Yale University
Department of Computer Science

Lower Bounds on the VC Dimension of Unions of Concept Classes

Lev Reyzin

YALEU/DCS/TR-1349
April 2006

Abstract

In this paper we consider bounds on the VC dimension of a class, $T_k(C)$, that consists of unions of k concepts from class C of VC dimension d. Blumer et al. [1] show that the VC dimension of $T_k(C)$ cannot exceed $2dk \log_2 3k$. By considering grids of points and the class of line segments, we show that it is possible for $T_k(C)$ to have VC dimension $\frac{8}{5}kd$. We also demonstrate that our method cannot produce classes that asymptotically match Blumer et al. upper bound, leaving the gap open for future research.

1 Introduction

Given any concept class C over X, we may form the class $T_k(C)$ consisting of unions of k concepts from C, where $T_k(C) = \{\cup_{i=1}^{k} c_i : c_i \in C, 1 \leq i \leq k\}$. We define $f(k, d)$ to be the maximum value for the VC dimension of $T_k(C)$ over all C of VC dimension $d \geq 1$. We wish to find lower and upper bounds on the asymptotic behavior of $f(k, d)$ in terms of k and d. Understanding the behavior of $f(k, d)$ would give us greater theoretical understanding of the VC dimension.

In [1], Blumer et al. bound the number of labelings $T_k(C)$ can achieve on set S of size m by a function $\Phi_d(|S|)^k$, where $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$ if $m \geq d$ and $\Phi_d(m) = 2^m$ otherwise. They argue that if $\Phi_d(m)^k < 2^m$, S cannot be shattered by $T_k(C)$ and the VC dimension of $T_k(C) < m$. They then show that $\Phi_d(m) \leq (\frac{em}{d})^d$ for all $m \geq d \geq 1$. Since $(\frac{em}{d})^{dk} < 2^m$ for $m = 2dk \log_2 3k$, they prove $f(k, d)$ is bounded from above by $2kd \log_2(3k)$.

The object of this paper is therefore to find lower bounds on $f(k, d)$ through geometric constructions. While we cannot asymptotically match the upper-bound, in Section 2 we present constructions that generate a lower found of $\frac{8}{5}kd$. Finally, in Section 3 we show the limitations of our method and find that our constructions cannot yield lower bounds greater than $2kd$.

2 Lower Bounds on $f(k, d)$

Proposition 1. *If $\exists C$ with VC dimension $= d$ and $T_k(C) \geq ckd$, then $f(k, d) \geq ckd$.*

Proof. For any n, we can consider a new class C' that contains n copies of concepts from C each in a different "universe." The VC dimension of $C' = nd$ and the VC dimension of $T_k(C') = (n$ times the VC dimension of

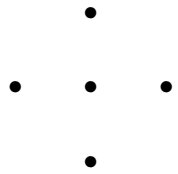

Figure 1: 5 points that can be shattered by $T_2(L)$

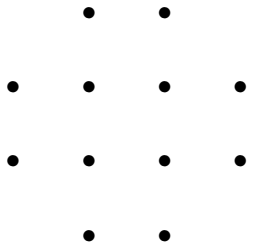

Figure 2: 12 points that can be shattered by $T_4(L)$

$T_k(C)$). We can similarly scale results linearly with k. Hence, if the VC dimension of $T_k(C) \geq ckd$, then this linear scaling shows that $\exists$ infinitely many d, k such that $f(k,d) \geq ckd$, giving a lower found on $f(k,d)$. □

Proposition 2. $f(k,d) \geq kd$

Proof. Follows directly from Proposition 1. □

Proposition 3. $f(k,d) \geq \frac{5}{4}kd$

Proof. Let L be the class of line segments. L has VC Dimension of 2. It is easy to see that $T_2(L)$ shatters the 5 points arranged as in Figure 1. For $k = 2$ and $d = 2$, this gives us a construction for the VC dimension of $T_k(C) \geq \frac{5}{4}kd$. By Proposition 1, $f(k,d) \geq \frac{5}{4}kd$.

□

Proposition 4. $f(k,d) \geq \frac{3}{2}kd$

Proof. First, We will show that $T_4(L)$ has VC dimension ≥ 12 by showing that $T_4(L)$ can shatter the 12 points in Figure 2.

Accounting for symmetry, Figure 3 shows how $T_4(L)$ can label all possible labelings of the innermost with the outermost points being labeled positive.

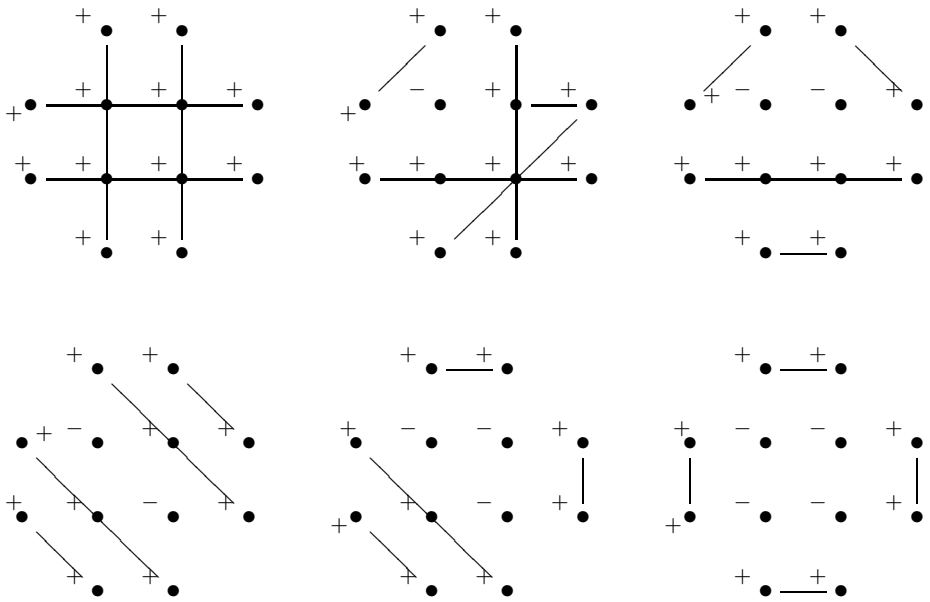

Figure 3: Accounting for symmetry, $T_4(L)$ achieving all possible latelings of the interior points with the exterior points labeled positive

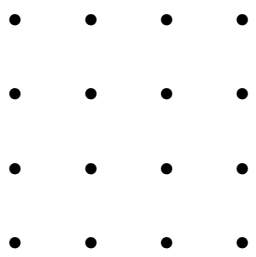

Figure 4: 16 points arranged in a grid that can be shattered by $T_5(L)$

For any labeling of the twelve points in Figure 2, we can find the corresponding line segment arrangement in Figure 3 by matching the labeling of the interior points and then shortening any line segment to avoid any of the exterior points labeled negative. Hence, these points are shattered by $T_4(L)$, showing the VC dimension is at least 12.

For $k = 4$ and $d = 2$, this gives us a construction for the VC dimension of $T_k(C) \geq \frac{3}{2}kd$. By Proposition 1, $f(k, d) \geq \frac{3}{2}kd$.

□

Proposition 5. $f(k, d) \geq \frac{8}{5}kd$

Proof. We will show that $T_5(L)$ has VC dimension ≥ 16 by showing that $T_5(L)$ can shatter the 16 points in Figure 4.

We call "case n" to be a proof that $T_5(L)$ can achieve any labeling with exactly n points labeled positive. To prove all 16 points are shattered, we will present cases 0 through 16.

Case 16: Obvious (uses only 4 vertical or horizontal line segments).

Case 15: At worst, the negatively labeled vertex will split a line segment from Case 16 into two, but Case 16 uses only 4 line segments, so we have an extra line segment for the split.

Case 14: If one of the two negatively labeled vertices is a corner, we shorten one of the 4 line segments from Case 16 and use the proof for $Case15$. If one of the two negatively labeled points is on an outside edge, we orient our 4 line segments in such a way as to cover the positive points with 4 line segments and apply Case 15. If both negatively labeled points are in the middle 4, if they are on the same horizontal or vertical line, then they only split one of the 4 line segments (properly oriented) into 2, and case 15 applies. This leaves only 1 case, shown in Figure 5.

Case 13: If one of the three negatively labeled line segments is a corner, we use Case 14 with shortened segments as needed to avoid the corner. If

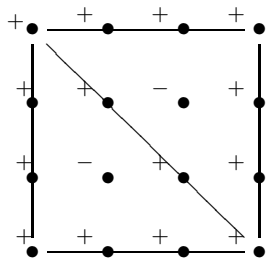

Figure 5: A labeling of the 16 points by $T_5(L)$ for Cases 14 and 12

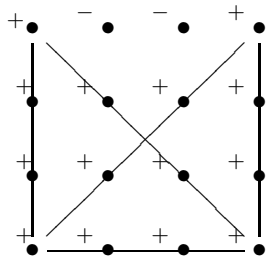

Figure 6: A labeling of the 16 points by $T_5(L)$ for Case 13

only one is on an outside edge, we use Case 14 (with proper orientation) and since every corner is covered by 2 segments (the bad case), we can shorten any of the line segments to avoid the negatively labeled point on an outside edge and still cover the corners. The same holds true if all three negative points are in the middle: we can shorten the diagonal line in Figure 5 to accommodate for the extra negative.

If all 3 negative points are on edges and any two are on the same edge, we can cover them as shown in Figure 6 and shorten any line segment as needed for the third negative point.

So, for the all negative points on edges case, we have the case left of each negative point being on a different edge, as shown in Figure 7. Three of ?s are negative, and three are positive. And we have 2 more edges to label the positive ? marks. Considering the graph where two points above share an edge if and only if they can be connected by a line segment that does not go through any other points, it can be easily seen that no three of the ? points form an independent set. Hence, two more line segments can cover any combination in Figure 7.

This leaves us with the case when 2 of the negative points are on an edge,

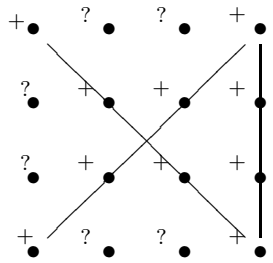

Figure 7: A partial labeling of the 16 points by $T_5(L)$ for Case 13

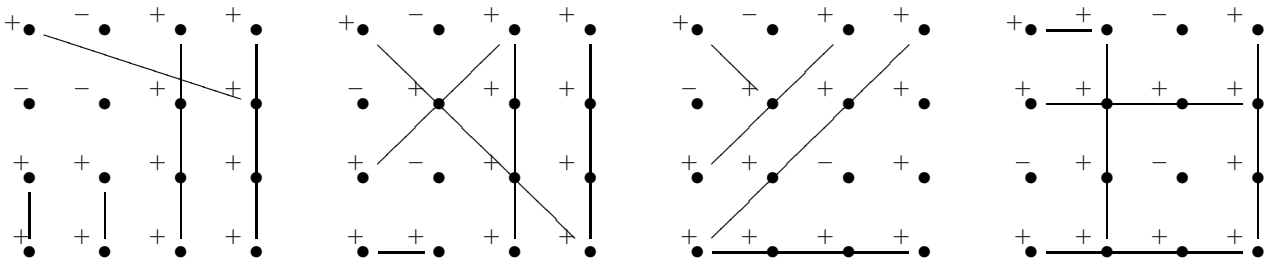

Figure 8: A partial labeling of the 16 points by $T_5(L)$ for Case 13

and one is in the middle. If the two negative points are on the same edge, then we can shorten two line segments and divide one for the middle point. In fact, we can do this if the two negative points are on opposite edges and are on the same vertical or horizontal line. Hence, the only cases left are: two negatives near the same corner, Figure 8, two negatives on adjacent sides but not near same corner and a negative middle point not next to to either of the negatives (becuase of the "shortening argument", Figure 9, and two negatives on opposite sides, but not on the same horizontal or vartical line (only 1 case without two adjacent negatives), Figure 10.

This finishes Case 13.

Case 12: We can again forget about the corners being negative since any line segment covering them can be shortened.

If all 4 negatives are in the center, we can make a 4 line segment box around the center. If all 4 negatives are on the sides, we can make an cross through the middle and cover one of the sides containing a positive point with a line segment. We left with the same situation as Figure 7, and we can cover that.

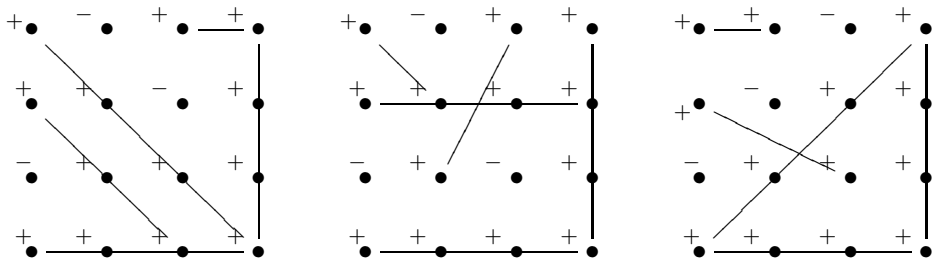

Figure 9: A partial labeling of the 16 points by $T_5(L)$ for Case 13

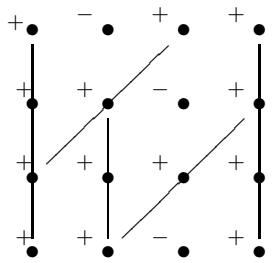

Figure 10: A partial labeling of the 16 points by $T_5(L)$ for Case 13

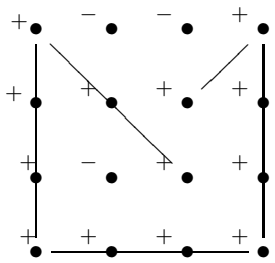

Figure 11: A partial labeling of the 16 points by $T_5(L)$ for Case 12

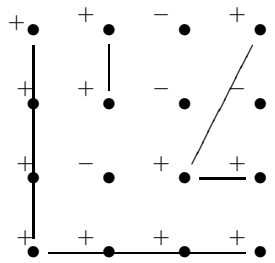

Figure 12: A partial labeling of the 16 points by $T_5(L)$ for Case 12

If three of the negatives are in the center, we can make a square around the outside points and cover the positive center point with its own line segment. Since every corner in a square is covered by two line segments, any one can be shortened to make an edge point negative.

If only one of the negative points is in the center, we can see in Figure 12 that again since all corners are covered by two line segments, we can shorten some line segment to avoid the 4th negative point.

This leaves the case of two negatives in the center and two on the sides, which was covered in Figure 5. We have to advoid two more edge points, and since each corner is covered twice, we can do this by shortening the appropriate line segments. This leaves one last case, where both points to avoid are near the same corner, which needs to still be covered. This is handled in Figure 12 and finishes Case 12.

Cases 0 through 5: Obvious since we can use one line segment per point.

Cases 6 and 7: We define the graph G to be the graph where verteces represent the points in Figure 4 and edges connect any pair of points between which a line can pass without crossing any other points. The maximum

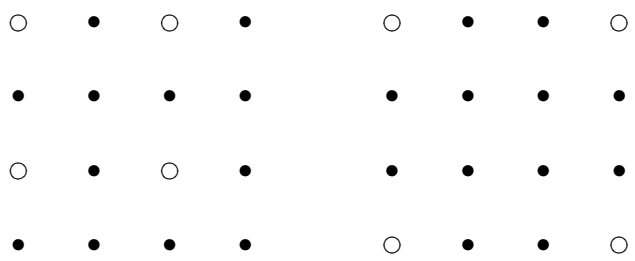

Figure 13: Maximum independent sets in G

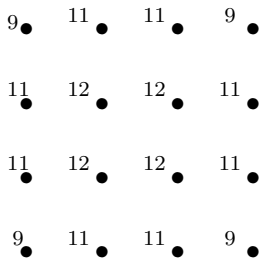

Figure 14: vertex degrees in G

independent set in G is of size 4. Both possibilities are shown in Figure 13.

In the graph represented by Figure 13 each point has edge to at least two vertices in the independent set. Hence one or two more points can share a line segment with one of the points in the independent set.

Case 8: We can see the in Figure 13, no two vertices only have edges to the same points in the independent set. Hence, we can cover a third point by a shared line segment, plus one more with its own.

Cases 9 and 10: If we consider the degrees of the vertices in G, in Figure 14, we can see all vertices have degree ≥ 3. In a graph for Case 10, 6 vertices are negative and are "removed," so each vertex has degree at least 3. And there are at most only 4 such vertices. Since we can see it is not true that each of the vertices labeled 9 (which will be the ones with degree ≥ 3 after the negative points are removed) have edges to only vertices in the independent set in Figure 13, at least two will have edges to each other. Hence, two pairs of those can be paired and the rest can be covered like in Case 8. Case nine is easier since the last pair is just one vertex.

Case 11: All we have to show is that a line segment thats involved in covering any of the 10 positively labeled vertices can extend to cover the

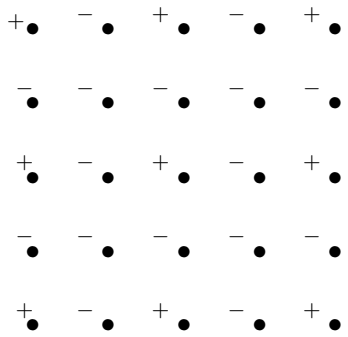

Figure 15: A labeling of points with both positive coordinates positive and the rest negative

eleventh. All we want to show is that in any choice of 11 points, three will be collinear without in intervening negative point and thereby coverable by a line segment of their own. This is easy to show since the four central points and the four corners all participate in different sets of tree collinear points, and at most 5 of the 6 points can be labeled negative since we are in Case 11. The rest of the case follows from Case 10.

Hence, all cases are covered, so at least 16 points are shattered by a union of 5 pairs of line segments, giving a $\frac{8}{5}kd$ ratio for $T_5(L)$. This shows by Proposition 1 that $f(k,d) \geq \frac{8}{5}kd$

□

3 Limitations of Grids for $T_k(L)$

These results start to give hope that as k increases, we can keep adding points to an n by n grid, and the fraction preceeding the kd term in our lower bounds will increase as well, giving us the $\log_2(k)$ term that we find in the upper bound. Unfortunately, this turns out not to be the case.

We can show that an $n \times n$ grid requires at least $\frac{1}{4}(n^2)$ unions of line segments to shatter. In this grid, we can assign coordinates to the all points, with $(0,0)$ being the bottom left. Take the labeling of the points where points with both even coordinates are positive and the rest are negative, as shown in Figure 15. Imagine a line segment connecting two points, point 1: (x_1, y_1) and point 2: (x_2, y_2), each with both positive coordinates. This line will have a slope of $m = \frac{y_2 - y_1}{x_2 - x_1}$. Before reducing the fraction, both the top and bottom of the fraction are even. But, after the fraction is reduced,

to $m = \frac{\delta_y}{\delta_x}$, either δ_x or δ_y is odd (or both). Take the bottom-most of the two points, say point 1. Then, the point $(x_1 + \delta_x, y_1 + \delta_y)$ is on the line segment and has at least one positive coordinate. Hence, any line segment that passes through two points with both even coordinates must also pass through a point with at least one odd coordinate. (The case when both even-coordinate points are on one vertical or horizontal line is not covered by the slope argument, but is easy to see) So achieve this labeling, we would need as many line segments as positively labeled points, which is $\geq \frac{1}{4}n^2$.

This implies that to shatter a grid of n^2 points, we need a union of $\frac{1}{4}n^2$ line segments. Therefore, this method of construction will give linear lower bounds on $f(k, d)$ with respect to k, at best $2kd$.

4 Discussion

This leaves a $\log k$ asymptotic gap between the upper and lower bounds on $f(k, d)$. While we may still possible improve the lower bound from $\frac{8}{5}kd$ to $2kd$ using grids of points like in this paper, if we want our lower bounds to match the $2kd \log_2(3k)$ upper bound, we would need another technique. One possible approach to matching the upper bound would be to consider grids in higher dimensions. Yet, we must also consider that the upper bound may not be tight and the actual behavior of $f(k, d)$ may linear with respect to both k and d.

5 Ackgnowledgements

I would like to thank Dana Angluin for many helpful discussions and ideas, Samuel Daitch for helpful suggestions, and Kevin Yuk-Lap Yip for sharing his non-geometric construction for $f(k, d) \geq \frac{3}{2}kd$.

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.