The Parameterized Complexity of Computing the VC-Dimension

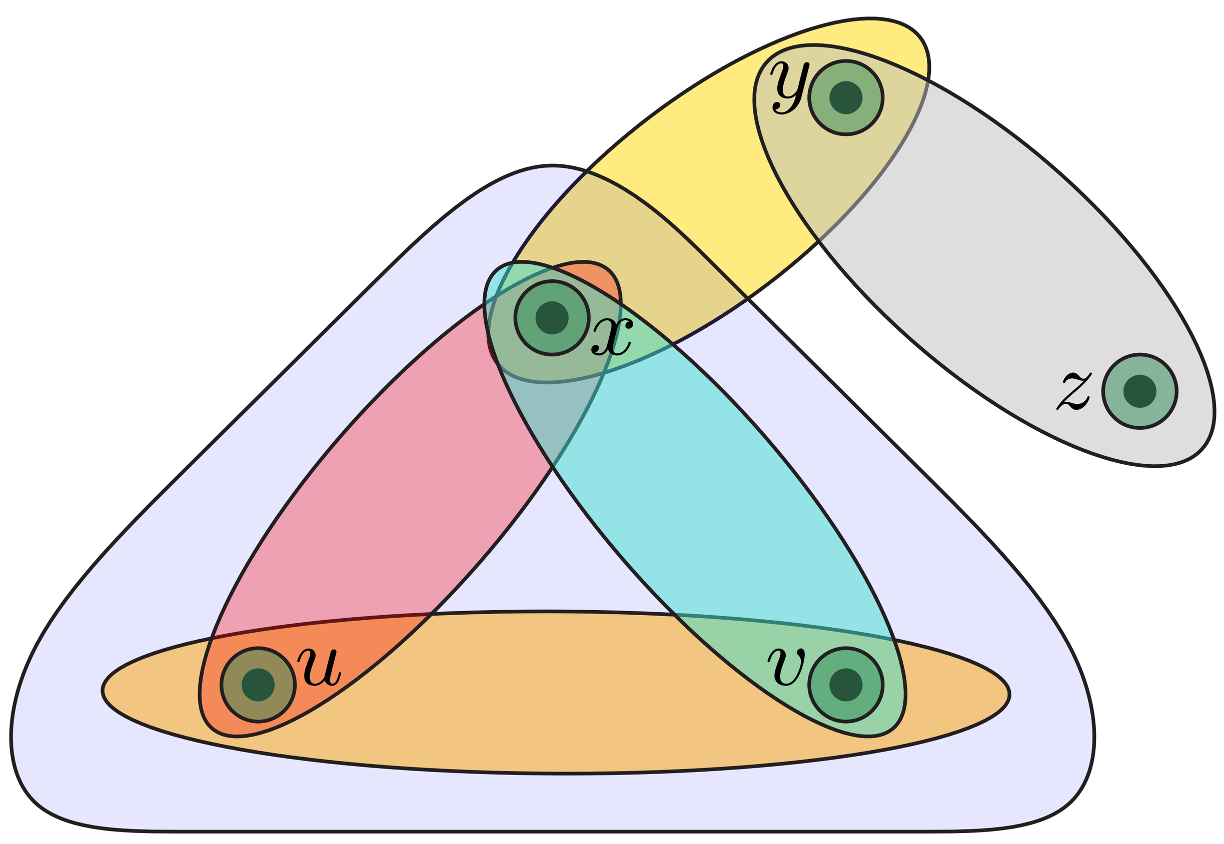
Florent Foucaud[1], Harmender Gahlawat[1], Fionn Mc Inerney[2], and Prafullkumar Tale[3]

[1]Université Clermont Auvergne (France), [2]Telefónica Scientific Research (Spain), [3]IISER Pune (India)

VC-Dimension

Given a non-empty finite set $\mathcal{V}$ and a set system $\mathcal{C} \subseteq 2^{\mathcal{V}}$, the **VC-dimension** of $\mathcal{C}$ is the size of a **largest** subset $S \subseteq \mathcal{V}$ that is **shattered** by $\mathcal{C}$, i.e., such that $\{C \cap S : C \in \mathcal{C} = 2^S\}$.

Can be represented by a **hypergraph** $\mathcal{H} = (\mathcal{V}, \mathcal{E})$:



The VC-dimension of $\mathcal{H}$ is 3 and $S = \{u, v, x\}$ is a shattered set:

$\{y, z\} \cap S = \emptyset$ $\quad$ $\{u, v, x\} \cap S = \{u, v, x\}$

$\{u\} \cap S = \{u\}$ $\quad$ $\{v\} \cap S = \{v\}$ $\quad$ $\{x\} \cap S = \{x\}$

$\{u, v\} \cap S = \{u, v\}$ $\quad$ $\{u, x\} \cap S = \{u, x\}$ $\quad$ $\{x, v\} \cap S = \{x, v\}$

The VC-dimension is a fundamental complexity measure of set systems that is central to many areas of machine learning such as **ϵ-nets, sample compression schemes, and machine teaching.**

Known results for VC-Dimension:

- LogNP-hard and, assuming the Exponential Time Hypothesis (ETH), cannot be solved in $|\mathcal{H}|^{o(\log |\mathcal{H}|)}$ time (and this is tight) [Papadimitriou, Yannakakis, 1996];
- assuming the Gap-ETH, cannot be $o(\log |\mathcal{H}|)$-approximated in polynomial time [Manurangsi, 2023];
- W[1]-hard parameterized by k [Downey et al., 1993] or the degeneracy of $\mathcal{H}$ [Drange et al., 2023].

Problem Statements

VC-Dimension
Input: A hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and $k \in \mathbb{N}$.
Question: Does there exist a subset $S \subseteq \mathcal{V}$ such that $|S| \geq k$ and $\{S \cap e : e \in \mathcal{E}\} = 2^S$?

Graph VC-Dimension
Input: A graph $G = (V, E)$ and $k \in \mathbb{N}$.
Question: Does there exist a subset $S \subseteq V$ such that $|S| \geq k$ and $\{S \cap N(v) : v \in V\} = 2^S$?

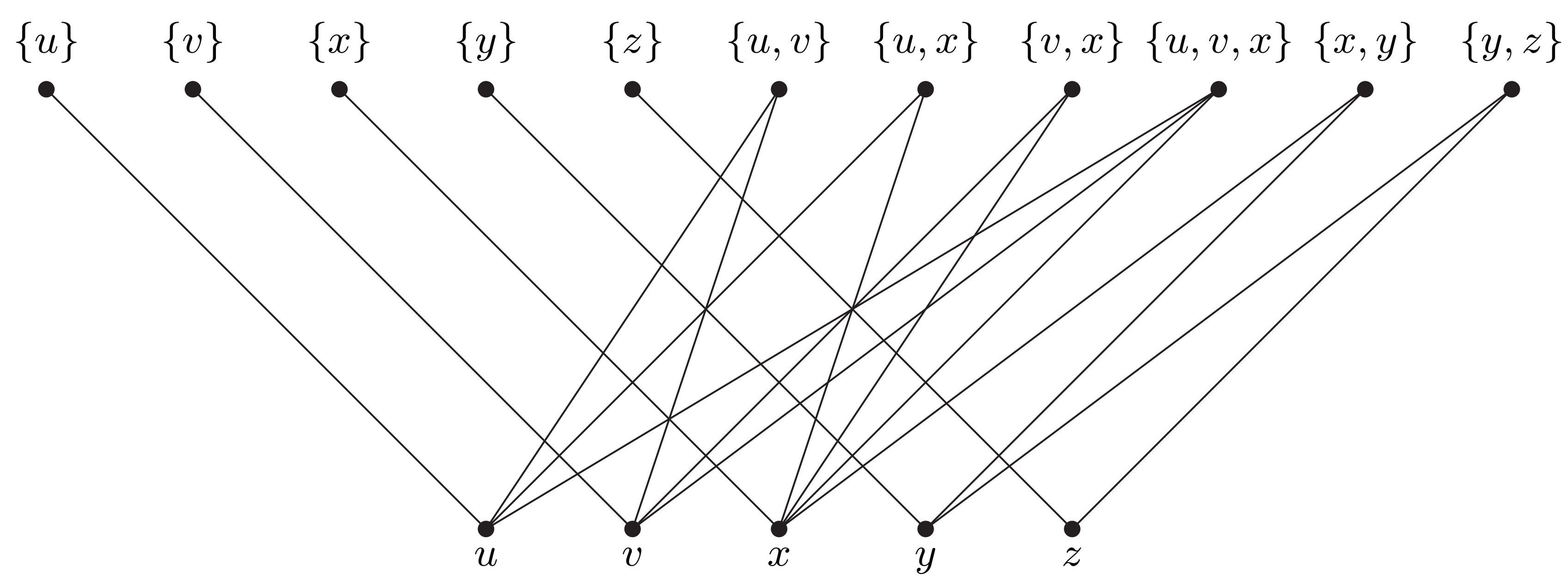
Generalized VC-Dimension (Gen-VC-Dim)
Input: A graph $G = (V, E)$, two subsets $X, Y \subseteq V$, and $k \in \mathbb{N}$.
Question: Does there exist a subset $S \subseteq X$ such that $|S| \geq k$ and $\{S \cap N(y) : y \in Y\} = 2^S$?

VC-Dimension in Graphs

Any finite set system or hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ can be **equivalently represented** by a set of **open neighborhoods** in a **graph** G:

- for all $e \in \mathcal{E}$, there is a vertex x_e
- for all $v \in \mathcal{V}$, there is a vertex v
- x_e is adjacent to v if and only if $v \in e$
- the set of open neighborhoods is $\{N(x_e) : e \in \mathcal{E}\}$

For example, the hypergraph on the left can be equivalently represented by the open neighborhoods of the vertices in the top part of the following bipartite graph:



Results

- Assuming the ETH, VC-Dimension does not admit an algorithm running in $2^{o(|\mathcal{V}|)} \cdot |\mathcal{H}|^{\mathcal{O}(1)}$ time (and this is tight).
- $2^{\mathcal{O}(\Delta \log \Delta)} \cdot |\mathcal{H}|^{\mathcal{O}(1)}$ time 1-additive FPT approximation algorithm for VC-Dimension.
- $2^D \cdot |\mathcal{H}|^{\mathcal{O}(1)}$ time FPT algorithm for VC-Dimension.
- VC-Dimension is LogNP-hard, even if $\mathcal{H}$ is a hypertree with transversal number 1.
- $2^{\mathcal{O}(\mathtt{tw} \log \mathtt{tw})} \cdot |\mathcal{V}|$ time FPT algorithm for Gen-VC-Dim.
- Assuming the ETH, Graph VC-Dimension does not admit an algorithm running in $2^{o(\mathtt{vcn}+k)} \cdot |\mathcal{V}|^{\mathcal{O}(1)}$ time.

ETH: Exponential Time Hypothesis $\quad$ $|\mathcal{H}| := |\mathcal{V}| + |\mathcal{E}|$

Δ: maximum degree of $\mathcal{H}$ $\quad$ D: dimension of $\mathcal{H}$

$\mathtt{tw}$: treewidth of G $\quad$ $\mathtt{vcn}$: vertex cover number of G

Future Directions

- Improve the FPT 1-additive approximation algorithm to an FPT algorithm for VC-Dimension parameterized by Δ?
- Close the gap between the $2^{\mathcal{O}(\mathtt{tw} \log \mathtt{tw})} \cdot |\mathcal{V}|$ time FPT algorithm and the $2^{o(\mathtt{vcn}+k)} \cdot |\mathcal{V}|^{\mathcal{O}(1)}$ ETH-based lower bound for Gen-VC-Dim.
- Consider the setting in which the set system is defined by a circuit, which allows the input size to be dependent only on the size of the domain in some cases.

Acknowledgements

CAP 20-25 I-SITE Clermont Auvergne Project Université Clermont Auvergne · France 2030 · anr agence nationale de la recherche · ReLaX Research Lab in Computer Science · TaRDIS · inspire