Deciding the VC Dimension is $\mathbf{\Sigma_3^p}$-complete, II

Marcus Schaefer

School of CTI
DePaul University
243 South Wabash Avenue
Chicago, Illinois 60604, USA
schaefer@cs.depaul.edu

October 30, 2000

Abstract

The path VC-dimension of a graph G is the size of the largest set U of vertices of G such that each subset of U is the intersection of U with a subpath of G. The VC-dimension for graphs was introduced by Kranakis, *et al.* [KKR$^+$97], building on an idea of Haussler and Welzl [HW87]. We show that computing the path VC-dimension of a graph is $\mathbf{\Sigma_3^p}$-complete. This adds a rare natural $\mathbf{\Sigma_3^p}$-complete problem to the repertoire.

1 Introduction

A set system $\mathcal{C}$ is said to *shatter* a set A if for each subset S of A there is a set $C \in \mathcal{C}$ such that $S = A \cap C$. If $\mathcal{C}$ shatters a set of cardinality k we say that $\mathcal{C}$ has Vapnik-Červonenkis-dimension (VC-dimension) at least k. The VC-dimension was introduced by Vapnik and Červonenkis in their study of uniform convergence of relative frequencies. It has become a successful tool in areas such as computaional learning theory, and computational geometry [AB92, HW87]. Roughly speaking, it is used to measure the complexity of set systems. For example, the VC-dimension of a concept class is finite precisely if the class is learnable in the PAC-learning model, and the VC-dimension of the class can be used to determine the necessary sample size. It is not surprising then that the complexity of *computing* the VC-dimension has received some attention. The complexity depends on the model, i.e. how the set system is represented as an input. Several different models have been studied in the literature; to mention just two examples: if the set system $\mathcal{C}$ is represented as a matrix, then determining the VC dimension is **LOGNP**-complete [PY93], if it the set system is represented by a circuit, the problem is $\mathbf{\Sigma_3^p}$-complete [Sch99b].

Haussler and Welzl [HW87] introduced the VC-dimension of a graph; the set to be shattered is a set of vertices of the graph, and $\mathcal{C}$ is the collection of neighborhoods of vertices in the graph [HW87, ABC95]. In a recent paper, Kranakis, Krizanc, Ruf, Urrutia, and Woeginger [KKR$^+$97] generalized this definition from neighborhoods of vertices to allow arbitrary collections of subgraphs. They determined the computational complexity of several of these problems, showing, for example, that deciding whether at least k vertices in a graph are shattered by subtrees of the graph is **NP**-complete. The main open problem left by that paper was the case in which $\mathcal{C}$ is the set of all subpaths of the graph, the *path VC-dimension* of the graph. The paper showed that the decision problem is **NP**-hard leaving a large gap to $\mathbf{\Sigma_3^p}$ which is a natural upper bound for the problem. We close the gap by showing the problem $\mathbf{\Sigma_3^p}$-complete.

Theorem 1.1 *Deciding the path Vapnik-Červonenkis-dimension of a graph is $\mathbf{\Sigma_3^p}$-complete.*

This is a companion result to the earlier result that determining the VC-dimension in the circuit model is $\mathbf{\Sigma_3^p}$-complete [Sch99b]. Despite the similarity in the statements of the two problems, the proofs are completely independent. The reason is the difference in models. A direct reduction from one problem to the other would probably be much more complicated than the approach through quantified Boolean formulas taken here.

Completeness proofs for higher levels of the polynomial hierarchy are surprisingly rare. Recent progress includes problems in learning theory, logic and Ramsey theory [Uma98, Sch99b, Sch99a, Uma99].

2 VC-Dimension of Graphs

Let $F = (V, E)$ be a graph, and $\mathcal{C}$ a collection of subgraphs of F. We say that $\mathcal{C}$ *shatters* a set of vertices W of F if for all $W' \subseteq W$ there is a graph $G \in \mathcal{C}$ such that G contains all the vertices in W', but none of the vertices in $W \setminus W'$. We define the *VC-dimension of $\mathcal{C}$ with respect to F* as

$$\mathrm{V}C_{\mathcal{C}}(F) = \max\{|W| : W \text{ is shattered by } \mathcal{C}\},$$

if $\mathcal{C} \neq \emptyset$, and let $\mathrm{V}C_{\emptyset}(F) = -1$, otherwise. Note that the definition of $\mathrm{V}C$ makes sense for both directed, and undirected graphs.

For any property $\mathcal{P}$ of graphs we will write $\mathrm{V}C_{\mathcal{P}}(F)$ for $\mathrm{V}C_{\mathcal{C}}(F)$ where $\mathcal{C} = \{G : G$ is a subgraph of F with property $\mathcal{P}\}$. Thus we have $\mathrm{V}C_{path}$, $\mathrm{V}C_{cycle}$, for example, and the corresponding decision problems. For directed graphs F, we also require the paths and cycles to be directed.

GRAPH VC$_{path}$ DIMENSION
Instance: (Finite) graph F, number k.
Question: $\mathrm{V}C_{path}(F) \geq k$?

DIGRAPH VC$_{path}$ DIMENSION
Instance: (Finite) directed graph F, number k.
Question: $\mathrm{V}C_{path}(G) \geq k$?

The complexity of GRAPH VC$_P$ DIMENSION depends on the property $\mathcal{P}$. It is simple to construct properties $\mathcal{P}$ for which GRAPH VC$_P$ DIMENSION is undecidable, but such properties are hardly natural properties of graphs.

Lemma 2.1 *$\mathrm{V}C_{\mathcal{P}}$ can be decided in $\mathbf{\Sigma_3^p}$ if $\mathcal{P}$ can be verified in* **NP**.

Proof. $\mathrm{V}C_{\mathcal{P}}(F) \geq k$ is equivalent to saying that there is a set W of k vertices of F such that for every subset S of W there is a subgraph G of F such that $\mathcal{P}(F)$ is true, and G contains all vertices in S, but no vertex in $W - S$. $\diamondsuit$

We conclude that GRAPH VC$_{path}$ DIMENSION, DIGRAPH VC$_{path}$ DIMENSION, GRAPH VC$_{cycle}$ DIMENSION, and DIGRAPH VC$_{cycle}$ DIMENSION all lie in $\mathbf{\Sigma_3^p}$. Table 1 summarizes the known complexity result.

For trees $\mathrm{V}C_{tree} = \mathrm{V}C_{connected}$ implies **NP**-completeness [KKR$^+$97]. We obtain a nonapproximability result from a simple construction.

Theorem 2.2 *If f is a function for which $|f(F) - \mathrm{V}C_{path}(F)| = o(|F|^{1/2})$, then $\mathrm{V}C_{path}$ can be computed from f in polynomial time with one query.*

Property $\mathcal{P}$	Computational Complexity	Reference
star	in **P**	Kranakis, *et al.* [KKR+97]
vertex neighborhoods	**LOGNP**-complete	Kranakis, *et al.* [KKR+97]
connected	**NP**-complete	Kranakis, *et al.* [KKR+97]
paths	$\mathbf{\Sigma_3^p}$-complete	Corollary 3.4
cycles	$\mathbf{\Sigma_3^p}$-complete	Corollary 3.3

Figure 1: Complexity of $\mathrm{V}C_{\mathcal{P}}$

This implies that approximating $\mathrm{V}C_{path}$ to within an additive constact of $o(n^{1/2})$ is still $\mathbf{\Sigma_3^p}$-complete. The proof follows immediately from the following lemma.

Lemma 2.3 *Given a graph F we can construct a graph G such that $\mathrm{V}C_{path}(G) = |F|\ \mathrm{V}C_{path}(F)$, and $|G| = O(|F|^2)$.*

Proof. Take $|F|$ copies $F_1, \ldots, F_{|F|}$ of F, and $|F|-1$ new vertices $v_1, \ldots, v_{|F|-1}$. We add edges from v_i to all vertices of F_i, and F_{i+1} $(1 \leq i < |F|)$. We claim that the resulting graph G fulfills $\mathrm{V}C_{path}(G) = |F|\ \mathrm{V}C_{path}(F)$. We immediately have $\mathrm{V}C_{path}(G) \geq |F|\ \mathrm{V}C_{path}(F)$ from the construction. To show the other direction, assume we have a set W of $|F|\ \mathrm{V}C_{path}(F)+1$ vertices shattered by paths. First note that W cannot contain any of the vertices v_i (since they disconnect G, hence all of W would have to lie on one side of it). Hence W contains $\mathrm{V}C_{path}(F)+1$ vertices in some F_i. However, this implies that these $\mathrm{V}C_{path}(F)+1$ are shattered by paths within F_i which contradicts $\mathrm{V}C_{path}(F_i) = \mathrm{V}C_{path}(F)$. $\diamondsuit$

3 Paths and Cycles

Our goal is to show that computing the path VC-dimension is $\mathbf{\Sigma_3^p}$-complete. We approach this goal in two steps: we first show that computing the (directed) cycle VC-dimension for directed graphs is $\mathbf{\Sigma_3^p}$-complete, and we then show how to modify the proof for paths, and then for undirected graphs.

Lemma 3.1 DIGRAPH VC_{cycle} DIMENSION *is $\mathbf{\Sigma_3^p}$-complete.*

Proof. We will show how to reduce QSAT_3, the standard $\mathbf{\Sigma_3^p}$-complete problem to DIGRAPH VC_{cycle} DIMENSION. Combined with Lemma 2.1 this proves the result.

Suppose we are given an instance of QSAT_3, that is a formula Ψ of the form $(\exists a)(\forall b)(\exists c)\Phi(a,b,c)$, where Φ is a boolean expression in CNF and the variables in Φ are partitioned into three sets X_1, X_2, and X_3. The strings a, b and c are assignments to variables in these sets, rsp.

We build on a construction that shows that deciding whether a directed graph contains a Hamiltonian cycle is $\mathbf{\Sigma_1^p}$-complete. We refer the reader to the construction in Hopcroft and Ullmann's book [HU79, Section 3.2] which gives us the following: for each quantifier-free formula Φ we can construct in polynomial time a graph $F = F_\Phi$ with the following properties:

- for each variable x of Φ, F_Φ contains three vertices u_x, v_x, and w_x and edges (u_x, v_x), (u_x, w_x), (v_x, w_x), and (w_x, v_x) (and no other edges between these three vertices),
- if the truth-assignment t (mapping variables of Φ to $\{\perp, \top\}$) makes Φ true, then there is a Hamiltonian cycle which includes the edge (u_x, v_x) if $t(x) = \perp$, and the edge (u_x, w_x) otherwise, and

- if Φ has no satisfying truth-assignment, then F_Φ does not contain a Hamiltonian cycle.

We will now modify F to fit our purposes in two steps.

In a first step we change the triangle associated with each variable. Let G be the graph obtained from F as follows: for each $x \in X_1$ we add two new vertices v'_x, and w'_x together with edges (u_x, v'_x), (v'_x, v_x), (u_x, w'_x), (w'_x, w_x). For each $x \in X_2$ we add one new vertex v'_x, remove the edge (u_x, v_x), and add two edges (u_x, v'_x), and (v'_x, v_x). We do not make any changes to the triangles associated with variables from X_3.

In the graph G let us distinguish between the vertices V_1 originally from F, and the vertices V_2 which were added to F in the construction of G. Let $\mathcal{C}$ be the collection of cycles in G that pass through all vertices in V_1. We claim that

$$\Psi \text{ is true if and only if } \mathrm{V}C_{\mathcal{C}}(G) \geq n,$$

where $n = |X_1| + |X_2|$. Let us verify the claim. If Ψ is true, then there is an assignment a to the variables in X_1 such that for all assignments b to variables in X_2 there is an assignment c to variables in X_3 such that $\Phi(a, b, c)$ is true. Fix such an a. Let

$$U = \{v'_x : a(x) = \top, x \in X_1\} \cup \{w'_x : a(x) = \bot, x \in X_1\} \cup \{v'_x : x \in X_2\}.$$

First notice that $|U| = |X_1| + |X_2|$. We claim that U is shattered by $\mathcal{C}$ which immediately implies $\mathrm{V}C_{\mathcal{C}}(G) \geq n$. Let $U' \subseteq U$ be an arbitrary subset of U. Define a truth assignment b for $x \in X_2$ by $b(x) = \top$ if $v'_x \in U'$, and $b(x) = \bot$ otherwise. Having a and b there is a truth assignment c to the variables in X_3 such that $\Phi(a, b, c)$ is true. Corresponding to the truth assignment is a Hamiltonian cycle through F that for each variable x passes through v_x if and only if x has been assigned the value true (otherwise it passes through w_x). We can now extend this Hamiltonian cycle to G in such a way that it contains all vertices in U' and no vertex in $U \setminus U'$. Since U' was chosen as an arbitrary subset of U, and $|U| = n$ this shows that $\mathrm{V}C_{\mathcal{C}}(G) \geq n$.

To show the other direction assume that $\mathrm{V}C_{\mathcal{C}}(G) \geq n$. Let U be a set of n vertices of G that is shattered by $\mathcal{C}$. Note that for each vertex $x \in X_1$ at most one of v'_x and w'_x can be in U (there is no path from one to the other). Since $|X_1| + |X_2| = n$, and $U \subseteq \{v'_x, w'_x : x \in X_1\} \cup \{v'_x : x \in X_2\}$ we can conclude that U contains all of $\{v'_x : x \in X_2\}$ and exactly one vertex from each pair $\{v'_x, w'_x\}$ where $x \in X_1$. Define $a(x) = \top$ if $v'_x \in U$ for $x \in X_1$, and $a(x) = \bot$ otherwise. Let b be an arbitrary truth assignment to the vertices in c. Now define $U' = \{v'_x : a(x) = \top, x \in X_1\} \cup \{w'_x : a(x) = \bot, x \in X_1\} \cup \{v'_x : b(x) = \top, x \in X_2\}$. Then U' is a subset of U (by the choice of a). Hence there is a cycle C through G which contains all vertices in V_1 and U', and no vertex from $U \setminus U'$. If C passes through v'_x for some $x \in X_1$ it also has to pass through v_x, and similarly it will pass through w_x if it passes through w'_x. If C passes through v'_x for some $x \in X_2$ it will pass through v_x, and if it does not pass through v'_x it will have to pass through w_x. Hence if we restrict C to F we get a Hamiltonian cycle that corresponds to a truth assignment to Φ that extends a and b to the variables of X_3. Since b was chosen arbitrarily this implies that Ψ is true.

We have showed that Ψ is true if and only if $\mathrm{V}C_{\mathcal{C}}(G) \geq n$, where $n = |X_1| + |X_2|$.

Remember that the vertices of G are partitioned into V_1 (those contained in all $\mathcal{C}$) and V_2 (subsets of which are shattered by $\mathcal{C}$). We now obtain H from G by splitting each vertex $v \in V_1$ of G into two vertices v_1 and v_2 such that v_1 has only incoming edges, and v_2 only outgoing edges. For each vertex $v \in V_1$ we add a new copy of a clique K_m (where $m = 2|G|$) and add edges from v_1 to each vertex of

K_m, and edges from each vertex of K_m to v_2. Finally we include an edge from v_1 to v_2. We say that the clique is associated with v. No changes to the vertices in V_2 are necessary.

If $\mathrm{V}C_{\mathcal{C}}(G) \geq 0$, then $\mathrm{V}C_{cycle}(H) \geq m|V_1| + \mathrm{V}C_{\mathcal{C}}(G)$. This is immediate from the construction of H: let U be a set of $\mathrm{V}C_{\mathcal{C}}(G)$ vertices shattered by $\mathcal{C}$ in G. Obviously $U \subseteq V_2$ (since the vertices in V_1 belong to each path in $\mathcal{C}$. Then the $|V_1|$ cliques K_m that are associated with vertices in V_1 together with the $\mathrm{V}C_{\mathcal{C}}(G)$ in U (as vertices in H) are shattered by cycles in H. This follows from the construction: for the cliques in V_1 note that we can skip them by the edge from v_1 to v_2, and we can skip a clique associated with a vertex v'_x or w'_x in V_2 because we can avoid that vertex in G.

Now suppose that $\mathrm{V}C_{cycle}(H) > m|V_1| + \mathrm{V}C_{\mathcal{C}}(G)$. Let U be a set of $\mathrm{V}C_{cycle}(H)$ vertices shattered in H by cycles. We first note that U contains vertices in all cliques K_m associated with vertices in V_1. If that was not the case, the size of U would be at most $m(|V_1|-1)+2|V_1|+|V_2| \leq m(|V_1|-1)+2|G| \leq m|V_1|$ contradicting $|U| = \mathrm{V}C_{cycle}(H) > m|V_1| + \mathrm{V}C_{\mathcal{C}}(G)$. Let $\mathcal{D}$ be the set of cycles in H that pass through every vertex in every clique associated with a vertex in V_1. Since U contains vertices in all these cliques, we can conclude that

$$\mathrm{V}C_{\mathcal{D}}(H) > \mathrm{V}C_{\mathcal{C}}(G).$$

Let U' be the set of $\mathrm{V}C_{\mathcal{D}}(H)$ vertices shattered by $\mathcal{D}$ in H. First note that $U' \subseteq V_2$. This is because $\mathcal{D}$ contains all the vertices in the cliques associated with V_1, and hence it cannot contain any of the vertices v_1, v_2 for $v \in V_1$ (if it contained v_1, then we cannot avoid v_1 and contain a vertex in the corresponding clique, if it contained v_2 it has to contain a vertex in the corresponding clique). But this implies that $\mathrm{V}C_{\mathcal{D}}(H) = \mathrm{V}C_{\mathcal{C}}(G)$, a contradiction, establishing $\mathrm{V}C_{cycle}(H) \leq m|V_1| + \mathrm{V}C_{\mathcal{C}}(G)$.

We conclude that $\mathrm{V}C_{cycle}(H) = m|V_1| + \mathrm{V}C_{\mathcal{C}}(G)$ if $\mathrm{V}C_{\mathcal{C}}(G) \geq 0$. Combining the two steps of the construction yields that Ψ is true if and only if $\mathrm{V}C_{\mathcal{C}}(G) \geq n$ if and only if $\mathrm{V}C_{cycle}(H) \geq m|V_1| + n$ (note that $n \geq 0$). This shows that QSAT_3reduces to DIGRAPH VC_{cycle} DIMENSION. $\diamondsuit$

Corollary 3.2 DIGRAPH VC_{path} DIMENSION *is* $\mathbf{\Sigma_3^p}$*-complete.*

Proof. This is a corollary to the proof of Lemma 3.1. The original graph F contains an edge (u, v) that all Hamiltonian cycles of F pass through. This edge is still present in H, and all the cycles in a set of cycles shattering $\mathrm{V}C_{cycle}(H)$ vertices pass through this edge. Construct a new graph H' from H by removing the edge (u, v), and adding two copies of a K_n (with $n > |H|$). Furthermore add edges from one of the K_n to u, and edges from v to the other. Then $\mathrm{V}C_{path}(H') = \mathrm{V}C_{cycle}(H) + 2n$, hence computing $\mathrm{V}C_{path}$ on directed graphs is $\mathbf{\Sigma_3^p}$-complete. $\diamondsuit$

Corollary 3.3 GRAPH VC_{cycle} DIMENSION *is* $\mathbf{\Sigma_3^p}$*-complete.*

Proof. Consider the graph H constructed in the proof of Lemma 3.1, and let H' be the undirected graph we get from H by removing the direction of edges in H. We claim that

$$\mathrm{V}C_{cycle}(H) = \mathrm{V}C_{cycle}(H').$$

This, of course, is not true in general and depends on the particular structure of the graph. We have to show that $\mathrm{V}C_{cycle}(H) \geq \mathrm{V}C_{cycle}(H')$. Assume for a contradicion that $\mathrm{V}C_{cycle}(H') > m|V_1| + \mathrm{V}C_{\mathcal{C}}(G)$. As in Lemma 3.1 we conclude that

$$\mathrm{V}C_{\mathcal{E}}(H') > \mathrm{V}C_{\mathcal{C}}(G),$$

where $\mathcal{E}$ now is the set of undirected cycles in H' that pass through each vertex in every clique associated with vertices in V_1 (this was a pure counting argument, not depending on the direction of edges). Now consider a cycle C in $\mathcal{E}$. For any vertex $v \in V_1$ the cycle C has to contain both v_1 and v_2 since it contains all vertices in the associated clique, and it has to enter and exit the clique, for which only v_1

and v_2 are available. Furthermore vertices $v \in V_2$ all of degree two (one incoming, one outgoing edge in H). This means that to C corresponds a directed cycle D in H containing all $v \in V_1$, and $v \in V_2$ if and only if $v \in C$. Therefore $\mathrm{V}C_{\mathcal{E}}(H') = \mathrm{V}C_{\mathcal{D}}(H) = \mathrm{V}C_{\mathcal{C}}(G)$ (the latter equation by the proof of Lemma 3.1) which contradicts our assumption. $\diamondsuit$

Repeating the argument of Corollary 3.2 we obtain the result we were looking for.

Corollary 3.4 GRAPH VC_{path} DIMENSION *is* $\mathbf{\Sigma_3^p}$*-complete.*

References

[AB92] Martin Anthony and Norman Biggs. *Computational Learning Theory.* Cambridge University Press, Cambridge, UK, 1992.

[ABC95] Martin Anthony, Graham Brightwell, and Colin Cooper. The Vapnik-Chervonenkis dimension of a random graph. *Discrete Mathematics*, 138:43–56, 1995.

[HU79] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata and Formal Languages.* Addison-Wesley, Reading, MA, 1979.

[HW87] David Haussler and Emo Welzl. Epsilon-nets and simplex range queries. *Discrete and Computational Geometry*, 2:127–151, 1987.

[KKR+97] Evangelos Kranakis, Danny Krizanc, Berthold Ruf, Jorge Urrutia, and Gerhard Woeginger. The VC-dimension of set systems defined by graphs. *Discrete Applied Mathematics and Combinatorial Operations Research and Computer Science*, 77:237–257, 1997.

[PY93] Christos H. Papadimitriou and Mihalis Yannakakis. On limited nondeterminism and the complexity of the V-C dimension. In *Proceedings of the 8th Annual Conference on Structure in Complexity Theory (SCTC '93)*, pages 12–18, San Diego, CA, USA, May 1993. IEEE Computer Society Press.

[Sch99a] Marcus Schaefer. Graph ramsey theory and the polynomial hierarchy. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 592–601. ACM, 1999.

[Sch99b] Marcus Schäfer. Deciding the Vapnik-Cervonenkis dimension is Σ_3^P-complete. *Journal of Computer and System Sciences*, 58:177–182, 1999.

[Uma98] Christopher Umans. The minimum equivalent DNF problem and shortest implicants. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 556–563, 1998.

[Uma99] Christopher Umans. Hardness of Approximating Σ_2^p Minimization Problems. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 465–474, 1999.