THE VC-DIMENSION OF A CLASS OF MULTIPLES OF THE PRIMES, AND A CONNECTION TO ADABOOST

ANDREW M. THOMAS

ABSTRACT. We discuss the VC-dimension of a class of multiples of integers and primes (equivalently indicator functions) and demonstrate connections to prime counting functions. Additionally, we prove limit theorems for the behavior of an empirical risk minimization rule as well as the weights assigned to the output hypothesis in AdaBoost for these "prime-identifying" indicator functions, when we sample m_n i.i.d. points uniformly from the integers $\{2, \ldots, n\}$.

1. THE VC-DIMENSION OF A CLASS OF MULTIPLES OF PRIMES

Let $\mathbb{N}$ denote the positive integers. Recall that for functions $f, g : \mathbb{N} \to (0, \infty)$ we say $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = \infty$. Consider a learning problem with domain $\mathcal{X} = \{n \in \mathbb{N} : n \geq 2\}$, label set $\{0, 1\}$ and hypothesis class $\mathcal{H}' := \{h_p : p \in \mathcal{X}, p \text{ is prime}\}$, where

$$h_p(x) = \begin{cases} 1 & \text{if } x \leq p \text{ or } p \nmid x, \\ 0 & \text{otherwise.} \end{cases}$$

We have $h_p(x) = 0$ if and only if x is divisible by p—i.e., $p \mid x$ and $x > p$. Note that if x is prime then $h_p(x) = 1$. Another characterization is that $h_p(x) = 0$ if and only if $x \in p\mathbb{N} \setminus \{p\} = \{2p, 3p, 4p, \ldots\}$. In this way, our efforts can be seen as assessing the sample complexity of a hypothesis class of indicator functions associated to the Sieve of Erastosthenes. A notable study containing results related to our investigations of VC-dimension is [2].

A class of indicator functions $\mathcal{G} = \big\{g : \mathcal{X} \to \{0, 1\}\big\}$ is said to *shatter* a set $C = \{c_1, \ldots, c_\ell\} \subset \mathcal{X}$ if the cardinality

$$\Big|\big\{(g(c_1), \ldots, g(c_\ell)) : g \in \mathcal{G}\big\}\Big| = 2^\ell.$$

The *VC-dimension* of $\mathcal{G}$, denoted $\text{VCdim}(\mathcal{G})$ is the size of the largest set that $\mathcal{G}$ shatters. See [4] for more details. To begin, we will show that $\mathcal{H}'$ can shatter a set of size 2. Let $C = \{6, 10\}$. Then $(h_2(6), h_2(10)) = (0, 0)$, $(h_3(6), h_3(10)) = (0, 1)$, $(h_5(6), h_5(10) = (1, 0)$ and $(h_7(6), h_7(10)) = (1, 1)$. Notice there are $2^2 - 1 = 3$ unique prime factors in $6 = 2 \cdot 3$ and $10 = 2 \cdot 5$. Thus, $\text{VCdim}(\mathcal{H}') \geq 2$. We will now show that $\text{VCdim}(\mathcal{H}') \geq n$

2010 *Mathematics Subject Classification.* 11B25, 68Q32, 60F15.

for any $n \geq 3$. Let us enumerate the first 2^n primes $p_1 < p_2 < \cdots < p_{2^n}$ and recognize that

$$2^n = \sum_{k=1}^{n} \binom{n}{k}.$$

Each c_i will be the product of 2^{n-1} primes. Label all 2^n subsets of $\{1, \ldots, n\}$ in some fashion, e.g. $A_1, \ldots, A_{2^n}$. We may now form c_i, $i = 1, \ldots, n$ as follows:

$$c_i := \prod_{k=1}^{2^n} p_k^{\mathbf{1}_{A_k}(i)}. \tag{1}$$

We may take a second to see that

$$\sum_{k=1}^{2^n} \mathbf{1}_{A_k}(i) = \sum_{j=0}^{n-1} \binom{n-1}{j} = 2^{n-1},$$

as i must be in each set and there are $\binom{n-1}{j}$ other options for a set containing index i of cardinality $j+1$. For simplicity, denote $h_{p_k} = h_k$, $k = 1, \ldots, 2^n$. We aim to prove that $h_k(c_i) = 1 - \mathbf{1}_{A_k}(i)$. By definition, $h_k(c_i) = 0$ if and only if $p_k \mid c_i$ and $c_i > p_k$. However, when $n \geq 2$ then c_i is the product of $2^{n-1} \geq 2$ primes, so that if $p_k \mid c_i$, then $c_i > p_k$. Hence, $h_k(c_i) = 0$ if and only if $p_k \mid c_i$. However, $p_k \mid c_i$ if and only if $i \in A_k$, as then and only then will p_k appear in c_i's factorization. Thus, $h_k(c_i) = 0$ if and only if $\mathbf{1}_{A_k}(i) = 1$, or $h_k(c_i) = 1 - \mathbf{1}_{A_k}(i)$. We have thus proved that $\mathcal{H}'$ shatters a set of size n for all $n \geq 2$, hence

$$\mathrm{VCdim}(\mathcal{H}') = \infty.$$

All finite classes have finite VC-dimension, hence the infinitude of the set of primes is equivalent to $\mathrm{VCdim}(\mathcal{H}') = \infty$. Thus, we have established a theorem which relates Euclid's second theorem to the VC-dimension of the class $\mathcal{H}'$.

Theorem 1.1. *The following statements are equivalent:*

(i) The set of primes is infinite

(ii) $\mathrm{VCdim}(\mathcal{H}') = \infty$

We have two interesting corollaries as a result of the above. We state the first here.

Corollary 1.2. *Suppose that $\mathcal{H}'$ shatters a set $\{c_1, \ldots, c_n\}$. Then c_i is divisible by the product of 2^{n-1} distinct primes and $\prod_{i=1}^{n} c_i$ has at least $2^n - 1$ distinct prime factors*

Proof. If $\mathcal{H}'$ shatters $\{c_1, \ldots, c_n\}$, then we have for each subset A_k of $\{1, \ldots, n\}$ that

$$\big(h(c_1), \ldots, h(c_n)\big) = \Big(1 - \mathbf{1}_{A_k}(1), \ldots, 1 - \mathbf{1}_{A_k}(n)\Big),$$

for some $h \in \mathcal{H}'$, which we assign a unique prime p_{j_k}. Therefore, $h(c_i) = 0$, i.e. $p_{j_k} \mid c_i$ and $c_i > p_{j_k}$ if and only if $i \in A_k$. As a result of the shattering,

$$c_i > \max_k p_{j_k}^{\mathbf{1}_{A_k}(i)}, \tag{2}$$

and

$$\prod_{k=1}^{2^n} p_{j_k}^{\mathbf{1}_{A_k}(i)} \mid c_i. \tag{3}$$

The last statement in the corollary follows from the fact that when $A_k \neq \emptyset$, then there exists some $i \in A_k$ such that $p_{j_k} \mid c_i$. □

Note that (2) is automatically satisfied by (3), as all primes are at least 2 and $\prod_{k=1}^{2^n} p_{j_k}^{\mathbf{1}_{A_k}(i)}$ consists of precisely 2^{n-1} primes. Hence, if there exists primes $p_{j_1}, \ldots, p_{j_{2^n}}$ such that (3) holds, then (2) is satisfied, so that $h_{p_{j_k}}(c_i) = 0$ for all k such that $i \in A_k$.

Now suppose that we have the class $\mathcal{H}'_{\leq n} = \{h_p : p \leq n\}$, and a set C of size $j(n) = \lfloor \log_2 \pi(n) \rfloor$, where $C = \{c_1, \ldots, c_{j(n)}\}$ and each c_i defined as at (1). We may shatter C as there are $2^{j(n)} \leq \pi(n)$ primes available to us. Additionally $|\mathcal{H}'_{\leq n}| \leq \pi(n)$. Thus, we have an interesting corollary to Theorem 1.1.

Corollary 1.3. *For the hypothesis class $\mathcal{H}'_{\leq n} = \{h_p : p \leq n\}$, we have*

$$\mathrm{VCdim}(\mathcal{H}'_{\leq n}) = \lfloor \log_2 \pi(n) \rfloor.$$

One thing that we may conclude from Theorem 1.1 is that the class $\mathcal{H}'$ of functions which cross out the multiples of primes (resp. positive integers greater than 1), is not *uniformly learnable* (in a probably approximately correct sense)—cf. [4].

2. The hopelessness of boosting the primes

The original consideration of the problem described above was under the pretense that one might be able to "learn" the primes if one could leverage a knowledge of divisibility by a certain integer. That is, could you use some combination of elementary heuristics in $\mathcal{H}'$ to determine whether or not a number is prime? The premise of using simple "rules of thumb" underlies the idea behind AdaBoost (cf. [5] for a detailed account). How good is the output of AdaBoost using a base hypothesis class $\mathcal{H}'$? As it turns out, when one has a sufficiently large uniform random sample from $\mathcal{X}_n$, AdaBoost is powerless to try to predict which numbers are primes and which aren't. The primes are simply too rich for $\mathcal{H}'$ to "understand".

Before continuing to the results, let us show that there is 1) no loss in generality in considering only the divisibility by primes and that 2) weak learning isn't possible. We address the first point by considering the larger hypothesis class $\mathcal{H} = \{h_d : d \in \mathcal{X}\}$ for h_d defined analogously to h_p above, though we allow our d to be *any* integer greater than or equal to 2, not just a prime. If our instances $X_1, \ldots, X_m$ are labelled by whether or not they are prime—i.e. $r(x) = 1$ if x is prime and 0 otherwise—then any algorithm that returns an *empirical risk minimization* (ERM) hypothesis will return only a hypothesis in $\mathcal{H}'$. We will denote our sample as $S = (X_1, r(X_1)), \dots, (X_m, r(X_m))$. Choose nonnegative constants $\mathbf{a} = (a_1, \dots, a_m)$ such that $\sum_{i=1}^m a_i = 1$, and define the weighted empirical risk of h_d as

$$\begin{aligned} L_{\mathbf{a}}(S,d) &:= \sum_{i=1}^{m} a_i \mathbf{1}\{h_d(X_i) \neq r(X_i)\} \\ &= \sum_{i\colon r(X_i)=0} a_i \mathbf{1}\{h_d(X_i) = 1\} \end{aligned}$$

as only composite numbers factor into the error. For ease of exposition, define

$$D(S,d) := \sum_{i\colon r(X_i)=0} a_i \mathbf{1}\{X_i \text{ is divisible by } d\} \mathbf{1}\{X_i > d\},$$

so that $L_{\mathbf{a}}(S,d) = 1 - D(S,d)$. Now define

$$d_S = \min\Big\{d \in \mathcal{X} : D(S,d) = \max_{k\in\mathcal{X}} D(S,k)\Big\}.$$

It is straightforward to show that $h_S := h_{d_S}$ is an ERM rule, as the training error for h_d with respect to the prime labeling function is $1 - D(S,d)$. We aim to show that d_S is prime. To show this, suppose that $p \mid d_S$. Then $p \leq d_S$ and if $d_S \mid x_i$ and $x_i > d_S$ then $p \mid x_i$ and $x_i > p$. Thus

$$D(S, d_S) \leq D(S,p),$$

but since $D(S,d_S)$ is maximal, $D(S,d_S) = D(S,p)$, and hence $p \geq d_S$, thus $p = d_S$. However, and perhaps no surprise due to the VC-dimension of $\mathcal{H}'$, the error $L_{\mathbf{a}}(S,d)$ can be arbitrarily large (depending on $\mathbf{a}$). Consider a sample of m values consisting of products of consecutive primes, e.g. $x_1 = p_1p_2, x_2 = p_3p_4, \dots, x_m = p_{2m-1}p_{2m}$. All of these numbers are composite so that $L_{\mathbf{a}}(S,d) \geq 1 - \max_i a_i$. With this established, we cannot necessarily find what is called a "weak learner" h_d such that $L_{\mathbf{a}}(S,d) \leq 1/2 - \gamma$, where $\gamma > 0$. There is clearly no hope of boosting this hypothesis class to discover the primes—see [5] for more information on weak learning.

There is something more quantitatively forceful we can say about the hopelessness of our situation. We consider the stochastic distribution of the weights derived by running AdaBoost with $\mathcal{H}'$ (or $\mathcal{H}$) as the base hypothesis class and where the underlying distribution is uniform on $\{2, \dots, n\}$. Let $\mathcal{X}_n := \{2, \dots, n\} \subset \mathcal{X}$ and for a probability distribution $\mathcal{D}$ define the generalization error to be $L_{\mathcal{D}}(h_d) := \mathbb{P}(h_d(X) \neq r(X))$ and the empirical risk $L_S(h_d) := L_{\mathbf{a}}(S,d)$ where $\mathbf{a} = (1/m, \dots, 1/m)$.

Proposition 2.1. *Consider the hypothesis class $\mathcal{H}$ or $\mathcal{H}'$. Suppose the $\mathcal{D}_n$ is the uniform distribution on $\{2, \dots, n\}$ and we have the prime labeling function $x \mapsto r(x)$ for the instances $X_1, \dots, X_{m_n}$ in our sample S_n. Then if $m_n = \omega\big((\log n)^2\big)$, we have the following convergence in probability,*

$$L_{\mathcal{D}_n}(h_{S_n}) \xrightarrow{P} 1/2,$$

as $n \to \infty$. Furthermore, if we have $m_n = \omega\big((\log n)^3\big)$, then we have convergence of the generalization error

$$L_{\mathcal{D}_n}(h_{S_n}) \overset{\text{a.s.}}{\to} 1/2, \quad n \to \infty,$$

where $\overset{\text{a.s.}}{\to}$ denotes almost sure convergence, or convergence with probability 1.

Proof. We begin by denoting

$$t(n) := \big|\{x \in \mathcal{X} : x \leq n,\, x \text{ is even}\}\big| = \begin{cases} (n-1)/2 & \text{if } n \text{ is odd },\\ n/2 & \text{if } n \text{ is even.}\end{cases}$$

It follows that $|t(n)/(n-1) - 1/2| \leq \frac{1}{2(n-1)}$. Notice that

$$\begin{aligned} L_{\mathcal{D}_n}(h_2) &= \mathbb{P}(h_2(X) = 1, r(X) = 0)\\ &= \mathbb{P}(r(X) = 0) - P(h_2(X) = 0, r(X) = 0)\\ &= \mathbb{P}(r(X) = 0) - P(h_2(X) = 0)\\ &= 1 - \frac{\pi(n)}{n-1} - \frac{t(n)-1}{n-1}, \end{aligned}$$

hence $|L_{\mathcal{D}_n}(h_2) - 1/2| \leq (3/2 + \pi(n))/(n-1) \to 0$, $n \to \infty$. So there exists some N such that if $n \geq N$ then

$$|L_{\mathcal{D}_n}(h_2) - 1/2| \leq (3/2 + \pi(n))/(n-1) \leq \epsilon/2.$$

Elementary union bounds yield that

$$\begin{aligned} &\mathbb{P}(|L_{\mathcal{D}_n}(h_{S_n}) - 1/2| > \epsilon)\\ &\quad\leq \mathbb{P}(|L_{\mathcal{D}_n}(h_{S_n}) - L_{\mathcal{D}_n}(h_2)| > \epsilon/2) + \mathbf{1}\big\{|L_{\mathcal{D}_n}(h_2) - 1/2| > \epsilon/2\big\}\\ &\quad\leq \mathbb{P}(d_{S_n} \neq 2), \end{aligned}$$

for $n \geq N$. We aim to show that $\mathbb{P}(d_{S_n} \neq 2) \to 0$, as $n \to \infty$. Define a sequence of random variables Y_i, $i = 1, 2, \ldots$ where $Y_i = 1$ if X_i is even and not prime, $Y_i = 0$ if X_i is prime and $Y_i = -1$ if X_i is odd and not prime.

$$\mathbb{P}(Y_i = 1) = \frac{t(n)-1}{n-1} > \mathbb{P}(Y_i = -1) = \frac{n - t(n) - \pi(n)}{n-1}$$

Therefore, we can define a random walk $T_n := \sum_{i=1}^{m_n} Y_i$. If $T_n \geq 0$, then $d_{S_n} = 2$, and vice versa. Hence, $\mathbb{P}(d_{S_n} \neq 2) = \mathbb{P}(T_n < 0)$. We will use an approximation to this probability, via Hoeffding's inequality. We note that

$$\mu_n := \mathbb{E}[Y_i] = \frac{2\big(t(n)-1\big) + \pi(n)}{n-1} - 1,$$

Additionally, standard properties of $\pi(n)$ imply that there exists some $N_0 \geq N$ such that

$$\mu_n \geq \frac{1}{\log n} - \frac{1}{n} \geq \frac{1}{2\log n},$$

for all $n \geq N_0$. We see that as $\mu_n > 0$, then

$$\mathbb{P}(T_n < 0) = P(T_n/m_n < \mu_n - \mu_n) \leq e^{-\frac{m_n \mu_n^2}{2}},$$

by Hoeffding's inequality and we know that for $n \geq N_0$ we have

$$m_n \mu_n^2 \geq \frac{m_n}{8(\log n)^2} \to \infty, \quad n \to \infty,$$

by the assumption on m_n. If $m_n = \omega\big((\log n)^3\big)$, then in particular there exists some $\delta > 0$ such that $m_n/(8 \log n)^2 \geq (1+\delta) \log n$ for large enough n. As $\sum_{n=1}^{\infty} n^{-(1+\delta)} < \infty$, we have almost sure convergence by the Borel-Cantelli lemma. □

In AdaBoost, given a sample S, we have at round $t \in \mathbb{N}$ a weight w_t for our chosen hypothesis $h_t \in \mathcal{H}'$. We denote w_t as W_t when S is seen as random as opposed to fixed/observed. Suppose that our instances are generated according to $\mathcal{D}_n$ and are labelled according to the prime labeling function $x \mapsto r(x)$. *We suppose that* 0 *is redefined as* -1 *in our indicator functions* to comport with the typical setting. One question is: as $n \to \infty$ (supposing that $m_n \to \infty$ as well) what is the stochastic behavior of the random weights $(W_{t,n}, t \in \mathbb{N})$? We have already established that the ERM threshold d_{S_n} for the first round of AdaBoost becomes 2 for large enough n (with probability 1) in the proof of Proposition 2.1. In the first round of AdaBoost, we set each $a_i = 1/m_n$ with $\epsilon_{1,n} = L_{S_n}(h_{S_n})$ and hence we get that

$$W_{1,n} = \frac{1}{2} \log\left(\frac{1}{\epsilon_{1,n}} - 1\right) \overset{\text{a.s.}}{\to} 0, \quad n \to \infty,$$

because $\big|L_{S_n}(h_{S_n}) - L_{\mathcal{D}_n}(h_{S_n})\big| \overset{\text{a.s.}}{\to} 0, n \to \infty$ when $m_n = \omega\big((\log n)^3\big)$, by Hoeffding's inequality, the triangle inequality and Proposition 2.1. We can extend this result quite a bit further. Before beginning, note that

$$\epsilon_{t,n} := \min_{h \in \mathcal{H}'} \sum_{i=1}^{m_n} D_i^{(t)} \mathbf{1}\{h(X_i) \neq r(X_i)\},$$

where $D_i^{(t)}$ are weights summing to 1 (defined at (4)) and h_t is the hypothesis in $\mathcal{H}'$ which minimizes $\epsilon_{t,n}$. Furthermore, $W_{t,n}$ is defined as

$$W_{t,n} := \frac{1}{2} \log\Big(\frac{1}{\epsilon_{t,n}} - 1\Big).$$

Theorem 2.2. *Suppose that $m_n = \omega\big((\log n)^3\big)$ and that $W_{t,n}$ are the weights for the t^{th} round AdaBoost hypothesis based on the sample S_n and uniform distribution $\mathcal{D}_n$. Then we have for all $t \in \mathbb{N}$ that*

$$W_{t,n} \overset{a.s.}{\to} 0, \quad n \to \infty.$$

Proof. We will define $D_i^{(t)}$ below. We have already established that $W_{1,n} \to 0$ a.s. as $n \to \infty$. We will now show that $\mathbb{P}\big(\lim_{n\to\infty} W_{t,n} = 0\big) = 1$ for each $t \in \mathbb{N}$. We proceed by induction (we have already demonstrated the base case) and suppose that for all

$s < t+1$ that $\mathbb{P}\big(\lim_{n\to\infty} W_{s,n} = 0\big) = 1$. We begin by noting that for any $1 \le i \le m_n$ that

$$D_i^{(t+1)} := \frac{D_i^{(t)} e^{-W_{t,n} r(X_i) h_t(X_i)}}{\sum_{i=1}^{m_n} D_i^{(t)} e^{-W_{t,n} r(X_i) h_t(X_i)}}, \tag{4}$$

where $D_i^{(1)} = 1/m_n$. Hence, we establish that

$$D_i^{(t)} e^{-2W_{t,n}} \le D_i^{(t+1)} \le D_i^{(t)} e^{2W_{t,n}}.$$

Therefore,

$$\frac{1}{m_n} \exp\Big(-2\sum_{j=1}^{t} W_{j,n}\Big) \le D_i^{(t+1)} \le \frac{1}{m_n} \exp\Big(2\sum_{j=1}^{t} W_{j,n}\Big),$$

for all n, and

$$\exp\Big(-2\sum_{j=1}^{t} W_{j,n}\Big) L_{S_n}(h_{S_n}) \le \epsilon_{t+1,n} \le \exp\Big(2\sum_{j=1}^{t} W_{j,n}\Big) L_{S_n}(h_{S_n}),$$

which implies that $\epsilon_{t+1,n} \overset{\text{a.s.}}{\to} 1/2$, by the induction hypothesis. Therefore, $W_{t+1,n} \overset{\text{a.s.}}{\to} 0$ as desired. □

What Theorem 2.2 says is that the output hypothesis of AdaBoost,

$$H_n^T(x) := \operatorname{sign}\left(\sum_{t=1}^{T} W_{t,n} h_t(x)\right),$$

for any fixed $T \in \mathbb{N}$ and a sufficiently large sample size n, is pretty much useless in learning the primality of the integers. At the very least, it is no better than a function that predicts all odd numbers are prime.

3. VC-dimension of finite arithmetic progressions

We conclude by returning a variation of our original problem and considering the class $\mathcal{H}_k = \{h_{d,k} : d \in \mathcal{X}\}$, $k \in \mathcal{X}$ where $h_{d,k}(x) = 0$ if and only if $x \in \{2d, \ldots, kd\}$. Suppose that $\mathcal{H}'_k$ is the restriction of $\mathcal{H}_k$ where d is prime. These classes also cannot misidentify the primes, as with $\mathcal{H}$ and $\mathcal{H}'$. The VC-dimension of $\mathcal{H}_2$ and $\mathcal{H}'_2$ both equal 1 and this can be seen by noting that if $c_1 = 2d$ and $c_1 \ne c_2$, then $c_2 \ne 2d$ so that if $h_{d,2}(c_1) = 1$ then $h_{d,2}(c_2) = 0$. Consider the set $C = \{12, 18\}$. Then $(h_{6,3}(12), h_{6,3}(18)) = (0,0)$, $(h_{4,3}(12), h_{4,3}(18)) = (0,1)$, $(h_{9,3}(12), h_{9,3}(18)) = (1,0)$ and $(h_{7,3}(12), h_{7,3}(18)) = (1,1)$. Thus, $\mathrm{VCdim}(\mathcal{H}_3) = 2$. Suppose that $\mathcal{H}_4$ shatters a set $\{c_1, c_2, c_3\}$ of size 3. Then, potentially relabeling instances, $c_1 = 2d$, $c_2 = 3d$ and $c_4 = 4d$ for some d. However, we must also have $c_1 = 3d'$ and $c_3 = 4d'$ or $c_1 = 2d'$ and $c_3 = 3d'$, which is impossible. Hence, $\mathrm{VCdim}(\mathcal{H}_4) = 2$.

As a final example, let us look at the VC-dimension of $\mathcal{H}'_3$. To shatter a set $\{c_1, c_2\}$, we must have $c_1 = 2p = 3q$ and $c_2 = 3p = 2r$ for some primes $q < p < r$. Therefore, $2 \mid p$ and $3 \mid p$, which is impossible as p is prime. Hence $\mathrm{VCdim}(\mathcal{H}'_3) = 1$. As usual, we will denote the number of primes less than or equal to x as $\pi(x)$. It will help to state a short lemma, similar to Corollary 1.2.

Lemma 3.1. *For $\mathcal{H}_k$ to shatter a set $C = \{c_1, \ldots, c_\ell\}$, it must be the case that each c_i has at least $2^{\ell-1}$ distinct factors $d_1, \ldots, d_{2^{\ell-1}}$, $d_i \in \mathcal{X}$ for $i = 1, 2, \ldots, 2^{\ell-1}$. As result of the structure of $\mathcal{H}_k$, if shattering occurs, we must also have that $2^{\ell-1} \leq k-1$.*

Proof. If $\mathcal{H}_k$ shatters C, then $h_{d,k}(c_i) = 0$ for $2^{\ell-1}$ distinct values of d. In other words $c_i = a_{i,1}d_1 = \cdots = a_{i,m_i}d_m$, where $m_i \geq 2^{\ell-1}$ where all $a_{i,j} \leq k$ for $1 \leq j \leq m_i$. As the d_j are distinct, so must the $a_{i,j}$ be. Therefore, we must have $m_i \leq k-1$. □

Our final result gives a bound on the VC-dimension of $\mathcal{H}_k$ and $\mathcal{H}'_k$. This result is conceptually similar to Theorem 3.1 in [2], which establishes a formula for the VC-dimension of subsets consisting of all multiples of d lying between $-n$ and n.

Proposition 3.2.

$$\lfloor \log_2 \pi(\eta_k) \rfloor \leq \mathrm{VCdim}(\mathcal{H}'_k) \leq \mathrm{VCdim}(\mathcal{H}_k) \leq \lceil \log_2(k-1) \rceil + 1$$

where

$$\eta_k := \lfloor (\log_2 k)/2 \rfloor.$$

Proof. Suppose first that $\ell \geq \lceil \log_2(k-1) \rceil + 2$. if $\mathcal{H}_k$ were to shatter $C = \{c_1, \ldots, c_\ell\}$, then Lemma 3.1 implies that $2^{\ell-1} \leq k-1$, or equivalently that

$$\begin{aligned} \ell - 1 &\leq \log_2(k-1) \\ \Rightarrow \lceil \log_2(k-1) \rceil + 1 &\leq \log_2(k-1), \end{aligned}$$

a contradiction. Hence, $\mathrm{VCdim}(\mathcal{H}_k) \leq \lceil \log_2(k-1) \rceil + 1$.

Now consider a set of size

$$\ell \leq \lfloor \log_2 \pi(n) \rfloor,$$

where $n \leq k$ so that there are at least $2^\ell \leq \pi(n)$ distinct primes less than or equal to n in the set $\{2, \ldots, k\}$. Enumerate these primes $p_1 < p_2 < \cdots < p_{2^\ell} \leq n \leq k$ and define

$$c_i := \prod_{j=1}^{2^\ell} p_j^{\mathbf{1}_{A_j}(i)},$$

where similar to the above A_j are an enumeration of all the subsets of $\{1, \ldots, \ell\}$. Now, consider some subset $A_m \subset \{1, \ldots, \ell\}$. Then we have for $i \in A_m$ that

$$c_i = p_m \prod_{\substack{1 \leq j \leq 2^\ell \\ j \neq m}} p_j^{\mathbf{1}_{A_j}(i)},$$

so we are set if we can show that

$$\prod_{\substack{1 \leq j \leq 2^\ell \\ j \neq m}} p_j^{\mathbf{1}_{A_j}(i)} \leq k.$$

We have that

$$\prod_{\substack{1\leq j\leq 2^{\ell}\\ j\neq m}} p_j^{\mathbf{1}_{A_j}(i)} \leq 2^{2n},$$

by Theorem 415 in [6]. Thus, taking any $n \leq \lfloor \log_2(k)/2 \rfloor$ will do. □

References

[1] Rosser, J. B. and Schoenfeld, L., *Approximate formulas for some functions of prime numbers*, Illinois Journal of Mathematics, vol. 6 **1** 64–94 (1962).

[2] Helmbold, D. Sloan, R., and Warmuth, M. K., *Learning Integer Lattices*, SIAM Journal on Computing, vol. 21 **2**, 240–266 (1992).

[3] P. Erdős and M. Kac, *The Gaussian Law of Errors in the Theory of Additive Number Theoretic Functions*, American Journal of Mathematics, vol. 62 **1** 738–742, (1940).

[4] Shalev-Shwartz, Shai and Ben-David, Shai, *Understanding machine learning: From theory to algorithms*, Cambridge University Press, (2014).

[5] Schapire, R. E. and Freund, Y. *Boosting: Foundations and Algorithms*, MIT Press, (2012).

[6] Hardy, G. H. and Wright, E.M., *An Introduction to the Theory of Numbers 6th Ed.*, Oxford University Press, (2008).

CENTER FOR APPLIED MATHEMATICS, CORNELL UNIVERSITY
Email address: amt269@cornell.edu