal{F}_N}(m) = 2^m\}$. Clearly, $tr(\mathcal{F}_N) = VC(\mathcal{D}_N) \le \mathrm{VC}(\mathcal{D}) \equiv \mathrm{VC}_\Delta(\mathcal{H}) = d$ (where $\mathcal{D}$ is defined in (2)).

The proof proceeds as in the proof of Sauer's lemma [2, Theorem 3.6] which is based on the shifting method (see [9], Ch. 17, Theorems 1 & 4, and [12,13,15]). The idea is to transform $\mathcal{F}_N$ into an *ideal* family $\mathcal{F}'_N$ of sets E, i.e., if $E \in \mathcal{F}'_N$ then $S \in \mathcal{F}'_N$ for every $S \subset E$, and such that $|\mathcal{F}_N| = |\mathcal{F}'_N| \le \beta_d^{(N)}(n)$.

Start by defining the operator T_x on $\mathcal{F}_N$ which removes an element $x \in [n]$ from every set $A \in \mathcal{F}_N$ provided that this does not duplicate any existing set. It is defined as follows:

$$T_x(\mathcal{F}_N) = \{A \setminus \{x\} : A \in \mathcal{F}_N\} \cup \{A \in \mathcal{F}_N : A \setminus \{x\} \in \mathcal{F}_N\}.$$

Consider now

$$\mathcal{F}'_N = T_1(T_2(\cdots T_n(\mathcal{F}_N)\cdots))$$

and denote the corresponding function class by $\mathcal{D}'_N$. Clearly, $|\mathcal{D}'_N| = |\mathcal{F}'_N|$.

We have $|\mathcal{F}'_N| = |\mathcal{F}_N|$ since the only time that the operator T_x changes an element A into a different set $A^* = T_x(A)$ is when A^* does not already exist in the class, so while new elements can be created they are replacing existing ones, hence no additional element can be created. It is also clear that for all $x \in [n]$, $T_x(\mathcal{F}'_N) = \mathcal{F}'_N$ since for each $E \in \mathcal{F}'_N$ there exists a G that differs from it on exactly one element hence it is not possible to remove any element $x \in [n]$ from all sets without creating a duplicate. Applying this repeatedly implies that $\mathcal{F}'_N$ is an ideal. Furthermore, since for all $A \in \mathcal{F}_N$, $A \models P_N$ then removing an element x from A still leaves $A \setminus \{x\} \models P_N$. Hence for all $E \in \mathcal{F}'_N$ we have $E \models P_N$.

Now, from Lemma 3 ([9], p. 133) we have $G_{\mathcal{F}'_N}(k) \le G_{\mathcal{F}_N}(k)$, for all $1 \le k \le n$. Since $tr(\mathcal{F}_N) \le d$ then $tr(\mathcal{F}'_N) \le d$. Together with $\mathcal{F}'_N$ being an ideal it follows that for all $E \in \mathcal{F}'_N$, $|E| \le d$. Now, for all $E \in \mathcal{F}'_N$, $E \models P_N$ hence the corresponding class $\mathcal{D}'_N$ satisfies the following: for all $\delta \in \mathcal{D}'_N$, δ has at most d 1's and every run of consecutive 1's is of length no larger than $2N + 1$ except possibly for a run which starts at $x = 1$ which may be as large as $2(N + 1)$. By Lemma 3 above, we therefore have $|\mathcal{D}'_N| \le \beta_d^{(N)}(n)$. We conclude that $|\mathcal{H}_N| = |\mathcal{D}_N| = |\mathcal{F}_N| = |\mathcal{F}'_N| = |\mathcal{D}'_N|$ and hence $|\mathcal{H}_N| \le \beta_d^{(N)}(n)$. This bound is tight since $\hat{\mathcal{H}}$ is considered whose corresponding class $\hat{\mathcal{D}}$ has *all* functions on $[n]$ with at most d 1's. Clearly, $\mathrm{VC}_\Delta(\hat{\mathcal{H}}) = \mathrm{VC}(\hat{\mathcal{D}}) = d$. The cardinality of $\hat{\mathcal{H}}_N$ equals that of $\hat{\mathcal{D}}_N$ which consists of all $\delta \in \hat{\mathcal{D}}$ that satisfy the above condition on runs of 1's. Clearly, $|\hat{\mathcal{D}}_N| = \beta_d^{(N)}(n)$. □

In the following section we aim at obtaining a closed-form approximation of $\beta_d^{(N)}(n)$.

3.2. *Approximation of* $\beta_d^{(N)}(n)$

We start with a result that estimates the function

$$S(k, m, N) \equiv w_{m,N}(k - m + 1)$$

where $w_{m,v}(n)$ is defined in Remark 2. Henceforth, denote by $\Phi(x)$ and $\phi(x)$ the normal cumulative and density probability distributions, respectively, with zero mean and unit variance.

Theorem 2. *Let μ_N and σ_N^2 be defined as in (7). Let*

$$\rho_N = \frac{1}{N+1}\left(\frac{(\mu_N+1)^{\underline{4}}}{2} + 2(\mu_N+1)^{\underline{3}} + (\mu_N+1)^{\underline{2}}\right).$$

Then for some constant $0 < \epsilon < 1$,

$$\hat{S}(k, m, N) = \frac{(N+1)^m}{\sigma_N\sqrt{m}}\phi\left(\frac{k - m(\mu_N+1) + \epsilon}{\sigma_N\sqrt{m}}\right)\mathbb{I}(k - m + 1 \geq 0) \tag{18}$$

estimates $S(k, m, N)$ to within an error which is bounded as

$$|S(k, m, N) - \hat{S}(k, m, N)| \leq (N+1)^m \frac{1.531\rho_N}{\sigma_N^3\sqrt{m}}$$

uniformly over $0 \leq k \leq mN$.

We proceed with the proof of Theorem 2.

Proof. We study the function $w(k) \equiv w_{m,N}(k)$ whose generating function is

$$W(x) = \sum_{k\geq 0} w(k)x^k.$$

As in the proof of Lemma 2, we have

$$W(x) = \left(\frac{1 - x^{N+1}}{1 - x}\right)^m \tag{19}$$

since the coefficient of x^k in the right-hand side of (19) equals the number of monomials $x^{i_1+i_2+\cdots+i_m}$ where $i_1, \ldots, i_m \in \Lambda$ with

$$\Lambda = \{0, 1, \ldots, N\}$$

and $i_1 + \cdots + i_m = k$.

Consider the m random variables

$$X_i, \quad 1 \leq i \leq m$$

which are drawn independently according to the same uniform probability distribution on Λ. Denote their sum by

$$S_X = \sum_{i=1}^{m} X_i.$$

The cumulative probability distribution $F_{S_X}(k)$ is defined for k running over all possible exponent values of the monomial $x^{i_1+\cdots+i_m}$, $i_j \in \Lambda$, $1 \leq j \leq m$. That is, $k \in \{0, 1, \ldots, mN\}$.

For any vector $v \in \Lambda^m$, due to independence, we have as its probability,

$$P(v) = \frac{1}{(N+1)^m}.$$

Then the simple relation

$$w(k) = (N+1)^m P(S_X = k) \tag{20}$$

clearly follows. Hence in order to estimate $w(k)$ we can now try to estimate the probability that $S_X = k$.

For any $1 \leq i \leq m$, the expected value is clearly

$$\mu \equiv \mu_N = \mathbb{E}(X_i) = \frac{N}{2} \tag{21}$$

and the variance

$$\sigma^2 \equiv \sigma_N^2 = \frac{1}{N+1} \sum_{k=0}^{N} (k-\mu)^2.$$

It is easy to show that $k^2 = k^{\underline{2}} + k$ where, again, for $i > 0$, $k^{\underline{i}} \equiv k(k-1)\cdots(k-i+1)$. So we have

$$\sum_{k=0}^{N} (k-\mu)^2 = 2\sum_{k=0}^{\mu} k^2 = 2\left(\sum_{k=0}^{\mu} k^{\underline{2}} + \sum_{k=0}^{\mu} k\right) = 2\left(\left.\frac{k^{\underline{3}}}{3}\right|_{k=\mu+1} + \left.\frac{k^{\underline{2}}}{2}\right|_{k=\mu+1}\right).$$

Hence it follows that

$$\sigma^2 = \frac{2}{N+1}\left(\frac{(\mu+1)^{\underline{3}}}{3} + \frac{(\mu+1)^{\underline{2}}}{2}\right). \tag{22}$$

Next, consider

$$\rho = \rho_N \equiv \mathbb{E}|X_i - \mu_N|^3.$$

As above we have

$$\begin{aligned}
\sum_{k=0}^{N} |k-\mu|^3 &= 2\sum_{k=0}^{\mu} k^3 = 2\left(\sum_{k=0}^{\mu} k^{\underline{3}} + 3k^{\underline{2}} + k\right) \\
&= 2\left(\left.\frac{k^{\underline{4}}}{4}\right|_{k=\mu+1} + \left.k^{\underline{3}}\right|_{k=\mu+1} + \left.\frac{k^{\underline{2}}}{2}\right|_{k=\mu+1}\right) \\
&= \frac{(\mu+1)^{\underline{4}}}{2} + 2(\mu+1)^{\underline{3}} + (\mu+1)^{\underline{2}}.
\end{aligned}$$

Hence

$$\rho = \frac{1}{N+1}\left(\frac{(\mu+1)^{\underline{4}}}{2} + 2(\mu+1)^{\underline{3}} + (\mu+1)^{\underline{2}}.\right) \tag{23}$$

Define the zero-mean random variables

$$Y_i = X_i - \mu, \quad 1 \leq i \leq m.$$

Clearly, the variance of Y_i is σ^2 and its third moment $\mathbb{E}|Y_i|^3 = \rho$. Consider the normalized sum

$$S_Y = \frac{\sum\limits_{i=1}^{m} Y_i}{\sqrt{m}\sigma}.$$

Then we have the following relationship between the cumulative distribution $F_{S_X}(k)$ of S_X and the cumulative distribution $F_{S_Y}(k)$ of S_Y:

$$F_{S_X}(k) = P(S_X \leq k) = P\left(\sum_i (X_i - \mu) \leq k - m\mu\right)$$

$$
\begin{aligned}
&= P\left(\sum_i Y_i \leq k - m\mu\right)\\
&= P\left(\frac{\sum_i Y_i}{\sqrt{m}\sigma} \leq (k - m\mu)/(\sqrt{m}\sigma)\right)\\
&= P\left(S_Y \leq (k - m\mu)/(\sqrt{m}\sigma)\right)\\
&= F_{S_Y}((k - m\mu)/(\sqrt{m}\sigma)).
\end{aligned}
$$

Hence we have

$$F_{S_Y}(x) = F_{S_X}(\sqrt{m}\sigma x + m\mu).$$

By a classic result from Berry (1941) and Essen (1942) (see [11], Theorem 1, p. 542) which holds more generally for *any* independent and identically distributed random variables Y_i with zero mean, variance σ^2 and third moment ρ we have for all x and m,

$$|F_{S_Y}(x) - \Phi(x)| \leq \frac{3\rho}{\sigma^3\sqrt{m}}.$$

In [8] the constant 3 (in the above bound) was sharpened down to 0.7975 and (later) to 0.7655 by [22]. It is mentioned in [20] that this result is best thus far.

Consider the distribution $F_{S_X}(k)$ where again the probability-1 support is $\{0, 1, \ldots, mN\}$. Let $\epsilon(m, N)$ be the error in approximation of F_{S_X} by Φ, which from the above, holds uniformly for all $0 \leq k \leq mN$. Then from the above we have

$$F_{S_X}(k) = \Phi\left(\frac{k - m\mu}{\sigma\sqrt{m}}\right) + \epsilon(m, N) \tag{24}$$

with the error of approximation being

$$|\epsilon(k, m, N)| \leq \frac{0.7655\rho}{\sigma^3\sqrt{m}}. \tag{25}$$

By definition of probability distribution we have $P(S_X = k) = F_{S_X}(k) - F_{S_X}(k - 1)$. Using (24) and (25) we have

$$P(S_X = k) = \Phi\left(\frac{k - m\mu}{\sigma\sqrt{m}}\right) - \Phi\left(\frac{k - 1 - m\mu}{\sigma\sqrt{m}}\right) + \eta(k, m, N)$$

where $\eta(k, m, N)$ is some function whose absolute value is in the worst case double the error $\epsilon(m, N)$, i.e.,

$$|\eta(k, m, N)| \leq \frac{1.531\rho_N}{\sigma_N^3\sqrt{m}} \equiv \overline{\eta}(m, N) \tag{26}$$

where ρ is defined in (23). Consequently with (20) we have

$$w_{m,N}(k) = (N + 1)^m \left(\Phi\left(\frac{k - m\mu_N}{\sigma_N\sqrt{m}}\right) - \Phi\left(\frac{k - 1 - m\mu_N}{\sigma_N\sqrt{m}}\right) + \eta(k, m, N)\right). \tag{27}$$

To get an estimate for $S(k, m, N)$ we substitute $k - m + 1$ for k in (27) and assume henceforth that $k \geq m - 1$. Denote by

$$\delta \equiv \delta(k, m, N) = k - m(\mu_N + 1),$$

$$h \equiv h(m, N) = \frac{1}{\sigma_N\sqrt{m}},$$

$$\Delta(k, m, N) \equiv \Phi\left((\delta + 1)h\right) - \Phi\left(\delta h\right)$$

and obtain

$$S(k,m,N) = (N+1)^m \left(\Delta(k,m,N) + \eta(k,m,N)\right). \tag{28}$$

By Cauchy's mean value theorem there exists a $\xi \in (\delta h, (\delta+1)h)$ such that (see for instance, [10], p. 171)

$$\Delta(k,m,N) = h\phi(\xi)$$

where the right-hand side above equals the standard normal probability density function evaluated at ξ. Hence for some constant $0 < \epsilon < 1$ we have $\xi = (\delta+\epsilon)h$ and substituting for $\Delta(k,m,N)$ in (28) we obtain

$$\begin{aligned} S(k,m,N) &= (N+1)^m \left(h\phi((\delta+\epsilon)h) + \eta(k,m,N)\right) \\ &= (N+1)^m \left(\frac{1}{\sigma_N\sqrt{m}}\phi\left(\frac{k-m(\mu_N+1)+\epsilon}{\sigma_N\sqrt{m}}\right) + \eta(k,m,N)\right) \end{aligned}$$

where μ_N, σ_N are defined in (21) and (22) and η satisfies (26). Therefore as an estimate of S we have

$$\hat{S}(k,m,N) = \frac{(N+1)^m}{\sigma_N\sqrt{m}}\phi\left(\frac{k-m(\mu_N+1)+\epsilon}{\sigma_N\sqrt{m}}\right)$$

for some constant $0 < \epsilon < 1$ with an approximation error $|S - \hat{S}|$ bounded above by $(N+1)^m\overline{\eta}$, where $\overline{\eta}$ is defined in (26). □

Next we state a lemma that estimates $c(k, n-k; m, N)$ (defined in (4)) which is the number of two-dimensional valid ordered m-partitions of $[k, n-k]$ satisfying (10) where a valid partition is defined according to (9).

Lemma 5. *For $n \geq k \geq m-1 \geq 1$ we have*

$$\begin{aligned} c(k,n-k;m,N) = &\binom{n-k}{m-1}(1+\alpha(k,m,N))(2N+1)^m \\ &\times \left(\frac{1}{\sigma_{2N}\sqrt{m}}\phi\left(\frac{k-m(\mu_{2N}+1)+\epsilon}{\sigma_{2N}\sqrt{m}}\right) + \eta(k,m,2N)\right) \end{aligned}$$

for some constant $0 < \epsilon < 1$, $0 < \alpha(k,m,N) \leq 2\mathrm{e}^{-(2N+1)(m-1)/k}$ and $\eta(k,m,N)$ that satisfies $|\eta(k,m,N)| \leq \overline{\eta}(m,N)$ where $\overline{\eta}$ is defined in (26).

Proof. By definition, from (4) the quantity $c(k,n-k;m,N)$ involves a sum of three terms, $w_{m,2N}(k-m+1)$, $w_{m-1,2N}(k-m-2N-1)$ and $w_{m-1,2N}(k-m-2N)$. Using Remark 2 the first equals

$$w_{m,2N}(k-m+1) = \sum_{l=0}^{m}(-1)^l\binom{m}{l}\binom{k-l(2N+1)}{m-1}. \tag{29}$$

From the definition of $w_{m,v}(n)$ (Lemma 2) it is easy to show that $w_{m-1,2N}(k-m-2N) \leq w_{m,2N}(k-m-2N)$ and $w_{m-1,2N}(k-m-2N-1) \leq w_{m,2N}(k-m-2N-1)$. We have

$$w_{m,2N}(k-m-2N) = \sum_{l=0}^{m}(-1)^l\binom{m}{l}\binom{k-l(2N+1)-(2N+1)}{m-1} \tag{30}$$

and similarly for $w_{m,2N}(k-m-2N-1)$. Hence

$$c(k,n-k;m,N) = \binom{n-k}{m-1}\sum_{l=0}^{m}(-1)^l\binom{m}{l}\binom{k-l(2N+1)}{m-1}(1+\epsilon(m,k,N,l)) \tag{31}$$

where

$$0 < \epsilon(m,k,N,l) \leq \frac{\binom{k-l(2N+1)-(2N+1)}{m-1}}{\binom{k-l(2N+1)}{m-1}} + \frac{\binom{k-l(2N+1)-2(N+1)}{m-1}}{\binom{k-l(2N+1)}{m-1}}$$

which for all $0 \le l \le m$ is bounded from above by

$$\frac{\binom{k-(2N+1)}{m-1}}{\binom{k}{m-1}} + \frac{\binom{k-2(N+1)}{m-1}}{\binom{k}{m-1}}. \tag{32}$$

Next, this is shown to be exponentially small in N. Using the standard identity of

$$\binom{k}{m} = \frac{k}{k-m}\binom{k-1}{m}$$

we have for $0 \le a \le k$,

$$\binom{k-a}{m}\Big/\binom{k}{m} = \prod_{i=0}^{a-1}\frac{k-m-i}{k-i} \le \prod_{i=0}^{a-1} \mathrm{e}^{-m/(k-i)} \tag{33}$$

where we used $1 - x \le \exp(-x)$ which holds for all $x \in \mathbb{R}$. The right-hand side equals

$$\mathrm{e}^{-m\sum_{i=0}^{a-1} 1/(k-i)}$$

which is upper bounded by $\exp(-am/k)$. The sum in (32) is thus bounded from above by some function $\alpha(k, m, N) \le 2\exp(-(2N + 1)(m - 1)/k)$ and we have

$$c(k, n-k; m, N) = \binom{n-k}{m-1}(1 + \alpha(k, m, N))\sum_{l=0}^{m}(-1)^l\binom{m}{l}\binom{k - l(2N+1)}{m-1}. \tag{34}$$

The inner summation in (34) equals $S(k, m, 2N)$ hence we may apply Theorem 2 to it and obtain $S(k, m, 2N) = \hat{S}(k, m, 2N) + (2N + 1)^m \eta(k, m, 2N)$ where $\hat{S}(k, m, N)$ is defined in (18) and $\eta(k, m, N)$ is some function which is bounded from above by $\overline{\eta}(m, N)$. Simple substitution in (34) yields the result. □

We may now proceed to approximate the function $\beta_d^{(N)}(n)$. We will use the notation $a_n \ll b_n$, $a_n \sim b_n$ to denote that $\lim_{n\to\infty} a_n/b_n$ equals 0 and 1, respectively. We henceforth assume that $d \equiv d_n \ll n$. From (5) and by Lemma 5 we have

$$\begin{aligned}\beta_d^{(N)}(n) = \sum_{k=0}^{d}\sum_{m=1}^{n}\binom{n-k}{m-1}(1 + \alpha(k, m, N))(2N+1)^m \\ \times\left(\frac{1}{\sigma_{2N}\sqrt{m}}\phi\left(\frac{k - m(\mu_{2N}+1) + \epsilon}{\sigma_{2N}\sqrt{m}}\right) + \eta(k, m, 2N)\right).\end{aligned}$$

We use

$$\sum_{k=0}^{d}\sum_{m=1}^{n}\binom{n-k}{m-1}\frac{(2N+1)^m}{\sigma_{2N}\sqrt{m}}\phi\left(\frac{k - m(\mu_{2N}+1) + \epsilon}{\sigma_{2N}\sqrt{m}}\right) \tag{35}$$

as an estimate of $\beta_d^{(N)}(n)$. We begin by treating the sum on m. Define the function

$$h(m) = h(m, n, k, N) \equiv \frac{(N+1)^m}{\sigma_N\sqrt{m}}\frac{\phi\left(\frac{k-m(\mu_N+1)+\epsilon}{\sigma_N\sqrt{m}}\right)}{\binom{k}{m-1}}. \tag{36}$$

The sum we are interested in takes the form

$$\sum_{m=1}^{n}\binom{n-k}{m-1}\binom{k}{m-1}h(m). \tag{37}$$

Let us find the maximum of the sequence

$$a_m \equiv \binom{n-k}{m-1}\binom{k}{m-1}.$$

We solve for the first m at which

$$\frac{a_{m+1}}{a_m} < 1.$$

This ratio equals

$$\frac{(n-k-(m-1))(k-(m-1))}{m^2}$$

and letting

$$m^* \equiv 1 + k(n-k)/n$$

we have

$$\frac{k^2(n-k)^2}{n^2m^2} = \left(\frac{m^*-1}{m^*}\right)^2$$

which is smaller than 1, hence m^* is the maximal point of a_m. The sequence a_m is dominated by the maximal component a_{m^*}. Hence we may approximate the sum in (37) by the simpler sum

$$h(m^*)\sum_{m=1}^{n}\binom{n-k}{m-1}\binom{k}{m-1} = h(m^*)\binom{n}{k} \tag{38}$$

where the last equality follows from the next standard identity (see [14], (5.23))

$$\sum_m \binom{l}{r+m}\binom{s}{u+m} = \binom{l+s}{l-r+u}.$$

To compute $h(m^*)$ we first treat the denominator of the right-hand side of (36). Denote by $\delta = \delta_n \equiv k/n$. We have

$$\binom{k}{m^*-1} = \binom{k}{\frac{k(n-k)}{n}} = \binom{k}{k(1-\delta)}.$$

Using Sterling's formula one has the following standard approximation of the binomial coefficients (see [10], (2.4))

$$\binom{n}{k} \sim \sqrt{\frac{n}{2\pi k(n-k)}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k}.$$

Thus

$$\binom{k}{k(1-\delta)} \sim \sqrt{\frac{1}{2\pi k\delta(1-\delta)}}\left(\frac{1}{1-\delta}\left(\frac{1-\delta}{\delta}\right)^{\delta}\right)^k \tag{39}$$

with large n. To simplify we take the natural log of the right factor and obtain

$$k\left(\ln\left(\frac{1}{1-\delta}\right) + \delta\ln\left(\frac{1-\delta}{\delta}\right)\right).$$

From the outer summation in (35) we know that $0 \le k \le d$ and by assumption $d \ll n$ hence $\delta \ll 1$. Hence we approximate $\ln(1/(1-\delta))$ by δ. Therefore the above expression is approximated by $(1 + c_3 \ln n)k\delta$ for some constant $0 < c_3 < 1$. Using this and multiplying the right-hand side of (39) by $\sigma_N\sqrt{m^*}$ yields the following approximation for the denominator of (36):

$$\sigma_N\sqrt{\frac{1}{2\pi\delta}}\exp\{k(1 + c_3\ln n)\delta\}.$$

Continuing from (36), for $h(m^*)$ we have

$$\exp\left\{-k(1+c_3\ln n)\delta - \frac{(k-(1+k(1-\delta))(\mu_N+1)+\epsilon)^2}{2\sigma_N^2(1+k(1-\delta))} + (1+k(1-\delta))\ln(N+1) - \ln\left(\sigma_N\sqrt{\frac{1}{\delta}}\right)\right\}. \tag{40}$$

Recalling that $0<\epsilon<1$, the second term inside the exponent is approximated by

$$-\frac{(1-\delta)(1-c_2)(\mu_N+1)^2k}{2\sigma_N^2}$$

for some constant $0<c_2<1$. We hence have

$$\exp\left\{A(n,N)k+\ln\left(\frac{(N+1)c_4}{\sigma_N}\right)\right\}$$

as an approximation of (40) with

$$A(n,N) = -(1+c_3\ln n)c_1 - \frac{1}{2\sigma_N^2}(1-c_1)(1-c_2)(\mu_N+1)^2 + (1-c_1)\ln(N+1)$$

and constants $0<c_1,c_4<1$. Hence the right-hand side of (38) is approximated by

$$\binom{n}{k}\frac{(N+1)c_4}{\sigma_N}e^{kA(n,N)}.$$

Continuing from (35) and computing the sum on k yields

$$\hat{\beta}_d^{(N)}(n) = \frac{(2N+1)c_4}{\sigma_{2N}}\left(1+e^{A(n,2N)}\right)^n b(n,p(n,2N),d) \tag{41}$$

as an approximation for $\beta_d^{(N)}(n)$ where

$$p(n,N) \equiv \frac{e^{A(n,N)}}{1+e^{A(n,N)}}$$

and $b(n,p,d) \equiv \sum_{k=0}^{d}\binom{n}{k}p^k(1-p)^{n-k}$ is the left tail of the binomial probability distribution with parameters n and p.

3.3. *Proof of Proposition 1*

Fix any $(x,y)\in\zeta$. The condition $\omega_h(x,y)>N$ implies that h must have a constant value of y over all elements z, $x-N-1\le z\le x+N+1$. For this x, the uniquely corresponding δ_h has a constant value of 1 over the interval $I_N(x)\equiv\{z: x-N\le z\le x+N+1\}$. By definition of $\mathcal{H}_N(\zeta)$ this holds for any $(x,y)\in\zeta$. Denote by $\mathcal{D}_N(\zeta_+)=\{\delta_h : h\in\mathcal{H}_N(\zeta)\}$ where $\zeta_+=\{x_i:(x_i,y_i)\in\zeta, 1\le i\le\ell\}$. Clearly, $|\mathcal{D}_N(\zeta_+)|=|\mathcal{H}_N(\zeta)|$. Hence we seek an upper bound on $|\mathcal{D}_N(\zeta_+)|$ for any ζ_+ and $\mathcal{H}$ with $\mathrm{VC}_\Delta(\mathcal{H})=d$.

Let $R(\zeta_+)=\bigcup_{x\in\zeta_+}I_N(x)$. Since for every $\delta\in\mathcal{D}_N(\zeta_+)$, $\delta(z)=1$ for all $z\in R(\zeta_+)$ then the cardinality of the restriction $\mathcal{D}_N(\zeta_+)_{|R(\zeta_+)}$ of the class $\mathcal{D}_N(\zeta_+)$ on the set $R(\zeta_+)$ is one. Denote by $R^c(\zeta_+)\equiv[n]\setminus R(\zeta_+)$ then we have

$$|\mathcal{D}_N(\zeta_+)| = |\mathcal{D}_N(\zeta_+)_{|R^c(\zeta_+)}|.$$

Since $\mathrm{VC}(\mathcal{D}_N(\zeta_+))\le\mathrm{VC}_\Delta(\mathcal{H})=d$ then by Lemma 1 it follows that

$$|\mathcal{D}_N(\zeta_+)_{|R^c(\zeta_+)}| \le \mathbb{S}(|R^c(\zeta_+)|,d). \tag{42}$$

We also have

$$\max\{|R^c(S)| : S\subset[n], |S|=\ell\} = n-\ell-2N-1 \tag{43}$$

which is achieved for instance by a set $S' = \{N+3, \dots, N+\ell+2\}$ with $R(S') = \{3, \dots, 2(N+1)+\ell+1\}$. Hence for any ζ_+ as above we have

$$|\mathcal{D}_N(\zeta_+)| \le \mathbb{S}(n-2N-\ell-1, d). \tag{44}$$

Since the bound of Lemma 1 is tight, there exists a class $\mathcal{D}_N(\zeta_+)$ (with a corresponding class $\mathcal{H}_N(\zeta)$) of this size. Proposition 1 follows. □

3.4. *Proof of Proposition 2*

The proof follows that of Proposition 1 up to (42) with $\mathcal{H}_N^*(\zeta)$ instead of $\mathcal{H}_N(\zeta)$. By Theorem 1 we have

$$|\mathcal{D}_N^*(\zeta_+)_{|R^c(\zeta_+)}| \le \beta_d^{(N)}(|R^c(\zeta_+)|).$$

From (43) and by the tightness of the bound in Theorem 1 it follows that there exists a class $\mathcal{D}_N^*(\zeta_+)$ and hence $\mathcal{H}_N^*(\zeta)$ of this size. Hence the statement of the proposition follows. □

Acknowledgements

The author is grateful for the comments made by the referees and to Vitaly Maiorov for commenting on an earlier draft.

References

[1] G.E. Andrews, The Theory of Partitions, Cambridge University Press, 1998.

[2] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.

[3] M. Anthony, J. Ratsaby, Maximal width learning of binary functions. Technical Report LSE-CDAM-2006-11, (Centre for Discrete and Applicable Mathematics), Department of Mathematics, London School of Economics and Political Science, October 2006.

[4] M. Anthony, G. Brightwell, C. Cooper, The Vapnik-Chervonenkis dimension of a random graph, Discrete Mathematics 138 (1–3) (1995) 43–56.

[5] R. Arratia, L. Goldstein, L. Gordon, Poisson approximation and the Chen-Stein method, Statistical Science 5 (1990) 403–434.

[6] N. Balakrishnan, M.V. Koutras, Runs and Scans with Applications, Wiley-Interscience, 2001.

[7] A.D. Barbour, O. Chryssaphinou, Compound Poisson approximation: A user's guide, The Annals of Applied Probability 11 (3) (2001) 964–1002.

[8] P. Van Beck, An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality, Probability Theory and Related Fields 23 (3) (1972) 187–196.

[9] B. Bollobás, Combinatorics: Set Systems, Hypergraphs, Families of Vectors, and Combinatorial Probability, Cambridge University Press, 1986.

[10] W. Feller, An Introduction to Probability Theory and Its Applications, second ed., vol. 1, Wiley, New York, 1957.

[11] W. Feller, An Introduction to Probability Theory and Its Applications, second ed., vol. 2, Wiley, New York, 1971.

[12] P. Frankl, On the trace of finite sets, Journal of Combinatorial Theory(A) 34 (1983) 41–45.

[13] P. Frankl, The shifting technique in extremal set theory, in: C. Whitehead (Ed.), Surveys in Combinatorics, Cambridge University Press, 1987, pp. 81–110.

[14] R.L. Graham, D.E. Knuth, O. Patashnik, Concrete Mathematics, Addison-Wesley, 1994.

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

[16] D. Haussler, E. Welzl, Epsilon-nets and simplex range queries, Discrete Computational Geometry 2 (1987) 127–151.

[17] J. Pach, P.K. Agarwal, Combinatorial Geometry, Wiley-Interscience Series, 1995.

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

[19] N. Sauer, On the density of families of sets, Journal of Combinatorial Theory (A) 13 (1972) 145–147.

[20] V.V. Senatov, Normal Approximation: New Results, Methods, and Problems, Utrecht, The Netherlands, 1998.

[21] S. Shelah, A combinatorial problem; stability and order for models and theories in infinitary languages, Pacific Journal of Mathematics 41 (1972) 247–261.

[22] I.S. Shiganov, Refinement of the upper bound of the constant in the central limit theorem, Journal of Soviet Mathematics (1986) 2545–2550.

[23] V.N. Vapnik, Statistical Learning Theory, Wiley, 1998.

[24] V.N. Vapnik, A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory of Probability and its Applications 16 (1971) 264–280.