Query Containment for Highly Expressive Datalog Fragments

Pierre Bourhis
CNRS LIFL University of Lille 1
& INRIA Lille Nord Europe
Lille, FR

Markus Krötzsch
Fakultät Informatik
Technische Universität
Dresden, DE

Sebastian Rudolph
Fakultät Informatik
Technische Universität
Dresden, DE

ABSTRACT
The containment problem of Datalog queries is well known to be undecidable. There are, however, several Datalog fragments for which containment is known to be decidable, most notably monadic Datalog and several “regular” query languages on graphs. Monadically Defined Queries (MQs) have been introduced recently as a joint generalization of these query languages.

In this paper, we study a wide range of Datalog fragments with decidable query containment and determine exact complexity results for this problem. We generalize MQs to (Frontier-)Guarded Queries (GQs), and show that the containment problem is 3ExpTime-complete in either case, even if we allow arbitrary Datalog in the sub-query. If we focus on graph query languages, i.e., fragments of linear Datalog, then this complexity is reduced to 2ExpSpace. We also consider nested queries, which gain further expressivity by using predicates that are defined by inner queries. We show that nesting leads to an exponentially increasing hierarchy for the complexity of query containment, both in the linear and in the general case. Our results settle open problems for (nested) MQs, and they paint a comprehensive picture of the state of the art in Datalog query containment.

1. INTRODUCTION
Query languages and their mutual relationships are a central topic in database research and a continued focus of intensive study. It has long been known that first-order logic expressions over the database relations (represented by extensional database predicates, EDBs) lack the expressive power needed in many scenarios. Higher-order query languages have thus been introduced, which allow for the recursive definition of new predicates (so called intensional database predicates, IDBs). Most notably, Datalog has been widely studied as a very expressive query language with tractable query answering (w.r.t. the size of the database).

On the other hand, Datalog has been shown to be too expressive a language for certain tasks which are of crucial importance in database management. In particular, the query containment problem that, given two queries \( Q_1 \) and \( Q_2 \), asks if every answer to \( Q_1 \) is an answer to \( Q_2 \) in every possible database, is undecidable for full Datalog [21]. However, checking query containment is an essential task facilitating query optimization, information integration and exchange, as well as database integrity checking. It comes handy for utilizing databases with materialized views and, as part of an offline preprocessing technique, and it may help accelerating online query answering.

This motivates the question for Datalog fragments that are still expressive enough to satisfy their purposes but exhibit decidable query containment. Moreover, once decidability is established, the precise complexity of deciding containment provides further insights. The pursuit of these issues has led to a productive and well-established line of research in database theory, which has already produced numerous results for a variety of Datalog fragments.

Non-recursive Datalog and unions of conjunctive queries. A non-recursive Datalog program does not have any (direct or indirect) recursion and it is equivalent to a union of conjunctive queries (UCQ) (and thus expressible in first-order logic). The problem of containment of a Datalog program (in the following referred to as Dlog) in a union of conjunctive queries is 2ExpTime-complete [14]. Due to the succinctness of non-recursive Datalog compared to UCQs, the problem of containment of Dlog in non-recursive Datalog is 3ExpTime-complete [14]. Some restrictions for decreasing the complexity of these problems have been considered. Containment of linear Datalog programs (LinDlog), i.e., one where rule bodies contain at most one IDB in a UCQ, is ExpSpace-complete; complexity further decreases to PSpace when the linear Datalog program is monadic (LinMDlog, see below) [13, 14].

The techniques to prove the upper bounds in these results are based on the reduction to the problem of containment of tree automata for the general case, and to the containment of word automata in the linear case.

Monadic Datalog. A monadic Datalog (MDlog) program is a program containing only unary intensional predicates. The problem of containment for MDlog is 2ExpTime complete. The upper bound is well known since the 80’s [15], while the lower bound has been established only recently [6]. Finally, the containment of Dlog in a monadic MDlog...
is also decidable. It is a straightforward application of Theorem 5.5 of [16]. So far, however, tight bounds have not been known for this result.

Guarded Datalog. Guarded Datalog (GDlog) allows the use of intensional predicates with unrestricted arities, however for each rule, the variables of the head should appear in a single extensional atom appearing in the body of the rule. While this notion of (frontier-)guarded rules is known for a while [8, 3], the first use of GDlog as a query language seems to be only recent [4]. GDlog is a proper extension of MDlog, since monadic rules can always be rewritten into guarded rules [4]. It is known that query containment for GDlog is \(2\text{ExpTime}\)-complete, a result based on the decidability of the guarded negation fixed point logic [5].

Navigational Queries. Conjunctive two-way regular path queries (C2RPQs) generalize conjunctive queries (CQs) by regular expressions over binary predicates [18, 9]. Variants of this type of queries are used, e.g., by the XPath query language for querying semi-structured XML data. Recent versions of the SPARQL 1.1 query language for RDF also support some of regular expressions that can be evaluated under a similar semantics. Intuitively, C2RPQ is a conjunction of atoms of the form \(xLy\) where \(L\) is a two-way regular expression. A pair of nodes \(\langle n_1, n_2\rangle\) is a valuation of the pair \(\langle x, y\rangle\) if and only if there exists a path between \(n_1\) and \(n_2\) matching \(L\). The containment of queries in this language was shown to be \(\text{ExpSpace-complete}\) [18, 10, 2, 17]. The containment of Dlog in C2RPQ is \(2\text{ExpTime}\)-complete [11].

Monadically Defined Queries. More recently, Monadically Defined Queries (MQs) and their nested version (MQ+s) have been introduced [19] as a proper generalization of MDlog which also captures (unions of) C2RPQs. At the same time, they are conveniently expressible both in Dlog and monadic second-order logic. Yet, as opposed to these two, MQs and MQ+s have been shown to have a decidable containment problem, but no tight bounds were known so far.

In spite of these continued efforts, the complexity of query containment is still unclear for many well-known Datalog fragments, especially for the most expressive ones. In this paper, we thus study a variety of known and new query languages in more detail. Figure 1 gives an overview of all Datalog fragments we consider, together with their respective query-answering complexities.

We provide a detailed complexity analysis of the mutual containment between queries of the aforementioned (and some new) formalisms. This analysis is fine-grained in the sense that—in the case of query formalisms that allow for nesting—precise complexities depending on the nesting depth are presented. Moreover, we consider the case where the used rules are restricted to linear Datalog.

- We introduce guarded queries (GQs) and their nested versions (GQ+s), Datalog fragments that properly generalize MQs and MQ+s, respectively, while featuring the same data and combined complexities for query answering. On the other hand, already unnested GQs subsume GDlog. We also consider the restrictions of all these queries to the linear Datalog case and observe that this drops data complexities to \(\text{NLogSpace}\) whereas it does not affect combined complexities.
- By means of sophisticated automata-based techniques involving iterated transformations on alternating two-way automata, we show a generic upper bound stating that containment of Dlog in nested guarded queries

---

\(^1\)We thank Michael Benedikt for this observation.
of depth $k$ ($\text{GQ}^k$) can be decided in $(k + 2)\text{ExpTime}$. Additionally we show that going down to $\text{GDlog}$ on the containment’s right-hand side allows deciding it in $2\text{ExpTime}$.

- Inductively defining alternating Turing machine simulations on tapes of $(k + 1)$-exponential size, we provide a matching generic lower bound by showing that containment of $\text{MDlog}$ in $\text{MQ}^k$ is $(k + 2)\text{ExpTime}$-hard. Together with the upper bound, this provides precise complexities for all cases, where the left-hand side of the containment is any fragment between $\text{MDlog}$ and $\text{Dlog}$ (cf. Fig. 1) and the right-hand side is any of $\text{MQ}$, $\text{GQ}$, $\text{MQ}^k$, $\text{GQ}^k$, $\text{GQ}^+$. In particular, this solves the respective open questions from [19]: $\text{MQ}$ containment is $3\text{ExpTime}$-complete and $\text{MQ}^+$ containment is $\text{NonElementary}$.

- We next investigate the situation in case only linear rules are allowed in the definition of the Datalog fragment used on the left hand side of the containment problem (this distinction generally makes no difference for the right-hand side). We find that in most of these cases, the complexities mentioned above drop to $(k + 1)\text{ExpSpace}$.

In summary, our results settle open problems for (nested) MQs, and they paint a comprehensive and detailed picture of the state of the art in Datalog query containment.

2. PRELIMINARIES

We consider a standard language of first-order predicate logic, based on an infinite set $C$ of constant symbols, an infinite set $P$ of predicate symbols, and an infinite set $V$ of first-order variables. Each predicate $P \in P$ is associated with a natural number $\text{ar}(P)$ called the arity of $P$. The list of predicates and constants forms the language’s signature $\mathcal{F} = (P, C)$. We generally assume $\mathcal{F} = (P, C)$ to be fixed, and only refer to it explicitly if needed.

Formulae, Rules, and Queries. A term is a variable $x \in V$ or a constant $c \in C$. We use symbols $s, t$ to denote terms, $x, y, z, v, w$ to denote variables, $a, b, c$ to denote constants. Expressions like $t, x, c$ denote finite lists of such entities. We use the standard predicate logic definitions of atom and formula, using symbols $\varphi, \psi$ for the latter.

Datalog queries are defined over an extended signature with additional predicate symbols, called IDB predicates; all other predicates are called EDB predicates. A Datalog rule is a formula of the form $\forall x, y. \varphi[x, y] \rightarrow \psi[x]$ where $\varphi$ and $\psi$ are conjunctions of atoms, called the body and head of the rule, respectively, and where $\psi$ only contains IDB predicates. We usually omit universal quantifiers when writing rules. Sets of Datalog rules will be denoted by symbols $P, R, S$. A set of Datalog rules $P$ is

- **monadic** if all IDB predicates are of arity one;
- **frontier-guarded** if the body of every rule contains an atom $p(t)$ such that $p$ is an EDB predicate and $t$ contains all variables that occur in the rule’s head;
- **linear** if every rule contains at most one IDB predicate in its body.

A conjunctive query (CQ) is a formula $Q[x] = \exists y. \psi[x, y]$ where $\psi[x, y]$ is a conjunction of atoms; a union of conjunctive queries (UCQ) is a disjunction of such formulae. A Datalog query $(P, Q)$ consists of a set of Datalog rules $P$ and a conjunctive query $Q$ over IDB or EDB predicates ($Q$ could be expressed as a rule in Datalog, but not in all restrictions of Datalog we consider). We write $\text{Dlog}$ for the language of Datalog queries. A monadic Datalog query is one where $P$ is monadic, and similarly for other restrictions. We use the query languages $\text{MDlog}$ (monadic), $\text{GDlog}$ (frontier-guarded), $\text{LinDlog}$ (linear), and $\text{LinMDlog}$ (linear, monadic).

**Databases and Semantics.** We use the standard semantics of first-order logic (FOL). A database instance $I$ consists of a set $\Delta^I$ called domain and a function $\delta^I$ that maps constants $c$ to domain elements $c^I \in \Delta^I$ and predicate symbols $P$ to relations $P^I \subseteq (\Delta^I)^{\text{ar}(P)}$, where $P^I$ is the extension of $P$.

Given a database instance $I$ and a formula $\varphi[x]$ with free variables $x = (x_1, \ldots, x_m)$, the extension of $\varphi[x]$ is the subset of $(\Delta^I)^m$ containing all those tuples $(\delta_1, \ldots, \delta_m)$ for which $I, \{x_i \mapsto \delta_i \mid 1 \leq i \leq m\} \models \varphi[x]$. We denote this by $(\delta_1, \ldots, \delta_m) \in \varphi^I$ or by $I \models \varphi[\delta_1, \ldots, \delta_m]$; a similar notation is used for all other types of query languages. Two formulae $\varphi[x]$ and $\psi[x]$ are called equivalent if their extensions coincide for every database instance $I$.

The set of answers of a UCQ $Q[x]$ over $I$ is its extension. The set of answers of a Datalog query $(P, Q)$ over $I$ is the intersection of the extensions of $Q$ over all extended database instances $I'$ that interpret IDB predicates in such a way that all rules of $P$ are satisfied. Datalog [1] can also be defined as the least fixpoint on the inflationary evaluation of $Q$ on $I$.

Note that we do not require database instances to have a finite domain, since all of our results are valid in either case. This is due to the fact that every entailment of a Datalog program has a finite witness, and that all of our query languages are positive, i.e., that their answers are preserved under homomorphisms of database instances.

3. GUARDED QUERIES

Monadically defined queries have been introduced in [19] as a generalization of monadic Datalog ($\text{MDlog}$) and conjunctive two-way regular path queries (C2RPQs) for which query containment is still decidable. The underlying idea of this approach is that candidate query answers are checked by evaluating a monadic Datalog program, i.e., in contrast to the usual evaluation of Datalog queries, we start with a “guessed” answer that is the input to a Datalog program.\footnote{The queries were called MODEQ in [19]; we shorten this to MQ.}
To implement this, the candidate answer is represented by special constants \( \lambda \) that the Datalog program can refer to. This mechanism was called \textit{flag & check}, since the special constants act as flags to indicate the answer that should be checked.

**Example 1.** A query that computes the transitive closure over a relation \( p \) can be defined as follows.

\[
p(A_1, y) \rightarrow U(y) \\
U(y) \land p(y, z) \rightarrow U(z) \\
U(I_2) \rightarrow \text{hit}
\]

One defines the answer of the query to contain all pairs \( \langle \delta_1, \delta_2 \rangle \) for which the rules entail \( \text{hit} \) when interpreting \( \lambda_1 \) as \( \delta_1 \) and \( \lambda_2 \) as \( \delta_2 \).

The approach used monadic Datalog for its close relationship to monadic second-order logic, which was the basis for showing decidability of query containment. In this work, however, we develop new techniques for showing the decidability (and exact complexity) of this problem directly. It is therefore suggestive to consider other types of Datalog programs to implement the "check" part. The following definition therefore introduces the general technique for arbitrary Datalog programs, and defines interesting fragments by imposing further restrictions.

**Definition 1.** Consider a signature \( \mathcal{S} \). An FCP ("flag & check program") of arity \( m \) is a set of Datalog rules \( P \) with \( \forall i \geq 0 \) IDB predicates \( U_1, \ldots, U_k \) that may use the additional constant symbols \( \lambda_1, \ldots, \lambda_m \notin \mathcal{S} \) and an additional nullary predicate symbol \( \text{hit} \). An FCQ ("flag & check query") \( P \) is of the form \( \exists y.P(z) \), where \( P \) is an FCP of arity \( |z| \) and all variables in \( y \) occur in \( z \). The variables \( x \) that occur in \( z \) but not in \( y \) are the free variables of \( P \).

Let \( I \) be a database instance over \( \mathcal{S} \). The extension \( P^I \) of \( P \) is the set of all tuples \( \langle \delta_1, \ldots, \delta_m \rangle \in (\Lambda^I)^m \) such that every database instance \( I' \) that extends \( I \) to the signature of \( P \) and that satisfies \( \langle \lambda_1^I, \ldots, \lambda_m^I \rangle = \langle \delta_1, \ldots, \delta_m \rangle \) also entails \( \text{hit} \). The semantics of FCQs is defined in the obvious way based on the extension of FCPs.

A GQ is an FCQ \( \exists y.P(z) \) such that \( P \) is frontier-guarded. Similarly, we define MQ (monadic), LinMQ (linear, monadic), and LinGQ (linear, frontier-guarded) queries.

In contrast to [19], we do not define monadic queries as conjunctive queries of FCPs, but we merely allow existential quantification to project some of the FCP variables. Proposition 2 below shows that this does not reduce expressiveness.

We generally consider monadic Datalog as a special case of frontier-guarded Datalog. Monadic Datalog rules do not have to be frontier-guarded. A direct way to obtain a suitable guard is to assume that there is a unary domain predicate that contains all (relevant) elements of the domain of the database instance. However, it already suffices to require safety of Datalog rules, i.e., that the variable in the head of a rule must also occur in the body. Then every element that is inferred to belong to an IDB relation must also occur in some EDB relation. We can therefore add single EDB guard atoms to each rule in all possible ways without modifying the semantics. This is a polynomial operation, since all variables in the guards are fresh, other than the single head variable that we want to guard. We therefore find, in particular, the GQ captures the expressiveness of MQ. The converse is not true, as the following example illustrates.

**Example 2.** The following 4-ary LinGQ generalizes Example 1 by checking for the existence of two parallel \( p \)-chains of arbitrary length, where each pair of elements along the chains is connected by a relation \( q \), like the steps of a ladder.

\[
q(A_1, A_2) \rightarrow U_q(A_1, A_2) \\
U_q(x, y) \land p(x, x') \land p(y, y'), q(x', y') \rightarrow U_q(x', y') \\
U_q(A_1, A_2) \rightarrow \text{hit}
\]

One might assume that the following MQ is equivalent:

\[
q(A_1, A_2) \rightarrow U_l(A_1) \\
q(A_1, A_2) \rightarrow U_l(A_2) \\
U_l(x) \land U_l(y) \land p(x, x') \land p(y, y'), q(x', y') \rightarrow U_l(x') \\
U_l(x) \land U_l(y) \land p(x, x') \land p(y, y'), q(x', y') \rightarrow U_l(y') \\
U_l(A_1) \land U_l(A_2) \rightarrow \text{hit}
\]

However, the latter query also matches structures that are not ladders. For example, the following database yields the answer \( (a, b, c, d) \), although there is no corresponding ladder structure: \( q(a, b), p(a, c), p(b, e), q(c, e), p(a, e'), p(b, d), q(e', d) \). One can extend the MQ to avoid this case, but any such fix is "local" in the sense that a sufficiently large ladder-like structure can trick the query.

It has been shown that monadically defined queries can be expressed both in Datalog and in monadic second-order logic [19]. While we lose the connection to monadic second-order logic with GQs, the expressibility in Datalog remains. The encoding is based on the intuition that the choice of the candidate answers for \( \lambda \) "contextualizes" the inferences of the Datalog program. To express this without special constants, we can store this context information in predicates of suitably increased arity.

**Example 3.** The 4-ary LinGQ of Example 2 can be expressed with the following Datalog query. For brevity, let \( y \) be the variable list \( \langle y_1, y_2, y_3, y_4 \rangle \), which provides the context for the IDB facts we derive.

\[
q(y_1, y_2) \rightarrow U_q(y_1, y_2, y) \\
U_q(x, y, y) \land p(x, x') \land p(y, y'), q(x', y') \rightarrow U_q(x', y', y) \\
U_q(y_3, y_4, y) \rightarrow \text{goal}(y)
\]

This result is obtained by a straightforward extension of the translation algorithm for MQs [19], which may not produce
the most concise representation. Also note that the first rule in this program is not safe, since y_3 and y_5 occur in the head but not in the body. According to the semantics we defined, such variables can be bound to any element in the active domain of the given database instance (i.e., they behave as if bound by a unary domain predicate).

This observation justifies that we consider MQs, GQs, etc. as Datalog fragments. It is worth noting that the translation does not change the number of IDB predicates in the body of rules, and thus preserves linearity. The relation to (linear) Datalog also yields some complexity results for query answering; we will discuss these at the end of the next section, after introducing nested variants of our query languages.

4. NESTED QUERIES

Every query language gives rise to a nested language, where we allow nested queries to be used as if they were predicates. Sometimes, this does not lead to a new query language (like for CQ and Dlog), but often it affects complexities and/or expressiveness. It has been shown that both are increased when moving from MQs to their nested variants [19]. We will see that nesting also has strong effects on the complexity of query containment.

**Definition 2.** We define k-nested FCPs inductively. A 1-nested FCP is an FCP. A k + 1-nested FCP is an FCP that may use k-nested FCPs of arity m instead of predicate symbols of arity m in rule bodies. The semantics of nested FCPs is immediate based on the extension of FCPs. A k-nested FCQ is of the form \( \exists y. Q(z) \), where \( Q \) is a k-nested FCP of arity |z| and all variables in y occur in z.

A k-nested GQ query is a k-nested frontier-guarded FCQ. For the definition of frontier-guarded, we still require EDB predicates in guards: subqueries cannot be guards. The language of k-nested GQ queries is denoted GQ^k; the language of arbitrarily nested GQ queries is denoted GQ^\infty. Similarly, we define languages MQ^k and MQ^\infty (monadic), LinMQ^k and LinMQ^\infty (linear, monadic), and LinGQ^k and LinGQ^\infty (linear, frontier-guarded).

Note that nested queries can use the same additional symbols (predicates and constants); this does not lead to any semantic interactions, however, as the interpretation of the special symbols is “private” to each query. To simplify notation, we assume that distinct (sub)queries always contain distinct special symbols. The relationships of the query languages we introduced here are summarized in Figure 1, where upwards links denote increased expressiveness. An interesting observation that is represented in this figure is that linear Datalog is closed under nesting:

**Theorem 1.** LinDlog = LinDlog^+.

Another kind of nesting that does not add expressiveness is the nesting of FCQs in UCQs. Indeed, it turns out that (nested) FCQs can internalize arbitrary conjunctions and disjunctions of FCQs (of the same nesting level). This even holds when restricting to linear rules.

**Proposition 2.** Let \( P \) be a positive query, i.e., a Boolean expression of disjunctions and conjunctions, of LinMQ^k queries with \( k \geq 1 \). Then there is a LinMQ^k query \( P' \) of size polynomial in \( P \) that is equivalent to \( P \). Analogous results hold when replacing LinMQ^k by MQ^k, GQ^k, or LinMQ^\infty queries.

Query answering for MQs has been shown to be NP-complete (combined complexity) and P-complete (data complexity). For MQ^+, the combined complexity increases to PSPACE while the data complexity remains the same. These results can be extended to frontier-guarded queries. We also note the query complexity for frontier-guarded Datalog, for which we are not aware of any published result.

**Theorem 3.** The combined complexity of evaluating GQ queries over a database instance is NP-complete. The same holds for GDlog queries. The combined complexity of evaluating GQ^+ queries is PSpace-complete. The data complexity is P-complete for GDlog, GQ, and GQ^+.

The lower bounds in the previous case are immediate from known results for monadically defined queries. In particular, the hardness proof for nested MQs also shows that queries of a particular fixed nesting level can encode the validity problem for quantified boolean formulae with a certain number of quantifier alternations; this explains why we show the combined complexity of MQ^k to be in the Polynomial Hierarchy in Figure 1. A modification of this hardness proof from [19] allows us to obtain the same results for the combined complexities in the linear cases; matching upper bounds follow from Theorem 3.

**Theorem 4.** The combined complexity of evaluating LinMQ queries over a database instance is NP-complete. The same holds for LinGDlog and LinGQ. The combined complexity of evaluating LinMQ^+ queries is PSpace-complete. The same holds for LinGQ^+. The data complexity is NLogSpace-complete for all of these query languages.

5. DECIDING QUERY CONTAINMENT WITH AUTOMATA

We first recall a general technique of reducing query containment to the containment problem for (tree) automata [14], which we build our proofs on. An introduction to tree automata is included in the appendix.

A common way to describe the answers of a Dlog query \( P = (P, p) \) is to consider its expansion trees. Intuitively speaking, the goal atom \( p(x) \) can be rewritten by applying rules of \( P \) in a backward-chaining manner until all IDB predicates have been eliminated, resulting in a CQ. The answers
of \( P \) coincide with the (infinite) union of answers to the CQs obtained in this fashion. The rewriting itself gives rise to a tree structure, where each node is labeled by the instance of the rule that was used in the rewriting, and the leaves are instances of rules that contain only EDB predicates in their body. The set of all expansion trees provides a regular description of \( P \) that we exploit to decide containment.

To formalize this approach, we describe the set of all expansion trees as a tree language, i.e., as a set of trees with node labels from a finite alphabet. The number of possible labels of nodes in expansion trees is unbounded, since rules are instantiated using fresh variables. To obtain a finite alphabet of labels, one limits the number of variables and thus the overall number of possible rule instantiations [14].

**Definition 3.** Given a Datalog query \( P = \langle \mathcal{P}, p \rangle \), \( \mathcal{R}_P \) is the set of all instantiations of rules of \( \mathcal{P} \) using only the variables \( \mathcal{V}_P = \{v_1, \ldots, v_n\} \), where \( n \) is twice the maximal number of variables occurring in any rule of \( \mathcal{P} \).

A proof tree for \( P \) is a tree with labels from \( \mathcal{R}_P \), such that

(a) the root is labeled by a rule with \( p \) as its head predicate;
(b) if a node is labeled by a rule \( p \) with an EDB atom \( B \) in its body, then it has a child node that is labeled by \( \rho' \) with head atom \( B \). The label of a node \( e \) is denoted \( \pi(e) \).

Consider two nodes \( e_1 \) and \( e_2 \) in a proof tree with lowest common ancestor \( e \). Two occurrences of a variable \( v \) in \( \pi(e_1) \) and \( \pi(e_2) \) are connected if \( v \) occurs in the head of \( \pi(f) \) for all nodes \( f \) on the shortest path between \( e_1 \) and \( e_2 \), with the possible exception of \( e \).

A proof tree encodes an expansion tree where we replace every set of mutually connected variable occurrences by a fresh variable. Conversely, every expansion tree is represented by a proof tree that replaces fresh body variables by variables that do not occur in the head; this is always possible since proof trees can use twice as many variables as any rule of \( \mathcal{P} \). The set of proof trees is a regular tree language that can be described by an automaton.

**Proposition 5.** (Proposition 5.9 [14].) For a Datalog query \( P = \langle \mathcal{P}, p \rangle \), there is a tree automaton \( \mathcal{A}_P \) of size exponential in \( P \) that accepts exactly the set of all proof trees of \( P \).

In order to use \( \mathcal{A}_P \) to decide containment of \( P \) in another query \( P' \), we construct an automaton \( \mathcal{A}_{P \subseteq P'} \) that accepts all proof trees of \( P \) that are “matched” by \( P' \). Indeed, every proof tree induces a witness, i.e., a minimal matching database instance, and one can check whether or not \( P' \) can produce the same query answer on this instance. If this is the case for all proof trees of \( P \), then containment is shown.

### 6. DECIDING GUARDED QUERY CONTAINMENT

Our first result provides the upper bound for deciding containment of GQ queries. In fact, the result extends to arbitrary Datalog queries on the left-hand side.

**Theorem 6.** Containment of Datalog queries in GQ queries can be decided in \( 3\text{ExpTime} \).

To prove this, we need to construct the tree automaton \( \mathcal{A}_{P \subseteq P'} \) for an arbitrary GQ \( P' \). As a first step, we construct an alternating 2-way tree automaton \( \mathcal{A}_{P \subseteq P'}^* \) that accepts the proof trees that we would like \( \mathcal{A}_{P \subseteq P'} \) to accept, but with nodes additionally being annotated with information about the choice of \( \lambda \) values to guide the verification.

We first construct automata to verify the match of a single, non-rerecursive rule that may refer to \( \lambda \) constants. The rule does not have to be monadic or frontier-guarded. Our construction is inspired by a similar construction for CQs by Chaudhuri and Vardi [14], with the main difference that the answer variables in our case are not taken from the root of the tree but rather from one arbitrary node that is marked accordingly.

To define this formally, we introduce trees with additional annotations besides their node labels. Clearly, such trees can be viewed as regular labelled trees by considering annotations to be components of one label; our approach, however, leads to a more readable presentation.

**Definition 4.** Consider a Datalog program \( \mathcal{P} \), a rule \( \rho = \varphi \rightarrow p(x) \), and \( n \geq 0 \) special constants \( \lambda_1, \ldots, \lambda_n \). The proof-tree variables \( \mathcal{V}_\mathcal{P} \) used in \( \mathcal{R}_\mathcal{P} \) are as in Definition 3.

A proof tree for \( \mathcal{P} \) is \( \lambda \)-annotated if every node has an additional \( \lambda \)-label that is a partial mapping \( \{\lambda_1, \ldots, \lambda_n\} \rightarrow \mathcal{V}_\mathcal{P} \), such that: every special constant \( \lambda_i \) occurs in at least one \( \lambda \)-label, and whenever a constant \( \lambda_i \) occurs in two \( \lambda \)-labels, it is mapped to the same variable and both variable occurrences are connected.

A proof tree for \( \mathcal{P} \) is \( p \)-annotated if exactly one node has an additional \( p \)-label of the form \( p(v) \), where \( v \) is a list of variables from \( \mathcal{V}_\mathcal{P} \).

A matching tree \( T \) for \( \rho \) and \( \mathcal{P} \) is a \( \lambda \)-annotated and \( p \)-annotated proof tree for \( \mathcal{P} \) for which there is a mapping \( \nu : \text{Var}(\rho) \cup \{\lambda_1, \ldots, \lambda_n\} \rightarrow \mathcal{V}_\mathcal{P} \) such that

(a) \( (p(x)) = p(v) \);
(b) for every atom \( \alpha \) of \( \varphi \), there is a node \( e_\alpha \) in \( T \) such that the rule instance that \( e_\alpha \) is labeled with contains the EDB atom \( \nu(\alpha) \) in its body;
(c) if \( \lambda_i \) occurs in \( \alpha \), then the \( \lambda \)-label maps \( \lambda_i \) to the occurrence of \( \nu(\lambda_i) \) in \( e_\alpha \);
(d) if \( \alpha, \alpha' \in \varphi \) share a variable \( x \), then the occurrences of \( \nu(x) \) in \( e_\alpha \) and \( e_{\alpha'} \) are connected.

**Proposition 7.** There is an automaton \( \mathcal{A}_{P \subseteq P} \) that accepts exactly the annotated matching trees for \( \rho \) and \( \mathcal{P} \), and which is exponential in the size of \( \rho \) and \( \mathcal{P} \).

We want to use the automata \( \mathcal{A}_{P \subseteq P} \) to verify the entailment of a single rule within a Datalog derivation. We would like an automaton to check whether a whole derivation is possible. Unfortunately, we cannot check these derivations using
automata of the form $A_P$, which each need to be run on
a $p$-annotated tree which has the unique entailment of the
rule marked. The length of a derivation is unbounded, and
we would not be able to distinguish an unbounded amount
of $p$-markers. To overcome this problem, we create a modi-
fied automaton $A_{p,p}$ that simulates the behavior of $A_P$ on
a tree with annotation $p(v)$. For $A_{p,p}$, to know which node
the annotation $p(v)$ refers to, it has to be started at this node.
This is a non-standard notion of run, where we do not start at
the root of the tree. Moreover, starting in the middle of the
tree makes it necessary to consider both nodes below and
above the current position, and $A_{p,p}$ therefore needs to be
an alternating 2-way tree automaton.

**Proposition 8.** There is an alternating 2-way tree au-
tomaton $A_{p,p}$ that is polynomial in the size of $A_P$ such
that, whenever $A_{p,p}$ accepts a matching tree $T$ that has the
$p$-annotation $p(v)$ on node $v$, then $A_{p,p}$ has an accepting
run that starts from the corresponding node $v'$ on the tree $T'$
that is obtained by removing the $p$-annotation from $T$.

Using the automata $A_{p,p}$, we can now obtain the claimed
alternating 2-way automaton $A_{p,p}$ for a GQ $P'$. Intuitively
speaking, $A_{p,p}$ concatenates the automata $A_{p,p}$ using
alternation: whenever a derivation requires a (recursive) IDB
atom, a suitable process $A_{p,p}$ is initiated, starting from a
node in the middle of the tree. The construction relies
on guardedness, which ensures that we can always find a
suitable start node (corresponding to the node that was $p-$
annotated earlier), by finding a suitable guard EDB atom in
the tree.

**Proposition 9.** For a Dlog query $P$ and a GQ query $P'$
with special constants $\lambda$, there is an alternating 2-way
automaton $A_{p,p}$ of exponential size that accepts the $\lambda-$
annotated proof trees of $P$ that encode expansion trees with
$\lambda$ assignments for which $P'$ has a match.

We are now ready to prove Theorem 6. The automaton
$A_{p,p}$ allows us to check the answers of $P'$ on a proof tree
that is $\lambda$-annotated to assign values for answer constants.
We can transform this alternating 2-way automaton into a
tree automaton $A_{p,p}$ that is exponentially larger, i.e.,
doubly exponential in the size of the input. To remove the need
for $\lambda$-labels, we modify the automaton $A_{p,p}$ so that it can
only perform a transition from its start state if it finds that the
constants in $\lambda$ are assigned to the answer variables of $P$ in
the root. Finally, we obtain $A_{p,p}$ by projecting to the alphabet
$\mathcal{R}$ without $\lambda$-annotations; this is again possible in poly-
nomial effort. The containment problem $P \subseteq P'$ is equivalent
by deciding the containment of $A_P$ in $A_{p,p}$, which is possi-
ble in exponential time w.r.t. to the size of the automata.
Since $A_P$ is exponential and $A_{p,p}$ is double exponential,
we obtain the claimed triple exponential bound.

Our proof of Theorem 6 can be used to obtain another
interesting result for the case of frontier-guarded Datalog. If $P$
is a GDL query, which does not use any special constants
$\lambda$, then the $\lambda$-annotations are not relevant and $A_{p,p}$ can
be constructed as an alternating 2-way automaton on proof
trees. For this, we merely need to modify the construction in
Proposition 9 to start in start states of automata for rules that
entail the goal predicate of $P'$ with the expected binding of
variables to answer variables of $P$. We can then omit the pro-
jection step, which required us to convert $A_{p,p}$ into a tree
automaton earlier. Instead, we can construct from $A_{p,p}$ a
complement tree automaton $A_{p,p}$ that is only exponentially
larger than $A_{p,p}$, i.e., doubly exponential overall [15][The-
orem A.1]. Containment can then be checked by checking the
non-emptiness of $A_P \cap A_{p,p}$, which is possible in poly-
nomial time, leading to a 2ExpTim algorithm overall.

**Theorem 10.** Containment of Dlog queries in gDlog
queries can be decided in 2ExpTim.

This generalizes an earlier result of Cosmadakis et al. for

Finally, we can lift our results to the case of nested
queries. Using Proposition 2, we can make the simplifying
assumption that rules with some nested query in their body
contain only one nested query and a guard atom as the only
other atom. Thus all rules with nested queries have the form
$g(s) \land Q(t) \rightarrow p(u)$, where $g$ is an EDB predicate, $Q$
is a nested query, and the variables $u$ occur in $s$.

In Proposition 8, we constructed alternating 2-way au-
tomata $A_{p,p}$ that can check the entailment of a particular
atom $p(v)$ starting from a node within the tree. Analogously,
we now construct automata $A_{p,p}$ that check that the nested
query $Q$ matches partially, where $\theta$ is a substitution that in-
terprets query variables in terms of proof-tree variables on
the current node of the tree. Only the variables that occur
in $Q(s)$ and $Q(t)$ are mapped by $\theta$; the remaining variables
can be interpreted arbitrarily, possibly in distant parts of the
proof tree.

To construct $A_{p_0}$, we use the alternating 2-way automa-
ton $A_{p_0}$ constructed in Proposition 9 (assuming, for a
start, that $Q$ is not nested). This automaton is extended to
an alternating 2-way automaton $A_{p_0}$ that accepts trees with
a unique annotation of the form $(Q, \theta)$, for which we check
that it is consistent with the $\lambda$-annotation (i.e., for each query
variable $x$ mapped by $\theta$, the corresponding constant $\lambda$ is
assigned to $\theta(x)$ at the node that is annotated with $(Q, \theta)$).
We then obtain a (top-down) tree automaton $A_{p_0}$ by transfor-
mation $A_{p_0}$ into a tree automaton (exponential), and projecting
away the $\lambda$-annotations (polynomial). The automaton $A_{p_0}$
is analogous to the tree automaton $A_{p_0}$ of Proposition 7. Us-
ning the same transformation as in Proposition 8, we obtain
an alternating 2-way automaton $A_{p_0}$ for each $\theta$.

The automaton $A_{p_0}$ is constructed for a nested query $P'$ as
constructed in Proposition 9, but using the automata $A_{p_0}$ instead
of automata $A_{p_0}$ to check the entailment of a subquery $Q$.

The size of $A_{p_0}$ is increased by one exponential since the
size of $A_{p_0}$ is exponentially increased when projecting out
λ-labels for Q. Applying this construction inductively, we obtain the following result.

**Theorem 11.** Containment of DLog queries in GQ^k queries can be decided in (k + 2)^\text{ExpTime}.

7. **SIMULATING ALTERNATING TURING MACHINES**

To show the hardness of query containment problems, we generally provide direct encodings of Alternating Turing Machines (ATMs) with a fixed space bound [12]. To simplify this encoding, we assume without loss of generality that every universal ATM configuration leads to exactly two successor configurations. The following definition defines ATM encodings formally. Rather than requiring concrete structures to encode ATMs, we abstract the encoding by means of queries that find suitable structures in a database instance; this allows us to apply the same definition for increasingly complex encodings. The following definition is illustrated in Figure 2.

**Definition 5.** Consider an ATM M = (Q, \Sigma, \Delta, q_0, q_f) and queries FirstConf[x, y], NextConf[x, y] for all \delta \in \Delta, LastConf[x, y]. State_q[x] for all q \in Q, Head[x, y], ConfCell[x, y], FirstCell[x, y], NextCell[x, y], LastCell[x, y], and Symbol[x, y]. To refer to tape symbols, we consider constants c_\sigma for all \sigma \in \Sigma, and to refer to positions of the head, we use constants h (here), l (left), and r (right).

With respect to these queries, an element c \in dom(I) in a database instance I encodes an M quasi-configuration of size s if I contains a structure

\[
\begin{align*}
\text{State}_q(c), & \text{FirstCell}(c, d_1), \\
\text{ConfCell}(c, d_1), & \text{Symbol}(d_1, c_2), \text{Head}(d_1, p_1), \text{NextCell}(d_1, d_2), \\
\text{ConfCell}(c, d_2), & \text{Symbol}(d_2, c_2), \text{Head}(d_2, p_2), \ldots, \text{NextCell}(d_{n-1}, d_n), \\
\text{ConfCell}(c, d_n), & \text{Symbol}(d_n, c_2), \text{Head}(d_n, p_1), \text{LastCell}(d_n),
\end{align*}
\]

where q \in Q, \sigma \in \Sigma, and p_i \in \{h, l, r\}. We say that c encodes an M configuration of size s if, in addition, the sequence (p_i)_{i=1}^{n} has the form l, ..., l, h, r, ..., r with zero or more occurrences of r and l, respectively.

An element c in I encodes a (quasi-)configuration tree of M in space s if

- I \models \text{FirstConf}(c, d_1) for some d_1,
- d_1 is the root of a tree with edges defined by NextConf,
- every node in this tree encodes an M (quasi-)configuration of size s,
- if there is a transition I \models \text{NextConf}_{\delta_1}(e, e_1), where \delta_1 = (q, \sigma, q', \sigma', d) and q is a universal state, then there is also a transition I \models \text{NextConf}_{\delta_2}(e, e_2) with \delta_1 \neq \delta_2.
- if e is a leaf node, then I \models \text{LastConf}(c).

If the tree is an accepting run, then c encodes an accepting run (of M in space s).

A same-cell query is a query SameCell[x, y] such that, if c_1, c_2 \in dom(I) encode two quasi-configurations, and d_1, d_2 \in dom(I) represent the same tape cell in the encodings c_1 and c_2, respectively, then (d_1, d_2) \in \text{SameCell}^p.

Two queries P_1[x] and P_2[x] containment-encode accepting runs of M in space s if, for every database instance I and element c \in P_1[x], c encodes an accepting run of M in space s, and every accepting run of M in space s is encoded by some c \in P_2[x] for some I.

Note that elements c may encode more than one configuration (or configuration tree). This is not a problem in our arguments.

The conditions that ensure that a quasi-configuration tree is an accepting run can be expressed by a query, based on the queries given in Definition 5. More specifically, one can construct a query that accepts all elements that encode a quasi-configuration sequence that is not a run. Together with a query that accepts only encodings of quasi-configurations, this allows us to containment-encode accepting runs of an ATM. Only linear queries, possibly nested, will be needed to perform the required checks, even in the case of ATMs.

**Theorem 13.** Deciding containment of MDLog queries in MQ^k queries is hard for (k + 2)^\text{ExpTime}.

Note that the statement includes the 3ExpTime-hardness for containment of MQs as a special case. To prove this result, we first construct an ExpSpace ATM that we then use to construct tapes of double exponential size.
Lemma 14. For any ATM $M$, there is an MDlog query $P_1[x], a \text{LinMQ}\ P_2[x]$, queries as in Definition 5 that are LinMQs, and a same-cell query that is a UCQ, such that $P_1[x]$ and $P_2[x]$ containment-encode accepting runs of $M$ in exponential space.

Figure 3 illustrates the encoding that we use to prove Lemma 14. While it resembles the structure of Figure 2, the labels are now EDB predicates rather than (abstract) queries. The encoding of tapes attaches to each cell an $\ell$-length predicate that counts from 0 to $2^\ell$. The query on the left-hand side of the containment then checks that there are cells with bit addresses, not that they actually contain, even the exact length of the tape is unspecified. The query on the right-hand side of the containment then checks that consecutive cells (in all tapes that occur in the configuration tree) represent successor addresses, and that the first and last address is as expected.

Another difference from Figure 2 is that we now treat configurations as linear structures, with a beginning and an end. In our representation of the configuration tree, we next configuration therefore connects to the last cell of the previous configuration’s tape, rather than its start. We do this to ensure that the encoding works well even when restricting to linear queries. Indeed, the only non-linear rules in $P_1$ are used to enforce multiple successor configurations for universal states of an ATM. For normal TMs, even $P_1$ is in LinMDlog. The rules of the $P_1$ are as follows:

Note that we do not enforce any structure to define the query ConfCell; this query is implemented by a LinMQ that navigates over an arbitrary number of cells within one configuration. This is the main reason why we need LinMQs rather than UCQs here.

We now use the exponential space ATM of Lemma 14 to encode the tape of $2EXSPACE$ ATM. The following result shows, that one can always obtain an exponentially larger tape by nesting linear queries on the right-hand side.

Lemma 15. Assume that there is some space bound $s$ such that, for every ATM $M$, there is an MDlog query $P_1[x]$ and an MQ$^{k+1}$ query $P_2[x]$ with $k \geq 0$, such that $P_1[x]$ and $P_2[x]$ containment-encode accepting runs of $M$ in $s$, where the queries required by Definition 5 are MQ$^{k+1}$ queries. Moreover, assume that there is a suitable same-cell query that is in MQ$^{2}$.

Then, for every ATM $M'$, there is an MDlog query $P_1'[x], an MQ^{k+1}$ $P_2'[x],$ and MQ$^{k+1}$ queries as in Definition 5, such that $P_1'[x]$ and $P_2'[x]$ containment-encode an accepting run of $M'$ in space $s' \geq 2^s$. Moreover, the size of the queries for this encoding is polynomial in the size of the queries for the original encoding.

We show this result by using a deterministic space-$s$ Turing machine $M$ to count from 0 to $2^s$, which takes a fixed number $s' > 2^s$ of steps. We then use the encodings of accepting runs of $M$ as encodings for tapes of the ATM $M'$, where every configuration of $M$ becomes a cell of $M'$. All tapes simulated in this way are of equal length $s'$. Some queries required by Definition 5 are easy to obtain: for example, the new query NextCell$'[x, y]$ is the query NextConf$[x, y]$ of the encoding of $M$. The most difficult to express is the new same-cell query, for which we use the following MQ$^{k+1}$:

$$U_{1}(x) \land \text{NextCell}(x, y) \rightarrow U_{2}(y)$$

$$U_{1}(x) \land \text{NextCell}(x, y') \rightarrow U_{2}(y')$$
where FirstCell, Symbol, SameCell, and LastCell are the queries from the encoding of $M$. The first two rules simply mark the tape starting at $\lambda_1$ with $U_1$. The next two rules then compare the two (potentially very long) tapes from configurations of $M$ to check if they contain exactly the same symbols at each position, and the last rule finishes. Since the tapes are not connected in any known way, we have to be careful to ensure never to lose the connection to either of the tapes, to avoid comparing random cells from other parts of the database. Indeed, the last two rules do not mention $\lambda_1$ or $\lambda_2$ at all. We need two IDB predicates to achieve this, which carefully mark the two tapes cell by cell.

Another important thing to note is that the query SameCell is only used exactly once in exactly one rule. Indeed, if we were using it twice, then the length of our queries would grow exponentially when applying the construction inductively. This is the reason why we encode symbols and head positions with constants, rather than using unary predicates like for states. In the latter case, we need many rules, one for each predicate, as can be seen in the third rule above. One could try to avoid the use of constants by more complex encodings that encode information using paths of different lengths as done by Björklund et al. [7]. However, some additional device is needed to ensure that database instances are sufficiently closely connected in this case, which may again require constants, IDBs of higher arity, or a greater nesting level of LinMQ queries to navigate larger distances.

With the previous results, Theorem 13 can be proved by an easy induction: for the base case $k = 1$ we apply Lemma 15 to the result of Lemma 14; for the induction step we use Lemma 15 again.

9. LINEAR DATALOG

Not only query answering, but also containment checking is often slightly simpler in fragments of linear Datalog. Intuitively, this is so because derivations can be represented as words rather than as trees. Thus, the automata theoretic techniques that we have used in Section 6 can be applied with automata on words where some operations are easier. In particular, containment of (nondeterministic) automata on words can be checked in polynomial space rather than in exponential time. This allows us to establish the following theorems, which reduce the 2ExpTime upper bound of Theorem 10 to ExpSpace and the $(k + 2)\ ExpTime$ upper bound of Theorem 11 to $(k + 1)\ ExpSpace$.

Theorem 16. Containment of LinDlog queries in GDlog queries can be decided in ExpSpace.

Theorem 17. Containment of LinDlog queries in GQ* queries can be decided in $(k + 1)\ ExpSpace$.

Establishing matching lower bounds for the complexity turns out to be more difficult. In general, we loose the power of alternation, which explains the reduction in complexity. The general approach of encoding (non-alternating) Turing machines is the same as in Section 7, where Definition 5 is slightly simplified since we do not need to consider universal states, so that configuration trees turn into configuration sequences. Moreover, Lemma 12 applies to this case as well, since it only requires linear queries. Likewise, our general inductive step in Lemma 15 uses deterministic (non-alternating) TMs to construct exponentially long tapes. Moreover, it turns out that the construction of an initial exponential space TM in Lemma 14 leads to linear queries if the TM has no universal states.

Yet it is challenging to lift the exact encodings of Lemma 14 and Lemma 15. The same-cell query that we constructed in Lemma 15 for our inductive argument is non-linear. As explained in Section 8, the use of two IDBs to mark both sequences of tape cells is essential there to ensure correctness. The main problem is that we must not loose connection to either of the sequences during our checks. As an alternative to using IDBs on both sequences, one could use the ConfCell query to ensure that the compared cells belong to the right configurations. This leads to the following same-cell query:
While this works in principle, it has the problem that the ConffCell query of Lemma 14 is a LinMQ, not a UCQ. Therefore, if we construct a same-cell query for the 2ExpSpace case, we obtain LinMQ^2 queries, which yields the following result:

**Theorem 18.** Deciding containment of LinMDlog queries in LinMQ^2 queries is hard for kExpSpace.

In order to do better, one can try to express ConffCell as a UCQ. In general, this is not possible on the database instances that the left-hand query in Lemma 14 recognizes, since cells may have an exponential distance to their configuration while UCQs can only recognize local structures. To make ConffCell local, we can modify the left-hand query to ensure that every cell is linked directly to its configuration with a binary predicate inConf. Using binary IDB predicates, we can do this with the following set of frontier-guarded rules:

\[
\begin{align*}
\text{firstConf}(x, y) & \land U_{\text{conf}}(y) \rightarrow U_{\text{head}}(x) \\
\text{state}(x) & \land \text{nextCell}(x, y) \land \text{inConf}(y, x) \land U_{\text{conf}}(y, x) \land U_{\text{conf}}(x) \\
\text{blt}(x, 0) & \land U_{\text{conf}}(y, z) \land \text{inConf}(x, z) \land U_{\text{conf}}(y, x) \land U_{\text{conf}}(z) \\
\text{symbol}(x, c, \sigma) & \land U_{\text{symbol}}(x, z) \land \text{inConf}(x, z) \land U_{\text{conf}}(x) \\
\text{head}(x, h) & \land U_{\text{head}}(x, z) \land \text{inConf}(x, z) \land U_{\text{conf}}(x) \\
\text{head}(x, f) & \land U_{\text{head}}(x, z) \land \text{inConf}(x, z) \land U_{\text{conf}}(x) \\
\text{nextCell}(x, y) & \land U_{\text{conf}}(y, z) \land \text{inConf}(x, z) \land U_{\text{conf}}(x) \\
\text{nextConf}(x, y) & \land U_{\text{conf}}(y) \land \text{inConf}(x, z) \land U_{\text{conf}}(x) \\
\text{lastConf}(x) & \land \text{inConf}(x, z) \land U_{\text{conf}}(x, z)
\end{align*}
\]

Structures matched by this query provide direct links from each element to their configuration element, and we can thus formulate ConffCell as a UCQ and obtain the following.

**Theorem 19.** Deciding containment of LinGDlog queries in LinMQ^2 queries is hard for (k + 1)ExpSpace.

It is not clear if this result can be extended to containments of LinMQ in LinMQ^2; the above approach does not suggest any suitable modification. In particular, the propagation of inConf in the style of a transitive closure does not work, since elements may participate in many inConf relations. On the other hand, the special constants \( \lambda \) in LinMQs cannot be used to refer to the current configuration, since there can be an unbounded number of configurations but only a bounded number of special constants. It is possible, however, to formulate a LinMQ Conff[x] that generates the required structure for a single configuration, since one can then represents the configuration by \( \lambda \). We can generate arbitrary sequences of such structures by using Conff[x] as a nested query to that matches a regular expression firstConf (Conff NextConf^* Conff LastConf, where we use NextConf to express the disjunction of all nextConf relationships. This proves the following statement.

**Theorem 20.** Deciding containment of LinMQ^2 queries in LinMQ^2 queries is hard for (k + 1)ExpSpace.

Finally, we can also continue to use the same approach for encoding SameCell as in Section 8, without using ConffCell, while still restricting to linear Datalog (and thus to non-alternating TMs) on the left-hand side. This leads us to the following result.

**Theorem 21.** Deciding containment of LinMDlog queries in MQ^2 queries is hard for (k + 1)ExpSpace.

We have thus established tight complexity bounds for the containment of nested GQs, while there remains a gap (of one exponential or one nesting level) for MQs.

10. CONCLUSIONS

We have studied the most expressive fragments of Datalog for which query containment is still known to be decidable today, and we have provided exact complexities for most of their query answering and query containment problems. While containment for nested queries tends to be non-elementary for unbounded nesting depth, we have shown tight exponential complexity hierarchies for the main cases that we studied. As part of our results, we have also settled a number of open problems for known query languages: the complexity of query containment for MQ and MQ^*, the complexity of query containment of Datalog in GDlog, and the expressivity of nested LinDlog.

Moreover, we have built on the recent “flag & check” approach of monadically defined queries to derive various natural extensions, which lead to new query languages with interesting complexity results. In most cases, we observed that the extension from monadic to frontier-guarded Datalog does not affect any of the complexities, whereas it might have an impact on expressivity. In contrast, the restriction to linear Datalog has the expected effects, both for query answering and for query containment.

The only case for which our results for containment complexity are not tight is when we restrict rules to be both linear and monadic: while small variations in the involved query languages lead to the expected tight bounds, this particular combination eludes our analysis. This case could be studied as part of a future program for analyzing the behavior of (nested) conjunctive regular path queries, which are also a special form of monadic, linear Datalog.

Another interesting open question is the role of constants. Our hardness proofs, especially in the nested case, rely on
the use of constants to perform certain checks more efficiently. Without this, it is not clear how an exponential blow-up of our encoding (or the use of additional nesting levels) could be avoided. Of course, constants can be simulated if we have either predicates of higher arity or special constants as in "flag & check" queries. However, for the case of (linear) monadic Datalog without constants, we conjecture that containment complexities are reduced by one exponential each when omitting constants.

An additional direction of future research is to study problems where we ask for the existence of a containing query of a certain type rather than merely check containment of two given queries. The most prominent instance of this scenario is the boundedness problem, which asks whether a given Datalog program can be expressed by some (yet unknown) UCQ. It has been shown that this problem can be studied using tree-automata-based techniques as for query containment [15], though other approaches have been applied as well [4]. Besides boundedness, one can also ask more general questions of rewritability, e.g., whether some Datalog program can be expressed in monadic Datalog or in a regular path query.

11. REFERENCES


Table 1: Summary of the known complexities of query containment for several Datalog fragments; sources for each claim are shown in square brackets, using \ to separate sources for lower and upper complexity bounds, respectively
APPENDIX

A. TREE AUTOMATA

We use standard definitions for two-way alternating tree automata as introduced in [15]. A regular (one-way, non-alternating) tree automaton is obtained by restricting this definition.

A tree automaton runs over ranked, labelled trees of some maximal arity (out-degree) \( f \). A ranked tree can be seen a function \( t \) mapping sequences of positive natural numbers (encoding nodes in the tree) to symbols from a fixed finite alphabet (the labels of each node). Each letter of the alphabet is ranked, i.e., associated with an arity that defines how many child nodes a node labeled with this symbol should have. The domain of \( t \), denoted \( \text{Nodes}(t) \), satisfies the following closure property: if \( i \cdot j \in \text{Nodes}(t) \), then \( i \in \text{Nodes}(t) \) and \( i \cdot k \in \text{Nodes}(t) \) for all \( 1 \leq k \leq j \). Given a ranked tree \( t \), we write \( i \in \text{Nodes}(t) \) to denote an arbitrary node of \( t \) and \( t(i) \) to denote the label of \( i \) in \( t \). We denote by \( \text{Trees}(\Sigma) \) the set of trees over the alphabet \( \Sigma \).

A two-way alternating tree automaton \( \mathcal{A} \) is a tuple \( (\Sigma, Q, Q_0, \delta, Q_1) \) where

- \( \Sigma \) is a tree alphabet;
- \( Q \) is a set of states;
- \( Q_0 \subseteq Q \) is the set of initial states;
- \( Q_1 \subseteq Q \) is the set of accepting states;
- \( \delta \) is a transition function from \( Q \times \Sigma \): let \( q \in Q \) be a state and \( \sigma \in \Sigma \) be a letter of arity \( \ell \); then \( \delta(q, \sigma) \) is a positive boolean combination of elements in \( \{-1, 0, 1, \cdots, \ell\} \times Q \).

The numbers used in transitions encode directions, where \(-1\) is up and \(0\) is stay. For example \( \delta(q, \sigma) = ((1,s_1) \land (1,s_2) \lor ((-1,t_3) \land (2,t_4))) \) is an example of transition for a state \( q \) and a node labeled \( \sigma \): a node labeled by \( \sigma \) can be in the state \( q \) iff its first child can be in the states \( s_1 \) and \( s_2 \), or its parent and its second child can be in the states \( s_3 \) and \( s_4 \), respectively.

Let \( t \) be a tree over \( \Sigma \). A run of \( \mathcal{A} \) over \( t \) is a tree labeled by elements of \( Q \times \{-1, 0, 1, \cdots, f\} \times \text{Nodes}(t) \cup \{-1\} \). \( \tau \) satisfies the following properties:

- \( \tau \) is finite.
- The root of \( \tau \) is labelled by \( (q_0, i, n) \), where \( q_0 \) is in \( Q_0 \).
- If a node \( v \) is labelled by \( (q, i, n) \) and \( n \) is not a node of \( t \), then \( v \) is a leaf of \( \tau \).
- If a node \( v \) is labelled by \( (q, i, n') \), \( n \) is a node of \( \tau \) labelled by \( \sigma \) of arity \( \ell \) and \( v' \) is labelled by \( (q_1, j, n') \) then
- if \( j = -1 \) there exists \( u \leq k \) such that \( n = n'u \)
- if \( j = 0 \), then \( n = n' \)
- if \( j \leq k \), then \( n' = n, j \).
- If a node \( v \) is labelled by \( (q, i, n) \), \( n \in t \) labelled by \( \sigma \) and the children of \( v \) are labelled by \( (q_1, j_1, n_1) \cdots (q_k, j_k, n_k) \) then \( \delta(q, \sigma) \) is satisfied when interpreting the symbols \( \langle j_1, q_1, \cdots, j_k, q_k \rangle \) as true and all other symbols as false.

\( \tau \) is valid iff, for each leaf of \( \tau \) labelled by \( (q, i, n) \), \( q \) is in \( Q_1 \). \( \mathcal{A} \) accepts a tree \( t \) if there exists a valid run of \( \tau \) over \( t \). We denote by \( \text{Trees}(\mathcal{A}) \). The set of trees accepted by \( \mathcal{A} \).

A regular (one-way, non-alternating) tree automaton is a 2-way alternating tree automaton where all transitions for a symbol \( \sigma \) of rank \( \ell \) are boolean formulae of the form \( (((1,q_{11}) \land \cdots \land (\ell,q_{1\ell})) \lor \cdots \lor ((1,q_{n1}) \land \cdots \land (\ell,q_{n\ell})) \) for some \( n \geq 0 \). In particular, directions 0 and \(-1\) do not occur.

In this case, we can represent transitions as sets of lists of states \( \langle q_{11},\cdots,q_{1\ell},\cdots,q_{n1},\cdots,q_{n\ell}\rangle \).

Finally, we recall two useful theorems from [15].

**Theorem 22** (Theorem A.1 of [15]). Let \( \mathcal{A} \) be a two-way alternating automaton. Then there exists a tree automaton \( \mathcal{A} \) whose size is exponential in the size of \( \mathcal{A} \) such that \( \text{Trees}(\mathcal{A}) = \text{Trees}(\Sigma) \setminus \text{Trees}(\mathcal{A}) \).

**Theorem 23** (Theorem A.2 of [15]). Let \( \mathcal{A} \) be a two-way alternating automaton. Then there exists a tree automaton \( \mathcal{A} \) whose size is exponential in the size of \( \mathcal{A} \) such that \( \text{Trees}(\mathcal{A}) = \text{Trees}(\mathcal{A}) \).

B. PROOFS

**Proofs for Section 4**

**Theorem 1.** LinDlog = LinDlog+.

**Proof.** We will prove that any LinDlog+ query can be rewritten into a LinDlog query of polynomial size. We make simplifying assumptions on the structure of the nested query which can be easily obtained by polynomial transformations and make the presentation easier: we assume that every rule body of any query occurring at any nesting depth contains at most one subquery atom (using, e.g., Proposition 2). Second, we assume that all variables and IDB predicates that are not in the same scope are appropriately renamed apart.

In order to prove our claim, we will first show that any LinDlog2 can be rewritten into an equivalent LinDlog query. Applying the rewriting iteratively inside-out (and observing that even manfold application can be done in polynomial total time) then allows to conclude that there is a polynomial rewriting of any LinDlog+ query of arbitrary depth into a LinDlog query.

Consider a LinDlog2 query \( P = (P, p) \) and assume w.l.o.g. that every rule body of the rules contains at most one LinDlog1 subquery. Now, going through all rules of \( P \) we produce the rules \( P' \) of the unnested but equivalent version.

Consider a rule \( r \in P \) having the shape

\[ Q(x_1, \ldots, x_n) \land \Omega(y_1, \ldots, y_r) \land B_1 \land \cdots \land B_k \rightarrow H \]

where \( p \) is the body IDB predicate and where \( Q = (Q, q) \) is a LinDlog1 query. For any \( k \)-ary IDB predicate \( r \) inside \( Q \) we increase its arity by \( \ell \) and let \( P' \) contain all rules of \( Q' \) which is obtained from the rules \( P' \) of \( Q \) by
• replacing any (head or body) IDB atom \( r(z_1, \ldots, z_k) \) of
  \( \rho' \) by \( r(z_1, \ldots, z_k, y_1, \ldots, y_r) \) and
• in case \( \rho' \) does not contain any IDB body atom, add
  \( p(y_1, \ldots, y_r) \) to the body.

Further we let \( \mathcal{P}' \) contain the rule
\[
q(x_1, \ldots, x_n, y_1, \ldots, y_r) \land B_1 \land \ldots \land B_k \rightarrow H.
\]

In case of a rule \( \rho \in \mathcal{P} \) having the shape
\[
Q(x_1, \ldots, x_n) \land B_1 \land \ldots \land B_k \rightarrow H
\]
we add \( Q \) to \( \mathcal{P}' \) without change and let \( \mathcal{P}' \) contain the rule
\[
q(x_1, \ldots, x_n) \land B_1 \land \ldots \land B_k \rightarrow H.
\]

In case a rule \( \rho \in \mathcal{P} \) does not contain a subquery atom we simply add \( \rho \) to \( \mathcal{P}' \).

It can now easily verified that \( (\mathcal{P}, p) \) and \( (\mathcal{P}', p) \) are equivalent: first it is straightforward, that \( (\mathcal{P}, p) \) is equivalent to \( (\mathcal{P}', p) \) where \( \mathcal{P}' \) is obtained from \( \mathcal{P} \) by replacing every \( Q(x_1, \ldots, x_n) \) by \( q(x_1, \ldots, x_n) \) (that is, the according goal predicate) and then adding all rules from \( Q \) with no changes made to them. Second one can show that there is a direct correspondence between proof trees of \( (\mathcal{P}', p) \) and linearized proof trees of \( (\mathcal{P}, p) \) which yields the desired result. \( \square \)

Proposition 2. Let \( P \) be a positive query, i.e., a Boolean expression of disjunctions and conjunctions. of LinMQ\(^k \) queries with \( k \geq 1 \). Then there is a LinMQ\(^k \) query \( P' \) of size polynomial in \( P \) that is equivalent to \( P \). Analogous results hold when replacing LinMQ\(^k \) by MQ\(^k \), GQ\(^k \), or LinMQ\(^k \) queries.

Proof. We show the claim by induction, by expressing the innermost disjunctions and conjunctions of \( P \) with equivalent LinMQ\(^k \) queries of linear size. We consider positive queries without existential quantifiers (i.e., where all variables are answer variables), but the inner LinMQ\(^k \) may use existential quantifiers.

Let \( P(x) = \bigvee_{i=1}^n P_i[x] \) be a disjunction of LinMQ\(^k \) queries. Each query \( P_i \) is of the form \( \exists z_i. P'_i[x'_i] \), where \( x'_i \) is the list of free variables of \( P'_i \) (corresponding to constants \( \lambda_i \), and \( z_i \) contains exactly those variables of \( x'_i \) that do not occur in \( x_i \). We assume without loss of generality that \( z_i \) is disjoint from \( z_j \) if \( i \neq j \), and that each \( P'_i \) uses a unique set of IDBs that does not occur in other queries. We consider queries \( \tilde{P}_i \) obtained by replacing the special constant that represents a variable \( x_i \in x \) by the special constant \( \lambda_i \) (assumed to not occur in \( P \) yet). Thus, the queries \( \tilde{P}_i \) share special constants exactly where queries \( P_i \) share variables. We can now define the LinMQ\(^k \) \( P' \) as \( \exists z_1 \ldots z_n. \bigcup_i^\infty \bigcup_j P_i \), where we assume that the correspondence of special constants to free variables is such that the existential quantifiers refer to the same variables as before.

Let \( P(x) = \bigwedge_{i=1}^n P_i[x] \) be a conjunction of LinMQ\(^k \) queries. Let \( P'_i = \exists z_i. P'_i[x'_i] \) as before, and let \( \cup_i \) for \( i \in \{1, \ldots, n-1\} \) be fresh IDB predicates. The queries \( \tilde{P}_i \) are defined as before by renaming special constants to reflect shared variables. For each \( i \in \{1, \ldots, n\} \), the set of rules \( \tilde{P}_i \) is obtained from \( \tilde{P}_i \) as follows: if \( i < n \), then every rule \( \varphi \rightarrow \text{hit} \in \tilde{P}_i \) is replaced by the rule \( \varphi \rightarrow \text{hit}(A_1) \), where \( A_1 \) is a fixed special constant in the queries; if \( i > 1 \), then every rule \( \varphi \rightarrow \psi \in \tilde{P}_i \) where \( \varphi \) does not contain an IDB predicate is replaced by the rule \( \varphi \rightarrow \text{hit}(A_1) \rightarrow \psi \), where \( A_1 \) is as before. The LinMQ\(^k \) \( P' \) is defined as \( \exists z_1 \ldots z_n. \bigcup_i^\infty \bigcup_j P_i \).

These constructions lead to equivalent LinMQ\(^k \) queries of linear size, so the claim follows by inductions. The cases for MQ\(^k \), GQ\(^k \), and LinMQ\(^k \) follow from the same constructions (note that, without the requirement of linearity, a simpler construction is possible in the case of conjunctions).

Theorem 3. The combined complexity of evaluating GQ queries over a database instance is NP-complete. The same holds for GDLq queries. The combined complexity of evaluating GQ\(^+\) queries is PSpace-complete. The data complexity is P-complete for GDLq, GQ, and GQ\(^+\).

Proof. The lower bounds are immediate from the matching complexities for MQ and MQ\(^+\) queries, respectively [19].

First, we prove that checking if a tuple is an answer of a GQ over a database instance \( I \) is in NP for combined complexity. Let \( I \) be an instance, let \( P \) be a GQ with frontier guarded rules \( P \), and let \( \delta \) be be a candidate answer for \( P \) as in Definition 1.

Since each rule in \( P \) is frontier-guarded, each intentional fact that is derived when checking the answer follows from the application of one particular rule, instantiated to match one particular (guard) EDB fact in the body. Therefore, the number of IDB facts that can be derived is polynomially bounded in the size of \( I \) and \( P \).

Thus, for every derivation of \( P \) in frontier-guarded, each intentional fact that is derived when checking the answer follows from the application of one particular rule, instantiated to match one particular (guard) EDB fact in the body. Therefore, the number of IDB facts that can be derived is polynomially bounded in the size of \( I \) and \( P \).

Once, for each derivation of \( P \), only a polynomial number of rule applications are necessary, since it is enough to derive each IDB fact once. It is clear that one can guess such a derivation, where we guess, for each derivable IDB fact, one specific rule instance by which it is derived. The correctness of this guess can be checked in polynomial time, showing that the problem can be solved in NP.

We now show that checking an answer of a GQ\(^+\) over an instance \( I \) is in PSpace. Let \( I \) be an instance, let \( P \) be a GQ\(^+\) with frontier guarded rules \( P \) (that may contain subqueries), and let \( \delta \) be be a candidate answer for \( P \) as in Definition 1.

We demonstrate by induction on \( k \) that checking if \( \delta \) is a solution for \( P \) w.r.t. \( I \) is in PSpace. For the induction base, the claim follows from the above result for GQs.

For the induction step, using the same argument as before, we can see that the number of IDB facts that can be derived by \( P \) is still polynomial. Therefore, we can again guess a polynomial derivation as before, though the rule instances now may refer to subqueries of smaller nesting depth. By the induction hypothesis, whenever we need to verify the applicability of such a rule, we can use an NSpace algorithm for the nested query. The overall number of such checks is polynomial, yielding the overall NSpace algorithm. The result follows since NSpace = PSpace [20].
The fact that query evaluation is in P for data complexity is immediate from the fact our queries can be expressed in Datalog, which is known to have this data complexity. A direct proof is also obtained by observing that the number of possible derivation sequences that the above algorithms need to consider is in itself polynomial in \( I \) if \( P \) is fixed, so that the algorithms themselves are already in P for data complexity.

**Theorem 4.** The combined complexity of evaluating LinMQ queries over a database instance is NP-complete. The same holds for LinDlog and LinGQ. The combined complexity of evaluating LinMQ+ queries is PSPACE-complete. The same holds for LinGQ+.

The data complexity is NLogSpace-complete for all of these query languages.

**Proof.** The claimed NP-completeness is immediate. Hardness follows from the hardness of CQ query answering. Membership follows from the membership of \( \text{PS} \).

The claimed membership in PSPACE follows from the PSPACE-membership of LinDlog; note that this uses Theorem 1. Hardness for LinGQ+ follows from the hardness for LinMQ+, which we show by modifying the PSpace-hardness proof for monadically defined queries from [19].

We show the result by providing a reduction from the validity problem of quantified Boolean formulae (QBFs). We recall that for any QBF, it is possible to construct in polynomial time an equivalent QBF that has the specific shape

\[
\bigvee_{L \subseteq \ell} \bigwedge_{i \in \ell} \bigvee_{L \subseteq \ell} \ell,
\]

with \( Q_1, \ldots, Q_n \in \{\exists, \forall\} \) and \( L \) being a set of sets of literals over the propositional variables \( x_1, \ldots, x_n \). In words, we assume our QBF to be in prenex form with the propositional part of the formula in disjunctive normal form. For every literal set \( L = \{x_{i_1}, \ldots, x_{i_k}, \neg x_{i_{k+1}}, \ldots, \neg x_{i_{n}}\} \), we now define the n-ary FCP \( p_L = (\ell(x_{i_1}) \land \ldots \land \ell(x_{i_k}) \land f(A_{i_{k+1}}) \land \ldots \land f(A_{i_{n}})) \rightarrow \text{hit} \). Moreover, we define the n-ary FCP \( p_L = (p_L(A_{i_1}, \ldots, A_{i_n}) \rightarrow \text{hit} \mid L \in \mathbb{L}) \). Letting \( p_L = p_n \) we now define FCPs \( p_{n-1}, \ldots, p_0 \) in descending order. If \( Q_i = \exists \), then the \( i \)-ary FCP \( p_{i-1} \) is defined as the singleton rule set \( \{p_i(A_{i_1}, \ldots, A_{i_{i-1}}, y) \rightarrow \text{hit}\} \). In case \( Q_i = \forall \), we let \( p_{i-1} \) contain the rules

\[
\begin{align*}
f(x) &\rightarrow U_i(x) \\
U_i(x) \land f(x) \land \ell(y) &\rightarrow U_i(y) \\
U_i(x) \land \ell(y) &\rightarrow \text{hit} \\
U_i(x) \land p_i(A_{i_1}, \ldots, A_{i_{i-1}}, x) &\rightarrow U_i(x)
\end{align*}
\]

Note that \( p_0 \) is a Boolean LinMQ+ query the size of which is polynomial in the size of the input QBF.

Now, let \( D \) be the database containing the two individuals 0 and 1 as well as the facts \( f(0) \) and \( r(1) \). We now show that the considered QBF is true exactly if \( D \models p_0(0) \). To this end, we first note that by construction the extension of \( p_L \) contains exactly those \( n \)-tuples \( \langle \delta_1, \ldots, \delta_n \rangle \) for which the corresponding truth value assignment \( \text{val} \), sending \( x \) to \text{true} iff \( \delta_i = 1 \), makes the formula \( \bigwedge_{i \in \ell} \ell \) true. In the same way, the extension of \( p_L \) represents the set of truth value assignments satisfying \( \bigvee_{L \subseteq \ell} \bigwedge_{i \in \ell} \ell \). Then, by descending induction, we can show that the extensions of \( p_i \) encode the assignments to free propositional variables of the subformula \( Q_{i+1} x_{i+1} \ldots Q_n x_n \bigwedge_{L \subseteq \ell} \bigwedge_{i \in \ell} \ell \) that make this formula true. Consequently, \( p_0 \) has a nonempty extension if the entire considered QBF is true.

Finally, the NLogSpace-completeness for data complexity is again immediate, where the upper bound is obtained from LinDlog, and the lower bound follows from the well-known hardness of reachability queries, which can be expressed in LinMDlog.

**Proofs for Section 6**

**Proposition 7.** There is an automaton \( \mathcal{A}_\rho \) that accepts exactly the annotated matching trees for \( \rho \) and \( \mathbb{P} \), and which is exponential in the size of \( \rho \) and \( \mathbb{P} \).

**Proof.** We first construct an automaton \( \mathcal{A}_\rho \), that accepts matching trees where each node is additionally annotated by a partial mapping of the form \( \text{Var}(\rho) \rightarrow \mathcal{V}_\mathbb{P} \) (called \( \text{Var}(\rho) \)-label), such that: every special variable \( x \in \text{Var}(\rho) \) occurs in at least one \( \text{Var}(\rho) \)-label, and whenever a variable \( x \in \text{Var}(\rho) \) occurs in two, it is mapped to the same variable and both variable occurrences are connected. Note that this is essentially the same condition that we imposed for \( \lambda \)-annotations.

The intersection of tree automata can be computed in polynomial time. We can therefore construct automata to check part of the conditions for (annotated) matching trees to simplify the definitions. We first construct an automaton \( \mathcal{A}_\rho \) for checking the condition on \( \text{Var}(\rho) \)-labels for one variable \( x \in \text{Var}(\rho) \). We define \( \mathcal{A}_\rho = (\Sigma, Q_x, Q_x, Q_x, Q_x) \), where the alphabet \( \Sigma \) consists of quadruples of proof-tree labels (from \( \mathcal{R}_\rho \)), \( \lambda \)-labels, \( p \)-labels, and \( \text{Var}(\rho) \)-labels. The state set \( Q_x \) is \([a, b, \text{accept}] \cup \{q_i \mid v \in \mathcal{V}_\mathbb{P}\} \), signifying that the current node is above the first node annotated with a mapping for \( x \), or at a node where \( x \) is mapped to a variable \( v \). That start-state set is \( Q_x = [a] \cup \{q_i \mid v \in \mathcal{V}_\mathbb{P}\} \); the end-state set if \( Q_x = \{\text{accept}\} \).

Consider a rule \( \rho' \in \mathcal{R}_\rho \) of the form \( r_1(v_1) \land \ldots \land r_n(v_m) \land h_1(w_1) \land \ldots \land h_m(w_m) \rightarrow h(v) \), where \( r_i \) are EDB predicates and \( h_i \) are IDB predicates. For the case that \( m > 0 \), there is a transition \( \langle q_1, \ldots, q_m \rangle \in \delta(q, \rho', \ldots, \rho') \) exactly if the following conditions are satisfied:

- if \( q = a \) and \( v(x) \) is undefined, then \( q_i = a \) for one \( 1 \leq i \leq m \) and \( q_i = b \) for all \( 1 \leq j \leq m \) with \( i \neq j \);
- if \( q = q_i \) and \( v(x) = v \), then \( q_i = q_v \) for all \( 1 \leq i \leq m \) such that \( v \) occurs in \( w_i \) and \( q_i = b \) for all other \( i \).
• if \( q = b \) and \( \nu(x) \) is undefined, then \( q_i = b \) for all \( i \leq m \).

For the case \( m = 0 \), there is a transition \( \langle \text{accept} \rangle \in \delta(q, (q', \ldots, \nu)) \) exactly if:

• if \( q = q_e \) and \( \nu(x) = v \);

• if \( q = b \) and \( \nu(x) \) is undefined.

It is easy to check that the automaton \( \mathcal{A}_i \) satisfies the required condition. Now an automaton for checking the condition on \( \text{Var}(\rho) \)-labels can be constructed as the intersection \( \mathcal{A}_{\text{Var}(\rho)} = \bigcap_{v \in \text{Var}(\rho)} \mathcal{A}_v \). The automaton \( \mathcal{A}_i \) for checking the condition on \( \lambda \)-labels is constructed in a similar fashion. Likewise, an automaton \( \mathcal{A}_p \) for checking the condition on \( p \)-labels is easy to define.

It remains to construct an automaton for checking the conditions (a)–(d) of Definition 4. To do this, we interpret the \( \text{Var}(\rho) \)-labels and \( \lambda \)-labels as partial specifications of the required mapping \( \nu \). Condition (a) further requires that \( \nu(x) = v \), i.e., that the \( \text{Var}(\rho) \)-label at the unique node annotated with \( p(v) \) contains this mapping. It is easy to verify this with an automaton \( \mathcal{H}_\rho \). Together, \( \mathcal{H}_\rho \), \( \mathcal{A}_i \), and \( \mathcal{A}_{\text{Var}(\rho)} \) provide a consistent variable mapping that respects the \( p \)-label (a) and the connectedness of variable occurrences, i.e., (c) and (d). To check the remaining condition (b), we use an automaton \( \mathcal{H}_p \).

The automaton for (b) will use auxiliary markers to record which atoms have been matched in the current node and how exactly this was done. We record such a match as a partial function from atoms \( q(z) \in \varphi \) to instances \( q(w) \) of such atoms using variables \( w \subseteq \mathcal{V}_r \). The set of all such partial functions is denoted \( \text{Match}_{\varphi,p} \). Note that this set is exponential (not double exponential).

We now define \( \mathcal{H}_{p|\rho} = \langle \Sigma, Q, \delta, \lambda, Q_0 \rangle \) where \( \Sigma \) is as for \( \mathcal{A}_i \) above. The set of states \( Q \) is \( \{\text{accept}\} \cup (2^{\rho} \times \text{Match}_{\varphi,p}) \), where elements from \( 2^{\rho} \) encode the subset of \( \varphi \) that should be witnessed at or below the current node, and the elements from \( \text{Match}_{\varphi,p} \) encode atoms that must be matched at the current node with their respective instantiations. The start-state set \( Q_0 \) is \( \{\langle \varphi, \mu \rangle \mid \mu \in \text{Match}_{\varphi,p}\} \); the end-state set \( Q_\lambda \) is \( \{\text{accept}\} \). The transition function \( \delta \) is defined as follows. Consider a rule \( \rho' \in \mathcal{R}_\rho \) of the form

\[
\begin{align*}
&r_1(v_1) \land \ldots \land r_n(v_n) \land h_1(w_1) \land \ldots \land h_m(w_m) \rightarrow h(v),
\end{align*}
\]

where \( r_i \) are EDB predicates and \( h_i \) are IBD predicates. For the case \( m > 0 \), there is a transition \( \langle \text{accept} \rangle = \delta((\beta_1, \mu_1), (\rho', v_{k_1}, \ldots, v_{k_m})) \in \delta((\beta, \mu), (\rho', v_{k_1}, \ldots, v_{k_m})) \) exactly if the set \( \beta \subseteq \varphi \) can be partitioned into sets \( \beta_1, \beta_2, \ldots, \beta_m \) such that \( \langle v_1 \cup \text{Var}(\rho) \rangle (\beta') = \mu(\beta') \) and \( \mu(\beta') \subseteq \{r_1(v_1), \ldots, r_n(v_n)\} \). The element \( \mu_i \) of successor states can be chosen freely; the validity of the choice will be checked later. For the case \( m = 0 \), there is a transition \( \langle \text{accept} \rangle = \delta((\beta, \mu), (\rho', v_{k_1}, \ldots, v_{k_m})) \) exactly if \( (v_1 \cup \text{Var}(\rho))(\beta') = \mu(\beta') \) and \( \mu(\beta') \subseteq \{r_1(v_1), \ldots, r_n(v_n)\} \). In fact, the information from \( \text{Match}_{\varphi,p} \) is not strictly necessary to define the transition, since the relevant elements \( \mu \) are always determined by other choices in the transition. However, having this information explicit will be important in later proofs.

The automaton \( \mathcal{A}_{p|\rho} \) is obtained as the intersection \( \mathcal{A}_{\text{Var}(\rho)} \cap \mathcal{A}_i \cap \mathcal{A}_p \cap \mathcal{A}_{p|\rho} \cap \mathcal{A}_{p|\rho} \). It is easy to verify that it accepts exactly the \( \text{Var}(\rho) \)-annotated matching trees. Note that \( \mathcal{A}_{p|\rho} \) is exponential in size, already due to the exponentially large alphabet \( \Sigma \). Now the required automaton \( \mathcal{A}_{p|\rho} \) is obtained by “forgetting” the \( \text{Var}(\rho) \)-label in transitions of \( \mathcal{A}_{p|\rho} \). This projection operation for tree automata is possible with a polynomial increase in size: every state of \( \mathcal{A}_{p\rho} \) is a pair of a state of \( \mathcal{A}_{p|\rho} \) and a \( \text{Var}(\rho) \)-label; transitions of \( \mathcal{A}_{p\rho} \) are defined as for \( \mathcal{A}_{p|\rho} \), but keeping \( \text{Var}(\rho) \)-label information in states and introducing transitions for all possible \( \text{Var}(\rho) \)-labels in child nodes.

Proposition 8. There is an alternating 2-way tree automaton \( \mathcal{A}^+_{p|\rho} \) that is polynomial in the size of \( \mathcal{A}_{p|\rho} \) such that, whenever \( \mathcal{A}_{p|\rho} \) accepts a matching tree \( T \) that has the \( p \)-annotation \( p(v) \) on node \( e \), then \( \mathcal{A}^+_{p|\rho} \) has an accepting run that starts from the corresponding node on \( T \) that is obtained by removing the \( p \)-annotation from \( T \).

Proof. Using alternating 2-way automata, we can traverse a tree starting from any node, visiting each node once. To control the direction of the traversal, we create multiple copies of each state \( q \); states \( q_{\text{down}} \) are processed like normal states in \( \mathcal{A}_{p|\rho} \), states \( q_{\text{up}} \) use an inverted transition of \( \mathcal{A}_{p|\rho} \) to move up the tree into a state \( q_{\text{up},i} \); these auxiliary states are used to check that the label of the upper node is actually \( \sigma \) and to start new downwards processes for all child nodes other than the one (i) that we came from.

To ensure that the constructed automaton \( \mathcal{A}^+_{p|\rho} \) simulates the behavior of \( \mathcal{A}_{p|\rho} \) in case the annotation \( p(v) \) is found, we eliminate all transitions that mention other \( p \)-annotations. Moreover, we assume without loss of generality that the states of \( \mathcal{A}_{p|\rho} \) that allow a transition mentioning \( p(v) \) cannot be left through any other transition; this can always be ensured by duplicating states and using them exclusively for one kind of transition. Let \( Q_0 \) be the set of states of \( \mathcal{A}_{p|\rho} \) that admit (only) transitions mentioning \( p(v) \). Let \( \mathcal{A}_{p|\rho} = \langle \Sigma', Q_0, \delta', Q_0 \rangle \) denote the automaton over the alphabet \( \Sigma' \) of \( \lambda \)-annotated proof trees (without \( p \)-annotations), with the same (start/end) states as \( \mathcal{A}_{p|\rho} \), and \( \delta' \) is defined based on the transition function \( \delta \) of \( \mathcal{A}_{p|\rho} \) as follows: \( \delta'((\rho', M)) \) is the union of all sets of the form \( \delta((\rho', \lambda \text{-label}, p \text{-label})) \) where \( p \)-label is either \( p(v) \) or empty. By this construction, there is a correspondence between the accepting runs of \( \mathcal{A}_{p|\rho} \) over trees where one node \( e \) is annotated with \( p(v) \) and accepting runs of \( \mathcal{A}^+_{p|\rho} \) (on trees without \( p \)-annotations) for which the node \( e \) is visited in some state of \( Q_0 \).
Transitions of $\mathcal{A}_p^{++}$ are defined as follows:

- For all $\sigma \in \Sigma$, let $\delta^+(\text{start}, \sigma)$ be the disjunction of all formulae $(0, q_{up}) \land (0, q_{down})$ where $q \in Q_p$.
- For states $q_{down}$ and $\sigma \in \Sigma$, let $\delta^+(q_{down}, \sigma)$ be the disjunction of all formulae $(1, q_{down}^i) \land \ldots \land (m, q_{down}^m)$ for which $\mathcal{A}_p^+$ has a transition $(q^1, \ldots, q^m) \in \delta(q, \sigma)$.
- For states $q_{up}$ and $\sigma \in \Sigma$, let $\delta^+(q_{up}, \sigma)$ be the disjunction of all formulae $(-1, q_{up}^i) \land \ldots \land (i-1, q_{up}^{i-1}) \land (i+1, q_{down}^{i+1}) \land \ldots \land (m, q_{down}^m)$ for which $\mathcal{A}_p^+$ has a transition $(q^1, \ldots, q^i, \ldots, q^m) \in \delta(q, \sigma)$.
- For states $q_{\sigma,i,q'}$, let $\delta^+(q_{\sigma,i,q'}, \sigma)$ be the disjunction of all formulae $(0, q_{up}^i) \land (1, q_{down}^{i-1}) \land \ldots \land (i-2, q_{down}^{i-2}) \land (i, q_{up}^{i+1}) \land \ldots \land (m, q_{down}^m)$ for which $\mathcal{A}_p^+$ has a transition $(q^1, \ldots, q^{i-1}, q', q^{i+1}, \ldots, q^m) \in \delta(q, \sigma)$ and the current node is the $i$th child of its parent (we can assume that this information is encoded in the labels $\sigma_i$ even for basic proof trees, which increases the alphabet only linearly; we omit this in our definitions since it would clutter all other parts of our proof without need).

- For states $q_{\sigma,i,q'}$, let $\delta^+(q_{\sigma,i,q'}, \sigma)$ be the disjunction of all formulae $(0, q_{up}^i) \land (1, q_{down}^{i-1}) \land \ldots \land (i-2, q_{down}^{i-2}) \land (i, q_{up}^{i+1}) \land \ldots \land (m, q_{down}^m)$ for which $\mathcal{A}_p^+$ has a transition $(q^1, \ldots, q^{i-1}, q', q^{i+1}, \ldots, q^m) \in \delta(q, \sigma)$.

- For all starting states $q \in Q_s$, let $\delta(q_{up}, \sigma) = (0, \text{accept})$.

It is not hard to verify that $\mathcal{A}_p^{++}$ has the required properties.

Proposition 9. For a Dlog query $P$ and a GQ query $P'$ with special constants $\lambda$, there is an alternating 2-way automaton $\mathcal{A}_p^{++}$ of exponential size that accepts the $\lambda$-annotated proof trees of $P$ that encode expansion trees for $P'$ with a match.

Proof. Let $P'$ be the set $\{\rho_1, \ldots, \rho_t\}$. For every IDB predicate $p$, let $P'_p$ denote the set of rules in $P'$ with head predicate $p$ (possibly hit). Without loss of generality, we assume that distinct rules use distinct sets of variables. For every frontier-guarded rule $\rho'$, let $\text{guard}(\rho')$ be a fixed EDB atom that acts as a guard in this rule, i.e., an atom that refers to all variables in the head of $\rho'$.

Consider a rule $\rho' \in P'$ with IDB atoms $q_1(t_1), \ldots, q_m(t_m)$ in its body. We construct new rules from $\rho'$ by replacing each atom $q_i(t_i)$ with a guard atom $\text{guard}(\rho'_i)$, suitably unified. Formally, assume that there are rules $\rho'_i \in P'_p$ with head $q_i(t_i)$ and a substitution $\theta$ that is a most general unifier for the problems $t_i \theta = s_i \theta$, for all $i \in \{1, \ldots, m\}$, and that maps every variable in $\rho'$ that does not occur in the head to a globally fresh variable. Then the guard expansion of $\rho'$ for $(\rho'_i)_{i=1}^m$ and $\theta$ is the rule that is obtained from $\rho' \theta$ by replacing each body atom $q_i(t_i) \theta$ by $\text{guard}(\rho'_i) \theta$. By construction, two distinct atoms $\text{guard}(\rho'_i) \theta$ and $\text{guard}(\rho'_j) \theta$ do not share variables, unless at positions that correspond to head variables in rules $\rho'_i$ and $\rho'_j$. The atoms $\text{guard}(\rho'_i) \theta$ in a guard expansion are called replacement guards. We consider two guard expansions to be equivalent if they only differ in the choice of the most general unifier. Let $\text{Guard}(\rho')$ be the set of all guard expansions of $\rho' \in P'$, i.e., a set containing one representative of each class of equivalent guard expansions. $\text{Guard}(\rho')$ is exponential since there are up to $|P'|^m$ non-equivalent guard expansions for a rule with $m$ IDB atoms.

The automaton $\mathcal{A}_p^{++}$ is constructed as follows. For every guard expansion $\rho_p \in \bigcup_{\rho \in \rho'} \text{Guard}(\rho')$ and every list $\vec{v}$ of proof-tree variables of the arity of the head of $\rho_p$, consider the alternating 2-way tree automaton $\mathcal{A}_{\rho_p,v}$ of Proposition 8. We assume w.l.o.g. that the state sets of these automata are mutually disjoint. Let $\mathcal{A}_p^{++} = (\Sigma, Q, Q_s, \delta, Q_f)$. As before, $\Sigma$ consists of pairs of a rule instance from $\mathcal{R}_p$ and a partial mapping of $\lambda$ to $\mathcal{V}_p$. The state set $Q$ is the disjoint union of all state sets of the automata of form $\mathcal{A}_p^{++}$ for which $\rho_p$ is a guard expansion of a rule with head hit (and $\vec{v}$ is the empty list). The end-state set $Q_f$ is the disjoint union of all end-state sets of automata $\mathcal{A}_p^{++}$.

The transition function $\delta$ is defined as follows. By the construction in Proposition 7, each state $q$ in the automaton $\mathcal{A}_p$ encodes a partial mapping $\text{match}(q)$ from body atoms of $\rho$ to instantiated atoms that use variables from $\mathcal{V}_p$, which are matched at the current tree node. This information is preserved through alphabet projections, intersections, and even through the construction in Proposition 8. We can therefore assume that each state $q$ of $\mathcal{A}_p^{++}$ is associated with a partial mapping $\text{match}(q)$.

For every state $q \in Q_p^{++}$ and every $\sigma \in \Sigma$, we define $\delta(q, \sigma) = \delta_{\rho_p,v}(q, \sigma) \land \psi$, where $\psi$ defined as follows. For every replacement guard atom $\alpha$ of $\rho_p$ for which $\text{match}(q)(\alpha)$ is defined, we consider the formula $\psi_\alpha = (0, q_1) \lor \ldots \lor (0, q_t)$, where

- $\alpha = \text{guard}(\rho') \theta$ for some rule $\rho'$ and substitution $\theta$;
- $\text{match}(q)(\alpha) = \alpha \theta'$ for some substitution $\theta'$;
- $q_1, \ldots, q_t$ are the start states of the automaton $\mathcal{A}_p^{++}$, where $p(z)$ is the head of $\rho'$.

Now $\psi$ is the conjunction of all formulae $\psi_\alpha$ thus defined.

Proofs for Section 7

Lemma 12. Consider an ATM $M$, and queries as in Definition 5, including SameCell$(x, y)$, that are $MQ^2$ queries for some $k \geq 0$. There is a $MQ^2$ query $P[x]$, polynomial in the size of $M$ and the given queries, such that the following hold.

- For every accepting run of $M$ in space $s$, there is some database instance $I$ with some element $c$ that encodes the run, such that $c \notin P^T$.
- If an element $c$ of $I$ encodes a tree of quasi-configurations of $M$ in space $s$, and if $c \notin P^T$, then $c$ encodes an accepting run of $M$ in space $s$.

Moreover, if all input queries are in LinMQ$^2$, then so is $P$.

Proof. We construct $P$ from all (polynomially many) positive queries obtained by instantiating the query patterns in Figure 4. Since $P$ needs to be a unary query with variable $x$, we extend every positive query that does not contain $x$
(1) **Unique head marker and correct left/right head markers:**

\[
\text{Head}(y, p_1) \land \text{NextCell}(y, z) \land \text{Head}(z, p_2) \quad \text{where } \langle p_1, p_2 \rangle \in \{\langle h, h \rangle, \langle h, l \rangle, \langle r, h \rangle, \langle r, l \rangle\}
\]

\[
\text{Head}(y, h) \land \text{Head}(y, p) \quad \text{where } p \in \{r, l\}
\]

(2) **Unique start configuration:**

\[
\text{FirstConf}(x, y) \land \text{State}_{q}(y) \quad \text{where } q \neq q_e
\]

\[
\text{FirstConf}(x, y) \land \text{FirstConf}(y, z) \land \text{Head}(z, p) \quad \text{where } p \in \{l, r\}
\]

\[
\text{FirstConf}(x, y) \land \text{ConfCell}(y, z) \land \text{Symbol}(z, c_r) \quad \text{where } c_r \neq \square
\]

(3) **Valid, uniquely defined transitions:**

\[
\text{State}_{q}(y) \land \text{Head}(z, h) \land \text{ConfCell}(y, z) \land \text{Symbol}(z, c_r) \land \text{NextConf}_{q}(y, y', y) \land \text{State}_{q'}(y') \land \text{ConfCell}(y', z') \land \text{SameCell}(z', z) \land \text{Symbol}(z', c_r') \quad \text{where } \delta = \langle q_1, \sigma_1, q_2, \sigma_2, d \rangle
\]

\[
\text{with } q_1 \neq q \text{ or } \sigma_1 \neq \sigma \text{ or } q_2 \neq q' \text{ or } \sigma_2 \neq \sigma'
\]

(4) **Unique end state:**

\[
\text{LastConf}(y) \land \text{State}_{q}(y) \quad \text{where } q \neq q_e
\]

(5) **Memory:**

\[
\text{ConfCell}(y_1, x_1) \land \text{Head}(x_1, r) \land \text{Symbol}(x_1, c_r) \land \text{NextConf}_{q}(y_1, y_2) \land \text{ConfCell}(y_1, x_2) \land \text{SameCell}(x_1, x_2) \land \text{Symbol}(x_2, c_r) \quad \text{where } \sigma \neq \sigma'
\]

\[
\text{ConfCell}(y_1, x_1) \land \text{Head}(x_1, l) \land \text{Symbol}(x_1, c_r) \land \text{NextConf}_{q}(y_1, y_2) \land \text{ConfCell}(y_2, x_2) \land \text{SameCell}(x_1, x_2) \land \text{Symbol}(x_2, c_r) \quad \text{where } \sigma \neq \sigma'
\]

(6) **Head movement:**

\[
\text{ConfCell}(y_1, x_1) \land \text{Head}(x_1, h) \land \text{NextConf}_{q}(y_1, y_2) \land \text{ConfCell}(y_2, x_2) \land \text{SameCell}(x_1, x_2) \land \text{NextCell}(x_2, x'_2) \land \text{Head}(x'_2, p) \quad \text{where } \delta = \langle q_1, \sigma_1, q_2, \sigma_2, \text{right} \rangle \text{ and } p \in \{r, l\}
\]

\[
\text{ConfCell}(y_1, x_1) \land \text{Head}(x_1, h) \land \text{NextConf}_{q}(y_1, y_2) \land \text{ConfCell}(y_2, x_2) \land \text{SameCell}(x_1, x_2) \land \text{NextCell}(x_2, x'_2) \land \text{Head}(x'_2, p) \quad \text{where } \delta = \langle q_1, \sigma_1, q_2, \sigma_2, \text{right} \rangle \text{ and } p \in \{r, l\}
\]

\[
\text{ConfCell}(y_1, x_1) \land \text{Head}(x_1, h) \land \text{NextConf}_{q}(y_1, y_2) \land \text{ConfCell}(y_2, x_2) \land \text{SameCell}(x_1, x_2) \land \text{NextCell}(x_2, x'_2) \land \text{Head}(x'_2, p) \quad \text{where } \delta = \langle q_1, \sigma_1, q_2, \sigma_2, \text{left} \rangle \text{ and } p \in \{r, l\}
\]

\[
\text{ConfCell}(y_1, x_1) \land \text{Head}(x_1, h) \land \text{NextConf}_{q}(y_1, y_2) \land \text{ConfCell}(y_2, x_2) \land \text{SameCell}(x_1, x_2) \land \text{NextCell}(x_2, x'_2) \land \text{Head}(x'_2, p) \quad \text{where } \delta = \langle q_1, \sigma_1, q_2, \sigma_2, \text{left} \rangle \text{ and } p \in \{r, l\}
\]

\[
\text{ConfCell}(y_2, x_2) \land \text{SameCell}(x_1, x_2) \land \text{FirstCell}(z, x_2) \land \text{Head}(x_2, p) \quad \text{where } \delta = \langle q_1, \sigma_1, q_2, \sigma_2, \text{left} \rangle \text{ and } p \in \{r, l\}
\]

**Figure 4:** Queries to construct a containment encoding as in Lemma 12

with the atom FirstConf[\(x, x'\)] (omitted for space reasons in Figure 4). By Proposition 2 we can express the disjunctions of all the positive queries in Figure 4 as a LinMQ\(L^k\) \(P[x]\) of polynomial size (for \(k = 0\) it is a UCQ).

If an element \(e\) in a database instance \(I\) encodes an accepting run of \(M\) in space \(s\), and \(I\) contains no other structures, then none of the queries in Figure 4 matches. Hence \(e \notin P^I\).

Conversely, assume that \(e\) encodes a tree of \(M\) quasi-configurations in space \(s\) and \(e \notin P^I\). If none of the queries in Figure 4 (1) match, the head positions of every configuration must form a sequence \(l, l, l, \ldots, l, h, r, r, \ldots, r\); hence all quasi-configurations are actually configurations. Queries (2)–(4) ensure that the first and last configuration are in the start and end state, respectively, and that each transition is matched by suitable state and tape modifications. Queries (5) ensure that tape cells that are not at the head of the TM are not modified between configurations. Queries (6) ensure that the movement of the head is consistent with the transitions, and especially does not leave the prescribed space. Note that the queries allow transitions that try to move the head beyond the tape and require that the head stays in its current position in this case. This allows the ATM to recognize the end of the tape, which is important for the Turing machines that we consider below. With all these restrictions observed, \(e\) must encode a run of \(M\) in space \(s\). □

**Proofs for Section 8**

**Lemma 14.** For any ATM \(M\), there is an MDlog query \(P_1[x]\), a LinMQ \(P_2[x]\), queries as in Definition 5 that are LinMQs, and a same-cell query that is a UCQ, such that \(P_1[x]\) and \(P_2[x]\) containment-encode accepting runs of \(M\) in exponential space.

**Proof.** Let \(M = \langle Q, \Sigma, \Delta, q_0, q_f \rangle\) with \(Q\) partitioned into existential states \(Q_3\) and universal states \(Q_4\). In order to use Lemma 12, we first construct queries \(P'_1\) and \(P'_2\) that containment-encode quasi-configuration trees of \(M\) in space \(2^\ell\) for some \(\ell\) that is linear in the size of the queries (w.r.t. to suitable queries as in Definition 5).

Our signature contains the binary predicates (distinguished from the queries of Definition 5 by using lower case
letters) firstConf, nextConf, for all δ ∈ Δ, firstCell, nextCell, bit, for all i ∈ {1, ..., ℓ}, symbol, head, as well as the unary predicates lastConf, and stateδ for all q ∈ Q.

We define P’s to be the following MDLog query that has the goal predicate Ugoal and uses two further constants 0 and 1:

\[
\begin{align*}
\text{firstConf}(x, y) \land U_conf(y) &\rightarrow U_{goal}(x) \\
\text{state}_q(x) \land \text{firstConf}(x, y) \land U_{bit_0}(y) &\rightarrow U_{conf}(y) \\
\text{bit}_i(x, 0) \land U_{bit_0}(x) &\rightarrow U_{bit}_1(x) \\
\text{bit}_i(x, 1) \land U_{bit_0}(x) &\rightarrow U_{bit}_1(x) \\
\text{symbol}(x, c_r) &\land U_{symbol}(x) \\
\text{head}(x, h) &\land U_{head}(x) \\
\text{head}(x, l) &\land U_{head}(x) \\
\text{head}(x, r) &\land U_{head}(x) \\
\text{nextCell}(x, y) &\land U_{conf}(y) \\
\text{lastCell}(x) &\rightarrow U_{goal}(x)
\end{align*}
\]

for δ = (q, σ, q′, σ′, d), with q ∈ Q3, for δ1 = (q, σ, q′, σ′, d), with q ∈ Q4, and δ1 ≠ δ2.

Using these queries, we can construct a LinMQ P as in Lemma 12 such that P1 = P′ and P2 = P′ ∨ P containment-encode accepting runs of M. □

**Lemma 15.** Assume that there is some space bound s such that, for every DTM M, there is a MDLog query P1[x] and an MQ^{k+1} query P2[x] with k ≥ 0, such that P1[x] and P2[x] containment-encode accepting runs of M in s, where the queries required by Definition 5 are MQ^{k+1} queries. Moreover, assume that there is a suitable same-cell query that is in MQ^5.

Then, for every ATM M′, there is a MDLog query P1′[x], an MQ^{k+1} P2′[x], and MQ^{k+1} queries as in Definition 5, such that P1′[x] and P2′[x] containment-encode accepting runs of M′ in space s′ ≥ 2^s. Moreover, the size of the queries for this encoding is polynomial in the size of the queries for the original encoding.

**Proof.** There is a TM M = (Q, Σ, Δ, q_0, q_a) that counts from 0 to 2^s in binary (using space s) and then halts. M can be small (constant size) since our formalization of (A)TMs allows the TMs to recognize the last tape position to ensure that the maximal available space is used. The computation will necessarily take steps > 2^s steps to complete since multiple steps are needed to increment the counter by 1. Let P1[x] and P2[x] be queries that containment-encode accepting runs of M in s, and let ConfCell, SameCell, etc. denote the respective LinMQ as in Definition 5.

Let M′ = (Q′, Σ′, Δ′, q_0′, q_a′) be an arbitrary ATM. We use the signature of P1, extended by additional binary predicates firstConf′, nextConf′, for all δ ∈ Δ′, symbol′, head′, as well as unary predicates lastConf′, and stateδ′ for all q ∈ Q′. All of these are assumed to be distinct from predicates in P1.

Let Ugoal be the goal predicate of P1, and let U_{tape} be a new unary IDB predicate. We construct the program ̃P1 from P1 as follows. For every rule of P1 that does not contain an IDB atom in its body we add the atom U_{tape}(x) to the body, where x is any variable that occurs in the rule. Intuitively speaking, the IDBs U_{tape} and U_{goal} mark the start and end of tapes of M′, which are represented by runs of M. Moreover, we
modify $\tilde{P}_1$ to “inject” additional state and head information for $\mathcal{M}'$ into configurations of $\mathcal{M}$, i.e., we extend $P_1$ to ensure that every element $e$ with state $\delta(e)$ also occurs in some symbol $\rho(e,e',\bar{\sigma})$ and in some relation head $(e,\rho)$. This can always be achieved by adding a linear number of IDB predicates and rules.

Now $P'_2$ is defined to be a MDlog query with goal predicate $U'_{goal}$ (assumed, like all IDB predicates of form $U'$ below, to be distinct from any IDB predicate in $P_1$), which is obtained as the union of $P_1$ with the following rules:

\[
\begin{align*}
\text{firstConf'}(x,y) & \land U'_{goal}(x) \rightarrow U'_{goal}(x) \\
\text{state'}(x,y) & \land U_{goal}(x) \rightarrow U_{goal}(x) & \text{for } q \in Q \\
\text{nextCell'}(x,y) & \land U_{goal}(y) \rightarrow U_{tape}(x) & \text{for } q \in Q \\
\text{nextConf'}(x,y) & \land U'_{goal}(y) \rightarrow U_{tape}(x) & \text{for } \delta = (q, q', q'', d) \\
& \text{with } q \in Q, \text{ and } \delta_1 \neq \delta_2 \\
\text{lastConf'}(x,y) & \rightarrow U_{tape}(x)
\end{align*}
\]

$P'_2$ encodes trees of trees of $\mathcal{M}$ quasi-configurations in space $s$. The structures matched by $P'_2$ but not by $P_2$ encode trees of accepting runs of $\mathcal{M}$ in space $s$ (note that these runs are linear, since $\mathcal{M}$ is not alternating). Every such run consists of the same number $s' \geq 2^s$ of configurations; these configurations represent the tape cells of our encoding of $\mathcal{M}'$ sequences. This encoding is formalized by queries as follows. The queries $\text{FirstConf'}[x,y]$, $\text{State'}[x,y]$, $\text{Head'}[x,y]$, and $\text{Symbol'}[x,y]$ are directly expressed by singleton CQs that use the eponymous predicates $\text{firstConf'}(x,y)$, etc. To access cells of $\mathcal{M}'$, we can use the analogous queries to access configurations of $\mathcal{M}$: $\text{FirstCell}[x,y] = \text{FirstConf'}(x,y)$, $\text{NextCell'}[x,y] = \text{NextConf'}(x,y)$, and $\text{LastCell'}[x] = \text{LastConf'}(x)$.

The remaining queries can be expressed as LinMQ queries. To present these queries in a more readable way, we specify them in regular expression syntax rather than giving many rules for each. It is clear that regular expressions over unary and binary predicates can be expressed in LinMQ (it was already shown that MQs can express regular path queries, which is closely related [19]). We use abbreviation $\text{P1SYMBOL}$ to express the regular expression that is a disjunction of all predicate symbols that occur in $P_1$ (this allows us to skip over any structures generated by $P_1$; with the specific forms of $P_1$ that can occur in our proofs, one could make this more specific to use only certain binary predicates, but our formulation does not depend on internals of $P_1$). Moreover, let $\text{STATE}$ be the disjunction of all atoms state$\delta_q(x)$ and $\exists y. \text{head}(x,y)$ (both unary).

\[
\begin{align*}
\text{NextConf'}[x,y] & := \text{STATE P1SYMBOL}$ nextConf$' \\
\text{LastConf'}[x] & := \text{STATE P1SYMBOL}$ lastConf$' \\
\text{ConfCell'}[x,y] & := \text{STATE P1SYMBOL}$ HEAD
\end{align*}
\]

The unary query $\text{LastConf'}[x]$ uses the variable at the beginning of the expression as its answer. It is easy to verify that the elements accepted by $P'_2$ but not by $P_2$ encode sequences of quasi-configurations of $\mathcal{M}'$ in space $s'$ with respect to these queries. To apply Lemma 12, we need to specify an additional SameCell' query for this encoding.

**SameCell'** is expressed by an MQ$^{k+1}$ query that can in general not be expressed by a MQ$^2$ query:

\[
\begin{align*}
\text{FirstCell}(x,y) & \rightarrow U_1(x) \\
\text{U}_1(x) \land \text{NextCell}(x',y') & \rightarrow U_1(x') \\
\text{State} \land \text{FirstCell}(x,y) \land \text{Symbol}(y,z) \land \text{Head}(y,v) & \rightarrow U_2(y) \\
\text{for all } q \in Q
\end{align*}
\]

where FirstCell, Symbol, SameCell, and LastCell are the queries for which $P_1$ and $P_2$ containment-encode runs of $\mathcal{M}$. Note that our constructions already ensure that the sequences of $\mathcal{M}$-cells compared by SameCell' are of the same length.

To complete the proof, we apply Lemma 12 to construct an MQ$^{k+1}$ $P_3$. The MQ$^{k+1}$ $P_3$ is obtained by expressing the disjunction of $P_2$ and $P_3$ as an MQ$^{k+1}$ using Proposition 2. Then $P_1$, $P_2$, and $P_3$ containment encode accepting runs of $\mathcal{M}'$ in space $s'$.

**Theorem 13.** Deciding containment of MDlog queries in MQ$k$ queries is hard for $(k+2)\text{ExpTime}$.

**Proof.** The claim is shown by induction on $k$. For the base case, we show that deciding containment of MQ queries is EXPTime-hard. By Lemma 14, for any DTM $\mathcal{M}$, there is a MDlog query $P_0$, a LinMQ $P_0'$, LinMQs as in Definition 5, and a same-cell query that is a UCQ with respect to which $P_0$ and $P_0'$ containment-encode accepting runs of $\mathcal{M}$ in exponential space $s$. By applying Lemma 15, we obtain, for an arbitrary ATM $\mathcal{M}$, a MDlog query $P_1$, an MQ $P_1'$, and MQ queries as in Definition 5 (including a same-cell query), that containment-encode accepting runs of $\mathcal{M}$ in space $s' \geq 2^s$.

The induction step for $k > 1$ is immediate from Lemma 15. \[\square\]