COMP8601: Advanced Topics in Theoretical Computer Science
Lecture 12: Haussler's Theorem
Lecturer: Hubert Chan
Date: 21 Nov 2013

These lecture notes are supplementary materials for the lectures. They are by no means substitutes for attending lectures or replacement for your own notes!

1 Upper Bound for the ϵ-covering Number

The goal of this lecture is to prove the following theorem.

Theorem 1.1 (Haussler's Theorem [2]) *Let C be a class of boolean functions from the set $S = \{x_1, \ldots, x_m\}$ to $\{0,1\}$. Suppose (S, C) has VC-dimension d. Then, we have $N(\epsilon, C, L_2^S) \leq \left(\frac{c}{\epsilon}\right)^{2d}$ for some constant $c > 1$.*

Interpretation: Function as a Point. A boolean function from S to $\{0,1\}$ can be viewed as a point in $\{0,1\}^S$, and we call $x_i \in S$ the *i-th coordinate* for $i \in [m]$. To simplify notation, for a collection C of functions in $\{0,1\}^S$, we simply say the VC-dimension of C to mean the VC-dimension of (S, C).

Distances between Points. A metric L_2^S can be defined on C such that, for $f, g \in C$, the distance is $L_2^S(f,g) := \sqrt{\frac{1}{m}\sum_{i=1}^m (f(x_i) - g(x_i))^2}$.

ϵ-Cover. A subset $\widehat{C} \subseteq C$ is an ϵ-cover if every point in C is at distance at most ϵ from some point in $\widehat{C}$. The ϵ-covering number is the size of a smallest ϵ-cover. Hence, Haussler's Theorem means that if C has VC-dimension at most d, then it has an ϵ-cover of size at most $O(\frac{1}{\epsilon^2})^d$, which is independent of m.

ϵ-Packing. A subset $\widehat{C} \subseteq C$ is an ϵ-packing if every 2 distinct points in $\widehat{C}$ are more than ϵ apart.

General Proof Strategy. Observe that a maximum ϵ-packing must be an ϵ-cover, otherwise there must be a point that is not covered and hence can be included to get a larger ϵ-packing. Hence, it suffices to show the following: any ϵ-packing $V \subset C$ can contain at most $O(\frac{1}{\epsilon^2})^d$ points.

Assume that V is an ϵ-packing. This means that the distance between any two points $f, g \in V$ is $\sqrt{\frac{1}{m}\sum_{i=1}^m (f(x_i) - g(x_i))^2} > \epsilon$. Observing that $|f(x_i) - g(x_i)|$ is either 0 or 1, we conclude that any distinct $f, g \in V$ must differ by more than $\epsilon^2 m$ coordinates. For convenience, we write $\rho := \epsilon^2$.

Easy Case. If $m \leq \frac{4d}{\rho}$, then by Sauer's Lemma, $|C| \leq (\frac{me}{d})^d \leq (\frac{4e}{\rho})^d$. Hence, $|V|$ is also at most $(\frac{4e}{\rho})^d \leq (\frac{4e}{\epsilon^2})^d$. Therefore, we can assume that the number m of coordinates is larger than $\frac{4d}{\rho}$.

Setup. We assume V is an ϵ-packing, where $M = |V|$. Moreover, $m > \frac{4d}{\rho}$. The goal is to give an upper bound on M in terms of d and $\rho = \epsilon^2$ only. We first give an intuitive argument that does not immediately work, and explain how to refine it.

Projections. Let $n := \left\lceil \frac{4d}{\rho} \right\rceil \leq m$. We sample n distinct coordinates I from $[m]$ uniformly at random, e.g., by sampling without replacement. For a point $f \in V$, we let $f_{|I}$ be the projection of f on I, and let $V_{|I} := \{f_{|I} : f \in V\}$ be the projection of V on I.

A Counting Argument. A pair $\{f, g\}$ of points in V is *separated* by I if $f_{|I} \neq g_{|I}$. Since f and g differ by more than ρm coordinates, the probability that they are separated is at least $1-(1-\rho)^n \geq 1-e^{-4d}$. Hence, the expected number of separated pairs is at least $\binom{M}{2} \cdot (1-e^{-4d}) \approx \frac{M^2}{2} \cdot (1-e^{-4d})$. Intuitively, if M is large, then there are lots of such pairs.

On the other hand, if $B = |V_{|I}|$, then we can think of distributing M items into B boxes. Since $V_{|I}$ also has VC-dimension at most d, we have $B \leq (\frac{ne}{d})^d = (\frac{4e}{\rho})^d$. Since there is an upper bound on B, maybe we can argue that the number of pairs separated by different boxes is small. The number of pairs separated by different boxes is maximized when the M items are distributed evenly, in which case we have at most $\binom{B}{2}(\frac{M}{B})^2 \approx \frac{M^2}{2}(1-\frac{1}{B})$ pairs. (Observe that this very pessimistic, because we assume that the M items are distributed totally evenly.)

Hence, comparing the two quantities, we conclude that $1 - e^{-4d} \leq 1 - \frac{1}{B}$, which gives $B \geq e^{4d}$. However, this does not seem very useful because we only have $B \leq (\frac{4e}{\rho})^d$. Although the simple argument does not work immediately, we shall modify it so that instead of considering number of pairs $\{f, g\}$ such that $f|_I \neq g|_I$, we shall consider the number of pairs such that $f|_I$ and $g|_I$ differ by exactly one coordinate. We present the argument given by Chazelle [1].

We next introduce the 1-inclusion graph as a useful tool.

2 The 1-Inclusion Graph

For $C \subseteq \{0,1\}^S$, the *1-inclusion graph* $G(C, E)$ of (S, C) has vertex set C and edge set $E = \{\{f, g\} : f \text{ and } g \text{ differ in exactly one coordinate}\}$. That is, there is an edge between f and g if and only if there exists $x \in S$ such that $f(x) \neq g(x)$ and $f(x') = g(x')$ for all $x' \neq x$. Suppose the VC-dimension of (S, C) is d. Then we have the following property for 1-inclusion graph.

Lemma 2.1 *Let $G(C, E)$ be the 1-inclusion graph of (S, C), which has VC-dimension d. Then, $|E| \leq d \cdot |C|$. In particular, since every subset $C' \subseteq C$ has VC-dimension at most d, it follows that the number of edges in the induced subgraph $G[C']$ is at most $d \cdot |C'|$.*

Proof: We consider a *shifting* procedure. We represent points in C as rows in a table, where each row corresponds to a point in C and each column corresponds to a coordinate in S. The procedure proceeds round by round. In each round, we select an arbitrary column that has not been selected, and we repeatedly change 1's into 0's if the changing does not lead to a row that is already in the table. Let (C^*, E^*) be the new 1-inclusion graph corresponding to the rows in the table after the shifting operations.

Number of Edges Does Not Decrease After Shifting One Entry. Consider an operation at row f, where we change $f(x_i)$ from 1 to 0; let f' be the new point, which does not exist before. Suppose g is adjacent to f in (C_i, E_i). Then g and f differ in exactly one coordinate x_k, and this coordinate cannot be x_i, i.e., $k \neq 1$ (since otherwise $f' = g$). Then $g(x_i) = 1$.

Now consider $g(x_i)$. If $g(x_i)$ can be changed from 1 to 0 and let g' be the new point, then g' and f' become adjacent again. It follows that the edge $\{f, g\}$ is replaced by a new edge $\{f', g'\}$.

Suppose $g(x_i)$ cannot be changed. Then there is a point h such that $h(x_i) = 0$ and $h(x_j) = g(x_j)$ for all $j \neq i$. Now h and f' differ in exactly one coordinate x_k, so there is a new edge $\{f', h\}$ after the operation. There is no edge between h and f before the operation, since they differ in two coordinates. In both cases, the number of edges does not decrease.

The VC-dimension Does Not Increase After Shifting One Column. We consider the effect of shifting a particular column in the procedure. Let (C_i, E_i) be the 1-inclusion graph corresponding to the table at the beginning of the round when we consider the column x_i, and let (C'_i, E'_i) be the 1-inclusion graph corresponding to the table at the end of this round.

Next we show that the VC-dimension of (S, C'_i) is at most the VC-dimension of (S, C_i). In particular, we show that if a subset T of S is shattered by C'_i, then T is also shattered by C_i. If the column x_i is not in T, then we are done. Suppose $x_i \in T$. Note that $T \setminus \{x_i\}$ must be shattered by C_i. For each $f \in C'_i$, if $f(x_i) = 1$, then f is also in C_i. Moreover, there must exist $g \in C_i$ that agrees with f in all coordinates but x_i, since otherwise $f(x_i)$ should have been changed from 1 to 0. Hence T is also shattered by C_i.

Analyzing the Inclusion Graph after Shifting. Applying the above analysis for all columns in the table, we conclude $\frac{|E|}{|C|} \leq \frac{|E^*|}{|C^*|}$ and that the VC-dimension of (S, C^*) is at most d.

We claim that in the final table, if there is $T \subseteq S$ and a row h such that $h(x) = 1$ for all $x \in T$, then T must be shattered by C^*. In particular, if we let $f_{|T}$ be the projection of f on T, then for every $v \in \{0,1\}^T$ there exists $f \in C^*$ such that $f_{|T} = v$. Suppose on the contrary there is at least one $v \in \{0,1\}^T$ such that $f_{|T} \neq v$ for all $f \in C^*$; let v be one such point with a maximum number j of 1's. Note that $j < |T|$ since we have a row h with all 1's at T. Suppose $v(x) = 0$ for some $x \in T$. Let $v' \in \{0,1\}^T$ be a point that differs with v at exactly one coordinate x; that is, $v'(x) = 1$. There exists $f \in C^*$ such that $f_{|T} = v'$. Then $f(x)$ should have been changed from 1 to 0, since this change cannot lead to a row already in the table. This is a contradiction.

It follows that in the final table every row has at most d 1's. Then, every point f has at most d neighbors with less 1's than f. We direct each edge $\{f, g\}$ of E^* as (f, g) if f has more 1's than g. Let $\vec{E}^*$ be the set of directed edges. Then each vertex has an out-degree at most d in $(C^*, \vec{E}^*)$. Hence $|E^*| = |\vec{E}^*| \leq d|C^*|$. Then we get $\frac{|E|}{|C|} \leq \frac{|E^*|}{|C^*|} \leq d$. ■

Lemma 2.2 *Suppose $G = (C, E)$ is an undirected graph such that, for every subset $C' \subseteq V$, the induced subgraph $G[V']$ has at most $d \cdot |C'|$ edges. Then, the edges in G can be directed such that the maximum out-degree is at most d.*

Proof: We will use the Hall's theorem, which states that a class of sets $V_1, \ldots, V_k$ has a set of distinct representatives $r_1 \in V_1, \ldots, r_k \in V_k$ if and only if for all $\ell \leq k$, the union of any ℓ of the V_i's contains at least ℓ elements.

For each vertex $f \in C$, we let $f^{(1)}, \ldots, f^{(d)}$ be d copies of f. We can think of each copy as a "token". If we want to direct an edge $\{f, g\}$ from f to g, then the edge must get a token from f. Since each vertex f has at most d tokens, if each edge can obtain one token from one of its incident vertices, then the problem is solved.

For each edge $e = \{f, g\} \in E$, define a set $V_e := \{f^{(1)}, \ldots, f^{(d)}, g^{(1)}, \ldots, g^{(d)}\}$. Consider the class of sets $\{V_e : e \in E\}$. Every $E' \subseteq E$ corresponds to a subgraph (C', E'), where C' is the set of vertices whose copies appear in the sets $\{V_e : e \in E'\}$.

Let (C', E'') be the subgraph induced by C'. Then $E' \subseteq E''$. Then, from hypothesis, we have $|\cup_{e \in E'} V_e| = d|C'| \geq |E''| \geq |E'|$. By Hall's theorem there exist distinct representatives $r_e \in V_e$ for all $e \in E$. For each edge $e = \{f, g\}$, we direct it as (f, g) if the representative r_e is in $\{f^{(1)}, \ldots, f^{(d)}\}$ and as (g, f) otherwise. Then, every edge is directed. Moreover, since the representatives are distinct, the out-degree of a vertex is at most the number of its copies, which is exactly d. ■

3 Continue with Packing V and Random Projection on I

Recall that we have an ϵ-packing V, where $|V| = M$. In particular, this implies that any two points f, g in V differ by more than ρm coordinates, where $\rho = \epsilon^2$. Moreover, we consider the case when $m \geq n := \left\lceil \frac{4d}{\rho} \right\rceil$, and I is a random subset of n coordinates from $[m]$.

Defining Weighted Inclusion Graphs. For each $I \subseteq [m]$ and $V \subseteq C$, we consider a weighted version of the 1-inclusion graph $(V_{|I}, E(V_{|I}))$. We define weight functions $q = q(V, I)$ on the vertices, and $w = w(V, I)$ on the edges as follows. For each $u \in V_{|I}$, define $q(u) := |\{f \in V : f_{|I} = u\}|$ as the number of points in V whose projection on I is exactly u. Note that $\sum_{u \in V_{|I}} q(u) = |V|$. For each edge $e = \{u, v\} \in E(V_{|I})$, define its weight as $w(e) := \frac{q(u) \cdot q(v)}{q(u)+q(v)}$. Let $W = W(V, I)$ be the sum of the edge weights, i.e., $W := \sum_{e \in E(V_{|I})} w(e)$. Observe that q, w and W depend on V and I.

Hence, $W = W(V, I)$ is a random variable. We consider the expectation $\mathbf{E}[W]$, which corresponds to the number of separated pairs in our simple argument.

The lemma below is the analogue of our simple counting argument when we tried to give an upper bound on the number of pairs separated by the B boxes.

Lemma 3.1 *For all $I \subseteq [m]$ and $V \subseteq C$, we have $W \leq d \cdot |V|$.*

Proof: Since the VC-dimension of $(I, V_{|I})$ is at most d, the 1-inclusion graph $(V_{|I}, E(V_{|I}))$ can be directed such that every vertex has an out-degree at most d. Let $(V_{|I}, \vec{E}(V_{|I}))$ be such a directed graph. Using the inequality $yz \leq (y + z) \min\{y, z\}$ for all $y, z \geq 0$, we have

$$\begin{aligned} W &= \textstyle\sum_{e \in \vec{E}(V_{|I})} w(e) = \sum_{(u,v) \in \vec{E}(V_{|I})} \frac{q(u)q(v)}{q(u)+q(v)} \leq \sum_{(u,v) \in \vec{E}(V_{|I})} \min\{q(u), q(v)\} \\ &\leq \textstyle\sum_{(u,v) \in \vec{E}(V_{|I})} q(u) \leq d \sum_{u \in V_{|I}} q(u) = d|V|, \end{aligned}$$

as required. ■

Recall that in our simple argument, we tried to argue that since each distinct pair $f, g \in V$ differ by more than ρm coordinates, the expected number of pairs separated is large. The following lemma is its analogue.

Lemma 3.2 $\mathbf{E}[W] \geq 2d(M - (en/d)^d)$.

Proof: Partition the edges of the 1-inclusion graph $(V_{|I}, E(V_{|I}))$ into n subsets $E_1, \ldots, E_n$, where E_k is the set of edges whose end points differ in the i_k-th coordinate, i.e., $E_k = \{\{u, v\} \in$

$E(V_{|I}) : u(x_{i_k}) \neq v(x_{i_k})\}$. For $k \in [n]$, let $W_k := \sum_{e \in E_k} w(e)$. Observe that $W = \sum_{k=1}^{n} W_k$. By symmetry we have $\mathbf{E}[W_k] = \mathbf{E}[W_n]$ for all $k \in [n]$. Then by linearity of expectation we get $\mathbf{E}[W] = \sum_{k=1}^{n} \mathbf{E}[W_k] = n\mathbf{E}[W_n]$.

To obtain a lower bound for $\mathbf{E}[W_n]$, we first consider $\mathbf{E}[W_n|I_{-n}]$, that is, the expectation of W_n conditioned on I_{-n}. Set $J := I_{-n}$. Note that given J, the random variable i_n is uniformly distributed in $[m] \setminus J$. We partition V into $|V_{|J}|$ subsets $\{V_t : t \in V_{|J}\}$ according to $V_{|J}$, that is, two points of V are in the same subset V_t if and only if they have the same projection t on J.

For any $i_n \in [m] \setminus J$ and $t \in V_{|J}$, the set V_t can be further divided into two sets A_t and B_t, where $A_t = \{f \in V_t : f(x_{i_n}) = 1\}$ and $B_t = \{f \in V_t : f(x_{i_n}) = 0\}$. Let $\ell_t := |A_t|$. For every edge $e \in E_n$, since its two end points has a common projection on J, there exists $t \in V_{|J}$ such that one of the end point corresponds to A_t (that is, all points in A_t are projected to this end point), and the other to B_t; it follows that $w(e) = \frac{\ell_t(|V_t|-\ell_t)}{|V_t|}$. On the other hand, every V_t corresponds to an edge in E_n, with A_t projected to one of the end point and B_t to the other, that has a weight $\frac{\ell_t(|V_t|-\ell_t)}{|V_t|}$; in case either A_t or B_t is empty, we can assume V_t corresponds to a zero-weighted edge. Note that given J, the value of $|V_t|$ is determined and only ℓ_t is a random variable. Then we have

$$\begin{aligned}\mathbf{E}[W_n|J] &= \mathbf{E}\big[\textstyle\sum_{e \in E_n} w(e)|J\big] = \mathbf{E}\Big[\sum_{t \in V_{|J}} \frac{\ell_t(|V_t|-\ell_t)}{|V_t|}|J\Big] \\ &= \textstyle\sum_{t \in V_{|J}} \frac{1}{|V_t|}\mathbf{E}[\ell_t(|V_t| - \ell_t)|J].\end{aligned}$$

We fix $t \in V_{|J}$ and consider $\mathbf{E}[\ell_t(|V_t| - \ell_t)|J]$. The value $\ell_t(|V_t| - \ell_t)$ is the number of unordered pairs $\{f, g\}$ with $f, g \in V_t$ such that $f(x_{i_n}) \neq g(x_{i_n})$. Let $\mathcal{P}_t := \{\{f, g\} : f \neq g \text{ and } f, g \in V_t\}$ be the set of all unordered pairs of distinct functions in V_t. Then $|\mathcal{P}_t| = \frac{|V_t|(|V_t|-1)}{2}$. For each $\{f, g\} \in \mathcal{P}_t$, since V is an ρ-packing, f and g differ in at least ρm coordinates. Then if i_n is uniformly drawn from $[m] \setminus J$, the probability that f and g differ in i_n is at least $\frac{\rho m}{m-n+1} \geq \rho$. That is, every pair in $\mathcal{P}_t$ contributes at least ρ to the value $\ell_t(|V_t| - \ell_t)$ in expectation. It follows that

$$\mathbf{E}[\ell_t(|V_t| - \ell_t)|J] \geq \tfrac{|V_t|(|V_t|-1)}{2} \cdot \rho = \tfrac{\rho \cdot |V_t|(|V_t|-1)}{2}.$$

Now we have $\mathbf{E}[W_n|J] = \sum_{t \in V_{|J}} \frac{1}{|V_t|}\mathbf{E}[\ell_t(|V_t| - \ell_t)|J] \geq \sum_{t \in V_{|J}} \frac{\rho(|V_t|-1)}{2} = \frac{\rho(M-|V_{|J}|)}{2}$. Since $(J, V_{|J})$ has VC-dimension at most d, by Sauer's lemma we have $|V_{|J}| \leq \left(\frac{e(n-1)}{d}\right)^d \leq \left(\frac{en}{d}\right)^d$. It follows that $\mathbf{E}[W_n|J] \geq \frac{\rho(M-(en/d)^d)}{2}$. Now we have

$$\mathbf{E}[W] = n\mathbf{E}[\mathbf{E}[W_n|J]] \geq \tfrac{\rho n(M-(en/d)^d)}{2} \geq 2d \cdot (M - (en/d)^d). \tag{3.1}$$

■

Finishing the Proof. Combining Lemmas 3.1 and 3.2, we have:

$2d \cdot (M - (en/d)^d) \leq dM$, which yields $M \leq 2 \cdot (\frac{en}{d})^d \approx 2 \cdot (\frac{4e}{\rho})^d = O(\frac{1}{\epsilon^2})^d$, since $n = \left\lceil \frac{4d}{\rho} \right\rceil$. This completes the proof.

4 Homework Preview

1. **Alternative Proof of Sauer's Lemma.** Suppose C is a class of boolean functions on X and the VC-dimension of (X, C) is at most d. Use the shifting procedure to prove Sauer's Lemma: for every subset S of X such that $|S| = m$, the cardinality of the projection $C(S)$ is at most $\binom{m}{\leq d}$.

   (Hint: You can use any results about the shifting procedure proved in class.)

2. **Alternative Proof for the 1-inclusion Graph.** Let S be a set of size m. Let $C \subseteq \{0,1\}^S$ be a collection of points with VC-dimension d. Suppose E is the set of edges in the 1-inclusion graph of (S, C). We have proved in class that $|E| \leq d \cdot |C|$ using the shifting procedure. In this question, we prove the same result by induction on d and m, which is similar to the proof of Sauer's Lemma shown in Lecture 8.

   (a) Prove that $|E| \leq d \cdot |C|$ is true for the base cases $d = 0$ or $m = 1$.

   (b) Suppose $d \geq 1$ and $m > 1$. Let $x \in S$ and $S' := S \setminus \{x\}$. Let $C_1 := C_{|S'}$ be the projection of C on S'. Let $C_2 \subseteq C_1$ be the set of points f in C_1 such that there exist $f_1, f_2 \in C$, where f_1 and f_2 disagree on x and $f_1 \,|_{S'} = f_2 \,|_{S'} = f$. Let E_1 be the set of edges in the 1-inclusion graph G of (S', C_1), and $E_2 \subseteq E_1$ the set of edges in the induced subgraph $G[C_2]$.

      i. Give upper bounds for $|E_1|$ and $|E_2|$ in terms of d, $|C_1|$ and $|C_2|$.
      (Hint: Use the induction hypothesis.)

      ii. Prove that $|E| \leq |E_1| + |E_2| + |C_2|$.

      iii. Prove that $|E| \leq d \cdot |C|$.

References

[1] B. Chazelle. A note on haussler's packing lemma. Unpublished manuscript, Princeton, 1992.

[2] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *J. Comb. Theory Ser. A*, 69(2):217–232, February 1995.