The VC Dimension of k-fold Union

David Eisenstat
Princeton University
deisenst@cs.princeton.edu

Dana Angluin
Yale University
dana.angluin@yale.edu

Abstract

The known $O(dk \log k)$ bound on the VC dimension of k-fold unions or intersections of a given concept class with VC dimension d is shown to be asymptotically tight.

Keywords: combinatorial problems, VC dimension, concept classes, unions of concepts, intersections of concepts

1 Concept Classes and VC Dimension

A **concept** over a set of **points** X is a subset of X. A **concept class** is a set of concepts. A concept class $\mathcal{C}$ over X **shatters** a set of points $Y \subseteq X$ if $\{c \cap Y \mid c \in \mathcal{C}\} = 2^Y$, where 2^Y denotes the powerset of Y. The **VC dimension** of $\mathcal{C}$ [3] is $\sup\{|Y| \mid \mathcal{C} \text{ shatters } Y \subseteq X\}$, a quantity that may be used to bound the number of examples necessary to PAC learn any concept from $\mathcal{C}$ [1].

2 k-fold Unions

For each integer $k \geq 1$, we define $\mathcal{C}^k_\cup$ to be the set of concepts that are the unions of k concepts in $\mathcal{C}$, or k**-fold union** of $\mathcal{C}$. In symbols, $\mathcal{C}^k_\cup := \{c_1 \cup \cdots \cup c_k \mid c_1, \ldots, c_k \in \mathcal{C}\}$. Note that $c_1, \ldots, c_k$ need not be distinct, so $\mathcal{C}^{k'}_\cup \subseteq \mathcal{C}^k_\cup$ for all $k' \in \{1, \ldots, k\}$.

Blumer *et al.* [1] show that there exists real $\beta > 0$, independent of $\mathcal{C}$ and k, such that the VC dimension of $\mathcal{C}_{\cup}^{k}$ is at most $\beta \cdot dk \log k$, where d is the VC dimension of $\mathcal{C}$. This result simplifies the analysis of complex concept classes that may be expressed as k-fold unions or intersections. In the same paper, for example, Blumer *et al.* use it to bound the VC dimension of intersections of k halfspaces in $\mathbb{R}^n$.

Our main result is that $O(dk \log k)$ is asymptotically tight when no restrictions are placed on $\mathcal{C}$.

Theorem 1. *There exists real $\alpha > 0$ such that for all $d \geq 5$ and for all $k \geq 1$, there is a concept class $\mathcal{C}$ of VC dimension at most d such that $\mathcal{C}_{\cup}^{k}$ has VC dimension at least $\alpha \cdot dk \log k$.*

3 Preliminaries

Let x be a real number and n be a positive integer. $(x)_n := \prod_{j=0}^{n-1}(x-j)$ denotes the nth falling power of x. $\log x$ denotes the logarithm of x base 2. We use the following standard inequalities, which hold when $x \leq 1$ and k is a positive integer.

$$\binom{n}{k} \leq (e \cdot n/k)^k \qquad 1 - x \leq e^{-x} \qquad 1 - nx \leq (1-x)^n$$

The last of these is **Bernoulli's inequality**.

4 Desired Properties of the Construction

Let n be a positive integer and let $X := \{1, \ldots, n2^n\}$. In the next section, we construct a random concept class $\mathcal{C}$ that possesses the following properties with probability 1 in the limit as $n \to \infty$:

1. $\mathcal{C}$ contains every concept with cardinality at most 1.

2. For each $b = 1, \ldots, \lceil \log n \rceil$ and for every set $S \subseteq X$ such that $|S| > n2^{n-b}$, there exists a concept $c \in \mathcal{C}$ such that $c \subseteq S$ and $|c| \geq n/b$.

3. No two distinct $\mathcal{C}$-concepts (concepts in $\mathcal{C}$) intersect in 5 or more points.

The significance of these properties is given by the next two lemmas.

Lemma 2. *If $\mathcal{C}$ satisfies Properties 1 and 2, then for all $k \geq 3 \cdot 2^n$, $\mathcal{C}_\cup^k$ shatters X.*

Proof. It suffices to show that any set $S \subseteq X$ may be written as a union of at most $3 \cdot 2^n$ concepts in $\mathcal{C}$. If $S = \emptyset$, this is clearly true, so assume $S \neq \emptyset$ to avoid unseemly edge cases. The following greedy algorithm expresses S as a union of $\mathcal{C}$-concepts:

1. Find a $\mathcal{C}$-concept $c \subseteq S$ of maximal cardinality.

2. If $S \neq c$, express $S \setminus c \neq \emptyset$ as a union of $\mathcal{C}$-concepts and append c to the union.

Since $\mathcal{C}$ includes all one-point concepts (Property 1), c always exists, and this algorithm always terminates. Property 2 ensures that the number of terms in the union is not too large. Let

$$f(m) := \begin{cases} b/n, & \text{if } m > 2^n \text{ and } n2^{n-b} < m \leq n2^{n-b+1} \\ 1, & \text{if } m \leq 2^n \end{cases}$$

$$F(m) := \sum_{j=1}^{m} f(j).$$

Intuitively, if we assign cost 1 to each concept in the union and divide it equally among the points in that concept, $f(m)$ is the marginal cost of adding the mth point to S, and $F(m)$ is the total cost of the first m points. We first show by induction on $|S|$ that the greedy algorithm expresses S as a union of at most $F(|S|)$ concepts in $\mathcal{C}$, confirming this interpretation. We then bound $F(|S|)$.

When $1 \le |S| \le 2^n$, the conclusion is immediate, as $F(|S|) = |S|$. Otherwise, let b be the smallest integer such that $|S| > n2^{n-b}$. The concept c chosen by the algorithm satisfies $|c| \ge n/b$ by Property 2. By the inductive hypothesis, the algorithm proceeds to write $S \setminus c$ as a union of at most $F(|S| - |c|)$ concepts. Since $f(|S|) = b/n$ and f is nonincreasing, each of the $|c| \ge n/b$ terms $f(|S|-|c|+1), f(|S|-|c|+2), \ldots, f(|S|)$ is at least b/n. Thus, $F(|S|) \ge 1 + F(|S| - |c|)$, and the conclusion follows.

We now verify that $F(|S|) \le 3 \cdot 2^n$ for all S. Since F is increasing, assume $S = X$. We have

$$\begin{aligned} F(n2^n) = \sum_{j=1}^{n2^n} f(j) &= 2^n + \sum_{j=2^n+1}^{n2^n} f(j) \\ &\le 2^n + \sum_{b=1}^{\lceil \log n \rceil} n2^{n-b}(b/n) = 2^n + \sum_{b=1}^{\lceil \log n \rceil} b2^{n-b} \le 3 \cdot 2^n, \end{aligned}$$

with the last inequality following from the sum $\sum_{b=1}^{\infty} b2^{-b} = 2$. □

Lemma 3. *If $\mathcal{C}$ satisfies Property 3, then its VC dimension is at most* 5.

Proof. Suppose that $\mathcal{C}$ shatters a nonempty set $Y \subseteq X$. Let $y \in Y$. There exist two distinct concepts $c_1, c_2 \in \mathcal{C}$ such that $Y = c_1 \cap Y$ and $Y \setminus \{y\} = c_2 \cap Y$. Thus $Y \setminus \{y\} \subseteq c_1 \cap c_2$, and by Property 3, $|Y| \le |c_1 \cap c_2| + 1 < 5 + 1$. □

The construction alone will give us concept classes of VC dimension 5 whose k-fold unions have VC dimension $\Theta(k \log k)$, but Theorem 1 calls for classes of VC dimension d whose k-fold unions have dimension $\Theta(dk \log k)$. The following lemma allows us to increase the VC dimensions of a concept class $\mathcal{C}$ and its k-fold union proportionally.

Lemma 4. *If a concept class $\mathcal{C}$ over X has VC dimension d and $\mathcal{C}^k_\cup$ has VC dimension d', then for all integers $\ell \geq 1$, there exists a concept class $\mathcal{D}$ over $\{1,\dots,\ell\} \times X$ such that $\mathcal{D}$ has VC dimension ℓd, and $\mathcal{D}^k_\cup$ has VC dimension $\ell d'$.*

Proof. A similar argument appears in the proof of Proposition 1 of [2]. Let

$$\mathcal{D} := \{(\{1\} \times c_1) \cup \cdots \cup (\{\ell\} \times c_\ell) \mid c_1, \dots, c_\ell \in \mathcal{C}\},$$

the **tagged ℓ-fold union** of $\mathcal{C}$. A set $Y \subseteq \{1,\dots,\ell\} \times X$ is shattered by $\mathcal{D}$ if and only if for all $j \in \{1,\dots,\ell\}$, the set $\{x \mid (j,x) \in Y\}$ is shattered by $\mathcal{C}$. Thus, the VC dimension of $\mathcal{D}$ is ℓd. Since $\mathcal{D}^k_\cup$ is equal to the tagged ℓ-fold union of $\mathcal{C}^k_\cup$, its VC dimension is $\ell d'$ by a similar argument. $\square$

5 The Construction

We construct the random concept class $\mathcal{C} := \bigcup_{b=0}^{\lceil \log n \rceil} \mathcal{C}_b$, where each $\mathcal{C}_b$ is defined as follows. Take $\mathcal{C}_0 := \{\emptyset\} \cup \{\{x\} \mid x \in X\}$ to ensure that $\mathcal{C}$ satisfies Property 1. For $b = 1, \dots, \lceil \log n \rceil$, let $a := \lceil n/b \rceil$ and $s := n2^{n-b}$. Let t be a positive integer to be determined later and take $\mathcal{C}_b := \{c_1, \dots, c_t\}$, where $c_1, \dots, c_t$ are independent random a-point concepts. By choosing t to be large enough, we ensure that $\mathcal{C}$ satisfies Property 2 with probability 1 in the limit.

Let $S \subseteq X$ be any s-point set. We call S **covered** if there exists j such that $c_j \subseteq S$ and **uncovered** otherwise; if all s-point sets are covered, then $\mathcal{C}$ satisfies Property 2 for the b under consideration. The probability that $c_j \subseteq S$ is

$$p := \binom{s}{a} \Big/ \binom{s2^b}{a}.$$

We bound p as follows:

$$p = (s)_a/(s2^b)_a \geq (s-a)^a/(s2^b)^a = 2^{-ab}(1-a/s)^a \geq 2^{-ab}(1-a^2/s),$$

using Bernoulli's inequality for the last step. Noting that $ab \leq n + \lceil \log n \rceil$ and also that

$$a \leq n \qquad\qquad s \geq 2^{n-1} \qquad\qquad a^2/s \leq n^2 \cdot 2^{-n+1},$$

we have $p \geq 2^{-(1+o(1))n}(1 - 2^{-(1-o(1))n}) = 2^{-(1+o(1))n}$.

The probability that some s-point set is uncovered is no greater than the expected number of uncovered sets, or $\binom{s2^b}{s}(1-p)^t$. Thus, when $t \geq (n + \ln\binom{s2^b}{s})/p$, the former quantity is bounded by

$$\binom{s2^b}{s}(1-p)^t \leq \binom{s2^b}{s}e^{-pt} \leq e^{-n}.$$

We analyze the asymptotic order of the numerator with the standard binomial bound given in Section 3, remembering that $s \leq n2^{n-1}$:

$$n + \ln\binom{s2^b}{s} \leq n + s\ln(e \cdot 2^b) = n + s(1 + b\ln 2) = 2^{(1+o(1))n}.$$

Regardless of b, then, choosing $t := 2^{(2+o(1))n}$ ensures that the probability of some set being uncovered is at most e^{-n}. $\mathcal{C}$ contains at most $1 + n2^n + \lceil \log n \rceil t = 2^{(2+o(1))n}$ concepts in total and satisfies Property 2 with probability at least $1 - \lceil \log n \rceil e^{-n}$.

We show separately that $\mathcal{C}$ satisfies Property 3 with probability 1 in the limit. In particular, we do not condition the distribution of $\mathcal{C}$ on satisfying Property 2: the key fact here is that most of $\mathcal{C}$ is made up of a fixed number of independently chosen random concepts, all of which have prescribed cardinalities of at most n points. If $0 \leq n_1, n_2 \leq n$, the probability that a random n_1-point concept intersects an independently chosen random n_2-point concept in at least 5 points is at most

$$q := \sum_{j=5}^{n} \binom{n}{j} \binom{n2^n - n}{n - j} \Big/ \binom{n2^n}{n},$$

with equality when $n_1 = n_2 = n$. (The summand is the probability that the n-point concepts intersect in exactly j points.) For large n, the $j = 5$ term of the sum is the largest, and we have the inequalities

$$\begin{aligned} q &\le (n-4) \cdot \binom{n}{5} \binom{n2^n - n}{n - 5} \Big/ \binom{n2^n}{n} \le n \cdot n^5 \left(\frac{(n2^n - n)^{n-5}}{(n-5)!} \right) \left(\frac{n!}{(n2^n - n)^n} \right) \\ &\le n^6 (n2^n - n)^{-5} (n)_5 \le n(2^n - 1)^{-5} n^5 \le n^6 (2^n - 1)^{-5} = 2^{-(5 - o(1))n}. \end{aligned}$$

Ignoring the $1 + n2^n$ concepts in $\mathcal{C}_0$, which are too small to have a 5-point intersection, $\mathcal{C}$ consists of $2^{(2+o(1))n}$ random concepts, making $2^{(4+o(1))n}$ pairs that can potentially violate Property 3. The probability that a particular pair constitutes a violation is at most q, so the probability that any violation occurs is at most $q \cdot 2^{(4+o(1))n} = 2^{-(1-o(1))n}$. $\mathcal{C}$, therefore, satisfies Properties 1, 2, and 3 with probability 1 in the limit.

We are now ready to prove Theorem 1. Given d and k, we take n to be the largest integer such that $3 \cdot 2^n \le k$, namely $n := \lfloor \log(k/3) \rfloor$. If k is sufficiently large, the preceding construction yields a concept class $\mathcal{C}$ with VC dimension at most 5 such that $\mathcal{C}_{\cup}^{3 \cdot 2^n}$, and thus $\mathcal{C}_{\cup}^{k}$, has VC dimension $n2^n$. Otherwise, let $\mathcal{C} := \{\emptyset, \{1\}\}$. Chosen in this way, $\mathcal{C}$ has VC dimension $\Theta(k \log k)$.

Applying Lemma 4 to $\mathcal{C}$ with expansion parameter $\ell := \lfloor d/5 \rfloor$, we obtain a concept class $\mathcal{D}$ with VC dimension at most $5\ell \le d$ such that $\mathcal{D}$ has VC dimension $\ell \cdot \Theta(k \log k)$, which is within a constant factor α of $dk \log k$. This concludes our proof of Theorem 1.

6 Discussion

The construction described in this paper also yields lower bounds on the maximum VC dimension of intersections, by duality under complementation, and symmetric differences, since the greedy algorithm of Lemma 2 always produces disjoint unions.

It is natural to ask about the VC dimension of $\mathcal{C}_{\cup}^{k}$ when $\mathcal{C}$ has VC dimension $d < 5$. A lower bound of dk is easily obtained by applying Lemma 4 with $\ell := d$ to the class $\{\emptyset\} \cup \{\{x\} \mid x \in \{1, \ldots, k\}\}$. This bound is in fact tight for $d = 1$, since if $\mathcal{C}_{\cup}^{k}$ shatters a $(k+1)$-point set Y, some concept $c \in \mathcal{C}$ contains two points $y_1, y_2 \in Y$, implying that $\mathcal{C}$ shatters the set $\{y_1, y_2\}$ and thus that $d > 1$.

For $2 \leq d \leq 4$, there is a gap between the known $O(k \log k)$ upper bound and the known $\Omega(k)$ lower bounds. Reyzin [2] gives an explicit construction that achieves $(8/5)dk$ for $d \in \{2, 4\}$ and infinitely many k. Also open is whether the $O(nk \log k)$ bound obtained by Blumer *et al.* on the VC dimension of intersections of k halfspaces in $\mathbb{R}^n$ is tight.

7 Acknowledgments

The authors would like to thank Lev Reyzin and the anonymous referees for their helpful comments.

References

[1] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[2] Lev Reyzin. Lower bounds on the VC dimension of unions of concept classes. Technical Report 1349, Yale University Department of Computer Science, April 2006.

[3] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.