Logical Methods in Computer Science
Volume 21, Issue 3, 2025, pp. 9:1–9:40
https://lmcs.episciences.org/

Submitted Mar. 26, 2024
Published Jul. 25, 2025

LEARNING CONCEPTS DEFINABLE IN FIRST-ORDER LOGIC WITH COUNTING

STEFFEN VAN BERGEREM

Humboldt-Universität zu Berlin, Germany
e-mail address: steffen.van.bergerem@hu-berlin.de

Abstract. We study Boolean classification problems over relational background structures in the logical framework introduced by Grohe and Turán (TOCS 2004). It is known (Grohe and Ritzert, LICS 2017) that classifiers definable in first-order logic over structures of polylogarithmic degree can be learned in sublinear time, where the degree of the structure and the running time are measured in terms of the size of the structure. We generalise the results to the first-order logic with counting FOCN, which was introduced by Kuske and Schweikardt (LICS 2017) as an expressive logic generalising various other counting logics. Specifically, we prove that classifiers definable in FOCN over classes of structures of polylogarithmic degree can be consistently learned in sublinear time. This can be seen as a first step towards extending the learning framework to include numerical aspects of machine learning. We extend the result to agnostic probably approximately correct (PAC) learning for classes of structures of degree at most $(\log \log n)^c$ for some constant c. Moreover, we show that bounding the degree is crucial to obtain sublinear-time learning algorithms. That is, we prove that, for structures of unbounded degree, learning is not possible in sublinear time, even for classifiers definable in plain first-order logic.

1. Introduction

In this paper, we study Boolean classification problems, where the input elements for the task come from a set X, the *instance space*. A *classifier* on X is a function $c\colon X \to \{0,1\}$. Given a *training sequence* T of labelled examples $(x,\lambda) \in X \times \{0,1\}$, we want to find a classifier, called a *hypothesis*, that explains the labels given in T, and that can also be used to predict the labels of elements from X not given as examples.

Regarding the requirements we impose on the hypotheses, we consider the following classic scenarios from computational learning theory. In *consistent learning*, the examples

Key words and phrases: first-order logic with counting, agnostic PAC learning, polylogarithmic degree, supervised learning.

* An extended abstract of this paper appeared in the Proceedings of the 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS 2019).

The work on the extended abstract was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 389872375 (gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 389872375). The subsequent work on this full version was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 431183758 (gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 431183758).

DOI:10.46298/LMCS-21(3:9)2025

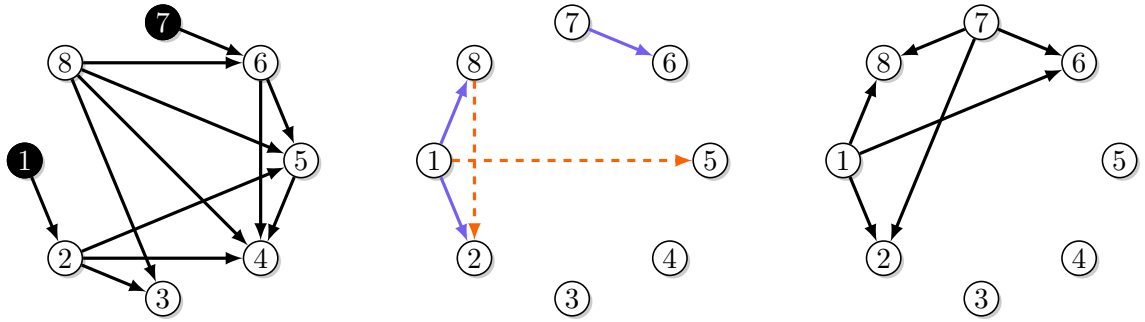

Figure 1: (Left) The database of an online encyclopedia viewed as a directed graph. Vertices represent pages, category pages are black, and edges represent hyperlinks. (Centre) Training examples. Positive examples, *i.e.*, tuples that should be contained in the relation, are shown as solid purple edges while negative examples are shown as dashed orange edges. (Right) The learned relation from Example 1.1.

are assumed to be generated using an unknown classifier, the *target concept*, from a known *concept class*. The task is to find a hypothesis that is *consistent* with the training sequence T, *i.e.* a function $h\colon X \to \{0,1\}$ such that $h(x) = \lambda$ for all $(x,\lambda) \in T$. In Haussler's model of *agnostic probably approximately correct (PAC) learning* [Hau92], a generalisation of Valiant's *PAC-learning* model [Val84], an (unknown) probability distribution $\mathcal{D}$ on $X \times \{0,1\}$ is assumed, and training examples are drawn independently from this distribution. The goal is to find a hypothesis that generalises well. That is, algorithms should return with high probability a hypothesis with a small expected error on new instances drawn from the same distribution. We discuss both models in more detail in Section 3.

We study learning problems in the framework that was introduced by Grohe and Turán [GT04] and further studied in [GR19, GLR17, GR17, vB19, vBS21, vBGR22, vB23]. There, the instance space X is a set of tuples from a relational structure, called the *background structure*, and classifiers are described using parametric models based on logics. Formally, we fix a number $k \in \mathbb{N}$ and, for a background structure $\mathcal{A}$, let the instance space be the set $X = \big(U(\mathcal{A})\big)^k$ of k-tuples of elements of $\mathcal{A}$.

Intuitively, in consistent learning, our goal is to learn a definition of a k-ary relation on the elements of $\mathcal{A}$ that is consistent with a given sequence of examples. That is, the relation shall contain all positive (*i.e.*, $\lambda = 1$) and none of the negative (*i.e.*, $\lambda = 0$) examples.

Example 1.1. Let $\mathcal{A}$ be the following relational structure representing a database of data from an online encyclopedia. The universe of the structure consists of all pages of the encyclopedia. There is a binary relation representing hyperlinks between pages and a unary relation representing category pages. Pages that are not category pages are article pages.

Let $k = 2$. That is, our task is to learn a definition of a binary relation containing tuples of pages. Suppose we want that the first element of the tuple is a category page, and the second element is a page that belongs to the category. The input for our task is a training sequence T of classified tuples, *e.g.*, tuples that have been classified by experts beforehand. That is, the training sequence T consists of pairs $\big((c,p),\lambda\big)$, where $\lambda \in \{0,1\}$, and $\lambda = 1$ if and only if p is a page that belongs to the category c.

Now suppose the database and the training examples are as depicted in Figure 1. Note that a definition of a consistent relation would be the following. The relation contains all

```
SELECT C.'page', CatLink.'to'
FROM Categories C, Links CatLink
WHERE CatLink.'from' = C.'page'
UNION
SELECT C.'page', L1.'from'
FROM Categories C, Links CatLink,
     Links L1, Links L2
WHERE CatLink.'from' = C.'page'
AND CatLink.'to' = L2.'from'
AND L1.'to' = L2.'to'
GROUP BY C.'page', L1.'from'
HAVING count(*) >= 2;
```

Figure 2: An SQL query that defines the relation learned in Example 1.1.

tuples, where the first element is a category page c, and the second element is a page p that fulfils at least one of the following two conditions.

(1) The page p is linked from the category page c.
(2) There is another page p' linked from the category page c, and both pages p, p' have at least two common linked pages.

The corresponding relation can also be seen in Figure 1. For example, the tuple $(1, 8)$ is contained in the relation since Page 1 is a category, Page 2 is linked from the category, and there are at least two pages (Pages 3 and 4) that both Page 2 and Page 8 link to.

Since the data are contained in a relational database, it would be convenient to learn an SQL query that defines the relation. Figure 2 shows an SQL query for the learned relation.

In [GR17], Grohe and Ritzert considered learning tasks where the hypotheses can be described using first-order logic. We are interested in learning hypotheses that can be expressed in SQL. While first-order logic can be seen as the "logical core" of SQL, there are aggregating operators in SQL, namely COUNT, AVG, SUM, MIN, and MAX, that do not have corresponding expressions in first-order logic. Motivated by this, we study the first-order logic with counting FOCN, which Kuske and Schweikardt introduced in [KS17] and which extends first-order logic by cardinality conditions similar to the COUNT operator in SQL. The logic depends on a collection $\mathbb{P}$ of numerical predicates, *i.e.*, functions $P\colon \mathbb{Z}^m \to \{0, 1\}$, that can be used in formulas to express restrictions on the results of counting terms.

Let $\mathcal{A} = (U(\mathcal{A}), L(\mathcal{A}), C(\mathcal{A}))$ be the background structure from Example 1.1, where the universe $U(\mathcal{A})$ is the set of all pages, $L(\mathcal{A})$ is the binary relation of links between pages, and $C(\mathcal{A})$ is the unary relation of category pages. The SQL query from Figure 2 can be expressed as the FOCN formula

$$\varphi(c, p) \coloneqq C(c) \wedge \Big(L(c, p) \vee \exists x \Big(L(c, x) \wedge \#(z).\big(L(x, z) \wedge L(p, z)\big) \geq 2 \Big) \Big),$$

where we assume that the numerical predicate $\geq$ is contained in $\mathbb{P}$. The counting term $\#(z).\big(L(x, z) \wedge L(p, z)\big)$ counts the number of pages z such that both x and p link to z. The formula $\#(z).\big(L(x, z) \wedge L(p, z)\big) \geq 2$ checks whether this number is at least 2. In a more

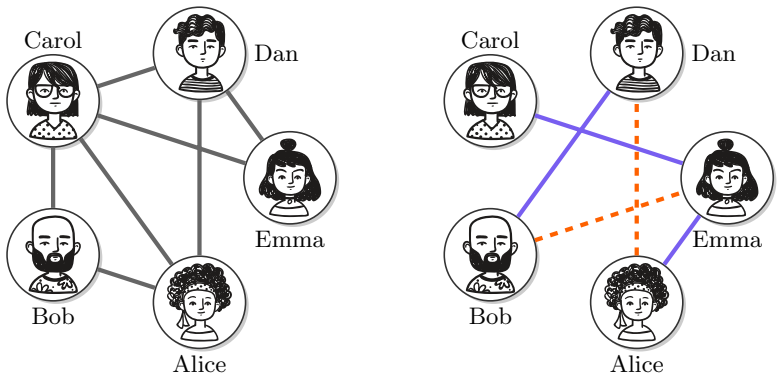

Figure 3: (Left) A friendship graph based on data from a social network[2]. An edge between two members indicates that they are friends. (Right) The training sequence from Example 1.2. Positive examples are shown as solid purple edges while negative examples are shown as dashed orange edges.

general approach, we may use the formula

$$\varphi'(c,p;\kappa) \coloneqq C(c) \wedge \Big(L(c,p) \vee \exists x\Big(L(c,x) \wedge \#(z).\big(L(x,z) \wedge L(p,z)\big) \geq \kappa\Big)\Big)$$

with the free number variable κ. When viewed as a parameter, for every assignment of κ, we obtain a new hypothesis.

In this paper, we specify hypotheses by triples $(\varphi(\bar{x};\bar{y},\bar{\kappa}),\bar{w},\bar{n})$. Here, $\varphi(\bar{x};\bar{y},\bar{\kappa})$ is an FOCN formula with free variables $\bar{x}=(x_1,\dots,x_k)$, $\bar{y}=(y_1,\dots,y_\ell)$, and $\bar{\kappa}=(\kappa_1,\dots,\kappa_m)$. Moreover, $\bar{w}=(w_1,\dots,w_\ell)\in\big(U(\mathcal{A})\big)^\ell$ and $\bar{n}=(n_1,\dots,n_m)\in\mathbb{Z}^m$ are tuples of elements of $U(\mathcal{A})$ and integers, respectively, called *parameter tuples*. The corresponding hypothesis is the mapping $h^{\mathcal{A}}_{\varphi,\bar{w},\bar{n}}\colon \big(U(\mathcal{A})\big)^k \to \{0,1\}$ which maps a tuple $\bar{v}=(v_1,\dots,v_k)\in\big(U(\mathcal{A})\big)^k$ to 1 if and only if φ is satisfied in $\mathcal{A}$ when interpreting the variables $x_1,\dots,x_k$ with $v_1,\dots,v_k$, interpreting $y_1,\dots,y_\ell$ with $w_1,\dots,w_\ell$, and $\kappa_1,\dots,\kappa_m$ with $n_1,\dots,n_m$. Otherwise, $\bar{v}$ is mapped to 0. For the training sequence T given in Example 1.1, the hypothesis $h^{\mathcal{A}}_{\varphi',\bar{w},\bar{n}}$ is consistent with T, where φ' is as specified above, $\bar{w}=()$ is the empty tuple, and $\bar{n}=(2)$. Here, we have $k=2$, $\ell=0$, and $m=1$.

Example 1.2. Let $G=\big(V(G),E(G)\big)$ be the friendship graph shown in Figure 3 based on data from a social network. We consider the instance space $X=\big(V(G)\big)^2$. Let the training sequence T contain (Alice, Emma), (Bob, Dan), and (Carol, Emma) as positive examples, and (Alice, Dan) and (Bob, Emma) as negative examples. The examples are also depicted in Figure 3. One hypothesis that is consistent with the labelled examples is the function $h\colon X\to\{0,1\}$ with $h(v_1,v_2)=1$ if and only if v_1 and v_2 have a common friend who is not Carol. This hypothesis can be defined as $h=h_{\varphi,\bar{w},\bar{n}}$ with $\varphi(x_1,x_2;y,\kappa)\coloneqq\Big(\#(z).\big(E(x_1,z)\wedge E(x_2,z)\wedge\neg(z{=}y)\big)\geq\kappa\Big)$, $\bar{w}\coloneqq(\text{Carol})$, and $\bar{n}\coloneqq(1)$. In this example, we have $k=2$, $\ell=1$, and $m=1$. The formula φ with parameters $\bar{w}$ and $\bar{n}$ counts the number of common friends of the vertices interpreted by x_1 and x_2 who are not Carol, and it checks that this number is at least 1.

[2]Avatars designed by pikisuperstar on Freepik

1.1. **Our Contributions.** We study learning problems for hypotheses that can be described using the first-order logic with counting FOCN. We analyse our algorithms under the *logarithmic-cost* measure and the *uniform-cost* measure. Under the logarithmic-cost measure, storing and accessing an element of the background structure takes time and space logarithmic in the size of the structure, whereas under the uniform-cost measure, both operations take constant time and space.

In Section 3, we show that bounding the degree of the structures is crucial to obtain sublinear-time learning algorithms, even for hypotheses that can be defined by pure first-order logic. More specifically, for classes of structures without a degree restriction, we show that there are no consistent-learning and no PAC-learning algorithms for first-order definable hypotheses that run in sublinear time.

For background structures that come from a class of bounded degree, we show that, under the logarithmic cost-measure, the consistent-learning problem can be solved in time polylogarithmic in the size of the background structure and polynomial in the number of training examples. Under the uniform-cost measure, we solve the problem in time polynomial in the number of training examples and independent of the size of the background structure. The hypotheses the algorithm returns can be evaluated in time polylogarithmic in the size of the background structure under the logarithmic-cost measure and in constant time under the uniform-cost measure. In addition, we extend this result to PAC-learning problems. We show all of these results in Section 6.

In Section 7, we consider classes of background structures where the degree is not uniformly bounded. For classes of structures where the degree of the structure is at most polylogarithmic in the size of the structure, our results imply that the consistent-learning problem can be solved in time sublinear in the size of the structure. For the PAC-learning problem, we obtain an analogous result on classes of structures where the degree of a structure $\mathcal{A}$ is bounded by $\big(\log(\log|U(\mathcal{A})|)\big)^c$ for some constant c.

1.2. **Related Work.** The learning framework that we consider in this paper was introduced in [GT04]. There, the authors proved information-theoretic learnability results for concepts that can be described using first-order and monadic second-order logic on restricted classes of background structures, such as the class of planar graphs and classes of graphs of bounded degree. Algorithmic aspects of the framework were first studied in [GR17]. The authors showed that first-order definable concepts are both consistent- and PAC-learnable in sublinear time over structures of at most polylogarithmic degree. In [GLR17], the authors examined the learnability of concepts definable in first-order and monadic second-order logic over simple structures of unbounded degree, namely ordered strings. Even in the unary case, *i.e.* for $X = U(\mathcal{A})$, they were able to show that there is no consistent-learning algorithm for first-order definable concepts running in sublinear time. However, by introducing a linear-time preprocessing phase to build an index for the background structure, concepts definable in monadic second-order logic can be learned in sublinear learning time. In [GR19], the results were extended from strings to trees.

Our consistent-learning result in Section 7 is a direct generalisation of the corresponding result for first-order logic [GR17], albeit with a running time that is quasi-polynomial in the degree, while the running time in [GR17] is polynomial in the degree. This generalisation is motivated by the fact that typical machine-learning models have numerical parameters; our results may be seen as a first step towards including numerical aspects in the learning framework. At least for background structures of small (say, logarithmic) but unbounded

degree, it is not obvious that an extension of the first-order result to FOCN holds at all. The reason is that FOCN loses its strong locality properties on structures of unbounded degree. For example, by comparing the degree sequences of the neighbours of nodes, one can establish quite complex relations that may range over long distances. Indeed, as shown in [GS18], various algorithmic meta theorems (with proofs based on locality properties) fail when extended from first-order logic to first-order logic with counting.

Thus, it is not surprising that, even though our result looks similar to the corresponding result for first-order logic, there are significant differences in the proofs. The proof of the first-order result in [GR17] is based on Gaifman's theorem, but there is no analogue of Gaifman's theorem for the counting logic FOCN. Instead, our proof is based on a variant of Hanf's theorem for FOCN [KS17]. This raises the technical difficulty that we have to deal with isomorphism types of local neighbourhoods in our structures. On classes of structures without a uniform bound on the degree, in contrast to the approach based on Gaifman's theorem in [GR17], this means that the size of the formulas we have to deal with may depend on the size of the structure. Hence, standard model-checking results (as used in [GR17]) do not yield the desired running-time bounds. Instead, we apply a recent graph isomorphism test running in time $n^{\text{polylog}(d)}$ for n-vertex graphs of maximum degree d [GNS23].

Our negative results for learning on structures of unbounded degree are similar to the ones given in [GLR17] for strings. There, however, the authors consider a more restrictive access model on the background structures.

The first-order logic with counting FOCN that we consider in this paper was introduced in [KS17]. The logic generalises logics such as FO(Cnt) from [Lib04] and FO+C from [Gro17]. In [vBS21], the authors introduced the new logic *first-order logic with weight aggregation* (FOWA) that operates on weighted structures and enables the aggregation of weights in terms similar to the counting terms of FOCN. The authors show that hypotheses definable in a fragment of FOWA can be learned in sublinear time on structures of at most polylogarithmic degree after a quasi-linear-time preprocessing step. The logic FOWA extends the fragment FOC of FOCN, but it is incomparable with FOCN.

Closely related to our learning framework is the framework of *inductive logic programming (ILP)* [Mug91, MR94, KD94, CJ95, CDEM22]. In both frameworks, we are in a passive supervised learning setting, where the learning algorithms are given labelled examples. These examples are labelled according to some target concept, and the algorithms should return a hypothesis that approximately matches this target concept. One of the main differences between both frameworks is that we encode the background knowledge in a relational structure, whereas in ILP, it is represented in a background theory, *i.e.*, a set of formulas. PAC-learning results for ILP have often been obtained by syntactically restricting the hypothesis classes (see, *e.g.*, [CJ95, KD94]), while we use restricted classes of background structures such as classes of small degree.

In the database literature, various approaches to learning queries from examples have been studied, both in passive (such as ours) and active learning settings. In passive learning settings, results often focus on conjunctive queries [Hau89, Hir00, BR17, KR18, BBDK21] or consider queries outside the relational database model [SW12, BCL15], while we focus on (extensions of) full first-order logic. In the *active learning* setting, as introduced by Angluin in [Ang87], learning algorithms are allowed to actively query an oracle. This includes membership queries that enable the learning algorithm to actively choose examples and obtain their corresponding labels. Results in this setting [AHHP98, SST10, AAP$^+$13, BCL15] again consider different types of queries, including conjunctive queries [tCD22]. Another

related subject in the database literature is the problem of learning schema mappings from examples [BCCT19, GS10, AtCKT11, tCDK13, tCKQT18]. In formal verification, related logical learning frameworks have been studied as well [GLMN14, LMN16, END$^+$18, ZMJ18, CCKS20].

Regarding PAC learning of concepts defined by logics, recent work has studied Occam algorithms for description-logic concepts [tCFJL23]. In such Occam algorithms (as introduced in computational learning theory, see [BEHW89]), the complexity of the returned concept should be bounded in terms of the complexity of the target concept. We, however, assume a fixed bound on the complexity of the target concept. On the other hand, we study concepts definable in (extensions of) first-order logic, whereas [tCFJL23] considers concepts definable in description logics.

2. Preliminaries

We let $\mathbb{Z}$, $\mathbb{N}$, and $\mathbb{N}_{\geq 1}$ denote the sets of integers, non-negative integers, and positive integers, respectively. For $m, n \in \mathbb{Z}$, we let $[m, n] \coloneqq \{\ell \in \mathbb{Z} \mid m \leq \ell \leq n\}$ and $[n] \coloneqq [1, n]$. For a k-tuple $\bar{v} = (v_1, \ldots, v_k)$, we write $|\bar{v}|$ to denote its *length* k.

2.1. Relational Structures.

A *(relational) signature* is a finite set of relation symbols. Every relation symbol R has an *arity* $\mathrm{ar}(R) \in \mathbb{N}$. Let σ be a signature. A *(relational) structure* $\mathcal{A}$ *over* σ, also called a σ*-structure*, is a tuple consisting of a finite set $U(\mathcal{A})$, called the *universe* of $\mathcal{A}$, and a relation $R(\mathcal{A}) \subseteq (U(\mathcal{A}))^{\mathrm{ar}(R)}$ for every $R \in \sigma$. The size of $\mathcal{A}$ is $|\mathcal{A}| \coloneqq |U(\mathcal{A})|$.

Let $\sigma' \supseteq \sigma$ be a signature. A σ'-structure $\mathcal{A}'$ is a σ'*-expansion* of a σ-structure $\mathcal{A}$ if $U(\mathcal{A}') = U(\mathcal{A})$ and $R(\mathcal{A}') = R(\mathcal{A})$ for all $R \in \sigma$. A σ-structure $\mathcal{B}$ is a *substructure* of a σ-structure $\mathcal{A}$ if $U(\mathcal{B}) \subseteq U(\mathcal{A})$ and $R(\mathcal{B}) \subseteq R(\mathcal{A})$ for every $R \in \sigma$. For a set $X \subseteq U(\mathcal{A})$, the *induced substructure of* $\mathcal{A}$ *on* X is the σ-structure $\mathcal{A}[X]$ with universe $U(\mathcal{A}[X]) = X$ and $R(\mathcal{A}[X]) = R(\mathcal{A}) \cap X^{\mathrm{ar}(R)}$ for every relation symbol $R \in \sigma$. The *union* of two σ-structures $\mathcal{A}$ and $\mathcal{B}$ is the σ-structure $\mathcal{A} \cup \mathcal{B}$ with universe $U(\mathcal{A} \cup \mathcal{B}) = U(\mathcal{A}) \cup U(\mathcal{B})$ and relations $R(\mathcal{A} \cup \mathcal{B}) = R(\mathcal{A}) \cup R(\mathcal{B})$ for all $R \in \sigma$.

A *graph* is a relational structure with signature $\{E\}$ where E is a binary relation symbol. The universe of a graph G is called the *vertex set* of G and is often denoted by $V(G)$; the relation $E(G)$ is called the *edge set* of G. The elements of the vertex set are called *vertices* and the elements of the edge set are called *edges*. All graphs in this paper are undirected and do not contain self-loops, *i.e.* E is symmetric and irreflexive. A unary relation symbol is called a *colour*. A *(σ-)coloured graph* is a σ-expansion of a graph where σ is a signature with $E \in \sigma$ and all other relation symbols in σ are colours.

Let G be a (coloured) graph. If $(v, w) \in E(G)$, then we say that v and w are *neighbours*. The *degree* $\deg(v)$ of a vertex $v \in V(G)$ is the number of neighbours of v and the degree $\deg(G)$ of G is the maximum degree of its vertices.

For $n \in \mathbb{N}$, a *path of length* n in G is a sequence $v_0, \ldots, v_n$ of distinct vertices in $V(G)$ such that $(v_i, v_{i+1}) \in E(G)$ for all $i \in [0, n-1]$. We say that $v_0, \ldots, v_n$ is a *path from* v_0 *to* v_n *in* G. The *distance* $\mathrm{dist}^G(v, w)$ between two vertices $v, w \in V(G)$ is the minimal length of a path from v to w in G; if no such path exists, we set $\mathrm{dist}^G(v, w) \coloneqq \infty$. For a tuple $\bar{v} = (v_1, \ldots, v_k) \in (V(G))^k$ and a vertex $w \in V(G)$, we let $\mathrm{dist}^G(\bar{v}, w) \coloneqq \min_{i \in [k]} \mathrm{dist}^G(v_i, w)$.

For a tuple $\bar{w} = (w_1, \dots, w_\ell) \in (V(G))^\ell$, we set $\operatorname{dist}^G(\bar{v}, \bar{w}) \coloneqq \min_{j \in [\ell]} \operatorname{dist}^G(\bar{v}, w_j)$. We omit the superscript ^{G} when it is clear from the context.

For $r \in \mathbb{N}$ and a tuple $\bar{v} \in (V(G))^k$ for some $k \in \mathbb{N}$, the *ball of radius* r (or r*-ball*) of $\bar{v}$ in G is the set $N_r^G(\bar{v}) \coloneqq \{w \in V(G) \mid \operatorname{dist}^G(\bar{v}, w) \leq r\}$. The *neighbourhood of radius* r (or r*-neighbourhood*) of $\bar{v}$ in G is the induced substructure $\mathcal{N}_r^G(\bar{v}) \coloneqq G[N_r^G(\bar{v})]$. Let $C_1, \dots, C_k$ be new colours not used in G. The *sphere of radius* r (or r*-sphere*) of $\bar{v}$ in G is the structure $\mathcal{S}_r^G(\bar{v})$ that is the expansion of $\mathcal{N}_r^G(\bar{v})$ by the colours $C_1, \dots, C_k$ with $C_i(\mathcal{S}_r^G(\bar{v})) = \{v_i\}$ for all $i \in [k]$.

The *Gaifman graph* $G_\mathcal{A}$ of a σ-structure $\mathcal{A}$ is the graph with vertex set $V(G_\mathcal{A}) = U(\mathcal{A})$ and edge set $E(G_\mathcal{A})$ that contains exactly those pairs of distinct vertices $a, b \in U(\mathcal{A})$ that appear in the same tuple of some relation of $\mathcal{A}$, *i.e.*, $a, b \in \bar{v}$ for some $\bar{v} \in R(\mathcal{A})$ and $R \in \sigma$.

We can generalise the graph-theoretic notions such as *degree*, *paths*, *connectivity*, *distance*, and *balls* from (coloured) graphs to general relational structures by applying the definitions to the corresponding Gaifman graphs. Using the generalised notion of balls, the notions of *neighbourhoods* and *spheres* also naturally generalise from (coloured) graphs to general relational structures.

2.2. **Logics.** In this section, we recapitulate the syntax and semantics of first-order logic as well as its extensions by counting and numerical predicates that we study in this paper.

Throughout this section, let σ be a relational signature. Let vars and nvars be fixed, disjoint, and countably infinite sets of *structure variables* and *number variables*, respectively. In the logics described in this section, structure variables from vars denote elements from the structure, and number variables from nvars denote integers.

A σ*-interpretation* $\mathcal{I} = (\mathcal{A}, \beta)$ consists of a σ-structure $\mathcal{A}$ and an *assignment* $\beta \colon \mathsf{vars} \cup \mathsf{nvars} \to U(\mathcal{A}) \cup \mathbb{Z}$ with $\beta(x) \in U(\mathcal{A})$ for every $x \in \mathsf{vars}$ and $\beta(\kappa) \in \mathbb{Z}$ for every $\kappa \in \mathsf{nvars}$. For $k, \ell \in \mathbb{N}$, k distinct structure variables $x_1, \dots, x_k \in \mathsf{vars}$, elements $v_1, \dots, v_k \in U(\mathcal{A})$, ℓ distinct number variables $\kappa_1, \dots, \kappa_\ell \in \mathsf{nvars}$, and integers $n_1, \dots, n_\ell \in \mathbb{Z}$, we write $\mathcal{I}\frac{v_1,\dots,v_k}{x_1,\dots,x_k}\frac{n_1,\dots,n_\ell}{\kappa_1,\dots,\kappa_\ell}$ for the interpretation $(\mathcal{A}, \beta\frac{v_1,\dots,v_k}{x_1,\dots,x_k}\frac{n_1,\dots,n_\ell}{\kappa_1,\dots,\kappa_\ell})$, where $\beta\frac{v_1,\dots,v_k}{x_1,\dots,x_k}\frac{n_1,\dots,n_\ell}{\kappa_1,\dots,\kappa_\ell}$ is the assignment β' with $\beta'(x_i) = v_i$ for every $i \in [k]$, $\beta'(\kappa_j) = n_j$ for every $j \in [\ell]$, and $\beta'(z) = \beta(z)$ for all $z \in (\mathsf{vars} \cup \mathsf{nvars}) \setminus \{x_1, \dots, x_k, \kappa_1, \dots, \kappa_\ell\}$.

Definition 2.1 ($\mathsf{FO}[\sigma]$)**.** The set of *formulas* for $\mathsf{FO}[\sigma]$ is built according to the following rules.

(1) $x_1{=}x_2$ and $R(x_1, \dots, x_k)$ are formulas for $x_1, x_2, \dots, x_k \in \mathsf{vars}$ and $R \in \sigma$ with $\operatorname{ar}(R) = k$.
(2) If φ and ψ are formulas, then $\neg\varphi$ and $(\varphi \vee \psi)$ are also formulas.
(3) If φ is a formula and $x \in \mathsf{vars}$, then $\exists x\, \varphi$ is a formula.

Let $\mathcal{I} = (\mathcal{A}, \beta)$ be a σ-interpretation. For a formula φ from $\mathsf{FO}[\sigma]$, the semantics $[\![\varphi]\!]^\mathcal{I} \in \{0, 1\}$ is defined as follows.

(1) $[\![x_1{=}x_2]\!]^\mathcal{I} = 1$ if $\beta(x_1) = \beta(x_2)$, and $[\![x_1{=}x_2]\!]^\mathcal{I} = 0$ otherwise; $[\![R(x_1, \dots, x_k)]\!]^\mathcal{I} = 1$ if $\big(\beta(x_1), \dots, \beta(x_k)\big) \in R(\mathcal{A})$, and $[\![R(x_1, \dots, x_k)]\!]^\mathcal{I} = 0$ otherwise.
(2) $[\![\neg\varphi]\!]^\mathcal{I} = 1 - [\![\varphi]\!]^\mathcal{I}$ and $[\![(\varphi \vee \psi)]\!] = \max\{[\![\varphi]\!]^\mathcal{I}, [\![\psi]\!]^\mathcal{I}\}$.
(3) $[\![\exists x\, \varphi]\!]^\mathcal{I} = \max\{[\![\varphi]\!]^{\mathcal{I}\frac{v}{x}} \mid v \in U(\mathcal{A})\}$.

The *quantifier rank* $\operatorname{qr}(\varphi)$ of an $\mathsf{FO}[\sigma]$ formula φ is the maximum nesting depth of constructs using rule (3) in order to construct φ. We write $(\varphi \wedge \psi)$ and $\forall x\, \varphi$ as shorthands for $\neg(\neg\varphi \vee \neg\psi)$ and $\neg\exists x\, \neg\varphi$.

Next, we consider the logic FOCN that Kuske and Schweikardt introduced in [KS17]. This logic allows building numerical statements based on counting terms as well as numerical predicates, and it includes number variables as well as quantification over numbers.

A *numerical predicate collection* is a triple $(\mathbb{P}, \mathrm{ar}, [\![.]\!])$ where $\mathbb{P}$ is a countable set of *predicate names*, and, to each $\mathsf{P} \in \mathbb{P}$, ar assigns an *arity* $\mathrm{ar}(\mathsf{P}) \in \mathbb{N}_{\geq 1}$ and $[\![.]\!]$ assigns a *semantics* $[\![\mathsf{P}]\!] \subseteq \mathbb{Z}^{\mathrm{ar}(\mathsf{P})}$. For the remainder of this paper, fix a numerical predicate collection $(\mathbb{P}, \mathrm{ar}, [\![.]\!])$. When analysing the running time of algorithms, we will assume that machines have access to oracles for evaluating the numerical predicates in constant time. That is, given a predicate $\mathsf{P} \in \mathbb{P}$ and a tuple $(i_1, \ldots, i_{\mathrm{ar}(\mathsf{P})})$ of integers, the oracle call "$(i_1, \ldots, i_{\mathrm{ar}(\mathsf{P})}) \in [\![\mathsf{P}]\!]$?" takes time $\mathcal{O}(1)$.

Definition 2.2 ($\mathsf{FOCN}[\sigma]$)**.** The set of *formulas* and *counting terms* for $\mathsf{FOCN}[\sigma]$ is built according to the rules (1)–(3) and the following rules.

(4) If φ is a formula and $\bar{x} = (x_1, \ldots, x_k)$ is a tuple of k pairwise distinct structure variables, then $\#\bar{x}.\varphi$ is a counting term.
(5) Every integer $i \in \mathbb{Z}$ is a counting term.
(6) If t_1 and t_2 are counting terms, then $(t_1 + t_2)$ and $(t_1 \cdot t_2)$ are also counting terms.
(7) If $\mathsf{P} \in \mathbb{P}$, $m = \mathrm{ar}(\mathsf{P})$ and $t_1, \ldots, t_m$ are counting terms, then $\mathsf{P}(t_1, \ldots, t_m)$ is a formula.
(8) Every number variable $\kappa \in \mathsf{nvars}$ is a counting term.
(9) If φ is a formula and $\kappa \in \mathsf{nvars}$ is a number variable, then $\exists\kappa\,\varphi$ is a formula.

Let $\mathcal{I} = (\mathcal{A}, \beta)$ be a σ-interpretation. For a formula or counting term ξ from $\mathsf{FOCN}[\sigma]$, the semantics $[\![\xi]\!]^{\mathcal{I}}$ is defined by the rules (1)–(3) and the following rules.

(4) $[\![\#\bar{x}.\varphi]\!]^{\mathcal{I}} = \Big|\big\{(v_1, \ldots, v_k) \in \big(U(\mathcal{A})\big)^k \mid [\![\varphi]\!]^{\mathcal{I}\frac{v_1,\ldots,v_k}{x_1,\ldots,x_k}} = 1\big\}\Big|$, where $\bar{x} = (x_1, \ldots, x_k)$.
(5) $[\![i]\!]^{\mathcal{I}} = i$ for $i \in \mathbb{Z}$.
(6) $[\![(t_1 + t_2)]\!]^{\mathcal{I}} = [\![t_1]\!]^{\mathcal{I}} + [\![t_2]\!]^{\mathcal{I}}$ and $[\![(t_1 \cdot t_2)]\!]^{\mathcal{I}} = [\![t_1]\!]^{\mathcal{I}} \cdot [\![t_2]\!]^{\mathcal{I}}$.
(7) $[\![\mathsf{P}(t_1, \ldots, t_m)]\!]^{\mathcal{I}} = 1$ if $([\![t_1]\!]^{\mathcal{I}}, \ldots, [\![t_m]\!]^{\mathcal{I}}) \in [\![\mathsf{P}]\!]$, and $[\![\mathsf{P}(t_1, \ldots, t_m)]\!]^{\mathcal{I}} = 0$ otherwise.
(8) $[\![\kappa]\!]^{\mathcal{I}} = \beta(\kappa)$ for $\kappa \in \mathsf{nvars}$.
(9) $[\![\exists\kappa\,\varphi]\!]^{\mathcal{I}} = \max\big\{[\![\varphi]\!]^{\mathcal{I}\frac{n}{\kappa}} \mid n \in \mathbb{Z}\big\}$.

For counting terms t_1 and t_2, we write $(t_1 - t_2)$ as a shorthand for $\big(t_1 + ((-1) \cdot t_2)\big)$.

Remark 2.3. The semantics of rule (9) differs from the semantics of the corresponding rule in [KS17], where quantified number variables may only range over $[0, |\mathcal{A}|]$, whereas we let them range over the integers. Hence, in contrast to the variant from [KS17], the variant used in the present paper has the full power of integer arithmetic. However, as remarked in [KS17], Theorem 4.1 (Theorem 3.2 in [KS17]) holds for both variants of the logic. Since this is the only result from [KS17] that we build upon, and since we do not rely on the exact definition of the semantics of rule (9), our results also hold for both variants of the logic.

An *expression* is a formula or a counting term. Let ξ be an expression. The *free variables* $\mathrm{free}(\xi)$ of ξ are inductively defined as follows.

(1) $\mathrm{free}(x_1{=}x_2) = \{x_1, x_2\}$ and $\mathrm{free}\big(R(x_1, \ldots, x_k)\big) = \{x_1, \ldots, x_k\}$.
(2) $\mathrm{free}(\neg\varphi) = \mathrm{free}(\varphi)$ and $\mathrm{free}(\varphi \vee \psi) = \mathrm{free}(\varphi) \cup \mathrm{free}(\psi)$.
(3) $\mathrm{free}(\exists x\,\varphi) = \mathrm{free}(\varphi) \setminus \{x\}$ for $x \in \mathsf{vars}$.
(4) $\mathrm{free}\big(\#(x_1, \ldots, x_k).\varphi\big) = \mathrm{free}\,\varphi \setminus \{x_1, \ldots, x_k\}$.
(5) $\mathrm{free}(i) = \emptyset$ for $i \in \mathbb{Z}$.
(6) $\mathrm{free}\big((t_1 + t_2)\big) = \mathrm{free}\big((t_1 \cdot t_2)\big) = \mathrm{free}(t_1) \cup \mathrm{free}(t_2)$.

(7) $\mathrm{free}\big(\mathsf{P}(t_1,\dots,t_m)\big) = \bigcup_{i=1}^{m} \mathrm{free}(t_i)$.
(8) $\mathrm{free}(\kappa) = \{\kappa\}$ for $\kappa \in \mathsf{nvars}$.
(9) $\mathrm{free}(\exists\kappa\,\varphi) = \mathrm{free}(\varphi) \setminus \{\kappa\}$ for $\kappa \in \mathsf{nvars}$.

We write $\xi(z_1,\dots,z_k)$ to indicate that $\mathrm{free}(\xi) \subseteq \{z_1,\dots,z_k\}$. A *sentence* is a formula without free variables, and a *ground term* is a counting term without free variables. The *binding rank* $\mathrm{br}(\xi)$ of ξ is the maximal nesting depth of constructs using rules (3) and (4), *i.e.* constructs of the form $\exists x$ or $\#\bar{x}$, to construct ξ. The *binding width* $\mathrm{bw}(\xi)$ of ξ is the maximal arity of an $\bar{x}$ of a term $\#\bar{x}.\psi$ in ξ. If ξ contains no such term, then $\mathrm{bw}(\xi) = 1$ if ξ contains a quantifier $\exists x$ with $x \in \mathsf{vars}$, and $\mathrm{bw}(\xi) = 0$ otherwise. Note that, for every FO formula φ, we have $\mathrm{br}(\varphi) = \mathrm{qr}(\varphi)$, $\mathrm{bw}(\varphi) = 1$ if and only if $\mathrm{qr}(\varphi) \geq 1$, and $\mathrm{bw}(\varphi) = 0$ otherwise.

For a formula φ and a σ-interpretation $\mathcal{I}$, we write $\mathcal{I} \models \varphi$ to indicate that $[\![\varphi]\!]^{\mathcal{I}} = 1$. Likewise, $\mathcal{I} \not\models \varphi$ indicates that $[\![\varphi]\!]^{\mathcal{I}} = 0$. For a formula $\varphi(x_1,\dots,x_k,\kappa_1,\dots,\kappa_m)$, a σ-structure $\mathcal{A}$, and tuples $\bar{v} = (v_1,\dots,v_k) \in \big(U(\mathcal{A})\big)^k$ and $\bar{n} = (n_1,\dots,n_m) \in \mathbb{Z}^m$, we write $\mathcal{A} \models \varphi[\bar{v},\bar{n}]$ or $(\mathcal{A},\bar{v},\bar{n}) \models \varphi$ to indicate that $(\mathcal{A},\beta) \models \varphi$ for all assignments β with $\beta(x_i) = v_i$ for all $i \in [k]$ and $\beta(\kappa_j) = n_j$ for all $j \in [m]$. Furthermore, we set $[\![\varphi(\bar{v},\bar{n})]\!]^{\mathcal{A}} \coloneqq 1$ if $\mathcal{A} \models \varphi[\bar{v},\bar{n}]$, and $[\![\varphi(\bar{v},\bar{n})]\!]^{\mathcal{A}} \coloneqq 0$ otherwise. Two expressions ξ,ξ' are *equivalent* if $[\![\xi]\!]^{\mathcal{I}} = [\![\xi']\!]^{\mathcal{I}}$ for all σ-interpretations $\mathcal{I}$. For $d \in \mathbb{N}$, the expressions are called *d-equivalent* if $[\![\xi]\!]^{\mathcal{I}} = [\![\xi']\!]^{\mathcal{I}}$ for all σ-interpretations $\mathcal{I} = (\mathcal{A},\beta)$ for all structures $\mathcal{A}$ of degree at most d. The *length* $|\xi|$ of an expression ξ is its length when viewed as a word over the alphabet $\sigma \cup \mathsf{vars} \cup \mathsf{nvars} \cup \mathbb{P} \cup \mathbb{Z} \cup \{,\} \cup \{=,\neg,\vee,(,),\exists,\#,.,+,\cdot\}$. By FOCN, we denote the union of all $\mathsf{FOCN}[\sigma]$ for arbitrary signatures σ. This applies analogously to FO.

Example 2.4. Let G be a graph, let $\sigma = \{E\}$, and let $\mathbb{P}$ contain the numerical predicate $\mathsf{P}_=$ with $[\![\mathsf{P}_=]\!] = \{(k,k) \mid k \in \mathbb{Z}\}$. We consider the $\mathsf{FOCN}[\sigma]$ sentence

$$\varphi = \exists\kappa\,\forall x\,\mathsf{P}_=\big(\#(y).E(x,y),\kappa\big).$$

The sentence has binding rank 2 and binding width 1. Note that the quantification of the number variable κ has no influence on the binding rank. The sentence holds in G (*i.e.* $G \models \varphi$ holds) if and only if G is a regular graph, *i.e.*, if there is some $k \in \mathbb{N}$ such that every vertex in G has degree k.

Let $r \in \mathbb{N}$. An $\mathsf{FOCN}[\sigma]$ formula $\varphi(\bar{x})$ with free variables $\bar{x} = (x_1,\dots,x_k)$ is *r-local (around $\bar{x}$)* if for every σ-structure $\mathcal{A}$ and every tuple $\bar{v} = (v_1,\dots,v_k) \in \big(U(\mathcal{A})\big)^k$, we have $\mathcal{A} \models \varphi[\bar{v}] \iff \mathcal{N}_r^{\mathcal{A}}(\bar{v}) \models \varphi[\bar{v}]$. A formula is *local* if it is r-local for some $r \in \mathbb{N}$. Intuitively, the evaluation of a local formula only depends on the neighbourhood around the free variables up to a certain radius.

Let $\mathrm{dist}^{\sigma}_{\leq r}(x,y)$ be an $\mathsf{FO}[\sigma]$ formula such that for every σ-structure $\mathcal{A}$ and all $v,w \in U(\mathcal{A})$, we have $\mathcal{A} \models \mathrm{dist}^{\sigma}_{\leq r}[v,w]$ if and only if $\mathrm{dist}^{\mathcal{A}}(v,w) \leq r$. Such a formula can be constructed recursively with quantifier rank at most $\mathcal{O}(\log r)$. To improve readability, we write $\mathrm{dist}^{\sigma}(x,y) \leq r$ instead of $\mathrm{dist}^{\sigma}_{\leq r}(x,y)$, and $\mathrm{dist}^{\sigma}(x,y) > r$ instead of $\neg\,\mathrm{dist}^{\sigma}_{\leq r}(x,y)$. We omit the superscript σ when it is clear from the context. For a tuple $\bar{x} = (x_1,\dots,x_k)$ of variables, $\mathrm{dist}(\bar{x};y) > r$ is a shorthand for $\bigwedge_{i=1}^{k} \mathrm{dist}(x_i,y) > r$, and $\mathrm{dist}(\bar{x};y) \leq r$ is a shorthand for $\bigvee_{i=1}^{k} \mathrm{dist}(x_i,y) \leq r$. For a tuple $\bar{y} = (y_1,\dots,y_\ell)$, we use $\mathrm{dist}(\bar{x};\bar{y}) > r$ as a shorthand for $\bigwedge_{j=1}^{\ell} \mathrm{dist}(\bar{x};y_j) > r$, and $\mathrm{dist}(\bar{x};\bar{y}) \leq r$ as a shorthand for $\bigvee_{j=1}^{\ell} \mathrm{dist}(\bar{x};y_j) \leq r$.

2.3. **Local Access and Complexity Measures.** Whenever we analyse the complexity of learning problems in this paper, we usually think of the background structures as being very large relational databases or huge graphs such as the web graph.

Hence, in case of relational databases, we would like to learn concepts from examples even if the database is too large to fit into the main memory. In case of the web graph, ideally our algorithms should also be able to explore only the regions of the web needed for learning, without having to rely on a previously gathered snapshot of the whole web graph saved to a hard disk.

Thus, the learning algorithms we consider do not obtain the full representation of a background structure as input. Instead, we provide algorithms *local access* to the background structures, *i.e.*, instead of having random access, algorithms may only retrieve the neighbours of vertices they already hold in memory, initially starting with the vertices given in the training examples. Formally, we give algorithms access to an oracle answering queries of the form "Is $\bar{v} \in R(\mathcal{A})$?" and "Return the ith neighbour of v in $\mathcal{A}$" in constant time. Often, instead of explicitly asking for neighbours of a vertex one after another, it will be convenient to use an oracle answering queries of the form "Return a list of all neighbours of v in $\mathcal{A}$" in time linear in the number of neighbours of v. In the context of learning, this local-access model has been introduced in [GR17]. Similar access models have also been studied in property testing for structures of bounded degree [GR02, AH18, AF23] and, more broadly, in the subject of local algorithms [RTVX11, EMR14, LRY17, LRR20]. In addition to granting only local access, we want to learn concepts even without looking at the entire structure. Hence, we are mainly interested in learning problems that can be solved in sublinear time.

As our machine model, we use a random-access machine (RAM) model. Usually, we consider running times under the uniform-cost measure. This allows us to store an element of the background structure in a single memory cell and access it in a single computation step. The uniform-cost RAM model is commonly used in the database theory literature as well as in the analysis of algorithmic meta-theorems [Gro01, FFG02, AGM13, DSS22, CZB+22]. For further details on this model, we refer to [FG06]. Additionally, we consider the logarithmic-cost measure, where storing an element of a structure $\mathcal{A}$ requires space $\mathcal{O}(\log|\mathcal{A}|)$, so accessing and storing takes $\mathcal{O}(\log|\mathcal{A}|)$ many steps.

In contrast to the large background structures, we usually consider formulas as being human-written and hence, rather short. This justifies that in our complexity analyses, we focus on the *data complexity* of a problem, that is, we consider formulas as fixed and measure running times in terms of the size of the background structure, *i.e.* the number of its elements. This approach is also common in database theory when analysing the complexity of the query-evaluation problem [Var82].

3. Learning First-Order Logic

In this section, we formally introduce the different types of learning problems that we consider in this paper. To exemplify this, we briefly describe the learnability results that Grohe and Ritzert obtained in [GR17] for concepts that can be described using first-order logic on structures of small degree within both learning scenarios considered in this paper. In Theorems 3.3 and 3.6, our main results of this section, we complement the results from [GR17] with lower bounds for learning on structures without a degree bound.

3.1. **Consistent Learning.** We start with the consistent-learning scenario. That is, as described in Section 1, we are given a sequence of training examples, and we assume that the examples have been generated using an unknown target concept from a known concept class. Our task is to find a hypothesis that is consistent with the training sequence.

To make this problem feasible at all, we only consider concept classes of limited complexity. Concepts should be definable, like the hypotheses that we learn, via formulas and tuples of parameters. We limit the complexity of the formulas, and we also bound the numbers of parameters. For the learning problem on a background structure $\mathcal{A}$ with k-tuples of elements given as examples, we require that the concept class can be defined as

$$\mathcal{H}_{\Phi^*,k,\ell}(\mathcal{A}) \coloneqq \left\{ h^{\mathcal{A}}_{\varphi,\bar{w}} \;\middle|\; \varphi \in \Phi^*, \bar{w} \in \big(U(\mathcal{A})\big)^{\ell} \right\}$$

for a set Φ^* of formulas $\varphi(\bar{x},\bar{y})$ with $|\bar{x}| = k$ and $|\bar{y}| = \ell$. To limit the complexity of the formulas in Φ^*, in case of first-order logic, the set will only contain formulas up to a certain quantifier rank.

Since we would like to use the learned hypothesis to predict the label of tuples we have not seen yet, we also limit our choice of hypotheses and require that the hypothesis comes from a *hypothesis class* of limited complexity. We do this mainly for two reasons. First, we want to make sure that we are able to evaluate the hypothesis on new tuples efficiently. Second, we want to avoid *overfitting*, where the hypothesis perfectly fits the training examples, but it does so by simply memorising the examples instead of learning an underlying rule. As we will see in Section 3.2, limiting the complexity of a hypothesis class is a key ingredient to finding hypotheses that generalise well. In the results of this paper, just as in the results of [GR17], we allow algorithms to return hypotheses that are more complex than the concepts contained in the concept class. Hence, we use a hypothesis class $\mathcal{H}_{\Phi,k,\ell}(\mathcal{A})$ with a set Φ of formulas that can be more complex than Φ^*.

Now, we consider the learning problem for first-order logic introduced in [GR17]. There, for fixed $k, \ell, q^* \in \mathbb{N}$ and a fixed signature σ, the authors considered concept classes based on first-order formulas of quantifier rank at most q^* with $k + \ell$ free variables, so

$$\Phi^* = \left\{ \varphi(\bar{x},\bar{y}) \in \mathsf{FO}[\sigma] \;\middle|\; \operatorname{qr}(\varphi) \leq q^*, |\bar{x}| = k, |\bar{y}| = \ell \right\}.$$

Note that, up to equivalence, there are only finitely many formulas in Φ^*. By Gaifman's Locality Theorem [Gai82], every single of those formulas is equivalent to a formula in Gaifman normal form. This shows that there is some $q \in \mathbb{N}$ such that every formula in Φ^* is equivalent to a formula in Gaifman normal form of quantifier rank at most q. Grohe and Ritzert use this q as the bound on the quantifier rank for Φ, *i.e.* they use $\Phi = \left\{ \varphi(\bar{x},\bar{y}) \in \mathsf{FO}[\sigma] \;\middle|\; \operatorname{qr}(\varphi) \leq q, |\bar{x}| = k, |\bar{y}| = \ell \right\}$. This allows them in their algorithms to only look for formulas in Gaifman normal form. Let $f\colon \mathbb{N}^2 \to \mathbb{N}$ be a function such that all $\mathsf{FO}[\sigma]$ formulas of quantifier rank at most q^* with $k + \ell$ free variables are equivalent to an $\mathsf{FO}[\sigma]$ formula of quantifier rank at most $f(k + \ell, q^*)$ in Gaifman normal form. The consistent-learning problem for first-order logic is defined as follows.

FO-Learn-Consistent(σ, k, ℓ, q^*)

Input: structure $\mathcal{A}$, training sequence $T \in \Big(\big(U(\mathcal{A})\big)^k \times \{0,1\}\Big)^m$ for some $m \in \mathbb{N}$

Task: Return a formula $\varphi(\bar{x}, \bar{y}) \in \mathsf{FO}[\sigma]$ of quantifier rank at most $f(k+\ell, q^*)$ with $k+\ell$ free variables and a tuple $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$ such that the hypothesis $h^{\mathcal{A}}_{\varphi,\bar{w}}$ is consistent with T. The algorithm may reject if there is no formula $\varphi^*(\bar{x}, \bar{y}) \in \mathsf{FO}[\sigma]$ of quantifier rank at most q^* and tuple $\bar{w}^* \in \big(U(\mathcal{A})\big)^\ell$ such that the hypothesis $h^{\mathcal{A}}_{\varphi^*,\bar{w}^*}$ is consistent with T.

Grohe and Ritzert [GR17] showed that the problem is solvable in time polynomial in the degree of the background structure and the number of examples in the training sequence.

Theorem 3.1 [GR17, Theorems I.1 and IV.3]**.** *Let σ be a relational signature and let $k, \ell, q^* \in \mathbb{N}$. There is an algorithm that solves* FO-Learn-Consistent(σ, k, ℓ, q^*) *in time $(\log n + d + m)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in time $(d+m)^{\mathcal{O}(1)}$ under the uniform-cost measure, where n is the size and d is the degree of the background structure, and m is the length of the training sequence.*

On classes of structures of polylogarithmic degree, that is, classes $\mathcal{C}$ for which there is some $c \in \mathbb{N}$ such that $\deg(\mathcal{A}) \in \mathcal{O}\big((\log |\mathcal{A}|)^c\big)$ for all structures $\mathcal{A}$ in $\mathcal{C}$, Theorem 3.1 implies that consistent learning is possible in sublinear time.

Corollary 3.2. *Let σ be a relational signature, let $k, \ell, q^* \in \mathbb{N}$, and let $\mathcal{C}$ be a class of structures of polylogarithmic degree. There is an algorithm that solves the problem* FO-Learn-Consistent(σ, k, ℓ, q^*) *on $\mathcal{C}$ in time sublinear in the size of the background structure and polynomial in the length of the training sequence, under the logarithmic-cost as well as the uniform-cost measure.*

In the proof of Theorem 3.1, Grohe and Ritzert [GR17] provide a brute-force algorithm that tests all combinations of certain formulas and parameters. Since the assumed target concept uses a formula of quantifier rank at most q^*, by Gaifman's Locality Theorem [Gai82], there is an $r^* = r(q^*) \in \mathbb{N}$ such that the used formula is r^*-local. Furthermore, there is an equivalent formula in Gaifman normal form of quantifier rank at most q. Hence, the algorithm in [GR17] tests all formulas from the (up to equivalence) finite set of r^*-local formulas in Gaifman normal form of quantifier rank at most q. Due to the locality of the considered formulas, it then suffices to limit the search for suitable parameters to a neighbourhood of a certain radius around the examples given in the training sequence. The size of the neighbourhood, and thus also the number of parameter tuples to test, is polynomial in the degree of the structure and the number of training examples. Finally, again due to the locality of the considered formulas, a single test of a hypothesis can be performed in time polynomial in the degree of the structure. All in all, this yields an algorithm with the desired running time bounds.

In this paper, we prove similar results for the extension FOCN of first-order logic with counting quantifiers in Sections 6 and 7. There, instead of using Gaifman locality and Gaifman normal forms, we use so-called Hanf locality and Hanf normal forms.

Prior to this, we provide a lower bound on the running time needed to learn first-order definable concepts on general structures. This shows that the degree bound imposed on the

class of background structures is crucial to be able to learn first-order definable concepts in sublinear time.

The result even holds for a stronger, *random-access* model. In this model, we give algorithms access to an oracle answering queries of the form "Return the ith element of $U(\mathcal{A})$", "Is $\bar{v} \in R(\mathcal{A})$?", "Return the ith tuple of $R(\mathcal{A})$", and "Return the ith tuple of $R(\mathcal{A})$ that contains v" in constant time.

Theorem 3.3. *Let σ be a signature that contains at least one relation symbol of arity at least 2. For all $k, \ell \in \mathbb{N}_{\geq 1}$ and $q^* \geq 2$, there is no algorithm with random access to the background structure that solves* FO-Learn-Consistent(σ, k, ℓ, q^*) *in time sublinear in the size of the background structure.*

Proof. First, we prove the statement for $k = \ell = 1$, $q^* = 2$, and $\sigma = \{E\}$ for a binary relation symbol E. Afterwards, we generalise this result.

We prove the statement by contradiction. Assume that there is an algorithm solving FO-Learn-Consistent(σ, k, ℓ, q^*) in sublinear time. Choose $n \in \mathbb{N}$ such that for all $n' \geq n$, the algorithm uses at most $\frac{n'-6}{16}$ many steps on background structures of size n' and training sequences of length 4. Now, we construct two almost identical background structures $\mathcal{A}_1$ and $\mathcal{A}_2$ of size $8n + 6$ with $16n + 4$ edges and corresponding training sequences T_1 and T_2 of length 4. Note that the algorithm, using at most $\frac{(8n+6)-6}{16} = \frac{n}{2}$ steps, can visit (that is, query) at most $\frac{n}{2}$ vertices or edges of the background structures. By construction of the background structures and training sequences, the algorithm will be unable to distinguish the two inputs. Hence, the algorithm has to return the same formula and the same parameter on both inputs. As we will see, the resulting hypothesis has to be inconsistent with at least one of the two inputs, which then contradicts our assumption that the algorithm solves the problem.

The background structure $\mathcal{A}_1$ is depicted in Figure 4. It is formally defined as the $\{E\}$-structure with

$$
\begin{aligned}
U_{i,j} &= \{z_{i,j,p} \mid p \in [n]\} \quad \text{for } i \in [2] \text{ and } j \in [4],\\
U(\mathcal{A}_1) &= \{x_1, x_2, x_3, x_4, y_1, y_2\} \cup \bigcup_{i \in [2],\ j \in [4]} U_{i,j},\\
R &= \big\{\{y_i, z_{i,j,p}\} \;\big|\; i \in [2], j \in [4], p \in [n]\big\}, && \text{(rows)}\\
C &= \big\{\{x_j, z_{i,j,p}\} \;\big|\; i \in [2], j \in [4], p \in [n]\big\}, && \text{(columns)}\\
E_1 &= \big\{\{z_{1,1,n-1}, z_{1,1,n}\}, \{z_{1,3,n-1}, z_{1,3,n}\},\\
&\qquad \{z_{2,1,n-1}, z_{2,1,n}\}, \{z_{2,4,n-1}, z_{2,4,n}\}\big\}, \text{ and}\\
E(\mathcal{A}_1) &= R \cup C \cup E_1,
\end{aligned}
$$

where $\{u, v\} \in E(\mathcal{A}_1)$ means that both (u, v) and (v, u) are contained in $E(\mathcal{A}_1)$. Intuitively, we can view the structure as eight sets of vertices $U_{i,j}$ being arranged in a table with two rows and four columns, and six additional vertices. The vertices y_1 and y_2 are used to indicate the first and second row. All vertices in a set in the ith row are connected to y_i via an R-edge. The vertices x_1 to x_4 are used to indicate the columns, and the vertices in the jth column are connected to x_j via a C-edge. Finally, there are four additional edges within the table. In the first row, there is one edge connecting two vertices in the first column and one edge connecting two vertices in the third column. In the second row, there is an edge in the first and fourth column.

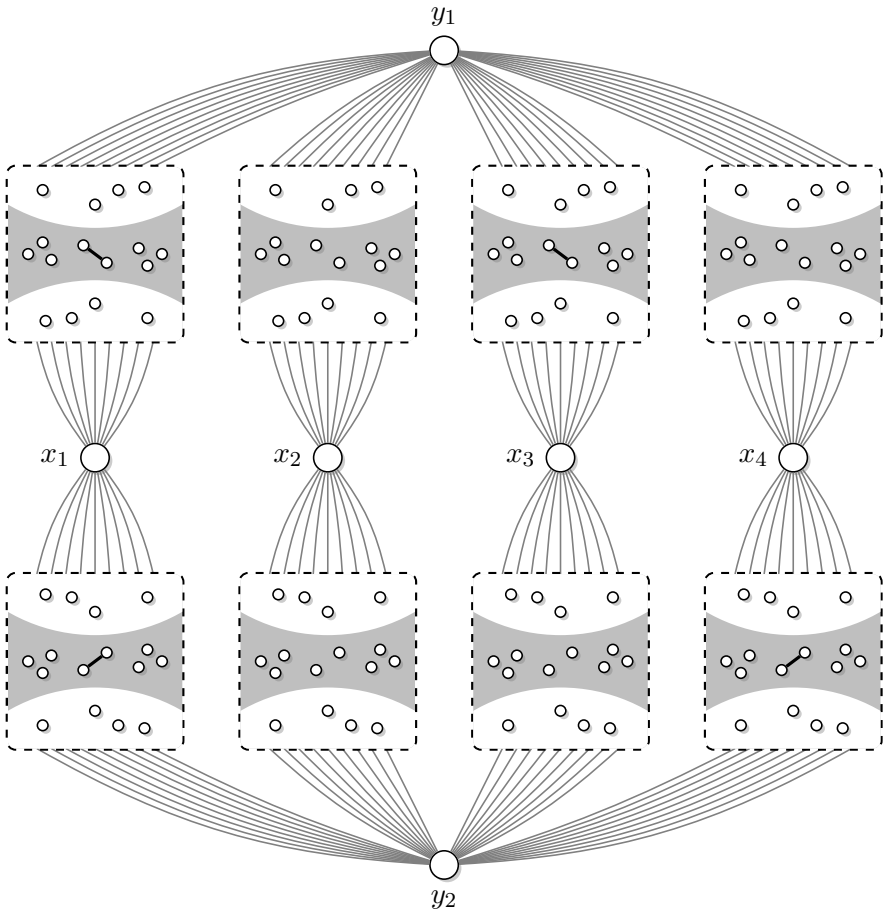

Figure 4: Background structure $\mathcal{A}_1$ from the proof of Theorem 3.3. Eight sets of vertices are placed in a table with two rows and four columns. The y_i vertices are connected to all vertices in the sets in the ith row, and the x_j vertices are connected to all vertices in the sets in the jth column. The vertices on the grey background are those parts of the background structure that the algorithm is unable to explore in sublinear time.

The structure $\mathcal{A}_2$ is almost identical to $\mathcal{A}_1$; only the additional edges differ. There, we have

$$E_2 = \big\{\{z_{1,1,n-1}, z_{1,1,n}\}, \{z_{1,4,n-1}, z_{1,4,n}\}, \{z_{2,1,n-1}, z_{2,1,n}\}, \{z_{2,3,n-1}, z_{2,3,n}\}\big\} \text{ and}$$
$$E(\mathcal{A}_2) = R \cup C \cup E_2,$$

i.e., the second edge in the first row is now in the fourth instead of the third column, and, in the second row, the edge is in the third instead of the fourth column.

For the target concept in both background structures, we use

$$\varphi^*(x, y) = \exists z_1 \exists z_2 \big(E(x, z_1) \wedge E(x, z_2) \wedge E(y, z_1) \wedge E(y, z_2) \wedge E(z_1, z_2)\big).$$

Both training sequences consists of the vertices x_1 to x_4 with corresponding labels, and we use y_1 or y_2 as a parameter. Hence, for the examples, the formula φ^* is satisfied if and only if there is an edge in the column indicated by x and the row indicated by y. For the first structure, we use y_1 as a parameter, so we select the first row. There, the first and third column contain an edge, so the resulting training sequence is

$$T_1 = \big((x_1, 1), (x_2, 0), (x_3, 1), (x_4, 0)\big).$$

For the second structure, we select y_2 as a parameter and hence the second row of $\mathcal{A}_2$. There, again the first and third column contain an edge, so $T_2 = T_1$.

As argued above, the algorithm can only visit at most $n/2$ vertices or edges of the background structure, and there are $8n + 6$ vertices and $16n + 4$ edges in total. Hence, for every such algorithm, there is a suitable ordering of the vertices and edges in the background structures (that defines which vertex is the ith vertex of $U(\mathcal{A}_1)$ resp. $U(\mathcal{A}_2)$ and which edge is the jth edge of $E(\mathcal{A}_1)$ resp. $E(\mathcal{A}_2)$) such that the algorithm will never find any edge from E_1 or E_2. Instead, due to the ordering, the algorithm will only visit edges from R and C. Hence, the algorithm is unable to distinguish the two inputs, and it will return the same formula φ and the same parameter w. Because the first and third column as well as the second and fourth column are indistinguishable for the algorithm, again, by choosing a suitable order on the vertices of the background structures, we can assume that the algorithm returns x_1, x_2, y_1, y_2, or some vertex from the first or second column as the parameter w.

We consider the isomorphism between $\mathcal{A}_1$ and $\mathcal{A}_2$ that keeps y_1, y_2 as well as the first and second column identical but swaps the third and fourth column (including x_3 and x_4). Note that the isomorphism also maps the parameter w to itself. The existence of such an isomorphism implies that the returned formula φ behaves in $\mathcal{A}_1$ on x_3 like it does in $\mathcal{A}_2$ on x_4, so $[\![\varphi(x_3, w)]\!]^{\mathcal{A}_1} = [\![\varphi(x_4, w)]\!]^{\mathcal{A}_2}$. However, in the training sequence $T_1 = T_2$, the vertices x_3 and x_4 have different labels. Hence, the algorithm cannot return on both $\mathcal{A}_1$ and $\mathcal{A}_2$ a consistent hypothesis, so it has to fail on at least one of them. This contradicts our assumption, so there is no algorithm solving FO-Learn-Consistent(σ, k, ℓ, q^*) in sublinear time for $\sigma = \{E\}$, $k = \ell = 1$ and $q^* = 2$.

Now, we generalise this result. Note that we did not use any bounds on the quantifier rank for the returned formula φ. Hence, our proof also works for larger values of q^*. If E is a relation symbol of higher arity, we can set the first two entries of the tuples like described above and then repeat the second entry to fill the rest of the tuple. Additional relation symbols have no influence on the argumentation presented above. Similarly, for $k > 1$, we can provide the same vertices as examples, but instead of using single vertices, we use tuples filled with the same vertex.

For $\ell > 1$, we use the disjoint union of ℓ copies of $\mathcal{A}_1$ as the first background structure and proceed analogously for the second background structure. The training sequence consists of the vertices x_1 to x_4 with their corresponding labels from every single of those ℓ copies. Then, the algorithm either puts exactly one parameter in each of the copies, or there is at least one copy without any parameters. Thus, in both cases, there is at least one copy with at most one parameter. Hence, the argumentation from above still applies for this copy, showing that the algorithm is unable to provide a consistent hypothesis for at least one of the inputs. □

3.2. **PAC Learning.** Next, we introduce Haussler's model of *agnostic probably approximately correct (PAC) learning* [Hau92], a generalisation of Valiant's *PAC-learning* model [Val84]. Moreover, to get familiar with this model within our logic learning framework, we discuss the agnostic PAC-learning results from [GR17] and describe techniques to prove these results based on the consistent-learning results.

Intuitively, in (agnostic) PAC learning, we are interested in hypotheses that generalise well, *i.e.* hypotheses that not only work well on the examples from the training sequence but also on tuples not given as examples.

In PAC learning, we assume an (unknown) probability distribution $\mathcal{D}$ on the instance space X and, as in consistent learning, a consistent target concept $c\colon X \to \{0,1\}$. The learner's goal is to find a hypothesis $h\colon X \to \{0,1\}$, based on a sequence of training examples randomly drawn from $\mathcal{D}$, such that h minimises the *generalisation error*

$$\operatorname{err}_{\mathcal{D},c}(h) \coloneqq \Pr_{x\sim\mathcal{D}}\big(h(x) \neq c(x)\big),$$

i.e. the probability of being wrong on a random instance. In practice, we want to find a hypothesis with a generalisation error below a certain threshold ε.

In agnostic PAC learning, we drop the assumption of having a consistent target concept. Instead, we assume an (unknown) probability distribution $\mathcal{D}$ on $X \times \{0,1\}$. Again, a learning algorithm should find a hypothesis h that minimises the generalisation error, which is now defined as

$$\operatorname{err}_{\mathcal{D}}(h) \coloneqq \Pr_{(x,\lambda)\sim\mathcal{D}}\big(h(x) \neq \lambda\big).$$

Here, since a generalisation error of 0 might not be possible, we want to find a hypothesis with a generalisation error close to the best possible one.

A hypothesis class $\mathcal{H}$ of hypotheses $h\colon X \to \{0,1\}$ is called *agnostically PAC-learnable* if there is a function $m_{\mathcal{H}}\colon (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathcal{L}$ with the following property: For all $\varepsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $X \times \{0,1\}$, when running $\mathcal{L}$ on a sequence T of m examples drawn i.i.d. from $\mathcal{D}$ with $m \geq m_{\mathcal{H}}(\varepsilon,\delta)$, it outputs a hypothesis $h \in \mathcal{H}$ such that, with probability of at least $1-\delta$ over the choice of training examples, it holds that

$$\operatorname{err}_{\mathcal{D}}(h) \leq \inf_{h'\in\mathcal{H}} \operatorname{err}_{\mathcal{D}}(h') + \varepsilon.$$

We call such an algorithm $\mathcal{L}$ an (agnostic) PAC-learning algorithm.

In this definition, we find two parameters, ε and δ. The first parameter ε, also called the accuracy parameter ("approximately correct"), describes how far the hypothesis returned by the algorithm is allowed to be from an optimal hypothesis. This allows the returned hypothesis to make a few mistakes, *e.g.* in case of outliers that are manually handled by an optimal solution but that we do not see in the limited number of training examples. The second parameter δ, also called the confidence parameter ("probably"), describes how confident we are to return a good hypothesis on a randomly chosen sequence of training examples. This refers to cases where the randomly chosen training sequence is not representative for $\mathcal{D}$, *e.g.* it consists only of positive examples or the same example is repeated over and over again. The function $m_{\mathcal{H}}$ determines, given the parameters ε and δ, the sample complexity of the problem, *i.e.* the number of examples needed to probably find an approximately correct hypothesis. For a more detailed discussion of (agnostic) PAC learning, we refer to [SB14].

Analogously to the results in the consistent-learning case, in [GR17], Grohe and Ritzert analysed a relaxed version of agnostic PAC learning. There, we want to approximately learn

concepts from a concept class, but we allow the algorithms to return hypotheses from a slightly more complex hypothesis class.

In addition to the previously defined membership and neighbourhood oracles for the background structure $\mathcal{A}$, we allow algorithms to query the size $|\mathcal{A}|$ of the structure. This information is needed to compute the sufficient length $m_{\mathcal{H}}(\varepsilon, \delta)$ of the training sequence. Furthermore, we give algorithms oracle access to the probability distribution $\mathcal{D}$ on $\big(U(\mathcal{A})\big)^k \times \{0,1\}$. That is, whenever an algorithm queries the oracle, it receives a labelled example from $\big(U(\mathcal{A})\big)^k \times \{0,1\}$ drawn from $\mathcal{D}$. The labelled examples are drawn independently of each other.

As in Section 3.1, let $f\colon \mathbb{N}^2 \to \mathbb{N}$ be a function such that all $\mathsf{FO}[\sigma]$ formulas of quantifier rank at most q^* with $k+\ell$ free variables are equivalent to an $\mathsf{FO}[\sigma]$ formula of quantifier rank at most $f(k+\ell, q^*)$ in Gaifman normal form.

The k-ary agnostic PAC-learning problem for first-order logic is defined as follows.

FO-Learn-PAC(σ, k, ℓ, q^*)

Input: structure $\mathcal{A}$, rational numbers $\varepsilon, \delta > 0$, probability distribution $\mathcal{D}$ on $\big(U(\mathcal{A})\big)^k \times \{0,1\}$

Task: Return a formula $\varphi(\bar{x}, \bar{y}) \in \mathsf{FO}[\sigma]$ with $\operatorname{qr}(\varphi) \leq f(k+\ell, q^*)$ and a tuple $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$ such that, with probability of at least $1-\delta$ over the choice of examples drawn i.i.d. from $\mathcal{D}$, it holds that

$$\operatorname{err}_{\mathcal{D}}\big(h^{\mathcal{A}}_{\varphi,\bar{w}}\big) \leq \varepsilon^* + \varepsilon,$$

where

$$\varepsilon^* \coloneqq \min_{\substack{\varphi^*(\bar{x},\bar{y}) \in \mathsf{FO}[\sigma] \\ \text{with } \operatorname{qr}(\varphi^*) \leq q^*, \\ \bar{w}^* \in (U(\mathcal{A}))^\ell}} \operatorname{err}_{\mathcal{D}}\big(h^{\mathcal{A}}_{\varphi^*,\bar{w}^*}\big).$$

To solve the problem algorithmically, we can follow the *Empirical Risk Minimisation (ERM)* rule [Vap91, SB14], that is, our algorithm should return a hypothesis h that minimises the *training error* (or *empirical risk*)

$$\operatorname{err}_T(h) \coloneqq \frac{1}{|T|} \cdot \big|\big\{(\bar{v}, \lambda) \in T \;\big|\; h(\bar{v}) \neq \lambda\big\}\big|$$

on the training sequence T of queried examples. Thus, in order to solve the PAC-learning problem FO-Learn-PAC(σ, k, ℓ, q^*), we first consider the following problem.

FO-Learn-ERM(σ, k, ℓ, q^*)

Input: structure $\mathcal{A}$, training sequence $T \in \Big(\big(U(\mathcal{A})\big)^k \times \{0,1\}\Big)^m$ for some $m \in \mathbb{N}$

Task: Return a formula $\varphi(\bar{x}, \bar{y}) \in \mathsf{FO}[\sigma]$ with $\operatorname{qr}(\varphi) \leq f(k+\ell, q^*)$ and a tuple $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$ such that

$$\operatorname{err}_T\big(h^{\mathcal{A}}_{\varphi,\bar{w}}\big) \leq \min_{\substack{\varphi^*(\bar{x},\bar{y}) \in \mathsf{FO}[\sigma] \\ \text{with } \operatorname{qr}(\varphi^*) \leq q^*, \\ \bar{w}^* \in (U(\mathcal{A}))^\ell}} \operatorname{err}_T(h^{\mathcal{A}}_{\varphi^*,\bar{w}^*}).$$

This problem is very similar to the consistent-learning problem. The only difference is that, instead of asking for a consistent hypothesis, we want to find a hypothesis that is at least as consistent as the best one from the concept class.

To solve FO-Learn-ERM, Grohe and Ritzert [GR17] use a brute-force algorithm similar to the one they present for the problem FO-Learn-Consistent. However, instead of checking whether a hypothesis is consistent, they count the number of errors the hypotheses make on the training sequence and return the hypothesis that minimises this number. They then show that this also yields an algorithm solving the PAC-learning problem.

Theorem 3.4 [GR17, Theorem V.7]**.** *Let σ be a relational signature and let $k, \ell, q^* \in \mathbb{N}$. There is an algorithm that solves* FO-Learn-PAC(σ, k, ℓ, q^*) *in time* $(\log n + d + 1/\varepsilon + 1/\delta)^{\mathcal{O}(1)}$ *under the logarithmic-cost and the uniform-cost measure, where n is the size and d is the degree of the background structure.*

Analogously to the consistent-learning problem, on classes of structures of polylogarithmic degree, Theorem 3.4 implies that probably approximately correct learning is possible in sublinear time.

Corollary 3.5. *Let σ be a relational signature, let $k, \ell, q^* \in \mathbb{N}$, and let $\mathcal{C}$ be a class of structures of polylogarithmic degree. There is an algorithm that solves* FO-Learn-PAC(σ, k, ℓ, q^*) *on $\mathcal{C}$ in time sublinear in the size of the background structure, under the logarithmic-cost as well as the uniform-cost measure.*

We prove similar PAC-learning results for FOCN in Sections 6 and 7. Theorem 3.4 and Corollary 3.5 show a strong connection between consistent and PAC learning. Only slight modifications are needed to turn the consistent-learning algorithm into an algorithm performing Empirical Risk Minimisation that can then be used within a PAC-learning algorithm. To conclude this section, we show that the strong connection also holds in the other direction. That is, analogously to a proof by Grohe, Löding, and Ritzert in [GLR17], we transform Theorem 3.3, our negative result for the consistent-learning problem, into a negative result for the PAC-learning problem.

Theorem 3.6. *Let σ be a signature that contains at least one relation symbol of arity at least 2. For all $k, \ell \in \mathbb{N}_{\geq 1}$ and $q^* \geq 2$, there is no algorithm with random access to the background structure that solves* FO-Learn-PAC(σ, k, ℓ, q^*) *in time sublinear in the size of the background structure.*

Proof. This proof is based on the proof of Theorem 3.3. We only consider the case $k = \ell = 1$, $q^* = 2$ and $\sigma = \{E\}$ for a binary relation symbol E. The generalisation can be done analogously to the original proof. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be the background structures and $T := T_1 = T_2$ be the training sequences from the proof. Let $\mathcal{D}$ be the uniform distribution over the examples from T, that is, $(x_1, 1)$, $(x_2, 0)$, $(x_3, 1)$, and $(x_4, 0)$ have probability $\frac{1}{4}$; all other $(v, \lambda) \in U(\mathcal{A}_1) \times \{0, 1\} = U(\mathcal{A}_2) \times \{0, 1\}$ have probability 0. By the choice of $\mathcal{D}$, if a hypothesis misclassifies at least one of the x_i, it has a generalisation error of at least $\frac{1}{4}$.

Assume that $\mathcal{L}$ is an algorithm that solves FO-Learn-PAC(σ, k, ℓ, q^*) in sublinear time. As we argued in the proof of Theorem 3.3, $\mathcal{L}$ is unable to distinguish $\mathcal{A}_1$ and $\mathcal{A}_2$ from each other (by choosing a suitable ordering on the vertices and edges). Furthermore, we argued that such an algorithm would also be unable to distinguish the first and third column as well as the second and fourth column of the background structures. In the proof of Theorem 3.3, we chose an ordering on the vertices and edges such that the parameter returned by the

algorithm is x_1, x_2, y_1, y_2, or some vertex from the first or second column. Here, the vertex returned by $\mathcal{L}$ may depend on the training sequence drawn from $\mathcal{D}$. However, by choosing a sufficient ordering on the vertices and edges, we can still make sure that the returned parameter is among the mentioned ones (*i.e.* among x_1, x_2, y_1, y_2, or some vertex from the first or second column) with probability at least $\frac{1}{2}$ over the choice of examples drawn from $\mathcal{D}$.

Now, we only consider those cases where the parameter is among the mentioned ones. For every fixed choice of examples, analogously to the proof of the consistent-learning case, the algorithm $\mathcal{L}$ has to return the same hypothesis on both background structures. Thus, the hypothesis returned by the algorithm has to misclassify at least one of the x_i on at least one of the two background structures. Hence, on one of the two background structures, it makes at least one error in at least half of the cases where the parameter is among the mentioned ones, so with (conditional) probability at least $\frac{1}{2}$.

Overall, including the probability that the chosen parameter is among the mentioned vertices, on at least one of the two background structures, $\mathcal{L}$ has to make at least one error on the x_i with probability at least $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Combined with our observation above, this means that, on one of two background structures, the algorithm has a generalisation error of at least $\frac{1}{4}$ with probability at least $\frac{1}{4}$ over the choice of examples drawn from $\mathcal{D}$. We choose $\varepsilon = \delta = \frac{1}{8}$. Then $\mathcal{L}$ does not meet the requirements of FO-Learn-PAC, which contradicts our assumption. All in all, this shows that there is no algorithm that solves the problem FO-Learn-PAC(σ, k, ℓ, q^*) in sublinear time. □

Remark 3.7. The strong connection between the consistent-learning problem and the PAC-learning problem mentioned in this section is specific to learning first-order definable concepts on structures on small degree. For example in Section 7, for FOCN-definable concepts, the proof of the PAC-learning result relies on a more restrictive degree bound than the one of the consistent-learning result. Moreover, when considering the parameterised complexity of learning, [vBGR22] shows that PAC learning of first-order definable concepts is fixed-parameter tractable on nowhere dense classes. For the consistent-learning problem, however, we are not aware of any such results.

For concepts definable in monadic second-order logic, [vBGR25] shows that both, the consistent-learning and the PAC-learning problem, are fixed-parameter tractable on classes of bounded clique-width in case $k = 1$, that is, if the examples are just labelled single vertices. However, in case $k > 1$, [vBGR25] shows that the PAC-learning problem remains fixed-parameter tractable, while the complexity bounds for consistent learning are weaker, and they are accompanied by a hardness result showing that this is optimal.

4. Locality of First-Order Logic with Counting

For the learnability results for first-order logic with counting that we prove in Sections 6 and 7, we rely on normal forms based on Hanf's locality theorem for first-order logic [Han65, FSV95]. This theorem implies that, to determine whether a finite structure satisfies a first-order sentence of quantifier rank at most q, it suffices to determine the number of realisations of neighbourhoods up to a certain radius within the structure. The version of the theorem provided by Fagin, Stockmeyer, and Vardi [FSV95] implies that on structures of degree at most d, it even suffices to determine the number of these realisations up to a certain threshold. Since, in structures of degree at most d, there are only finitely many types of

neighbourhoods of radius at most r, this condition can be expressed as a first-order sentence in so-called *Hanf normal form*.

In this paper, we use the Hanf normal form for the first-order logic with counting FOCN provided by Kuske and Schweikardt [KS17]. Before stating the exact result, we first introduce the basic building blocks.

Let $r \in \mathbb{N}$, $k \in \mathbb{N}_{\geq 1}$, let $\mathcal{A}$ be a relational structure, and let $\bar{v} = (v_1, \ldots, v_k) \in \big(U(\mathcal{A})\big)^k$. A *sphere formula with k centres of locality radius r* is a first-order formula $\mathrm{sph}^{\mathcal{A}}_{r,\bar{v}}(x_1, \ldots, x_k)$ such that for every structure $\mathcal{A}'$ and every tuple $\bar{v}' = (v'_1, \ldots, v'_k) \in \big(U(\mathcal{A}')\big)^k$, it holds that $\mathcal{A}' \models \mathrm{sph}^{\mathcal{A}}_{r,\bar{v}}[\bar{v}']$ if and only if there is an isomorphism between the two r-neighbourhoods of $\bar{v}$ and $\bar{v}'$ that maps the centres upon each other, *i.e.*, there is an isomorphism π between $\mathcal{N}^{\mathcal{A}}_r(\bar{v})$ and $\mathcal{N}^{\mathcal{A}'}_r(\bar{v}')$ with $\pi(v_i) = v'_i$ for all $i \in [k]$, or, equivalently, there is an isomorphism between $\mathcal{S}^{\mathcal{A}}_r(\bar{v})$ and $\mathcal{S}^{\mathcal{A}'}_r(\bar{v}')$. For a fixed signature σ, given a tuple $\bar{v}$, a radius r, and local access to a σ-structure $\mathcal{A}$, the time needed to construct the sphere formula $\mathrm{sph}^{\mathcal{A}}_{r,\bar{v}}(x_1, \ldots, x_k)$ is polynomial in the size of the r-neighbourhood of $\bar{v}$ [KS17]. Note that sphere formulas of locality radius at most r are r-local.

A *basic counting term* is a counting term of the form $\#(x).\varphi(x)$ in FOCN, where x is a structure variable in vars and φ is a sphere formula with a single centre. The *locality radius* of the basic counting term is the locality radius of the sphere formula.

A *numerical condition on occurrences of types with one centre* (or *numerical oc-type condition*) is an FOCN formula that is built from basic counting terms and rules (2) and (5)–(9) from Definitions 2.1 and 2.2, *i.e.*, using number variables and integers, and combining them by addition, multiplication, numerical predicates from $\mathbb{P} \cup \{\mathsf{P}_\exists\}$ (with $\mathrm{ar}(\mathsf{P}_\exists) = 1$ and $[\![\mathsf{P}_\exists]\!] = \mathbb{N}_{\geq 1}$), Boolean combinations, and quantification of number variables. Its locality radius is the maximal locality radius of the involved basic counting terms. Note that numerical oc-type conditions do not have any free structure variables.

A formula is in *Hanf normal form for* FOCN or an *hnf formula for* FOCN if it is a Boolean combination of numerical oc-type conditions and sphere formulas. The locality radius of an hnf formula is the maximal locality radius of the involved conditions and formulas.

The following result is due to Kuske and Schweikardt [KS17].

Theorem 4.1 [KS17, Theorem 3.2]**.** *For any relational signature σ, any degree bound $d \in \mathbb{N}$, and any* FOCN$[\sigma]$ *formula φ, there exists a d-equivalent hnf formula ψ for* FOCN$[\sigma]$ *of locality radius smaller than* $\big(2 \cdot \mathrm{bw}(\varphi) + 1\big)^{\mathrm{br}(\varphi)}$ *with* $\mathrm{free}(\psi) = \mathrm{free}(\varphi)$.

Next, analogously to the local types used in [GR17], we introduce local Hanf types, and we also provide similar locality results for them. Let $\mathcal{A}$ be a relational structure, $k \in \mathbb{N}_{\geq 1}$, $r \in \mathbb{N}$, and $\bar{v} \in \big(U(\mathcal{A})\big)^k$. The *local Hanf type (for* FOCN*) of $\bar{v}$ with locality radius at most r in $\mathcal{A}$* is

$$\mathrm{lhtp}^{\mathcal{A}}_r(\bar{v}) := \big\{\varphi(\bar{x}) \text{ hnf formula } \mid \mathcal{A} \models \varphi[\bar{v}], \text{ locality radius of } \varphi \text{ is at most } r\big\}.$$

We use Kuske's and Schweikardt's result to show that FOCN formulas are unable to distinguish tuples that have the same local Hanf type (of a certain locality radius).

Lemma 4.2. *Let $\mathcal{A}$ be a relational structure, let $\bar{x} = (x_1, \ldots, x_k)$ be a tuple of structure variables, let $\bar{\kappa} = (\kappa_1, \ldots, \kappa_\ell)$ be a tuple of number variables, and let $\varphi(\bar{x}, \bar{\kappa})$ be an* FOCN

formula. For all $\bar{v}, \bar{v}' \in \big(U(\mathcal{A})\big)^k$ *and* $\bar{n} = (n_1, \dots, n_\ell) \in \mathbb{Z}^\ell$*, if*

$$\operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}) = \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}') \qquad \text{for} \quad r = (2 \cdot \operatorname{bw}(\varphi) + 1)^{\operatorname{br}(\varphi)},$$

then

$$\mathcal{A} \models \varphi[\bar{v}, \bar{n}] \iff \mathcal{A} \models \varphi[\bar{v}', \bar{n}].$$

Proof. Let $\varphi'(\bar{x}) \coloneqq \varphi(\bar{x}, \bar{n})$, *i.e.* we replace every occurrence of the number variable κ_i in φ with the integer n_i for all i. Note that $\operatorname{br}(\varphi) = \operatorname{br}(\varphi')$ and $\operatorname{bw}(\varphi) = \operatorname{bw}(\varphi')$. Using Theorem 4.1, we obtain an hnf formula $\psi(\bar{x})$ of locality radius smaller than $r = \big(2 \cdot \operatorname{bw}(\varphi) + 1\big)^{\operatorname{br}(\varphi)}$ that is $\deg(\mathcal{A})$-equivalent to φ'. Let $\bar{v}$ and $\bar{v}'$ be k-tuples from $\mathcal{A}$ with $\operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}) = \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}')$. We show $\mathcal{A} \models \varphi[\bar{v}, \bar{n}] \implies \mathcal{A} \models \varphi[\bar{v}', \bar{n}]$, then the other direction follows by symmetry.

Assume that $\mathcal{A} \models \varphi[\bar{v}, \bar{n}]$ holds. This implies that $\mathcal{A} \models \varphi'[\bar{v}]$, and $\mathcal{A} \models \psi[\bar{v}]$ hold as well. Thus, since ψ is an hnf formula of locality radius smaller than r, we have $\psi \in \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}) = \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}')$, which implies that $\mathcal{A} \models \psi[\bar{v}']$. By the $\deg(\mathcal{A})$-equivalence between ψ and φ', this shows that $\mathcal{A} \models \varphi'[\bar{v}']$, which finally implies that $\mathcal{A} \models \varphi[\bar{v}', \bar{n}]$. □

The following results help us to reduce the formula and parameter spaces we have to consider to find consistent hypotheses. The first lemma states that two tuples have the same local Hanf type if and only if their spheres are isomorphic.

Lemma 4.3. *Let* $\mathcal{A}$ *be a relational structure,* $k \in \mathbb{N}_{\geq 1}$*,* $r \in \mathbb{N}$*, and* $\bar{v}, \bar{v}' \in \big(U(\mathcal{A})\big)^k$*. It holds that* $\operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}) = \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}')$ *if and only if* $\mathcal{S}_r^{\mathcal{A}}(\bar{v}) \cong \mathcal{S}_r^{\mathcal{A}}(\bar{v}')$*.*

Proof. For the forward direction, assume $\operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}) = \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}')$. We have $\operatorname{sph}_{r,\bar{v}}^{\mathcal{A}} \in \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v})$ and hence, $\operatorname{sph}_{r,\bar{v}}^{\mathcal{A}} \in \operatorname{lhtp}_r^{\mathcal{A}}(\bar{v}')$. Thus, $\mathcal{A} \models \operatorname{sph}_{r,\bar{v}}^{\mathcal{A}}[\bar{v}']$, which is equivalent to $\mathcal{S}_r^{\mathcal{A}}(\bar{v})$ and $\mathcal{S}_r^{\mathcal{A}}(\bar{v}')$ being isomorphic.

For the backward direction, assume the spheres $\mathcal{S}_r^{\mathcal{A}}(\bar{v})$ and $\mathcal{S}_r^{\mathcal{A}}(\bar{v}')$ are isomorphic. Let $\bar{x} = (x_1, \dots, x_k)$ and let $\varphi(\bar{x})$ be an hnf formula of locality radius at most r. Then, φ is a Boolean combination of numerical oc-type conditions and sphere formulas with locality radius at most r. We show that $\mathcal{A} \models \varphi[\bar{v}]$ if and only if $\mathcal{A} \models \varphi[\bar{v}']$.

The numerical oc-type conditions in φ do not have any free structure variables. Hence, their evaluation only depends on the structure and is independent of the assignment.

The free variables of the sphere formulas used in φ are a subset of $\operatorname{free}(\varphi)$. Let $\operatorname{sph}_{r',\bar{w}}^{\mathcal{A}'}(x_{i_1}, \dots, x_{i_\ell})$ be such a sphere formula used in φ for some relational structure $\mathcal{A}'$, an ℓ-tuple $\bar{w}$ from $\mathcal{A}'$, and some locality radius $r' \leq r$. It follows from our assumption that $\mathcal{S}_{r'}^{\mathcal{A}}(v_{i_i}, \dots, v_{i_\ell}) \cong \mathcal{S}_{r'}^{\mathcal{A}}(v'_{i_i}, \dots, v'_{i_\ell})$. Thus,

$$\begin{aligned}
&\mathcal{A} \models \operatorname{sph}_{r',\bar{w}}^{\mathcal{A}}[v_{i_1}, \dots, v_{i_\ell}] \\
\iff\ &\mathcal{S}_{r'}^{\mathcal{A}}(v_{i_i}, \dots, v_{i_\ell}) \cong \mathcal{S}_{r'}^{\mathcal{A}'}(w_1, \dots, w_\ell) \\
\iff\ &\mathcal{S}_{r'}^{\mathcal{A}}(v'_{i_i}, \dots, v'_{i_\ell}) \cong \mathcal{S}_{r'}^{\mathcal{A}'}(w_1, \dots, w_\ell) \\
\iff\ &\mathcal{A} \models \operatorname{sph}_{r',\bar{w}}^{\mathcal{A}}[v'_{i_1}, \dots, v'_{i_\ell}].
\end{aligned}$$

This holds for all sphere formulas in φ. Thus, we have $\mathcal{A} \models \varphi[\bar{v}]$ if and only if $\mathcal{A} \models \varphi[\bar{v}']$. □

The following result is a variant of the Local Composition Lemma for first-order logic from [GR17], translated to first-order logic with counting and local Hanf types. It allows

us to analyse the parameters we choose by splitting them into two parts with disjoint neighbourhoods.

Lemma 4.4 (Local Composition Lemma for FOCN)**.** *Let $\mathcal{A}$ be a relational structure, $k, \ell \in \mathbb{N}_{\geq 1}$, $r \in \mathbb{N}$, $\bar{v}, \bar{v}' \in \big(U(\mathcal{A})\big)^k$, and $\bar{w}, \bar{w}' \in \big(U(\mathcal{A})\big)^\ell$, such that* $\operatorname{dist}(\bar{v}, \bar{w}) > 2r + 1$, $\operatorname{dist}(\bar{v}', \bar{w}') > 2r + 1$, $\operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}) = \operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}')$, *and* $\operatorname{lhtp}^{\mathcal{A}}_r(\bar{w}) = \operatorname{lhtp}^{\mathcal{A}}_r(\bar{w}')$. *Then,* $\operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}\bar{w}) = \operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}'\bar{w}')$.

Proof. From $\operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}) = \operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}')$, using Lemma 4.3, it follows that $\mathcal{S}^{\mathcal{A}}_r(\bar{v})$ and $\mathcal{S}^{\mathcal{A}}_r(\bar{v}')$ are isomorphic. Similarly, we obtain that $\mathcal{S}^{\mathcal{A}}_r(\bar{w})$ and $\mathcal{S}^{\mathcal{A}}_r(\bar{w}')$ are isomorphic from $\operatorname{lhtp}^{\mathcal{A}}_r(\bar{w}) = \operatorname{lhtp}^{\mathcal{A}}_r(\bar{w}')$. Because of the lower bounds for the distances, we have $\mathcal{N}^{\mathcal{A}}_r(\bar{v}) \cup \mathcal{N}^{\mathcal{A}}_r(\bar{w}) = \mathcal{N}^{\mathcal{A}}_r(\bar{v}\bar{w})$ and $\mathcal{N}^{\mathcal{A}}_r(\bar{v}') \cup \mathcal{N}^{\mathcal{A}}_r(\bar{w}') = \mathcal{N}^{\mathcal{A}}_r(\bar{v}'\bar{w}')$. Hence, by combining the above-mentioned isomorphisms, we can deduce that $\mathcal{S}^{\mathcal{A}}_r(\bar{v}\bar{w}) \cong \mathcal{S}^{\mathcal{A}}_r(\bar{v}'\bar{w}')$. With Lemma 4.3, it follows that $\operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}\bar{w}) = \operatorname{lhtp}^{\mathcal{A}}_r(\bar{v}'\bar{w}')$. □

5. Learning Problems for FOCN

With the definition of the Hanf normal form at hand, we can now formalise the learning problems for the first-order logic with counting FOCN that we consider in this paper.

Recall the problem FO-Learn-Consistent(σ, k, ℓ, q^*), where target concepts only use formulas of quantifier rank at most q^*, and algorithms are only allowed to return hypotheses of quantifier rank at most $f(k + \ell, q^*)$ for some function f. In the remainder of this paper, for the logic FOCN, instead of bounding the quantifier rank, we bound the binding rank and the binding width of the formulas by constants c_{br} and c_{bw}. For $c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$, let $\mathsf{FOCN}[\sigma, c_{\mathrm{br}}, c_{\mathrm{bw}}]$ denote the set of all formulas in $\mathsf{FOCN}[\sigma]$ of binding rank at most c_{br} and binding width at most c_{bw}. Since formulas from FOCN may have free number variables, we allow concepts to use number parameters in addition to the parameters from the structure.

Let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$ and fix a signature σ. We consider concepts that can be defined using a formula from

$$\Phi^* = \big\{\varphi(\bar{x}, \bar{y}, \bar{\kappa}) \in \mathsf{FOCN}[\sigma, c_{\mathrm{br}}, c_{\mathrm{bw}}] \;\mid\; |\bar{x}| = k,\ |\bar{y}| = \ell\big\},$$

combined with parameters that are elements from the structure as well as number parameters. For a σ-structure $\mathcal{A}$, a formula $\varphi(\bar{x}, \bar{y}, \bar{\kappa}) \in \Phi^*$, and tuples $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$ and $\bar{n} \in \mathbb{Z}^{|\bar{\kappa}|}$, the resulting hypothesis is the mapping $h^{\mathcal{A}}_{\varphi, \bar{w}, \bar{n}}(\bar{x})\colon \big(U(\mathcal{A})\big)^k \to \{0, 1\}$ which maps a tuple $\bar{v} \in \big(U(\mathcal{A})\big)^k$ to $[\![\varphi(\bar{v}, \bar{w}, \bar{n})]\!]^{\mathcal{A}}$.

As it turns out, to describe these concepts on a fixed structure, it actually suffices to use a Boolean combination of sphere formulas up to a certain locality radius without any number variables or number parameters. Hence, the formulas that our algorithms return come from the set

$$\begin{aligned}\Phi = \big\{\varphi(\bar{x}, \bar{y}) \in \mathsf{FO}[\sigma] \;\mid\; |\bar{x}| = k,\ |\bar{y}| = \ell,\\ \varphi \text{ is a Boolean combination of sphere formulas}\\ \text{of locality radius at most } (2 \cdot c_{\mathrm{bw}} + 1)^{c_{\mathrm{br}}}\big\}.\end{aligned}$$

As we will see, with different techniques in Sections 6 and 7, the restriction of the locality radius of the returned formula still allows us to evaluate the hypothesis on new tuples efficiently. The consistent-learning problem for FOCN is formally defined as follows.

FOCN-Learn-Consistent$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$

Input: σ-structure $\mathcal{A}$, training sequence $T \in \Big(\big(U(\mathcal{A})\big)^k \times \{0,1\}\Big)^m$

Task: Return a first-order formula φ and a tuple $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$, where φ is a Boolean combination of sphere formulas of locality radius at most $(2 \cdot c_{\mathrm{bw}} + 1)^{c_{\mathrm{br}}}$, such that the hypothesis $h^{\mathcal{A}}_{\varphi,\bar{w}}$ is consistent with T.

The algorithm may reject if there is no combination of a formula $\varphi^*(\bar{x}, \bar{y}, \bar{\kappa}) \in \mathsf{FOCN}[\sigma, c_{\mathrm{br}}, c_{\mathrm{bw}}]$ and tuples $\bar{w}^* \in \big(U(\mathcal{A})\big)^\ell$, $\bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|}$ such that the hypothesis $h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}$ is consistent with T.

In Section 6, we show that this problem is solvable in sublinear time on classes of structures of bounded degree. In Section 7, with a different approach, we extend this result to classes of structures of polylogarithmic degree.

The ERM- and PAC-learning problems for FOCN that we study in this paper are defined as follows.

FOCN-Learn-ERM$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$

Input: structure $\mathcal{A}$, training sequence $T \in \Big(\big(U(\mathcal{A})\big)^k \times \{0,1\}\Big)^m$

Task: Return a first-order formula φ and a tuple $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$, where φ is a Boolean combination of sphere formulas of locality radius at most $(2 \cdot c_{\mathrm{bw}} + 1)^{c_{\mathrm{br}}}$, such that

$$\mathrm{err}_T\big(h^{\mathcal{A}}_{\varphi,\bar{w}}\big) \leq \min\Big\{ \mathrm{err}_T(h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}) \;\Big|\; \varphi^*(\bar{x},\bar{y},\bar{\kappa}) \in \mathsf{FOCN}[\sigma, c_{\mathrm{br}}, c_{\mathrm{bw}}], \; \bar{w}^* \in (U(\mathcal{A}))^\ell,\ \bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|} \Big\}.$$

FOCN-Learn-PAC$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$

Input: structure $\mathcal{A}$, rational numbers $\varepsilon, \delta > 0$, probability distribution $\mathcal{D}$ on $\big(U(\mathcal{A})\big)^k \times \{0,1\}$

Task: Return a first-order formula φ and a tuple $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$, where φ is a Boolean combination of sphere formulas of locality radius at most $(2 \cdot c_{\mathrm{bw}} + 1)^{c_{\mathrm{br}}}$, such that, with probability of at least $1 - \delta$ over the choice of examples drawn i.i.d. from $\mathcal{D}$, it holds that

$$\mathrm{err}_{\mathcal{D}}\big(h^{\mathcal{A}}_{\varphi,\bar{w}}\big) \leq \varepsilon^* + \varepsilon,$$

where

$$\varepsilon^* \coloneqq \min\Big\{ \mathrm{err}_{\mathcal{D}}(h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}) \;\Big|\; \varphi^*(\bar{x},\bar{y},\bar{\kappa}) \in \mathsf{FOCN}[\sigma, c_{\mathrm{br}}, c_{\mathrm{bw}}], \; \bar{w}^* \in (U(\mathcal{A}))^\ell,\ \bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|} \Big\}.$$

In Section 6, we modify the algorithm we give for the consistent-learning problem to show that FOCN-Learn-ERM is solvable in sublinear time on classes of structures of bounded degree. Afterwards, we use this result to show that also PAC learning is possible in sublinear time on these classes of structures. In Section 7, we extend the consistent-learning result on classes of structures of polylogarithmic degree to ERM learning. Furthermore, we

provide a PAC-learning algorithm that runs in sublinear time on classes of structures with a stricter (but still not constant) degree bound.

6. Learning on Structures of Bounded Degree

In this section, we present learning results for FOCN on classes of structures of bounded degree. We start with the consistent-learning problem.

Theorem 6.1. *Let σ be a relational signature, let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$, and let $\mathcal{C}$ be a class of structures of degree at most d for some $d \in \mathbb{N}$. There is an algorithm that solves* FOCN-Learn-Consistent$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$ *on $\mathcal{C}$ in time $(\log n + m)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in time $m^{\mathcal{O}(1)}$ under the uniform-cost measure, where n is the size of the background structure and m is the length of the training sequence.*

Furthermore, the hypotheses returned by the algorithm can be evaluated in time $(\log n)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in constant time under the uniform-cost measure.

The high-level proof idea is similar to the one Grohe and Ritzert [GR17] presented for the consistent-learning problem for first-order logic. We use a brute-force algorithm that checks all combinations of certain choices of formulas and certain choices of parameters. Hence, there are two main ingredients for our proof.

First, as we observe in Lemma 6.4, for fixed $\sigma, d, k, \ell, c_{\mathrm{br}}$, and c_{bw}, the number of formulas we need to check is constant.

Second, to bound the number of parameters to check, we show that it suffices to consider only parameters in a certain neighbourhood around the training examples. As shown in Figure 5, intuitively, this holds because parameters that are far away from the training examples do not help to distinguish positive from negative examples. The formal result is given in Lemma 6.2.

For the rest of this section, let σ be a fixed relational signature, $d, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$, let $r \coloneqq (2 \cdot c_{\mathrm{bw}} + 1)^{c_{\mathrm{br}}}$, let $\mathcal{C}$ be a class of σ-structures of degree at most d, and let $\mathcal{A}$ be a structure from $\mathcal{C}$. Let Φ^*, *i.e.* the set of formulas that our target concepts are based upon, be defined as in the last section, that is,

$$\Phi^* = \big\{\varphi(\bar{x}, \bar{y}, \bar{\kappa}) \in \mathsf{FOCN}[\sigma, c_{\mathrm{br}}, c_{\mathrm{bw}}] \;\mid\; |\bar{x}| = k,\ |\bar{y}| = \ell\big\}.$$

For Φ, that is, the set of formulas our algorithms are allowed to return in a hypothesis, we can even use a restriction of the set from the last section and set

$$\begin{aligned}\Phi_d \coloneqq \big\{\varphi(\bar{x}, \bar{y}) \in \mathsf{FO}[\sigma] \;\mid\; |\bar{x}| = k,\ |\bar{y}| = \ell,\\ \varphi \text{ is a Boolean combination of sphere formulas}\\ \text{of locality radius at most } r\\ \text{based on spheres of degree at most } d\big\}.\end{aligned}$$

For a training sequence $T = \big((\bar{v}_1, \lambda_1), \ldots, (\bar{v}_m, \lambda_m)\big)$ and a radius $r' \in \mathbb{N}$, let $N^{\mathcal{A}}_{r'}(T) \coloneqq \bigcup_{i \in [m]} N^{\mathcal{A}}_{r'}(\bar{v}_i)$.

Lemma 6.2. *Let $T \in \big((U(\mathcal{A}))^k \times \{0,1\}\big)^m$ be a training sequence and let $\varphi^* \in \Phi^*$, $\bar{w}^* \in \big(U(\mathcal{A})\big)^\ell$, and $\bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|}$ be such that the hypothesis $h^{\mathcal{A}}_{\varphi^*, \bar{w}^*, \bar{n}^*}$ is consistent with T. There is a formula $\varphi \in \Phi_d$ and a tuple $\bar{w} \in \big(N^{\mathcal{A}}_{(2r+1)\ell}(T)\big)^\ell$ such that the hypothesis $h^{\mathcal{A}}_{\varphi, \bar{w}}$ is consistent with T.*

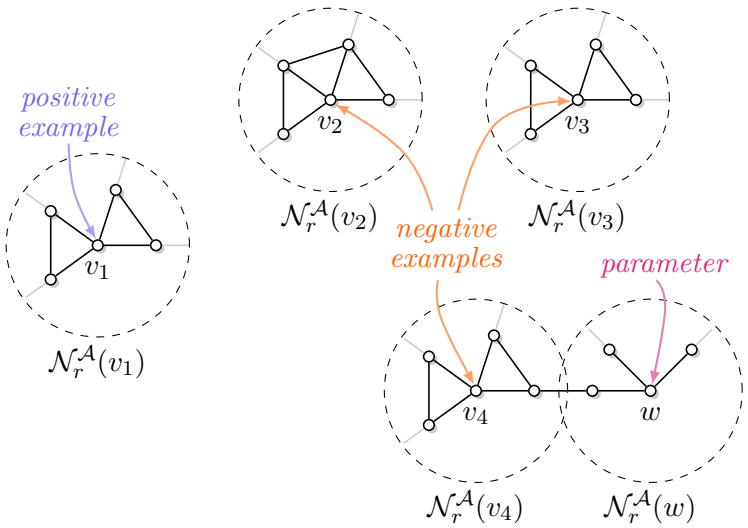

Figure 5: One positive and three negative examples from a training sequence as well as a parameter with their local neighbourhoods. The vertices v_1 and v_2 can easily be distinguished by a formula since they have different local types. The vertices v_1 and v_3 have the same local types, and even if we take the parameter w into consideration, the local types of the tuples (v_1, w) and (v_3, w) are still the same since the parameter is too far away from both vertices v_1 and v_3. Thus, there is no way to distinguish v_1 and v_3 using w and a formula with locality radius at most r. The only way to distinguish vertices of the same local type is to have a parameter close to one of the vertices, as shown for v_4. This argumentation is formalised in the proof of Lemma 6.2.

Proof. Let $T = \big((\bar{v}_1, \lambda_1), \dots, (\bar{v}_m, \lambda_m)\big)$, φ^*, $\bar{w}^* = (w_1^*, \dots, w_\ell^*)$, and $\bar{n}^*$ be as given in the lemma. We iteratively select vertices $w^{(i)}$ from the parameters $w_1^*, \dots, w_\ell^*$ that have distance at most $2r+1$ from the examples or the already selected vertices. This process is repeated for s steps until all remaining parameters are too far away (or all parameters have already been selected). For the tuple $\bar{w}$ that we are looking for in this proof, we use these selected parameters and omit the others.

Formally, to select the parameters, we start with the neighbourhood $N^{(0)} := N^{\mathcal{A}}_{2r+1}(T)$ of radius $2r+1$ around the examples and select a vertex $w \in \{w_1^*, \dots, w_\ell^*\} \cap N^{(0)}$. If there is no such vertex, we set $s := 0$ and stop this process. Otherwise, we set $w^{(1)} := w$, $N^{(1)} := N^{(0)} \cup N^{\mathcal{A}}_{2r+1}(w)$, and continue. For $i \geq 2$, we select a vertex $w \in \{w_1^*, \dots, w_\ell^*\} \setminus \{w^{(1)}, \dots, w^{(i-1)}\}$ that is contained in the neighbourhood $N^{(i-1)}$. If there is no such vertex, we set $s := i-1$ and stop. Otherwise, we set $w^{(i)} := w$, $N^{(i)} := N^{(i-1)} \cup N^{\mathcal{A}}_{2r+1}(w)$, and continue. W.l.o.g. let $w^{(i)} = w_i^*$ for $i \in [s]$. Let $\bar{w}^{\text{in}} := (w_1^*, \dots, w_s^*)$ and $\bar{w}^{\text{out}} := (w_{s+1}^*, \dots, w_\ell^*)$. We let $\bar{y}^{\text{in}} := (y_1, \dots, y_s)$ and choose

$$\varphi(\bar{x}, \bar{y}) := \bigvee_{i \in [m],\ \lambda_i = 1} \operatorname{sph}^{\mathcal{A}}_{r, \bar{v}_i \bar{w}^{\text{in}}}(\bar{x}, \bar{y}^{\text{in}}).$$

The formula φ is a Boolean combination of sphere formulas of locality radius at most r based on spheres of degree at most d and thus, $\varphi \in \Phi_d$. We turn $\bar{w}^{\text{in}} = (\bar{w}_1^*, \ldots, \bar{w}_s^*)$ into a tuple $\bar{w} \in \left(N^{\mathcal{A}}_{(2r+1)\ell}(T)\right)^\ell$ by choosing an arbitrary $w \in N^{\mathcal{A}}_{(2r+1)\ell}(T)$ and filling the missing $(\ell - s)$ positions with the vertex w.

It remains to show that the hypothesis $h^{\mathcal{A}}_{\varphi,\bar{w}}$ is consistent with T. If $\lambda_i = 1$, then by the construction of $h^{\mathcal{A}}_{\varphi,\bar{w}}$ (especially the construction of φ), it holds that $h^{\mathcal{A}}_{\varphi,\bar{w}}(\bar{v}_i) = 1$. For the other direction, we use the following claim.

Claim 6.3. Let $i, j \in [m]$ such that $\mathcal{A} \models \text{sph}^{\mathcal{A}}_{r,\,\bar{v}_i\bar{w}^{\text{in}}}[\bar{v}_j\bar{w}^{\text{in}}]$. Then $\lambda_i = \lambda_j$.

Proof. First, from $\mathcal{A} \models \text{sph}^{\mathcal{A}}_{r,\,\bar{v}_i\bar{w}^{\text{in}}}(\bar{v}_j\bar{w}^{\text{in}})$, it follows that $\mathcal{S}^{\mathcal{A}}_r(\bar{v}_i\bar{w}^{\text{in}}) \cong \mathcal{S}^{\mathcal{A}}_r(\bar{v}_j\bar{w}^{\text{in}})$. Using Lemma 4.3, we obtain $\text{lhtp}^{\mathcal{A}}_r(\bar{v}_i\bar{w}^{\text{in}}) = \text{lhtp}^{\mathcal{A}}_r(\bar{v}_j\bar{w}^{\text{in}})$.

Second, from the construction of $w^{(i)}$ and $N^{(i)}$, it follows that $N^{\mathcal{A}}_{2r+1}(\bar{v}_p) \subseteq N^{(0)} \subseteq N^{(s)}$ for all $p \in [m]$, $N^{\mathcal{A}}_{2r+1}(\bar{w}_p^*) \subseteq N^{(p)} \subseteq N^{(s)}$ for all $p \in [s]$, and $\bar{w}_p^* \notin N^{(s)}$ for all $p \in [s+1, \ell]$. Thus, $\text{dist}^{\mathcal{A}}(\bar{v}_p\bar{w}^{\text{in}}, \bar{w}^{\text{out}}) > 2r+1$ for every $p \in [m]$.

Using Lemma 4.4 and $\bar{w}^* = \bar{w}^{\text{in}}\bar{w}^{\text{out}}$, we obtain $\text{lhtp}^{\mathcal{A}}_r(\bar{v}_i\bar{w}^*) = \text{lhtp}^{\mathcal{A}}_r(\bar{v}_j\bar{w}^*)$. With Lemma 4.2 and our choice of the radius r, it then follows that

$$\mathcal{A} \models \varphi^*[\bar{v}_i, \bar{w}^*, \bar{n}^*] \iff \mathcal{A} \models \varphi^*[\bar{v}_j, \bar{w}^*, \bar{n}^*].$$

Since $h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}$ is assumed to be consistent with T, this implies $\lambda_i = \lambda_j$. ⌟

If $h^{\mathcal{A}}_{\varphi,\bar{w}}(\bar{v}_i) = 1$, then there is some $p \in [m]$ such that $\lambda_p = 1$ and $\mathcal{A} \models \text{sph}^{\mathcal{A}}_{r,\,\bar{v}_p\bar{w}^{\text{in}}}[\bar{v}_i\bar{w}^{\text{in}}]$. Using the claim, we obtain $\lambda_i = \lambda_p = 1$. Thus, all in all, $h^{\mathcal{A}}_{\varphi,\bar{w}}$ is consistent with T. □

This result shows that we only have to look for parameters in a local neighbourhood around the examples. In structures of bounded degree, this drastically reduces the number of parameters we have to check. Next, we bound the number of formulas we have to consider.

Lemma 6.4. *For fixed $\sigma, d, k, \ell, c_{\text{br}}$, and c_{bw}, up to equivalence, the number of formulas in Φ_d is constant.*

Proof. In σ-structures of degree at most d, for $r = (2 \cdot c_{\text{bw}} + 1)^{c_{\text{br}}}$, the number of elements in an r-sphere with $(k+\ell)$ centres can be bounded by $(k+\ell) \cdot \mu_d(r)$ with $\mu_0(r) \coloneqq 1$, $\mu_1(r) \coloneqq 2$, and $\mu_d(r) \coloneqq 1 + d \cdot \sum_{i=0}^{r}(d-1)^i$ for $d \geq 2$. We have $\mu_2(r) = 2r+1$ and, for $d > 2$, one can show that $\mu_d(r) \leq (d-1)^{r+1}$. Hence, since σ is fixed, there is a constant number of non-isomorphic spheres of radius at most r in such σ-structures. Thus, the number of sphere formulas based on those spheres, up to equivalence, is also constant. Since Φ_d consists of all Boolean combinations of these sphere formulas, the number of non-equivalent formulas in Φ_d is constant as well. □

With a bound on the number of parameters and a constant number of formulas, it remains to show that we can check every single hypothesis efficiently. For this, we use the following result due to Seese [See96].

Theorem 6.5 [See96]**.** *Let $d \in \mathbb{N}$ and let σ be a relational signature. There is a function $f\colon \mathbb{N} \to \mathbb{N}$ and an algorithm $\mathcal{A}_{MC}$ that, on input $(\mathcal{A}, \varphi)$ for a σ-structure $\mathcal{A}$ of degree at most d and an $\mathsf{FO}[\sigma]$-sentence φ, decides whether $\mathcal{A} \models \varphi$ holds in time $f(|\varphi|) \cdot |\mathcal{A}|$ under the uniform-cost measure and in time $f(|\varphi|) \cdot |\mathcal{A}| \log |\mathcal{A}|$ under the logarithmic-cost measure.*

We can now prove the consistent-learning result.

Require: local access to background structure $\mathcal{A}$,
training sequence $T = \big((\bar{v}_1, \lambda_1), \dots, (\bar{v}_m, \lambda_m)\big)$

1: $N \leftarrow N^{\mathcal{A}}_{(2r+1)\ell}(T)$
2: **for all** $\bar{w} \in N^\ell$ **do**
3: **for all** $\varphi \in \Phi_d$ **do**
4: *consistent* $\leftarrow$ **true**
5: **for all** $i \in [m]$ **do**
6: **if** $[\![\varphi(\bar{v}_i, \bar{w})]\!]^{\mathcal{N}^{\mathcal{A}}_r(\bar{v}_i\bar{w})} \neq \lambda_i$ **then**
7: *consistent* $\leftarrow$ **false**
8: **break**
9: **if** *consistent* **then**
10: **return** $(\varphi, \bar{w})$
11: **reject**

Figure 6: Learning algorithm $\mathcal{A}^d_{\mathrm{con}}$ for Theorem 6.1

Proof of Theorem 6.1. We show that the algorithm given in Figure 6 fulfils the requirements of the theorem. The algorithm goes through all tuples $\bar{w} \in \big(N^{\mathcal{A}}_{(2r+1)\ell}(T)\big)^\ell$ and all non-equivalent formulas $\varphi \in \Phi_d$. A hypothesis $h^{\mathcal{A}}_{\varphi,\bar{w}}$ is consistent with the training sequence T if and only if $[\![\varphi(\bar{v}_i, \bar{w})]\!]^{\mathcal{A}} = \lambda_i$ for all $i \in [m]$. Since Φ_d only contains Boolean combinations of sphere formulas of locality radius at most r, all formulas in Φ_d are r-local. Thus, $h^{\mathcal{A}}_{\varphi,\bar{w}}$ is consistent with T if and only if $[\![\varphi(\bar{v}_i, \bar{w})]\!]^{\mathcal{N}^{\mathcal{A}}_r(\bar{v}_i\bar{w})} = \lambda_i$ for every $i \in [m]$. Hence, if the algorithm returns a hypothesis, then it is consistent. Furthermore, if there is a consistent hypothesis $h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}$ using a formula $\varphi^*(\bar{x}, \bar{y}, \bar{\kappa}) \in \Phi^*$ and tuples $\bar{w}^* \in \big(U(\mathcal{A})\big)^\ell$, and $\bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|}$, then, by Lemma 6.2, there is a consistent hypothesis among the ones we check, so the algorithm returns a hypothesis.

It remains to show that the algorithm satisfies the running-time requirements while only using local access to the structure $\mathcal{A}$. For all $\bar{v} \in \big(U(\mathcal{A})\big)^k$ and $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$, as discussed in the proof of Lemma 6.4, the size of the neighbourhood $N^{\mathcal{A}}_r(\bar{v}\bar{w})$ can be bounded by $(k+\ell) \cdot \mu_d(r)$, so it is constant for fixed d, k, ℓ, r. Hence, under the logarithmic-cost measure, the neighbourhood can be computed in time $\mathcal{O}(\log n)$ using only local access. Under the uniform-cost measure, it takes constant time to compute the neighbourhood. By Theorem 6.5, on an already computed constant-size neighbourhood, the evaluation of the hypothesis in line 6 runs in constant time. The algorithm checks up to $|N|^\ell \cdot |\Phi_d| \in \mathcal{O}\big((m \cdot k \cdot d^{(2r+1)\ell+1})^\ell \cdot |\Phi_d|\big)$ hypotheses on m examples with $N = N^{\mathcal{A}}_{(2r+1)\ell}(T)$ and where $|\Phi_d|$ only considers non-equivalent formulas. All in all, since d, k, ℓ, r are considered constant, the running time of the algorithm is in $(m + \log n)^{\mathcal{O}(1)}$ under the logarithmic-cost measure, in $m^{\mathcal{O}(1)}$ under the uniform-cost measure, and it only uses local access to the structure $\mathcal{A}$.

To evaluate the hypothesis returned by the algorithm on a tuple $\bar{v}$, we only have to evaluate it within the neighbourhood $\mathcal{N}^{\mathcal{A}}_r(\bar{v}\bar{w})$, using only local access to $\mathcal{A}$. Analogously to the consistency check, the hypothesis can be evaluated in time $(\log n)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in constant time under the uniform-cost measure. □

In the proof, we rely on Φ_d being a constant-sized set of formulas which is expressive enough to describe every concept that can be described using a formula from Φ^*. We obtain

the expressiveness via formulas in Hanf normal form. However, to bound the number of these formulas, we need to bound the degree of the structures we consider in Lemma 6.4. Without this bound on the degree, even in structures of only logarithmic degree, the bound on the number of formulas in Φ_d would be superlinear in the size of the structure, so this would not yield a sublinear-time learning algorithm any more. Thus, in Section 7, we use a different technique to prove consistent learnability on structures of polylogarithmic degree.

Next, we extend Theorem 6.1 to the ERM problem.

Theorem 6.6. *Let σ be a relational signature, let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$, and let $\mathcal{C}$ be a class of structures of degree at most d for some $d \in \mathbb{N}$. There is an algorithm that solves* FOCN-Learn-ERM$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$ *on $\mathcal{C}$ in time $(\log n + m)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in time $m^{\mathcal{O}(1)}$ under the uniform-cost measure, where n is the size of the background structure and m is the length of the training sequence.*

Furthermore, the hypotheses returned by the algorithm can be evaluated in time $(\log n)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in constant time under the uniform-cost measure.

To prove this result, we use the following corollary of Lemma 6.2.

Corollary 6.7. *Let $T \in \big((U(\mathcal{A}))^k \times \{0,1\}\big)^m$ be a training sequence and let $\varphi^* \in \Phi^*$, $\bar{w}^* \in \big(U(\mathcal{A})\big)^\ell$, and $\bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|}$. There is a formula $\varphi \in \Phi_d$ and a tuple $\bar{w} \in \big(N^{\mathcal{A}}_{(2r+1)\ell}(T)\big)^\ell$ such that $\mathrm{err}_T\big(h^{\mathcal{A}}_{\varphi,\bar{w}}\big) \leq \mathrm{err}_T\big(h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}\big)$.*

Proof. Let $\varepsilon \coloneqq \mathrm{err}_T\big(h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}\big)$. There is a sequence S that is a subsequence of T of length $(1-\varepsilon)\cdot|T|$ such that $h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}$ is consistent with S. By Lemma 6.2, there is also a formula $\varphi \in \Phi_d$ and a tuple $\bar{w} \in \big(N^{\mathcal{A}}_{(2r+1)\ell}(T)\big)^\ell$ such that $h^{\mathcal{A}}_{\varphi,\bar{w}}$ is consistent with S. Thus, we have $\mathrm{err}_T\big(h^{\mathcal{A}}_{\varphi,\bar{w}}\big) \leq \varepsilon = \mathrm{err}_T\big(h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}\big)$. □

Using this corollary, we can now prove Theorem 6.6.

Proof of Theorem 6.6. We show that the algorithm given in Figure 7 fulfils the requirements of the theorem. The algorithm goes through all tuples $\bar{w} \in \big(N^{\mathcal{A}}_{(2r+1)\ell}(T)\big)^\ell$ and all non-equivalent formulas $\varphi \in \Phi_d$ and counts the number of errors that $h^{\mathcal{A}}_{\varphi,\bar{w}} = [\![\varphi(\bar{x},\bar{y})]\!]^{\mathcal{A}}(\bar{x},\bar{w})$ makes on T. Then, it returns a hypothesis with minimal training error. By Corollary 6.7, the hypothesis returned by the algorithm fulfils the requirements of the problem FOCN-Learn-ERM. The running-time analysis is analogous to the one presented in the proof of Theorem 6.1. □

To solve FOCN-Learn-PAC, the remaining missing ingredient is the following result that gives us a bound on the needed queried examples as well as a bound on the difference between the training and the generalisation error.

Lemma 6.8 (Uniform Convergence [SB14])**.** *Let $\mathcal{H}$ be a finite hypothesis class over the instance space X and let*

$$m^{UC}_{\mathcal{H}}(\varepsilon,\delta) \coloneqq \left\lceil \frac{\log(2\,|\mathcal{H}|\,/\delta)}{2\varepsilon^2} \right\rceil.$$

For all $\varepsilon, \delta > 0$ and for every distribution $\mathcal{D}$ over $X \times \{0,1\}$, if a training sequence T of length at least $m^{UC}_{\mathcal{H}}(\varepsilon,\delta)$ is drawn i.i.d. from $\mathcal{D}$, then, with probability at least $1-\delta$, the training sequence is ε-representative, that is, for all $h \in \mathcal{H}$,

$$\big|\,\mathrm{err}_T(h) - \mathrm{err}_{\mathcal{D}}(h)\big| \leq \varepsilon.$$

Require: local access to background structure $\mathcal{A}$,
training sequence $T = \big((\bar{v}_1, \lambda_1), \ldots, (\bar{v}_m, \lambda_m)\big)$

$N \leftarrow N^{\mathcal{A}}_{(2r+1)\ell}(T)$
$error_{\min} \leftarrow |T| + 1$
for all $\bar{w} \in N^\ell$ **do**
  for all $\varphi \in \Phi_d$ **do**
    $error \leftarrow 0$
    for all $i \in [m]$ **do**
      if $[\![\varphi(\bar{v}_i, \bar{w})]\!]^{\mathcal{N}_r^{\mathcal{A}}(\bar{v}_i\bar{w})} \neq \lambda_i$ **then**
        $error \leftarrow error + 1$
    if $error < error_{\min}$ **then**
      $error_{\min} \leftarrow error$
      $(\varphi_{\min}, \bar{w}_{\min}) \leftarrow (\varphi, \bar{w})$
return $(\varphi_{\min}, \bar{w}_{\min})$

Figure 7: Learning algorithm $\mathcal{A}^d_{\mathrm{ERM}}$ for Theorem 6.6

Finally, we obtain agnostic PAC learnability of FOCN via the ERM algorithm.

Theorem 6.9. *Let σ be a relational signature, let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$, and let $\mathcal{C}$ be a class of structures of degree at most d for some $d \in \mathbb{N}$. There is an algorithm that solves* FOCN-Learn-PAC$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$ *on $\mathcal{C}$ in time $\big(\log|\mathcal{A}| + \log\frac{1}{\delta} + \frac{1}{\varepsilon}\big)^{\mathcal{O}(1)}$, under the logarithmic-cost as well as the uniform-cost measure, where n is the size of the background structure.*

Furthermore, the hypotheses returned by the algorithm can be evaluated in time $(\log n)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in constant time under the uniform-cost measure.

Proof. Let $\mathcal{A} \in \mathcal{C}$ be a background structure of degree at most d. We consider the concept class

$$\mathcal{H}^* = \big\{h^{\mathcal{A}}_{\varphi,\bar{w},\bar{n}} \;\big|\; \varphi(\bar{x}, \bar{y}, \bar{\kappa}) \in \Phi^*,\ \bar{w} \in \big(U(\mathcal{A})\big)^\ell,\ \bar{n} \in \mathbb{Z}^{|\kappa|}\big\}$$

and the hypothesis class

$$\mathcal{H} = \big\{h^{\mathcal{A}}_{\varphi,\bar{w}} \;\big|\; \varphi(\bar{x}, \bar{y}) \in \Phi_d,\ \bar{w} \in \big(U(\mathcal{A})\big)^\ell\big\}.$$

Since, by Lemma 6.4, Φ_d contains (up to equivalence) only finitely many formulas, the number of hypotheses in $\mathcal{H}$ is bounded by $s \cdot |\mathcal{A}|^\ell$ for some constant s.

Claim 6.10. It holds that $\mathcal{H}^* \subseteq \mathcal{H}$.

Proof. Let $h^* := h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*} \in \mathcal{H}^*$. We consider the training sequence T that contains an example $\big(\bar{v}, h^*(\bar{v})\big)$ for every k-tuple $\bar{v}$ from $\mathcal{A}$. By Lemma 6.2, there is a formula $\varphi \in \Phi_d$ and a tuple $\bar{w} \in \big(U(\mathcal{A})\big)^\ell$ such that the hypothesis $h^{\mathcal{A}}_{\varphi,\bar{w}} \in \mathcal{H}$ is consistent with T. By the definition of T, we have $h^* = h^{\mathcal{A}}_{\varphi,\bar{w}}$, and thus $h^* \in \mathcal{H}$. ⌟

By using the claim, we can also bound the number of hypotheses in $\mathcal{H}^*$ by $s \cdot |\mathcal{A}|^\ell$. Our algorithm that solves FOCN-Learn-PAC works as follows.

Given local access to a background structure $\mathcal{A}$, oracle access to the size $|\mathcal{A}|$ of the structure, oracle access to a probability distribution $\mathcal{D}$ on $\big(U(\mathcal{A})\big)^k \times \{0,1\}$, and given

rational numbers $\varepsilon, \delta > 0$, our algorithm queries

$$m(|\mathcal{A}|, \varepsilon, \delta) \coloneqq \left\lceil \frac{2 \log(2s \cdot |\mathcal{A}|^\ell / \delta)}{\varepsilon^2} \right\rceil$$

many examples from $\mathcal{D}$. Then, it runs $\mathcal{A}^d_{\mathrm{ERM}}$ on the resulting training sequence.

Next, we show that this algorithm indeed solves the problem FOCN-Learn-PAC. Let $\mathcal{D}$ be a distribution over $\big(U(\mathcal{A})\big)^k \times \{0,1\}$ and let $h^* \in \mathcal{H}^*$ be a hypothesis that minimises the generalisation error, that is, $\mathrm{err}_{\mathcal{D}}(h^*) = \min_{h' \in \mathcal{H}^*} \mathrm{err}_{\mathcal{D}}(h')$. Let T be the training sequence of length $m(|\mathcal{A}|, \varepsilon, \delta)$ drawn i.i.d. from $\mathcal{D}$ by our algorithm, and let $h \in \mathcal{H}$ be the hypothesis returned by $\mathcal{A}^d_{\mathrm{ERM}}$ on input T. By Theorem 6.6, the hypothesis h fulfils $\mathrm{err}_T(h) \leq \mathrm{err}_T(h^*)$.

Furthermore, by the Uniform Convergence Lemma (Lemma 6.8), with probability at least $1 - \delta$, it holds that $\big|\mathrm{err}_T(h') - \mathrm{err}_{\mathcal{D}}(h')\big| \leq \frac{\varepsilon}{2}$ for all $h' \in \mathcal{H}$. This especially holds for h as well as for h^*. Hence,

$$\mathrm{err}_{\mathcal{D}}(h) \leq \mathrm{err}_T(h) + \frac{\varepsilon}{2} \leq \mathrm{err}_T(h^*) + \frac{\varepsilon}{2} \leq \mathrm{err}_D(h^*) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

with probability at least $1 - \delta$. This is exactly the requirement we have in the problem FOCN-Learn-PAC for the returned hypothesis.

The number $m(|\mathcal{A}|, \varepsilon, \delta)$ of queried examples can be bounded by $\mathcal{O}\left(\frac{\log(|\mathcal{A}|/\delta)}{\varepsilon^2}\right)$. Thus, by Theorem 6.6, we can bound the running time of our algorithm by $\big(\log|\mathcal{A}| + \log\frac{1}{\delta} + \frac{1}{\varepsilon}\big)^{\mathcal{O}(1)}$ under the logarithmic-cost as well as the uniform-cost measure. The evaluation time of the hypothesis given in the theorem follows directly from Theorem 6.6. □

7. Learning on Structures of Small Degree

In this section, we extend the sublinear-time results for consistent learning and the ERM problem of the previous section to classes of structures of at most polylogarithmic degree. At the end of this section, we give a bound on the degree of a structure in terms of its size such that PAC learning is still possible in sublinear time. We start with the extension of the consistent-learning result.

Theorem 7.1. *Let σ be a relational signature and let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$. There is an algorithm that solves* FOCN-Learn-Consistent$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$ *in time $(\log n + m)^{\mathcal{O}(1)} \cdot d^{\mathrm{polylog}\, d}$ under the logarithmic-cost measure and in time $m^{\mathcal{O}(1)} \cdot d^{\mathrm{polylog}\, d}$ under the uniform-cost measure, where n is the size of the background structure, d is the degree of the background structure, and m is the length of the training sequence. Furthermore, the hypotheses returned by the algorithm can be evaluated with the same time bound.*

On classes of structures of polylogarithmic degree, Theorem 7.1 implies that consistent learning is possible in sublinear time.

Corollary 7.2. *Let σ be a relational signature, let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$, and let $\mathcal{C}$ be a class of structures of polylogarithmic degree. There is an algorithm that solves the problem* FOCN-Learn-Consistent$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$ *on $\mathcal{C}$ in time sublinear in the size of the background structure and polynomial in the length of the training sequence, under the logarithmic-cost as well as the uniform-cost measure. The hypotheses returned by the algorithm can be evaluated with the same bound on the running time.*

Require: local access to background structure $\mathcal{A}$,
training sequence $T = \big((\bar{v}_1, \lambda_1), \ldots, (\bar{v}_m, \lambda_m)\big)$

$N \leftarrow N^{\mathcal{A}}_{(2r+1)\ell}(T)$
for all $\bar{w} = (w_1, \ldots, w_\ell) \in N^\ell$ **do**
  for all $s \in [0, \ell]$ **do**
    consistent $\leftarrow$ **true**
    $\bar{w}^{\text{in}} \leftarrow (w_1, \ldots, w_s)$
    for all $i \in [m]$ **do**
      $\mathcal{S}_i \leftarrow \mathcal{S}^{\mathcal{A}}_r(\bar{v}_i \bar{w}^{\text{in}})$
    for all $i, j \in [m]$ with $\lambda_i = 0$ and $\lambda_j = 1$ **do**
      if $\mathcal{S}_i \cong \mathcal{S}_j$ **then**
        consistent $\leftarrow$ **false**
        break
    if *consistent* **then**
      $\varphi(\bar{x}, \bar{y}) \leftarrow \bigvee_{i \in [m],\ \lambda_i = 1} \text{sph}^{\mathcal{A}}_{r, \bar{v}_i \bar{w}^{\text{in}}}(\bar{x}, y_1, \ldots, y_s)$
      return $(\varphi, \bar{w})$
reject

Figure 8: Learning algorithm $\mathcal{A}_{\text{con}}$ for Theorem 7.1

In contrast to the algorithms in Section 6 (and also in [GR17]), the length of the formulas returned by the algorithms in the present section will depend on the size of the structure. Hence, standard model-checking results (such as Theorem 6.5) do not yield the desired running-time bounds. Instead, in the proof of Theorem 7.1, to check the consistency of a hypothesis and to evaluate it on new tuples, we use the following result on isomorphism testing due to Grohe, Neuen, and Schweitzer [GNS23].

Theorem 7.3 [GNS23], [Neu19, Theorem 6.6.4]**.** *There is a constant c such that for all σ-structures $\mathcal{A}_1$ and $\mathcal{A}_2$, it can be decided in time $n^{\mathcal{O}(a \cdot (\log d)^c)}$ whether $\mathcal{A}_1$ and $\mathcal{A}_2$ are isomorphic, where $n \coloneqq \max\{|\mathcal{A}_1|, |\mathcal{A}_2|\}$, $d \coloneqq \max\{\deg(\mathcal{A}_1), \deg(\mathcal{A}_2)\}$, and $a \coloneqq \max_{R \in \sigma} \text{ar}(R)$.*

Whenever we evaluate a hypothesis on a given tuple, we assume that we are not only given the formula for the hypothesis, but also a description of the spheres, *i.e.* the relational structures, that are the basis for the sphere formulas used in the hypothesis. Then, to evaluate the hypothesis, for every sphere formula used in the hypothesis, we determine whether the sphere of the sphere formula is isomorphic to the sphere around the elements given to the sphere formula. The label defined by the hypothesis is then simply a Boolean combination of the determined truth values. We analyse the running time of this procedure in the following proof of the consistent-learning result.

Proof of Theorem 7.1. The pseudocode for our algorithm is shown in Figure 8. As in the last section, let $r \coloneqq (2 \cdot c_{\text{bw}} + 1)^{c_{\text{br}}}$. The algorithm is based on the proof of Lemma 6.2. It goes through all tuples $\bar{w} \in \big(N^{\mathcal{A}}_{(2r+1)\ell}(T)\big)^\ell$, and, for all $s \in [0, \ell]$, it considers the tuple consisting of the first s entries of $\bar{w}$. For these values, it checks whether the hypothesis $(\varphi, \bar{w})$ is consistent with the training sequence, where φ is the formula given in Lemma 6.2, that is, the disjunction of sphere formulas around the positive examples and the s-tuple derived from $\bar{w}$.

First, we show that every hypothesis returned by the algorithm is consistent with the training sequence. Let $(\bar{v}_i, \lambda_i) \in T$. By the construction of φ, we have $\mathcal{A} \models \varphi[\bar{v}_i, \bar{w}]$ (and thus $h^{\mathcal{A}}_{\varphi,\bar{w}}(\bar{v}_i) = 1$) if and only if there is some j with $\lambda_j = 1$ such that $\mathcal{A} \models \mathrm{sph}^{\mathcal{A}}_{r,\,\bar{v}_j\bar{w}^{\mathrm{in}}}[\bar{v}_i, \bar{w}^{\mathrm{in}}]$, or, equivalently, $\mathcal{S}^{\mathcal{A}}_r(\bar{v}_i\bar{w}^{\mathrm{in}}) \cong \mathcal{S}^{\mathcal{A}}_r(\bar{v}_j\bar{w}^{\mathrm{in}})$. If $\lambda_i = 1$, then this is trivially the case, so the hypothesis correctly classifies the tuple $\bar{v}_i$ as positive. If $\lambda_i = 0$, then the checks in lines 8–11 of the algorithm guarantee that there is no positive example with an isomorphic sphere, and hence the hypothesis correctly classifies the tuple $\bar{v}_i$ as negative. All in all, this shows that every hypothesis returned by the algorithm is consistent.

For the other direction, we assume that there is a formula $\varphi^* \in \Phi^*$ and tuples $\bar{w}^* \in \big(U(\mathcal{A})\big)^\ell$ and $\bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|}$ such that the hypothesis $h^{\mathcal{A}}_{\varphi^*,\bar{w}^*,\bar{n}^*}$ is consistent with T. Then it follows from the proof of Lemma 6.2 that there is a tuple $\bar{w}$ among the ones we check such that the resulting hypothesis is consistent with the training sequence. Thus, our algorithm returns a hypothesis in these cases.

It remains to show that the algorithm satisfies the running-time requirements while only using local access to the structure $\mathcal{A}$. Analogously to the proof of Theorem 6.5, for fixed k, ℓ, c_{br}, and c_{bw}, the size of the set N computed in line 1 is polynomial in m and d. It can be computed in time $(\log n + m + d)^{\mathcal{O}(1)}$ under the logarithmic-cost measure and in time $(m + d)^{\mathcal{O}(1)}$ under the uniform-cost measure, using only local access to the background structure. For every single choice of $\bar{w}$ and s, the size of a single sphere $\mathcal{S}_i$ is polynomial in d, and it can be computed in time polynomial in d and $\log n$ under the logarithmic-cost measure and polynomial in d under the uniform-cost measure. By Theorem 7.3, every single isomorphism test between the spheres runs in time $d^{\mathrm{polylog}\, d}$. All in all, the algorithm runs in time $(\log n + m)^{\mathcal{O}(1)} d^{\mathrm{polylog}\, d}$ under the logarithmic-cost measure and in time $m^{\mathcal{O}(1)} d^{\mathrm{polylog}\, d}$ under the uniform-cost measure, while only using local access.

To evaluate the hypothesis returned by the algorithm on a new tuple $\bar{v}$, we compute the r-sphere around $\bar{v}\bar{w}^{\mathrm{in}}$ and check whether it is isomorphic to one of the spheres used in the returned formula φ. Thus, we obtain the same running-time bounds as for the learning algorithm. □

Next, we extend this result to the ERM problem.

Theorem 7.4. *Let σ be a relational signature and let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$. There is an algorithm that solves* FOCN-Learn-ERM$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$ *in time $(\log n + m)^{\mathcal{O}(1)} \cdot d^{\mathrm{polylog}\, d}$ under the logarithmic-cost measure and in time $m^{\mathcal{O}(1)} \cdot d^{\mathrm{polylog}\, d}$ under the uniform-cost measure, where n is the size of the background structure, d is the degree of the background structure, and m is the length of the training sequence. Furthermore, the hypotheses returned by the algorithm can be evaluated with the same time bound.*

Proof. The pseudocode for our algorithm $\mathcal{A}_{\mathrm{ERM}}$ is shown in Figure 9. Let $r := (2 \cdot c_{\mathrm{bw}} + 1)^{c_{\mathrm{br}}}$. The algorithm goes through all tuples $\bar{w} \in \big(N^{\mathcal{A}}_{(2r+1)\ell}(T)\big)^\ell$. For all $s \in [0, \ell]$, it considers the tuple $\bar{w}^{\mathrm{in}}$ consisting of the first s entries of $\bar{w}$. For every sphere $\mathcal{S}_i = \mathcal{S}^{\mathcal{A}}_r(\bar{v}_i\bar{w}^{\mathrm{in}})$, the algorithm counts the number $error^+_i$ of errors the hypothesis would make on the training sequence if we would include the sphere formula for $\mathcal{S}_i$ in the hypothesis. Additionally, it also counts the number $error^-_i$ of errors the hypothesis would make on the training sequence if we would leave out the sphere formula for $\mathcal{S}_i$. The sphere formula is included in the hypothesis if $error^+_i \leq error^-_i$. For every combination of a tuple $\bar{w}$ and a number s, the algorithm sums up the number of errors the hypothesis would make on the training sequence. In the end, it returns the hypothesis with the minimum number of errors.

Require: local access to background structure $\mathcal{A}$,
training sequence $T = \big((\bar{v}_1, \lambda_1), \ldots, (\bar{v}_m, \lambda_m)\big)$

$N \leftarrow N^{\mathcal{A}}_{(2r+1)\ell}(T)$
$error_{\min} \leftarrow |T| + 1$
for all $\bar{w} = (w_1, \ldots, w_\ell) \in N^\ell$ **do**
  for all $s \in [0, \ell]$ **do**
    $error \leftarrow 0$
    $\bar{w}^{\text{in}} \leftarrow (w_1, \ldots, w_s)$
    for all $i \in [m]$ **do**
      $\mathcal{S}_i \leftarrow \mathcal{S}^{\mathcal{A}}_r(\bar{v}_i \bar{w}^{\text{in}})$
    for all $i \in [m]$ **do**
      $error^+_i \leftarrow |\{j \in [m] \mid \mathcal{S}_i \cong \mathcal{S}_j \text{ and } \lambda_j = 0\}|$
      $error^-_i \leftarrow |\{j \in [m] \mid \mathcal{S}_i \cong \mathcal{S}_j \text{ and } \lambda_j = 1\}|$
      $error \leftarrow error + \min\{error^+_i, error^-_i\}$
    if $error < error_{\min}$ **then**
      $error_{\min} \leftarrow error$
      $\bar{w}_{\min} \leftarrow \bar{w}$
      $\varphi_{\min}(\bar{x}, \bar{y}) \leftarrow \bigvee_{\substack{i \in [m], \\ error^+_i \le error^-_i}} \text{sph}^{\mathcal{A}}_{r, \bar{v}_i \bar{w}^{\text{in}}}(\bar{x}, y_1, \ldots, y_s)$
return $(\varphi_{\min}, \bar{w}_{\min})$

Figure 9: Learning algorithm $\mathcal{A}_{\text{ERM}}$ for Theorem 7.4

Claim 7.5. Let $(\varphi_{\min}, \bar{w}_{\min})$ be the hypothesis returned by the algorithm. For all formulas $\varphi^*(\bar{x}, \bar{y}, \bar{\kappa}) \in \mathsf{FOCN}[\sigma, c_{\text{br}}, c_{\text{bw}}]$ and tuples $\bar{w}^* \in (U(\mathcal{A}))^\ell$ and $\bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|}$, it holds that $\text{err}_T\big(h^{\mathcal{A}}_{\varphi_{\min}, \bar{w}_{\min}}\big) \le \text{err}_T\big(h^{\mathcal{A}}_{\varphi^*, \bar{w}^*, \bar{n}^*}\big)$.

Proof. Choose $\varphi^*(\bar{x}, \bar{y}, \bar{\kappa}) \in \mathsf{FOCN}[\sigma, c_{\text{br}}, c_{\text{bw}}]$, $\bar{w}^* \in (U(\mathcal{A}))^\ell$, and $\bar{n}^* \in \mathbb{Z}^{|\bar{\kappa}|}$ such that $\text{err}_T\big(h^{\mathcal{A}}_{\varphi^*, \bar{w}^*, \bar{n}^*}\big)$ is minimal. Let T^* be the subsequence of T that contains exactly those examples that are correctly classified by $h^* \coloneqq h^{\mathcal{A}}_{\varphi^*, \bar{w}^*, \bar{n}^*}$. It suffices to show that, for all examples $(\bar{v}_i, \lambda_i)$ in T^*, we have that $error^+_i \le error^-_i$ if $\lambda_i = 1$ and $error^+_i \ge error^-_i$ if $\lambda_i = 0$. If we had $error^+_i > error^-_i$ and $\lambda_i = 1$, then using $\varphi \coloneqq \varphi^* \wedge \neg \text{sph}^{\mathcal{A}}_{r, \bar{v}_i \bar{w}^{\text{in}}}$ would yield a hypothesis that is consistent with more examples than h^*, which contradicts the optimality of h^*. On the other hand, if we had $error^+_i < error^-_i$ and $\lambda_i = 0$, then we could use $\varphi \coloneqq \varphi^* \vee \text{sph}^{\mathcal{A}}_{r, \bar{v}_i \bar{w}^{\text{in}}}$. ⌟

The analysis of the running time of the algorithm $\mathcal{A}_{\text{ERM}}$ is analogous to the analysis of the algorithm $\mathcal{A}_{\text{con}}$ in the proof of Theorem 7.1, and it yields the same result. □

Analogously to the consistent-learning case, on classes of structures of polylogarithmic degree, Theorem 7.4 implies that the ERM problem is solvable in sublinear time.

Corollary 7.6. *Let σ be a relational signature, let $k, \ell, c_{\text{br}}, c_{\text{bw}} \in \mathbb{N}$, and let $\mathcal{C}$ be a class of structures of polylogarithmic degree. There is an algorithm that solves the problem* FOCN-Learn-ERM$(\sigma, k, \ell, c_{\text{br}}, c_{\text{bw}})$ *on $\mathcal{C}$ in time sublinear in the size of the background structure and polynomial in the length of the training sequence, under the logarithmic-cost as*

well as the uniform-cost measure. The hypotheses returned by the algorithm can be evaluated with the same bound on the running time.

To turn the algorithm $\mathcal{A}_{\mathrm{ERM}}$ into a sublinear-time PAC-learning algorithm, we want to find a sublinear bound on the number of examples needed to fulfil the probability bounds. In contrast to the approach in the last section, the formulas we use in the hypotheses do not come from a constant-sized set of formulas any more. Instead, the number of non-equivalent disjunctions of sphere formulas is exponential in the number of non-isomorphic spheres, which is again exponential in their size. This leads to the following result.

Theorem 7.7. *Let σ be a relational signature, let $k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}} \in \mathbb{N}$, let $a \coloneqq \max_{R\in\sigma} \mathrm{ar}(R)$, $r \coloneqq (2 \cdot c_{\mathrm{bw}} + 1)^{c_{\mathrm{br}}}$, and let $\mathcal{C}$ be a class of structures $\mathcal{A}$ of degree at most $\big(\log(\log|\mathcal{A}|)\big)^{\frac{1}{(r+1)\cdot a}}$. There is an algorithm that solves* FOCN-Learn-PAC$(\sigma, k, \ell, c_{\mathrm{br}}, c_{\mathrm{bw}})$ *on $\mathcal{C}$ in time sublinear in the size of the background structure and polynomial in $\log\frac{1}{\delta}$ and $\frac{1}{\varepsilon}$ under the logarithmic-cost as well as the uniform-cost measure.*

Furthermore, the hypotheses returned by the algorithm can be evaluated with the same bound on the running time.

Proof. Let $\mathcal{A} \in \mathcal{C}$ be a background structure of degree d with $d \leq \big(\log(\log|\mathcal{A}|)\big)^{\frac{1}{(r+1)\cdot a}}$. We consider the concept class

$$\mathcal{H}^* = \big\{ h^{\mathcal{A}}_{\varphi,\bar{w},\bar{n}} \mid \varphi(\bar{x},\bar{y},\bar{\kappa}) \in \Phi^*,\ \bar{w} \in \big(U(\mathcal{A})\big)^{\ell},\ \bar{n} \in \mathbb{Z}^{|\kappa|} \big\}.$$

Running on $\mathcal{A}$, the algorithm $\mathcal{A}_{\mathrm{ERM}}$ only returns formulas from the set

$$\begin{aligned}\Phi_d \coloneqq \big\{ \varphi(\bar{x},\bar{y}) \in \mathsf{FO}[\sigma] \ \mid\ & |\bar{x}| = k,\ |\bar{y}| = \ell, \\ & \varphi \text{ is a disjunction of sphere formulas} \\ & \text{of locality radius at most } r \\ & \text{based on spheres of degree at most } d \big\}.\end{aligned}$$

Thus, we consider the hypothesis class

$$\mathcal{H} = \big\{ h^{\mathcal{A}}_{\varphi,\bar{w}} \mid \varphi(\bar{x},\bar{y}) \in \Phi_d,\ \bar{w} \in \big(U(\mathcal{A})\big)^{\ell} \big\}.$$

As in the proof of Theorem 6.9, it holds that $\mathcal{H}^* \subseteq \mathcal{H}$.

Next, we bound number of non-equivalent hypotheses in $\mathcal{H}$ and thus also in $\mathcal{H}^*$. As discussed in Lemma 6.4, in a structure of degree at most d, a sphere of radius at most r with $(k+\ell)$ centres has size at most $s \coloneqq (k+\ell)\cdot\mu_d(r) \in \mathcal{O}\big(d^{r+1}\big) \subseteq \mathcal{O}\big((\log(\log(|\mathcal{A}|)))^{\frac{1}{a}}\big)$. Thus, over a signature σ, the number of non-isomorphic spheres of radius at most r with $(k+\ell)$ centres can be bounded by $\prod_{R\in\sigma} 2^{s^{\mathrm{ar}(R)}} = 2^{\left(\sum_{R\in\sigma} s^{\mathrm{ar}(R)}\right)} \leq 2^{|\sigma|\cdot s^a}$. The number of non-equivalent disjunctions of sphere formulas based on such spheres is at most exponential in the number of non-isomorphic spheres. Hence, the set Φ_d contains at most $\mathcal{O}\left(|\mathcal{A}|^{|\sigma|}\right)$ non-equivalent formulas, and the number of non-equivalent hypotheses in $\mathcal{H}$ and $\mathcal{H}^*$ is bounded by $c\cdot|\mathcal{A}|^{\ell+|\sigma|}$ for some constant c.

The remainder of this proof is analogous to the proof of Theorem 6.9. We use Lemma 6.8 to bound the number of examples needed for a PAC-learning algorithm by

$$m(|\mathcal{A}|,\varepsilon,\delta) \coloneqq \left\lceil \frac{2\log(2c\cdot|\mathcal{A}|^{\ell+|\sigma|}/\delta)}{\varepsilon^2} \right\rceil.$$

Then, it suffices to query $m(|\mathcal{A}|, \varepsilon, \delta)$ examples from the distribution $\mathcal{D}$ and run $\mathcal{A}_{\mathrm{ERM}}$ on the resulting training sequence. With the bound on the number of training examples, Theorem 7.4 yields the desired running time. □

8. Conclusion

In this paper, we have studied Boolean classification problems in the logical framework introduced by Grohe and Turán [GT04] over relational background structures. We have proved that, on the one hand, in general, hypotheses definable in first-order logic are not learnable in sublinear time. On the other hand, over classes of structures of at most polylogarithmic degree, we have given a sublinear-time consistent-learning algorithm, even for hypotheses definable in the extension FOCN of first-order logic with counting.

The extended abstract [vB19] of this paper gives an agnostic PAC-learning result for classes of structures with a fixed degree bound. The present paper proves that this result can be generalised to classes of structures where the degree is not bounded by any fixed constant but rather depends on the size of the structure. More precisely, for classes of structures of degree at most $(\log \log n)^c$ for some constant c, we have extended the consistent-learning result to agnostic PAC-learning problems.

Another question raised in [vB19] is whether similar learning results can be proved for logics that also include other means of aggregation. We have confirmed this in [vBS21], which introduces the first-order logic with weight aggregation FOWA. This logic is defined over weighted structures, which extend ordinary relational structures by assigning weights to tuples in the structure. In FOWA formulas, with concepts similar to counting terms in FOCN, these weights can be aggregated. For the fragment FOWA_1 of FOWA, [vBS21] proved sublinear-time learnability based on Gaifman-style locality results. However, for the full logic FOWA, which extends the fragment FOC of FOCN, there is no analogue of Gaifman's theorem. It remains open whether Hanf normal forms exist for FOWA and whether they can be used to obtain learnability results similar to the ones we have presented in this paper for FOCN.

References

[AAP+13] Azza Abouzied, Dana Angluin, Christos H. Papadimitriou, Joseph M. Hellerstein, and Avi Silberschatz. Learning and verifying quantified Boolean queries by example. In Richard Hull and Wenfei Fan, editors, *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA, June 22–27, 2013*, pages 49–60. ACM, 2013. doi:10.1145/2463664.2465220.

[AF23] Isolde Adler and Polly Fahey. Faster property testers in a variation of the bounded degree model. *ACM Trans. Comput. Log.*, 24(3):25:1–25:24, 2023. doi:10.1145/3584948.

[AGM13] Albert Atserias, Martin Grohe, and Dániel Marx. Size bounds and query plans for relational joins. *SIAM J. Comput.*, 42(4):1737–1767, 2013. doi:10.1137/110859440.

[AH18] Isolde Adler and Frederik Harwath. Property testing for bounded degree databases. In Rolf Niedermeier and Brigitte Vallée, editors, *35th Symposium on Theoretical Aspects of Computer Science, STACS 2018, February 28 to March 3, 2018, Caen, France*, volume 96 of *LIPIcs*, pages 6:1–6:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.STACS.2018.6.

[AHHP98] Howard Aizenstein, Tibor Hegedüs, Lisa Hellerstein, and Leonard Pitt. Complexity theoretic hardness results for query learning. *Comput. Complex.*, 7(1):19–53, 1998. doi:10.1007/PL00001593.

[Ang87] Dana Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1987. doi:10.1007/BF00116828.

[AtCKT11] Bogdan Alexe, Balder ten Cate, Phokion G. Kolaitis, and Wang-Chiew Tan. Characterizing schema mappings via data examples. *ACM Trans. Database Syst.*, 36(4):23:1–23:48, 2011. doi:10.1145/2043652.2043656.

[BBDK21] Pablo Barceló, Alexander Baumgartner, Victor Dalmau, and Benny Kimelfeld. Regularizing conjunctive features for classification. *J. Comput. Syst. Sci.*, 119:97–124, 2021. doi:10.1016/j.jcss.2021.01.003.

[BCCT19] Angela Bonifati, Ugo Comignani, Emmanuel Coquery, and Romuald Thion. Interactive mapping specification with exemplar tuples. *ACM Trans. Database Syst.*, 44(3):10:1–10:44, 2019. doi:10.1145/3321485.

[BCL15] Angela Bonifati, Radu Ciucanu, and Aurélien Lemay. Learning path queries on graph databases. In Gustavo Alonso, Floris Geerts, Lucian Popa, Pablo Barceló, Jens Teubner, Martín Ugarte, Jan Van den Bussche, and Jan Paredaens, editors, *Proceedings of the 18th International Conference on Extending Database Technology, EDBT 2015, Brussels, Belgium, March 23–27, 2015*, pages 109–120. OpenProceedings.org, 2015. doi:10.5441/002/edbt.2015.11.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989. doi:10.1145/76359.76371.

[vB19] Steffen van Bergerem. Learning concepts definable in first-order logic with counting. In *34th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2019, Vancouver, BC, Canada, June 24–27, 2019*, pages 1–13. IEEE Computer Society, 2019. doi:10.1109/LICS.2019.8785811.

[vB23] Steffen van Bergerem. *Descriptive Complexity of Learning*. PhD thesis, RWTH Aachen University, Germany, 2023. doi:10.18154/RWTH-2023-02554.

[vBGR22] Steffen van Bergerem, Martin Grohe, and Martin Ritzert. On the parameterized complexity of learning first-order logic. In Leonid Libkin and Pablo Barceló, editors, *PODS '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12–17, 2022*, pages 337–346. ACM, 2022. doi:10.1145/3517804.3524151.

[vBGR25] Steffen van Bergerem, Martin Grohe, and Nina Runde. The parameterized complexity of learning monadic second-order logic. In Jörg Endrullis and Sylvain Schmitz, editors, *33rd EACSL Annual Conference on Computer Science Logic, CSL 2025, February 10–14, 2025, Amsterdam, Netherlands*, volume 326 of *LIPIcs*, pages 8:1–8:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.CSL.2025.8.

[vBS21] Steffen van Bergerem and Nicole Schweikardt. Learning concepts described by weight aggregation logic. In Christel Baier and Jean Goubault-Larrecq, editors, *29th EACSL Annual Conference on Computer Science Logic, CSL 2021, January 25–28, 2021, Ljubljana, Slovenia (Virtual Conference)*, volume 183 of *LIPIcs*, pages 10:1–10:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPIcs.CSL.2021.10.

[BR17] Pablo Barceló and Miguel Romero. The complexity of reverse engineering problems for conjunctive queries. In Michael Benedikt and Giorgio Orsi, editors, *20th International Conference on Database Theory, ICDT 2017, March 21–24, 2017, Venice, Italy*, volume 68 of *LIPIcs*, pages 7:1–7:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPIcs.ICDT.2017.7.

[tCD22] Balder ten Cate and Víctor Dalmau. Conjunctive queries: Unique characterizations and exact learnability. *ACM Trans. Database Syst.*, 47(4):14:1–14:41, 2022. doi:10.1145/3559756.

[tCDK13] Balder ten Cate, Víctor Dalmau, and Phokion G. Kolaitis. Learning schema mappings. *ACM Trans. Database Syst.*, 38(4):28, 2013. doi:10.1145/2539032.2539035.

[tCFJL23] Balder ten Cate, Maurice Funk, Jean Christoph Jung, and Carsten Lutz. SAT-based PAC learning of description logic concepts. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th–25th August 2023, Macao, SAR, China*, pages 3347–3355. ijcai.org, 2023. doi:10.24963/ijcai.2023/373.

[tCKQT18] Balder ten Cate, Phokion G. Kolaitis, Kun Qian, and Wang-Chiew Tan. Active learning of GAV schema mappings. In Jan Van den Bussche and Marcelo Arenas, editors, *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems [PODS 2018], Houston, TX, USA, June 10–15, 2018*, pages 355–368. ACM, 2018. doi:10.1145/3196959.3196974.

[CCKS20] Adrien Champion, Tomoya Chiba, Naoki Kobayashi, and Ryosuke Sato. ICE-based refinement type discovery for higher-order functional programs. *J. Autom. Reason.*, 64(7):1393–1418, 2020. doi:10.1007/s10817-020-09571-y.

[CDEM22] Andrew Cropper, Sebastijan Dumancic, Richard Evans, and Stephen H. Muggleton. Inductive logic programming at 30. *Mach. Learn.*, 111(1):147–172, 2022. doi:10.1007/s10994-021-06089-1.

[CJ95] William W. Cohen and C. David Page Jr. Polynomial learnability and inductive logic programming: Methods and results. *New Gener. Comput.*, 13(3&4):369–409, 1995. doi:10.1007/BF03037231.

[CZB+22] Nofar Carmeli, Shai Zeevi, Christoph Berkholz, Alessio Conte, Benny Kimelfeld, and Nicole Schweikardt. Answering (unions of) conjunctive queries using random access and random-order enumeration. *ACM Trans. Database Syst.*, 47(3):9:1–9:49, 2022. doi:10.1145/3531055.

[DSS22] Arnaud Durand, Nicole Schweikardt, and Luc Segoufin. Enumerating answers to first-order queries over databases of low degree. *Log. Methods Comput. Sci.*, 18(2), 2022. doi:10.46298/lmcs-18(2:7)2022.

[EMR14] Guy Even, Moti Medina, and Dana Ron. Deterministic stateless centralized local algorithms for bounded degree graphs. In Andreas S. Schulz and Dorothea Wagner, editors, *22th Annual European Symposium on Algorithms, ESA 2014, September 8–10, 2014, Wroclaw, Poland*, volume 8737 of *Lecture Notes in Computer Science*, pages 394–405. Springer, 2014. doi:10.1007/978-3-662-44777-2_33.

[END+18] P. Ezudheen, Daniel Neider, Deepak D'Souza, Pranav Garg, and P. Madhusudan. Horn-ICE learning for synthesizing invariants and contracts. *Proc. ACM Program. Lang.*, 2(OOPSLA):131:1–131:25, 2018. doi:10.1145/3276501.

[FFG02] Jörg Flum, Markus Frick, and Martin Grohe. Query evaluation via tree-decompositions. *J. ACM*, 49(6):716–752, 2002. doi:10.1145/602220.602222.

[FG06] Jörg Flum and Martin Grohe. *Parameterized Complexity Theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2006. doi:10.1007/3-540-29953-X.

[FSV95] Ronald Fagin, Larry J. Stockmeyer, and Moshe Y. Vardi. On monadic NP vs. monadic co-NP. *Inf. Comput.*, 120(1):78–92, 1995. doi:10.1006/inco.1995.1100.

[Gai82] Haim Gaifman. On local and non-local properties. In *Proceedings of the Herbrand Symposium*, volume 107 of *Studies in Logic and the Foundations of Mathematics*, pages 105–135. North-Holland Publishing Company, 1982. doi:10.1016/S0049-237X(08)71879-2.

[GLMN14] Pranav Garg, Christof Löding, P. Madhusudan, and Daniel Neider. ICE: A robust framework for learning invariants. In Armin Biere and Roderick Bloem, editors, *26th International Conference on Computer Aided Verification, CAV 2014, Held as Part of the Vienna Summer of Logic, VSL 2014, July 18–22, 2014, Vienna, Austria*, volume 8559 of *Lecture Notes in Computer Science*, pages 69–87. Springer, 2014. doi:10.1007/978-3-319-08867-9_5.

[GLR17] Martin Grohe, Christof Löding, and Martin Ritzert. Learning MSO-definable hypotheses on strings. In Steve Hanneke and Lev Reyzin, editors, *28th International Conference on Algorithmic Learning Theory, ALT 2017, 15–17 October 2017, Kyoto University, Kyoto, Japan*, volume 76 of *Proceedings of Machine Learning Research*, pages 434–451. PMLR, 2017. URL: http://proceedings.mlr.press/v76/grohe17a.html.

[GNS23] Martin Grohe, Daniel Neuen, and Pascal Schweitzer. A faster isomorphism test for graphs of small degree. *SIAM J. Comput.*, 52(6):S18–1, 2023. doi:10.1137/19m1245293.

[GR02] Oded Goldreich and Dana Ron. Property testing in bounded degree graphs. *Algorithmica*, 32(2):302–343, 2002. doi:10.1007/s00453-001-0078-7.

[GR17] Martin Grohe and Martin Ritzert. Learning first-order definable concepts over structures of small degree. In *32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20–23, 2017*, pages 1–12. IEEE Computer Society, 2017. doi:10.1109/LICS.2017.8005080.

[GR19] Émilie Grienenberger and Martin Ritzert. Learning definable hypotheses on trees. In Pablo Barceló and Marco Calautti, editors, *22nd International Conference on Database Theory, ICDT 2019, March 26–28, 2019, Lisbon, Portugal*, volume 127 of *LIPIcs*, pages 24:1–24:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.ICDT.2019.24.

[Gro01] Martin Grohe. Generalized model-checking problems for first-order logic. In Afonso Ferreira and Horst Reichel, editors, *STACS 2001, 18th Annual Symposium on Theoretical Aspects of Computer*

Science, Dresden, Germany, February 15–17, 2001, Proceedings, volume 2010 of *Lecture Notes in Computer Science*, pages 12–26. Springer, 2001. doi:10.1007/3-540-44693-1_2.

[Gro17] Martin Grohe. *Descriptive Complexity, Canonisation, and Definable Graph Structure Theory*, volume 47 of *Lecture Notes in Logic*. Cambridge University Press, 2017. doi:10.1017/9781139028868.

[GS10] Georg Gottlob and Pierre Senellart. Schema mapping discovery from data instances. *J. ACM*, 57(2):6:1–6:37, 2010. doi:10.1145/1667053.1667055.

[GS18] Martin Grohe and Nicole Schweikardt. First-order query evaluation with cardinality conditions. In Jan Van den Bussche and Marcelo Arenas, editors, *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems [PODS 2018], Houston, TX, USA, June 10–15, 2018*, pages 253–266. ACM, 2018. doi:10.1145/3196959.3196970.

[GT04] Martin Grohe and György Turán. Learnability and definability in trees and similar structures. *Theory Comput. Syst.*, 37(1):193–220, 2004. doi:10.1007/s00224-003-1112-8.

[Han65] William Hanf. Model-theoretic methods in the study of elementary logic. In *The Theory of Models. Proceedings of the 1963 International Symposium at Berkeley*, Studies in Logic and the Foundations of Mathematics, pages 132–145. North-Holland Publishing Company, 1965. doi:10.1016/B978-0-7204-2233-7.50020-4.

[Hau89] David Haussler. Learning conjunctive concepts in structural domains. *Mach. Learn.*, 4:7–40, 1989. doi:10.1007/BF00114802.

[Hau92] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992. doi:10.1016/0890-5401(92)90010-D.

[Hir00] Kouichi Hirata. On the hardness of learning acyclic conjunctive queries. In Hiroki Arimura, Sanjay Jain, and Arun Sharma, editors, *11th International Conference on Algorithmic Learning Theory, ALT 2000, December 11–13, 2000, Sydney, Australia*, volume 1968 of *Lecture Notes in Computer Science*, pages 238–251. Springer, 2000. doi:10.1007/3-540-40992-0_18.

[KD94] Jörg-Uwe Kietz and Saso Dzeroski. Inductive logic programming and learnability. *SIGART Bull.*, 5(1):22–32, 1994. doi:10.1145/181668.181674.

[KR18] Benny Kimelfeld and Christopher Ré. A relational framework for classifier engineering. *ACM Trans. Database Syst.*, 43(3):11:1–11:36, 2018. doi:10.1145/3268931.

[KS17] Dietrich Kuske and Nicole Schweikardt. First-order logic with counting. In *32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20–23, 2017*, pages 1–12. IEEE Computer Society, 2017. doi:10.1109/LICS.2017.8005133.

[Lib04] Leonid Libkin. *Elements of Finite Model Theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2004. doi:10.1007/978-3-662-07003-1.

[LMN16] Christof Löding, P. Madhusudan, and Daniel Neider. Abstract learning frameworks for synthesis. In Marsha Chechik and Jean-François Raskin, editors, *22nd International Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS 2016, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2016, April 2–8, 2016, Eindhoven, The Netherlands*, volume 9636 of *Lecture Notes in Computer Science*, pages 167–185. Springer, 2016. doi:10.1007/978-3-662-49674-9_10.

[LRR20] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Local algorithms for sparse spanning graphs. *Algorithmica*, 82(4):747–786, 2020. doi:10.1007/s00453-019-00612-6.

[LRY17] Reut Levi, Ronitt Rubinfeld, and Anak Yodpinyanee. Local computation algorithms for graphs of non-constant degrees. *Algorithmica*, 77(4):971–994, 2017. doi:10.1007/s00453-016-0126-y.

[MR94] Stephen H. Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *J. Log. Program.*, 19/20:629–679, 1994. doi:10.1016/0743-1066(94)90035-3.

[Mug91] Stephen H. Muggleton. Inductive logic programming. *New Gener. Comput.*, 8(4):295–318, 1991. doi:10.1007/BF03037089.

[Neu19] Daniel Neuen. *The Power of Algorithmic Approaches to the Graph Isomorphism Problem*. PhD thesis, RWTH Aachen University, Germany, 2019. doi:10.18154/RWTH-2020-00160.

[RTVX11] Ronitt Rubinfeld, Gil Tamir, Shai Vardi, and Ning Xie. Fast local computation algorithms. In Bernard Chazelle, editor, *Innovations in Computer Science, ICS 2011, Tsinghua University, Beijing, China, January 7–9, 2011. Proceedings*, pages 223–238. Tsinghua University Press, 2011. URL: http://conference.iiis.tsinghua.edu.cn/ICS2011/content/papers/36.html.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi:10.1017/CBO9781107298019.

[See96] Detlef Seese. Linear time computable problems and first-order descriptions. *Math. Struct. Comput. Sci.*, 6(6):505–526, 1996. doi:10.1017/s0960129500070079.

[SST10] Robert H. Sloan, Balázs Szörényi, and György Turán. Learning Boolean functions with queries. In Yves Crama and Peter L. Hammer, editors, *Boolean Models and Methods in Mathematics, Computer Science, and Engineering*, pages 221–256. Cambridge University Press, 2010. doi:10.1017/cbo9780511780448.010.

[SW12] Slawek Staworko and Piotr Wieczorek. Learning twig and path queries. In Alin Deutsch, editor, *15th International Conference on Database Theory, ICDT 2012, March 26–29, 2012, Berlin, Germany*, pages 140–154. ACM, 2012. doi:10.1145/2274576.2274592.

[Val84] Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. doi:10.1145/1968.1972.

[Vap91] Vladimir Vapnik. Principles of risk minimization for learning theory. In John E. Moody, Stephen Jose Hanson, and Richard Lippmann, editors, *Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2–5, 1991]*, pages 831–838. Morgan Kaufmann, 1991. URL: https://proceedings.neurips.cc/paper_files/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf.

[Var82] Moshe Y. Vardi. The complexity of relational query languages (extended abstract). In Harry R. Lewis, Barbara B. Simons, Walter A. Burkhard, and Lawrence H. Landweber, editors, *Proceedings of the 14th Annual ACM Symposium on Theory of Computing [STOC 1982], May 5–7, 1982, San Francisco, California, USA*, pages 137–146. ACM, 1982. doi:10.1145/800070.802186.

[ZMJ18] He Zhu, Stephen Magill, and Suresh Jagannathan. A data-driven CHC solver. In Jeffrey S. Foster and Dan Grossman, editors, *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2018, Philadelphia, PA, USA, June 18–22, 2018*, pages 707–721. ACM, 2018. doi:10.1145/3192366.3192416.