Aggregate Operators in Constraint Query Languages

Michael Benedikt
Bell Laboratories
263 Shuman Blvd
Naperville, IL 60566
E-mail: benedikt@research.bell-labs.com

Leonid Libkin*
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 3H5, Canada
E-mail: libkin@cs.toronto.edu

Abstract

We investigate the problem of how to extend constraint query languages with aggregate operators. We deal with standard relational aggregation, and also with aggregates specific to spatial data, such as volume. We study several approaches, including the addition of a new class of approximate aggregate operators which allow an error tolerance in the computation. We show how techniques of [23, 25] based on VC-dimension can be used to give languages with approximation operators, but also show that these languages have a number of shortcomings. We then give a set of results showing that it is impossible to get constraint-based languages that admit definable aggregation operators, both for exact operators and for approximate ones. These results are quite robust, in that they show that closure under aggregation is problematic even when the class of functions permitted in constraints is expanded.

This motivates a different approach to the aggregation problem. We introduce a language FO + POLY +STM, which permits standard discrete aggregation operators to be applied to the outputs of range-restricted constraint queries. We show that this language has a number of attractive closure and expressivity properties, and that it can compute volumes of linear-constraint databases.

1 Introduction

New applications of database technology, such as Geographical Information Systems, have spurred a considerable amount of research into generalizations of the standard relational model to deal with the manipulation of geometric or spatial data. One common approach to modeling spatial databases is to consider input databases as given by a set of well-behaved relations in euclidean space – for example, by a set of semi-linear or semi-algebraic sets. There are a number of proposed query languages that extend classical relational algebra to this setting, languages that allow the use of various geometric operations in manipulating spatial databases. One of the most well-developed models for spatial queries is the constraint database model [22, 27]. In this model, spatial databases are represented as sets of linear or polynomial constraints. Databases are queried using standard relational calculus with linear (resp. polynomial) inequalities as selection criteria, see [3, 4, 5, 19, 20, 31, 36]. These languages, denoted by FO + LIN and FO + POLY, have become the dominant ones in the constraint database literature. They have a very important closure property: the application of a FO + LIN query to a linear constraint set yields a new set of linear constraints; similarly FO + POLY queries on sets definable with polynomial constraints produce sets that can still be defined with polynomial constraints.

*Contact author. Research affiliation: Bell Laboratories.
Constraint Query Languages, then, give a natural analog of relational calculus in the geometric context. A crucial question, though, concerns how to extend standard aggregation constructs from the relational model to the geometric setting. This question has two components. First, we would like our languages to be able to apply standard SQL operators such as TOTAL and AVG to spatial queries, whenever these operators make sense. Since the output of queries in constraint query languages (and in other spatial query languages) may be merely finitely representable (that is, representable by some finite means, e.g., a finite set of constraints) and not finite, the aggregation operators cannot be allowed to be applied to any constraint query output. One problem then, is to design a language that allows the safe application of classical aggregates.

The second component of the ‘aggregation question’ concerns aggregation notions that are specific to the spatial databases. Most commonly, given a database and the output of a query over it, one wishes to form new queries about the volume of this output. One may also extend standard aggregates such as AVG, and ask for the average value of a polynomial over a spatial object. Such aggregates arise both from practical concerns of GIS, and also as the natural continuous analogs of classical aggregation queries. Thus, we would like to extend constraint query languages to incorporate the ability to calculate volumes and other aggregates arising in the spatial setting.

In attempting to add aggregation to constraint query languages, one immediately encounters some daunting obstacles. While standard constraint databases are closed under first-order operations such as join and projection, they are clearly not closed under taking of volumes. This fact is well-known in the literature [22, 26, 11], and stems from the fact that neither the semi-linear nor semi-algebraic sets are closed under integrals. To take an example from the semi-algebraic setting, a query asking for the volume of initial slices of the epigraph of $1/x$ outputs the graph of the $\ln$ function, while iterating volume queries in this fashion would give as output transcendental functions that are not even expressible using field operations, logarithms and exponents. Thus, one cannot hope to add a general volume operator to existing first-order constraint query languages such as FO + POLY and get a closed language while still remaining within the domain of polynomial constraint databases.

There are several approaches to the volume problem mentioned above. First, one could weaken the requirement that volumes be computed exactly and instead aim only to compute approximate volumes. Thus a query might have a tolerance associated with each instance of a volume operator, with output required only to be correct within the given tolerance. There are a number of practical and theoretical motivations for this approach. While it is known that computing volumes of even simple geometric objects (convex polytopes) is a hard problem (\#P-hard, see [13]), approximation of volumes, at least of convex sets, can be done in polynomial time by a randomized algorithm [14]. Moreover, in contrast to the well-known fact that semi-algebraic and semi-linear sets are not closed under volume operators, the papers [23, 24, 25] show that volumes of sets definable with polynomial constraints can be approximated, for any given $\varepsilon > 0$, by a first-order formula with polynomial constraints. By giving up exact volume and settling for an approximation, one might hope to retain desirable closure properties.

A second approach to the aggregation problem would be to expand out of the domain of polynomial constraints, and add new functions to the signature of both the constraints and the query language. This would give the possibility of retaining a constraint-based representation of databases, while gaining closure under volume operators. Of course, in this approach one should expand the constraint set so that it still defines only topologically well-behaved objects.

A third approach to the volume problem is to search for languages which can compute or approximate the volumes of important classes of sets, but which may not be closed under iterative application of
volume operators. For example, one could allow volume and other aggregation operators to be applied only to a subclass of the input queries. Restrictions on the nesting of volume operators would then have to be imposed.

An example of this last approach in the existing literature is [10], where it is shown that polynomial constraint query languages can express the (exact) volume for any set that admits a special condition called ‘variable independence’. This condition means, informally, that in the constraint specification of sets in, say, \( \mathbb{R}^2 \), there is no interaction between \( x \) and \( y \). Unfortunately, this condition is too restrictive: it excludes many of the sets that arise most often in spatial applications. As for practical applications of aggregation in constraint databases, implemented systems normally do not have aggregate operations (the Dedale system, for example). One spatial extension of SQL, proposed in [28], does include aggregates, but a careful examination of the language shows that they are severely restricted: the only allowed aggregate operations are traditional relational ones applied to finite relations, and spatial union and intersection, which are first-order definable.

In this paper, we analyze the feasibility of each of the above approaches in detail. For the first approach, we show that techniques based on VC dimension, coming out of the work of [23, 24, 25] give us approximate volume operators that give semi-algebraic output on semi-algebraic input. However, we show a number of shortcomings of such an approach. Not only are the approximate volume operators obtained according to the technique of [23, 24, 25] sensitive to the input representation, but the blow-up in the size of the constraint databases produced in query evaluation precludes any possible use of these operators in practice.

Turning to the second approach, we show that it is completely infeasible. No first-order constraint language based on any reasonably well-behaved class of functions can express, or even approximate, volume. In the process of showing this, we develop a new set of techniques for proving inexpressibility results, techniques not based on the usual method of reduction to generic queries.

We then consider solutions that give up full closure under volume, and give a number of positive results. We present a higher-order language that allows one to calculate the volume of arbitrary semi-linear sets. Specifically, we give a language, called FO + POLY + SUM, that has attractive closure properties, remains within the domain of polynomial constraint databases, and allows the exact calculation of volumes for linear-constraint input databases. This language also has the pleasant feature that it is closed under the classical aggregation operators SUM and AVG. Since FO + POLY + SUM includes SQL aggregation, contains FO + POLY, and also allows one to make use of standard aggregation evaluation techniques in calculating volumes, it seems to be a good candidate for the constraint analog of classical aggregation languages.

We remark that another approach to the aggregation problem was considered in [11], which gave a new aggregate operator \( \mu \), under which FO + LIN is closed. However, \( \mu(X) = 0 \) for any bounded set \( X \); thus, this operator cannot be used to deal with volumes.

**Organization** Section 2 introduces the notation. Approximability is studied in Section 3. The method of defining approximate volumes of [23, 24, 25] is analyzed, and the main difficulties in applying the approximation operators coming from this work are outlined. Section 4 shows that approximate volume operators cannot be defined in first-order constraint languages, even when the signature is expanded. Section 5 defines an extension of FO + POLY with SQL-like aggregation (summation over finite sets) and shows that this extension can express volumes of semi-linear databases.

The extended abstract of this paper appeared in the Proceedings of the 18th ACM Symposium on
Principles of Database Systems [7].

2 Notation

Structures, instances, queries  Most notations are fairly standard in the literature on constraint databases, cf. [4, 5, 27, 31, 19]. Let $\mathcal{M} = \langle \mathcal{U}, \Omega \rangle$ be an infinite structure, where $\mathcal{U}$ is an infinite set, called a universe (in the database literature often called the domain), and $\Omega$ is a set of interpreted functions, constants, and predicates. In the field of constraint databases, most examples have $\mathcal{U} = \mathbb{R}$, the set of real numbers. Examples of signatures (and corresponding classes of constraints) that have been considered are:

- **Dense Order Constraints:** $\langle \mathbb{R}, < \rangle$;
- **Linear Constraints:** $\mathbf{R}_{\text{lin}} = \langle \mathbb{R}, +, -, 0, 1, < \rangle$;
- **Polynomial Constraints:** $\mathbf{R} = \langle \mathbb{R}, +, \ast, 0, 1, < \rangle$;
- **Exponential Constraints:** $\mathbf{R}_{\text{exp}} = \langle \mathbb{R}, +, \ast, e^x, < \rangle$.

A (relational) database schema $SC$ is a nonempty collection of relation names $\{S_1, \ldots, S_t\}$ with associated arities $p_1, \ldots, p_t > 0$. We shall consider both finite and finitely representable instances. Given $\mathcal{M}$, an finite instance of $SC$ over $\mathcal{M}$ is a family of finite sets, $\{R_1, \ldots, R_t\}$, where $R_i \subseteq \mathcal{U}^{p_i}$. That is, each schema symbol $S_i$ of arity $p_i$ is interpreted as a finite $p_i$-ary relation over $\mathcal{U}$. Given a finite instance $D$, $\text{adom}(D)$ denotes its active domain, that is, the set of all elements that occur in the relations in $D$.

A finitely-representable (f.r.) instance of $SC$ over $\mathcal{M}$ is a family of sets $\{X_1, \ldots, X_t\}$, with $X_i \subseteq \mathcal{U}^{p_i}$, such that for each $X_i$ there exists a quantifier-free formula $\alpha_i(x_1, \ldots, x_{p_i})$ in the language of $\mathcal{M}$ with $X_i = \{\bar{a} \in \mathcal{U}^{p_i} \mid M \models \alpha_i(\bar{a})\}$. Most applications of constraint databases consider f.r. instances defined over $\mathbf{R}_{\text{lin}}$ (these are called semi-linear sets) or over $\mathbf{R}$ (called semi-algebraic sets). For example, in the spatial setting, a f.r. instance interprets the schema predicates as semi-linear or semi-algebraic sets.

As our basic query language, we consider relational calculus, or first-order logic, FO, over the underlying structure and the database schema. In what follows, $L(SC, \Omega)$ stands for the language that contains all symbols of $SC$ and $\Omega$; $\text{FO}(SC, \Omega)$ is the class of all first-order formulae built up from the atomic $SC$ and $\Omega$ formulae by using Boolean connectives $\lor$, $\land$, $\neg$ and quantifiers $\forall, \exists$.

Regardless of whether we are in the ‘classical’ setting, where these queries are applied to finite databases, or in the constraint query setting, we will refer to the above syntactic query languages as relational calculus with $\Omega$ constraints. This will be denoted by $\text{FO} + \text{LIN}$, $\text{FO} + \text{POLY}$ and $\text{FO} + \text{EXP}$.

In the case of finite databases, we shall also use the active-domain quantifiers: for a formula $\varphi(x, \bar{y})$, one can form formulae $\exists x \in \text{adom} \varphi(x, \bar{y})$ and $\forall x \in \text{adom} \varphi(x, \bar{y})$. For a structure $\mathcal{M}$ and a $SC$-instance $D$, the notion of $\langle \mathcal{M}, D \rangle \models \varphi$ is defined in a standard way for $\text{FO}(SC, \Omega)$ formulae, where $\langle \mathcal{M}, D \rangle \models \exists x \varphi(x, \cdot)$ means for some $a \in \mathcal{U}$ we have $\langle \mathcal{M}, D \rangle \models \varphi(a, \cdot)$, and $\langle \mathcal{M}, D \rangle \models \exists x \in \text{adom} \varphi(x, \cdot)$ means that for some $a \in \text{adom}(D)$ we have $\langle \mathcal{M}, D \rangle \models \varphi(a, \cdot)$. If $\mathcal{M}$ is understood, we write $D \models \varphi$.

Given $\varphi(\bar{x}, \bar{y})$ and $\bar{a}$, we write $\varphi(\bar{a}, D)$ for $\{\bar{b} \mid D \models \varphi(\bar{a}, \bar{b})\}$; in the absence of $\bar{x}$ we just write $\varphi(D)$ for the output of $\varphi$ on $D$.

The class of subformulae of $\text{FO}$ that only use the active-domain quantification is denoted by $\text{FO}_{\text{act}}$. 
Adding aggregate operators We shall use $\text{Vol}(X)$ to denote the volume of a set $X \subseteq \mathbb{R}^n$. More precisely, $\text{Vol}(X)$ is the measure of any Lebesgue-measurable set $X \subseteq \mathbb{R}^n$. We shall not worry about dealing with non-measurable sets, as all bounded sets defined with constraints relevant for spatial applications (those listed above, plus some extensions) are measurable.

We shall consider adding volume to a query language as follows. If $\varphi(\vec{x}, \vec{y})$ is a formula, then the following is a formula with free variables $\vec{x}, z$:

$$[\text{Vol} \ y, \varphi(\vec{x}, \vec{y})](\vec{x}, z)$$

Assume that a structure $\mathcal{M} = (\mathbb{R}, \Omega)$ is fixed. Let an instance (finite or f.r.) $D$ be given. Then

$$D \models [\text{Vol} \ y, \varphi(\vec{x}, \vec{y})](\vec{a}, v) \iff v = \text{Vol}(\varphi(\vec{a}, D)).$$

Recall that $\varphi(\vec{a}, D) = \{\vec{b} \mid D \models \varphi(\vec{a}, \vec{b})\}$.

The extension of any query language $\mathcal{L}$ with Vol will be denoted by $\mathcal{L} + \text{Vol}$; for example, one can speak of $\text{FO} + \text{LIN} + \text{Vol}$ or $\text{FO} + \text{POLY} + \text{Vol}$. Of course we know that due to the nonclosure results mentioned in the introduction, $\text{FO} + \text{LIN} \subseteq \text{FO} + \text{LIN} + \text{Vol}$ and $\text{FO} + \text{POLY} \subseteq \text{FO} + \text{POLY} + \text{Vol}$.

As the next step, we restrict our attention to bounded sets. Without any loss of generality, we shall deal with subsets of $I^n \subset \mathbb{R}^n$, where $I$ throughout this paper denotes the interval $[0, 1]$. We define $\text{Vol}_I \ y, \varphi(\vec{x}, \vec{y})$ just as above, except that now we require that $v = \text{Vol}(\varphi(\vec{a}, D) \cap I^n)$. In particular, $0 \leq v \leq 1$. We similarly define languages $\mathcal{L} + \text{Vol}_I$. As with Vol, languages like $\text{FO} + \text{LIN}$ and $\text{FO} + \text{POLY}$ are not closed under $\text{Vol}_I$: for example, $\arctan(x) = \int_0^x \frac{dy}{y^2+1} = \text{Vol}_I(\{(y, z) \mid (0 \leq y \leq x) \wedge (0 \leq z \leq 1/(y^2+1))\})$, for $0 \leq x \leq 1$.

As standard languages are not closed under taking volume, we address the question of whether one can obtain closure by lowering one's demands. In particular, we would like to see if approximating the volume, rather than computing it directly, gives us a closed language. The hope that closure might be obtained in this way is motivated by the fact that for every formula $\varphi(\vec{x}, \vec{y})$ in $\mathcal{R}$ and for every $\epsilon > 0$, one can find a formula $\psi_\epsilon(\vec{x}, z)$ that gives $\epsilon$-approximation of volumes of sets $\varphi(\vec{a}, \mathbb{R}) = \text{Vol}_I(\{(\vec{b} \mid \models \varphi(\vec{a}, \vec{b}))\})$, see [23, 24, 25].

We have to explain what we mean by approximating volume in this context. Clearly, we cannot hope to find $\psi_\epsilon(\vec{x}, z)$ with $z$ defining an $\epsilon$-interval around the real value of the volume - then the volume itself would be definable as the center of the interval! Thus, we settle for less. Similar to [23, 24, 25], we say for every $\epsilon > 0$, that an operator $\text{Vol}_\epsilon$ is an $\epsilon$-approximation operator if for every f.r., over $\mathcal{M}$, set $A \subset \mathbb{R}^n \times \mathbb{R}^m$, given by a formula $\varphi(\vec{x}, \vec{y})$, $\text{Vol}_\epsilon$ returns a f.r. set in $\mathbb{R}^n \times \mathbb{R}$, given by $\psi_\epsilon(\vec{x}, z)$ such that :

1. For every $\vec{a} \in \mathbb{R}^n$, $\psi_\epsilon(\vec{a}, \cdot)$ must be satisfiable (that is, $\mathcal{M} \models \exists z. \psi_\epsilon(\vec{a}, z)$);
2. If $\mathcal{M} \models \psi_\epsilon(\vec{a}, v)$, then $v \geq 0$ and $|v - \text{Vol}(\varphi(\vec{a}, \mathbb{R}))| < \epsilon$.

Thus, $\text{Vol}_\epsilon$ must return a $\psi_\epsilon$ that is guaranteed to find an (absolute) $\epsilon$-approximation of the volume. We next say that a query language $\mathcal{L}$ defines $\text{Vol}_\epsilon$, if there is a query in $\mathcal{L}$ that defines such an operator. That is, for each query $\varphi(\vec{x}, \vec{y})$ in $\mathcal{L}$ and $\epsilon > 0$ there is a $\mathcal{L}$-query $\psi_\epsilon(\vec{x}, z)$ such that for any database $D$, and any $\vec{a}$, we have

- $D \models \exists z. \psi_\epsilon(\vec{a}, z)$, and

5
\begin{itemize}
  \item $D \models \psi_\varepsilon(\bar{a}, v)$ implies $v \geq 0$ and $|v - \text{Vol}(\varphi(\bar{a}, D))| < \varepsilon$.
\end{itemize}

Notice that in the last definition $\psi_\varepsilon$ is independent of $D$. Also notice that to show definability of approximate operators in standard query languages, it suffices to show that there is a query in the language returning the $\varepsilon$-approximate volume on every base relation of some fixed arity.

We also define $\varepsilon$-approximation operators to volume in the case where we restrict to bounded sets. As before, we use, w.l.o.g., $I^n$ as bounding set. An $\varepsilon$-approximation operator in the bounded setting is denoted by $\text{Vol}_\varepsilon$. Such an operator must satisfy the variant of condition 2 above: $|v - \text{Vol}(\varphi(\bar{a}, D) \cap I^n)| < \varepsilon$ and $0 \leq v \leq 1$.

These operators, and their definability in query languages, are studied in Sections 3 and 4.

**O-minimality, VC dimension** Many results that we prove extend beyond linear and polynomial constraints. To state them in greater generality, we shall use $o$-minimality [35], which plays an important role in the study of constraint query languages (cf. [4, 5, 6]).

A structure $\mathcal{M} = \langle \mathcal{U}, \Omega \rangle$ is $o$-minimal, if every definable set is a finite union of points and open intervals $(a, b) = \{x \mid a < x < b\}$, $(-\infty, a) = \{x \mid x < a\}$, and $(a, \infty) = \{x \mid x > a\}$ (we assume that $< \in$ is in $\Omega$). Defines sets are those of the form $\{x \mid \mathcal{M} = \varphi(x)\}$, where $\varphi$ is a first-order formula in the language of $\mathcal{M}$, possibly supplemented with symbols for constants from $\mathcal{M}$. All the structures on the reals we mentioned so far – $\text{Re}_{\text{lin}}, \mathbb{R}, \mathbb{R}_{\text{exp}}$ – are $o$-minimal (the first two by Tarski’s quantifier-elimination, the last one by [37]).

If $\mathcal{M} = \langle \mathbb{R}, \Omega \rangle$, we define $\mathcal{M}_{+,*}$ to be $\langle \mathbb{R}, \Omega, +, * \rangle$. We often require that not just $\mathcal{M}$ but also $\mathcal{M}_{+,*}$ be $o$-minimal.

We also consider structures having finite VC dimension of definable families [2, 29] (also known as structures without the independence property [34]). VC dimension, introduced in statistics to study uniform convergence of stochastic processes, has become central to computational learning theory [2, 9], and found application in other areas, e.g., complexity [30].

Suppose $X$ is an infinite set, and $C \subseteq 2^X$. Let $F \subseteq X$ be finite; we say that $C$ shatters $F$ if the collection $\{F \cap C \mid C \in C\}$ is $2^F$. The Vapnik-Chervonenkis (VC) dimension of $C$, $\text{VCdim}(C)$, is the maximal cardinality of a finite set shattered by $C$. If arbitrarily large finite sets are shattered by $C$, we let $\text{VCdim}(C) = \infty$.

Let $\mathcal{M} = \langle \mathcal{U}, \Omega \rangle$, and let $\varphi(\bar{x}, \bar{y})$ be a formula in the language of $\mathcal{M}$ with $|\bar{x}| = n, |\bar{y}| = m$. For each $\bar{a} \in \mathcal{U}^n$, define $\varphi(\bar{a}, \mathcal{M}) = \{\bar{b} \in \mathcal{U}^m \mid \mathcal{M} \models \varphi(\bar{a}, \bar{b})\}$, and let $F_\varphi(\mathcal{M})$ be $\{\varphi(\bar{a}, \mathcal{M}) \mid \bar{a} \in \mathcal{U}^n\}$. Families of sets arising in such a way are called definable families. We say that $\mathcal{M}$ is a structure with finite VC dimension if the VC dimension of each definable family is finite. Every $o$-minimal structure is a structure with finite VC dimension [29], and the latter class is in fact much larger than the class of $o$-minimal structures.
3 Approximating aggregates in constraint query languages

3.1 The VC dimension-based implementation of approximate volume operators

We now start our investigation of the expressibility of approximate volume operators. The results of [23, 24, 25] do immediately give a closed language for computing approximate volumes. From [23, 24, 25] we can easily derive:

**Theorem 1** Let $\epsilon > 0$, and let $\varphi(\bar{x}, \bar{y})$ be a FO + POLY query. Then for every semi-algebraic (resp. semi-linear) database instance $D$ there exists a formula $\varphi^*_D(\bar{x}, z)$ over the real ordered field $\mathbb{R}$ (resp. group $\mathbb{R}_{\text{lin}}$) such that $\varphi^*_D(\bar{a}, \cdot)$ is satisfiable for all $\bar{a}$, and $\varphi^*_D(\bar{a}, v)$ implies $|v - \text{Vol}_I(\varphi(\bar{a}, D))| < \epsilon$ and $0 \leq v \leq 1$. Hence, there is a collection of $\epsilon$-approximation operators $\text{Vol}_I^\epsilon$, $\epsilon > 0$, for $\mathbb{R}$ and $\mathbb{R}_{\text{lin}}$.

Since we want to examine those operators with regard to their efficiency, we now review the ideas of [23, 24, 25] that lead to this theorem.

**Pre-requisites** (see [2, 9, 23, 25]) The idea of the approximation technique can be traced back to the simplest randomized method for computing volumes. For a set $S \subseteq \mathbb{R}^n$, take $k$ points $x_1, \ldots, x_k$ from the uniform distribution on $\mathbb{R}^n$. Then $\text{Vol}(S)$ can be approximated as $v_S = \sum_{i=1}^k \chi_S(x_i)/k$, where $\chi_S$ is the characteristic function of $S$: $\chi_S(x) = 1$ if $x \in S$ and $\chi_S(x) = 0$ if $x \notin S$. Then for $\epsilon > 0$,

$$P(|v_S - \text{Vol}(S)| \geq \epsilon) < 2e^{-2k\epsilon^2};$$

this follows from Hoeffding’s inequality. There are two reasons why this is not sufficient for getting $\epsilon$-approximations to volume. First, the volume operators, as we defined them, may depend on parameters. Indeed, $\text{Vol} \varphi(\bar{x}, \bar{y})$ requires computing the volume for every instantiation of parameters $\bar{x}$. Secondly, the randomized method above only tells us that $|v_S - \text{Vol}(S)| < \epsilon$ with high probability, and thus the procedure must be derandomized to ensure a certain answer.

To overcome the first problem, we use techniques from statistics and machine learning to ensure that one sample will suffice to test multiple volumes. Let $\varphi(\bar{x}, \bar{y})$ be a first-order formula over the real field $\mathbb{R}$, with $|\bar{x}| = n$ and $|\bar{y}| = m$, and let $\epsilon, \delta > 0$. Define $\varphi(\bar{a}, \bar{R}) = \{\bar{b} \in \mathbb{R}^m \mid \bar{R} \models \varphi(\bar{a}, \bar{b})\}$. Let $M > 0$ be given, and assume that an $M$-point sample $C = \{\bar{c}_1, \ldots, \bar{c}_M\}$ is randomly chosen in $\mathbb{R}^m$. For each $\bar{a}$, let $v(\bar{a}, C)$ be the fraction of $C$ that falls into $\varphi(\bar{a}, \bar{R}) \cap \mathbb{R}^m$. Then one wants to achieve $|v(\bar{a}, X) - \text{Vol}_I(\varphi(\bar{a}, \bar{R}))| < \epsilon$ for all $\bar{a} \in \mathbb{R}^n$, with probability at least $1 - \delta$.

The classical results of learning theory [2, 9] say that this is possible when the VC dimension of the family $F_\varphi(\bar{R}) = \{\varphi(\bar{a}, \bar{R}) \mid \bar{a} \in \mathbb{R}^n\} \subseteq 2^{\mathbb{R}^m}$ is finite, and the size of the sample of $M$ is then proportional to the VC dimension. In the construction of approximating formulae, we shall use the following corollary of this result, that states the existence of so-called $\epsilon$-nets:

**Fact 1** ($\epsilon$-nets) Let $\varphi(\bar{x}, \bar{y})$ be a first-order formula over the real field $\mathbb{R}$, with $|\bar{y}| = m$, and let $\epsilon > 0$. Let $d = \text{VCdim}(F_\varphi(\bar{R}))$. If $M \geq \frac{\log \frac{1}{\delta}}{\epsilon^2}$, then there exists an $M$-element set $C = \{\bar{c}_1, \ldots, \bar{c}_M\} \subseteq \mathbb{R}^m$ such that for every $\bar{a}$ with $\text{Vol}(\varphi(\bar{a}, \bar{R}) \cap \mathbb{R}^m) \geq \epsilon$ it is the case that $\varphi(\bar{a}, \bar{R}) \cap \mathbb{R}^m \cap C \neq \emptyset$. □
**Approximation method** We now combine the existence of \( \epsilon \)-nets with a derandomization procedure.

Assume that we are given a \( \text{FO}+\text{POLY} \) query \( \psi(\vec{x}, \vec{y}) \) and a semi-algebraic database \( D \). Put the definition of \( D \) into \( \psi \), to obtain a new formula \( \varphi(\vec{x}, \vec{y}) \) in the language of the real field, such that \( R \models \varphi(\vec{a}, \vec{b}) \) iff \( D \models \psi(\vec{a}, \vec{b}) \). For example, if \( \psi(x, y) \equiv \exists u \ (S(x, y, u) \land x < 0) \) and \( S \) is defined as \( p(x, y, u) > 0 \), where \( p \) is a polynomial, then \( \psi(x, y) \) is \( \exists u \ (p(x, y, u) > 0 \land x < 0) \).

Thus, we have to define approximating formulae for a formula \( \varphi(\vec{x}, \vec{y}) \) over \( R \). To simplify notations, write \( \Phi(\vec{a}) \) for \( \varphi(\vec{a}, \vec{R}) \cap I^m \). Fix a number \( k \in \mathbb{N} \) and \( \nu, \delta \in I \). Define, for each \( \vec{a} \in \mathbb{R}^n \),

\[
S(\vec{a}, \nu, \delta) = \{ (\vec{c}_1, \ldots, \vec{c}_k) \in (I^m)^k \mid \frac{1}{k} \sum_{i=1}^{k} \chi_{\Phi(\vec{a})}(\vec{c}_i) - \nu \leq \delta \}.
\]

That is, \( S(\vec{a}, \nu, \delta) \) is the set of \( k \)-samples that produce an approximate volume of \( \Phi(\vec{a}) \) within \( \delta \) of \( \nu \). Note that for every fixed \( k \), this set is definable with parameters \( \vec{a}, \nu, \delta \).

Next, define an operation \( \oplus : I^n \times I^n \to I^n \) by \( x \oplus y = (x + y) \mod 1 \), with the mod 1 operation applied component-wise. Let \( \ominus \) be the inverse \( x \ominus y = z \) iff \( x = y \ominus z \). These operations can then be extended to sets: \( x \ominus S = \{ x \in y \mid y \in S \} \).

Let \( \vec{c} \in (I^m)^k \) be a \( k \)-element sample of points in \( I^m \). Define \( T(\vec{c}, \vec{a}, \nu, \delta) = \vec{c} \in S(\vec{a}, \nu, \delta) \). For a fixed \( k \), this is definable if \( \Phi \) is FO over the real field.

Fix now \( \vec{c} \in I^m \), and define the family \( \vec{c} \in \Phi(\vec{a}) \). As this is a definable family over \( R \), it has finite VC dimension \([15, 29]\), which we denote by \( d \). Then \([23, 24, 25]\) calculate an upper bound on the VC dimension of the family \( \mathcal{T} \) of all sets \( T(\vec{c}, \vec{a}, \nu, \delta) \) as \( 4d \cdot k \log k \), for each fixed \( \nu \) and \( \delta \).

Applying Fact 1, we obtain that for \( M \geq \frac{32dk \log k \log 13}{\epsilon^2} \), there is an \( \epsilon \)-net \( \{t_1, \ldots, t_M\} \) for \( \mathcal{T} \). As translation by \( \ominus \) does not change the volume, we see that all elements of \( \mathcal{T} \) have the same volume; thus, if this volume is \( \geq \epsilon \), then every member of \( \mathcal{T} \) contains one of the \( t \)'s. From this one derives that the sets \( \vec{t}_i \ominus S(\vec{a}, \nu, \delta) \) cover the entire \( (I^m)^k \), if the volume of \( S(\vec{a}, \nu, \delta) \) is at least \( \epsilon \).

By calculations based on Hoeffding’s inequality, \([23, 24, 25]\) show that the inequality \( \text{Vol}(S(\vec{a}, \nu, \epsilon/2)) > 2e^{-k\epsilon^2/2} \) implies \( |\nu - \text{Vol}(\Phi(\vec{a}))| \leq \epsilon \), and that \( |\nu - \text{Vol}(\Phi(\vec{a}))| \leq \epsilon/4 \) implies \( \text{Vol}(S(\vec{a}, \nu, \epsilon/2)) > 1 - 2e^{-k\epsilon^2/8} \). Using this, one arrives at the following.

**Proposition 1** (Karpinski, Macintyre, Koiran) Let

\[
0 < \epsilon \leq 1/2, \quad k \geq \frac{8 \cdot \ln 4}{\epsilon^2}, \quad M \geq \max\left(\frac{1}{2ek^2/2}, \frac{32dk \log k}{\epsilon^2} \cdot \log 13\right)
\]

where \( d \) is the VC dimension of the family \( \vec{c} \in \Phi(\vec{a}) \), \( \vec{c} \in I^m \). Then the formula saying that \( M \) translates (by \( \ominus \)) of \( S(\vec{a}, \nu, \epsilon/2) \) cover \( (I^m)^k \) defines \( \nu \) as an \( \epsilon \)-approximation of \( \text{Vol}(\Phi(\vec{a})) \). \( \square \)

As \( d \) is finite and depends only on \( \varphi \), the statement of the proposition can be converted into a \( \text{FO} \)-definition, which serves as an approximating definition of volumes \( \varphi(\vec{x}, R) \). Note that the resulting approximating formulae satisfy a rather strong condition: every \( \nu \) within \( \epsilon/4 \) of the real volume is returned by the approximating formulae. Also note that the approximating formulae have entirely semi-linear character - multiplication is never used except in the formula \( \varphi \) itself. We thus obtain Theorem 1 as a corollary of the above results.

8
3.2 Shortcomings of the approximation technique

We note here some shortcomings of the technique of Lemma 1 in the context of constraint databases. In the technique, one has to put the definition of a constraint database $D$ into a query $\varphi$, and then apply the method of [23, 24, 25] to the result. That method produces an output formula whose size is a polynomial in the input formula and $\frac{1}{\epsilon}$, theoretically, a nice bound. In attempting to apply this technique in practice, however, we find that the bounds obtained are rather unpleasant, even for modest values of $\epsilon$, as the size of the quantifier prefix is quite large. In the constraint database setting, those will have to be eliminated, via a quantifier-elimination procedure, which will be very costly. Let us illustrate this by a simple example.

Example: Let the schema contain one unary predicate $U$ interpreted as a subset of $[0,1]$. The query $\varphi(x_1,x_2; y_1,y_2)$ is given by

$$\forall \forall x_1 \forall x_2 \forall y_1 \forall y_2 \forall 0 \leq y_2 \leq y_1 \quad (x_1 \land x_2 \land x_1 < y_1 \land x_2 < x_1),$$

For $a, b \in U$, we have $\text{Vol}(\varphi(a, b, R)) = (b^2 - a^2)/2$.

Let $\epsilon = 1/10$. We want to evaluate the query

$$[\text{Vol}_\epsilon \varphi(x_1, x_2; y_1, y_2)](x_1, x_2, z)$$

saying that $z$ is an $\epsilon$-approximation to the volume of $\varphi(x_1, x_2, U) = \{ (y_1, y_2) \mid U \models \varphi(x_1, x_2; y_1, y_2) \}$, where $\text{Vol}_\epsilon$ is the operator obtained through the method above. Note that $\text{Vol}_1(\varphi(a, b, U)) = (b^2 - a^2)/2$, for $a < b$ in $U$, and 0 otherwise. To evaluate this query on a database where $U$ consists of $N$ elements, by applying Theorem 1, we would first plug $U$ in $\varphi$ to obtain a formula with $> 2N$ atomic subformulae that does not mention $U$.

We then use bounds of Proposition 1 and obtain, by simple calculations:

$$k > 1, 109 \quad M > 25, 206, 250.$$
4 Uniformly definable volume operators and expansion of the signature

We saw in the last section that the main shortcoming of all known examples of approximate volume operators was the blow-up in the size of the representation. It was also left open whether some volume approximation operators can be defined in standard languages, like FO + POLY, uniformly for all database instances. We now investigate whether we can find other approximation methods that can be expressed in nicely-behaved languages and that admit low complexity evaluation techniques. The main result is that one cannot capture approximate volume operators in a nice constraint language such as FO+POLY. That is,

**Inexpressibility of Approximate Operators** FO + LIN, FO + POLY and FO + EXP cannot express Vol^n for any ε < 1/2.

In fact, we prove a stronger result. Theorem 3 shows that even if one extends the constraint signature to include functions beyond FO + EXP, as long as we stay within a well-behaved structure, we cannot capture approximate volume. Furthermore, we show that in languages like FO + POLY, only trivial approximations are possible. An example of a trivial approximation is returning 1/2 for every subset of \( I^n \) - in this case we know that the difference between the real volume and its approximation is \( \leq 1/2 \).

Proving expressivity bounds such as Theorem 3 and Corollary 1 is not very simple. Almost all, if not all, existing expressivity bounds for constraint query languages either involve generic queries (e.g., the parity test, see [4, 5, 31, 3]) or are proved by reduction to generic queries (e.g., [20]). However, queries involving approximation defined as in Section 2 are extremely nongeneric. We introduce the main ideas for the proof in several steps. We first consider an easier case of the aggregate \( \text{Avg} \) for finite instances and prove that it can be neither defined nor approximated in languages like FO + POLY. The proof introduces the idea of reduction to what we call a \((c_1, c_2)\)-separating sentence, with \( c_1, c_2 \) being constant real numbers. We then show how the same reduction easily proves that FO + POLY and the likes cannot produce relative approximations of Vol. For the absolute approximation Vol^n, the reduction only works under very special assumptions on the input, and to conclude the proof we need to use results from circuit complexity.

This section gives further evidence that if one wants to stay within a reasonable (for spatial applications) class of constraints, one must give up uniform closure under any nontrivial approximation to the volume.

**Prerequisites: Collapse results** We shall need the following two results proved in [4, 5], stated here in a form most convenient for the proofs below.

**Fact 2** a) Given any ordered structure \( \mathcal{M} = \langle U, \Omega \rangle \), an infinite set \( X \subseteq U \), and an active-semantics query \( \varphi(\bar{x}) \) in \( \text{FO_{act}}(\text{SC}, \Omega) \), there exists an infinite set \( Y \subseteq X \) and a \( \text{FO_{act}}(\text{SC}, <) \) query \( \psi(\bar{x}) \) such that \( D \models \varphi(\bar{a}) \iff D \models \psi(\bar{a}) \) whenever \( \text{dom}(D) \cup \bar{a} \subseteq Y \).

b) Let \( \mathcal{M} = \langle U, \Omega \rangle \) be \( \omega \)-minimal, and \( \varphi(\bar{x}) \) and arbitrary natural-semantics \( \text{FO}(\text{SC}, \Omega) \) query. Then there exists an expansion \( \mathcal{M}' = \langle U', \Omega' \rangle \) and an active-semantics query \( \psi(\bar{x}) \) in \( \text{FO_{act}}(\text{SC}, \Omega') \) such that for every \( \text{SC} \)-database \( D \) over \( U \) and for every \( \bar{a}, D \models \varphi(\bar{a}) \iff D = \psi(\bar{a}) \). Furthermore, each relation in \( \Omega' - \Omega \) is interpreted as a set definable over \( \mathcal{M} \); thus, if \( \mathcal{M} \) admits quantifier-elimination,
one can take $\mathcal{M}'$ to be $\mathcal{M}$.

Separating sentences We shall consider a relational database schema $SC$ that consists of two unary relations, $U_1$ and $U_2$. Let $c_1, c_2 > 1$ be two real numbers. We say that $\Phi$ is a $(c_1, c_2)$-separating sentence if for any finite instance $D$ of $SC$, it is the case that $\text{card}(U_1) > c_1 \cdot \text{card}(U_2)$ implies $D \models \Phi$ and $\text{card}(U_2) > c_2 \cdot \text{card}(U_1)$ implies $D \models \neg \Phi$. Note that this definition says nothing about the case when $\frac{1}{c_2} \cdot \text{card}(U_2) \leq \text{card}(U_1) \leq c_1 \cdot \text{card}(U_2)$, and thus direct application of bounds on expressiveness of generic queries is impossible. Still, we can show:

**Proposition 2** Let $\mathcal{M} = (\mathcal{U}, \Omega)$ be o-minimal, $c_1, c_2 > 1$, and $SC$ as above. Then no $(c_1, c_2)$-separating sentence is definable in $\text{FO}(SC, \Omega)$.

**Proof.** Assume that there is a $(c_1, c_2)$-separating sentence $\Phi$. From Fact 2, b), we conclude that there is a $\text{FO}_{\text{act}}(SC, \Omega')$ $(c_1, c_2)$-separating sentence $\Phi'$ for some extension $\Omega' \supseteq \Omega$. From Fact 2, a), we obtain that there is an infinite set $Y \subseteq \mathcal{U}$ and a $\text{FO}_{\text{act}}(SC, <)$-sentence $\Psi$ such that for every instance $D$ with $\text{adom}(D) \subseteq Y$ it holds: $D \models \Phi'$ iff $D \models \Psi$. Thus, it remains to show that $\text{FO}_{\text{act}}(SC, <)$ cannot express a $(c_1, c_2)$-separating sentence $\Psi$, on instances over an infinite set. Assume it can; and let $q$ be the quantifier rank of $\Psi$. We now consider two instances over $Y$. In both instances $D_1$ and $D_2$ all elements of $U_1$ precede $U_2$ in the linear order $<$. In $D_1$, $\text{card}(U_1) = [c_1 (2^q + 1)]$ and $\text{card}(U_2) = 2^q + 1$; in $D_2$, $\text{card}(U_1) = 2^q + 1$ and $\text{card}(U_2) = [c_2 (2^q + 1)]$. Since $\Psi$ is a $(c_1, c_2)$-separating sentence, we must have $D_1 \models \Psi$ and $D_2 \models \neg \Psi$. We shall obtain contradiction by showing that $D_1 \models \Psi$ if $D_2 \models \Psi$.

To show the latter, we must prove that the duplicator can win in a $q$-move Ehrenfeucht-Fraïssé game on $D_1$ and $D_2$. This follows from the fact for every $n, m > 2^q$, the duplicator can win a $q$-move game on two ordered sets of cardinalities $n$ and $m$ [21]. Thus, for $D_1$ and $D_2$, the duplicator picks a separate strategy for $U_1$ and $U_2$, and whenever the spoiler plays in $U_1$, the duplicator forgets about the moves in $U_2$ and responds in $U_1$ using the strategy for $U_1$, and likewise in the case when the spoiler plays in $U_2$. Let $(a_1, b_1), \ldots, (a_l, b_1)$ be moves made in the $U_1$ part of $D_1$ and $D_2$, with $a_i$s played on $D_1$ and $b_i$s played in $D_2$. Similarly, let $(c_1, d_1), \ldots, (c_k, d_k)$ be moves made in the $U_2$ part of $D_1$ and $D_2$. Then both $a_i \mapsto b_i, i = 1, \ldots, l$ and $c_i \mapsto d_i, i = 1, \ldots, k$, are partial isomorphisms; since all elements of $U_1$ precede all elements of $U_2$, putting them together we get a partial isomorphism between $D_1$ and $D_2$. This shows that $D_1 \models \Psi$ if $D_2 \models \Psi$, and thus concludes the proof. 

4.1 Dealing with AVG

We assume that instances store elements of a numerical domain, for example $\mathbb{R}$. Given a query $\varphi(\vec{x}, \vec{y})$, we define AVG$_{\varphi}(\vec{x}, \vec{y})$ by letting $D \models \text{AVG}_{\varphi}(\vec{x}, \vec{y})$ iff $\text{card}(\varphi(\vec{a}, D)) < \infty$ and $v = \text{AVG}(\varphi(\vec{a}, D))$, where $\text{AVG}(C) = (\sum_{c \in C} c) / \text{card}(C)$. Note that the aggregate AVG is typically defined using the bag semantics; however, as we show inexpressibility results, dealing with this simplified version will suffice.

It can be easily shown (by reduction to equal cardinality) that AVG$_{\varphi}$ is not definable in $\text{FO} + \text{POLY}$, even if $D \models \varphi(\vec{a}, c)$ implies $0 \leq c \leq 1$. We now define $\epsilon$-approximation of AVG just as we did it for Vol. Assume a query $\varphi(\vec{x}, \vec{z})$ is given, and $|\vec{x}| = n$. An operator AVG$_{\varphi}$, when applied to $\varphi$, produces

\footnotesize

1We shall come back to the multiset semantics later.

11
a query $\psi_\epsilon(x, z)$ such that, for any instance $D$ and any $\bar{a}, D = \exists z. \varphi(\bar{a}, z)$, and if $D \models \varphi(\bar{a}, v)$, then $|v - \text{Avg}(\varphi(\bar{a}, D) \cap I)| < \epsilon$ and $0 \leq v \leq 1$. For convenience, we let $\text{Avg}(C) = 0$ for $C$ infinite.

For $\epsilon \geq 1/2$, $\text{Avg}_I^\epsilon$ is definable in $\text{FO}(\text{SC}, \Omega)$ if the input is finite or f.r. over $\Omega$, as long as the constants $0, 1/2$ and $1$ are definable. However, $\text{Theorem 2}$ Let $M = (\mathbb{R}, \Omega)$, and let $\mathcal{M}_{+, \cdot}$ be $a$-minimal. Let $\epsilon < 1/2$. Then $\text{Avg}_I^\epsilon$ is not definable in $\text{FO} + \Omega$, even over finite instances. In particular, $\text{Avg}_I^\epsilon$ is not definable in $\text{FO} + \text{POLY}$.

Proof. Assume $\text{Avg}_I^\epsilon$ is definable. Let the schema $\text{SC}$ consists of two unary predicates, $U_1$ and $U_2$. Let $\Delta = (1 - 2\epsilon)/16$. Given two finite sets $U_1$ and $U_2$, containing at least two elements each, we translate them into intervals $[0, \Delta]$ and $[1 - \Delta, 1]$. By translating a finite set $X$ with $\min X = c, \max X = d > c$ into an interval $[a, b]$ we mean the set $X'$ containing exactly the numbers of the form $a + \frac{(x-c)(b-a)}{d-c}$ where $x \in X$; clearly $X' \subset [a, b]$. As the next step, we define $U_1^0 = U_1 \cup \{4\Delta - x \mid x \in U_1\}$ and $U_2^0 = U_2 \cup \{2 - 4\Delta - x \mid x \in U_2\}$. One observes $U_1^0 \subseteq [0, 4\Delta]$ and $U_2^0 \subseteq [1 - 4\Delta, 1]$.

The preceding shows that $U_1^0$ and $U_2^0$ are $\text{FO} + \text{POLY}$-definable. Thus, the set $C = U_1^0 \cup U_2^0 \subseteq [0, 1]$ is definable in $\text{FO} + \text{POLY}$. Now easy calculations show that

$$\text{Avg}(C) = \frac{1}{8} - \frac{\epsilon}{4} + \frac{m}{n + m} \cdot \frac{3 + 2\epsilon}{4},$$

where $n$ is the cardinality of $U_1$ and $m$ is the cardinality of $U_2$.

We now define a Boolean query $\Phi$ by letting $D \models \Phi$ iff $\text{Avg}^\epsilon(C) = \text{Avg}_I^\epsilon(C) > 1/2$. More precisely, $C$ is defined by a $\text{FO} + \text{POLY}$ query $\alpha(x)$, and thus under the assumption that $\text{Avg}_I^\epsilon$ is definable, we have a satisfiable formula $\beta_\alpha(x)$ such that $D \models \beta_\alpha(a)$ implies that $|a - \text{Avg}(C)| < \epsilon$. We now let $\Phi$ be $\exists x. \beta_\alpha(x) \land (x > 1/2)$. Under the assumption that $\text{Avg}_I^\epsilon$ is definable in $\text{FO} + \Omega$, we would obtain that $\Phi$ is in $\text{FO} + \Omega$ as well.

Let $c_0 = 1 + \frac{3\epsilon}{\sqrt{8\epsilon}} > 1$. Assume $m > c_0 \cdot n$. Plugging this into the equation for $\text{Avg}(C)$, we derive $\text{Avg}(C) > 1/2 + \epsilon$; thus, in this case $\text{Avg}^\epsilon(C) > 1/2$ no matter which $\epsilon$-approximation of the average is picked, and thus $D = \Phi$. Similarly, if we assume $n > c_0 \cdot m$, we derive $\text{Avg}(C) < 1/2 - \epsilon$, and thus $\text{Avg}^\epsilon(C) < 1/2$ and $D = \neg \Phi$. Hence, $\Phi$ is a $(c_0, c_0)$-separating sentence, which is definable in $\text{FO} + \Omega$. This contradiction proves the theorem.

$\square$

4.2 Dealing with volume

We start with two easy results. First, for unbounded measures (no restriction to $\mu^n$) volume cannot be approximated in languages like $\text{FO} + \text{POLY}$.

$\text{Proposition 3}$ Let $\mathcal{M} = (\mathbb{R}, \Omega)$, and let $(\mathbb{R}, \Omega, +, \cdot)$ be $a$-minimal. Then no $\epsilon$-approximation operator $\text{Vol}^\epsilon$ is definable in $\text{FO} + \Omega$.

Proof. Let $m > 2\epsilon + 1$ be an integer. Consider a scheme with two unary symbols $U_1$ and $U_2$ and let $\psi^1(x) \equiv \exists y. U_1(y) \land |y - x| < m$ and $\psi^2(x) \equiv \exists y. U_2(y) \land |y - x| < m$. Assume that $\text{Vol}^\epsilon$ is definable; we then have queries $\beta_\epsilon^1(x)$ and $\beta_\epsilon^2(x)$ which give $\epsilon$-approximation for the measure of outputs of $\psi^1(x)$ and $\psi^2(x)$. Now let

$$\Psi \equiv \exists x_1 x_2 (\beta_\epsilon^1(x_1) \land \beta_\epsilon^2(x_2) \land |x_1 - x_2| < 2\epsilon)$$

12
Let $(\star)$ be the following condition on $U = U_1 \cup U_2$: for every $a, b \in U$, if $|a - b| \leq 2M$ then $a = b$. Then under $(\star)$ it holds: $D \models \Psi$ iff $\text{card}(U_1) = \text{card}(U_2)$. However, this is impossible: Fact 2, a), implies that any generic query definable in $\text{FO} + \Omega$ on databases over an infinite set must be definable in $\text{FO}_{\text{act}}(\langle \rangle)$, but it is well known that equal cardinality is not $\text{FO}_{\text{act}}(\langle \rangle)$-definable (cf. [21]).

Thus, $\text{FO} + \text{LIN}$ and $\text{FO} + \text{POLY}$ cannot define $\epsilon$-approximations of volumes. Note that the proof above is by reduction to equal cardinality, for sparse finite sets. It relies on the fact that there is no a priori bound on the outputs of queries. Thus, a different approach is needed to show inexpressibility of $\text{VOL}_F$.

For a query $\varphi(x, y)$ and two constants $0 < c_1 < c_2$, we say that $\psi(x, z)$ gives a $(c_1, c_2)$-relative approximation of the volume if for any $\bar{a}$, $\psi(\bar{a}, \cdot)$ is satisfiable, and

$$D \models \psi(\bar{a}, v) \implies c_1 < \left(\frac{v}{\text{VOL}(\varphi(\bar{a}, D))}\right) < c_2$$

By a reduction to separating sentences, we will now show:

**Proposition 4** Assume that $(\mathbb{R}, \Omega)$ is such that $(\mathbb{R}, \Omega, +, \cdot)$ is o-minimal. Then for any $0 < c_1 < c_2$, the $(c_1, c_2)$-relative approximation of the volume is not definable in $\text{FO} + \Omega$, for any dimension $k > 0$, even for queries restricted to $[0, 1]^k$.

**Proof.** Let $k = 1$ (extension to $k > 1$ is trivial by taking a product with $[0, 1]^{k-1}$) and let the schema contain two unary relations $U_1$ and $U_2$. We shall assume that their interpretations are subsets of $[0, 1]$. Let $n = \text{card}(U_1), m = \text{card}(U_2)$. Let $c' = \frac{c_1}{2c_2} < \frac{1}{2}$ and $c'' < (\frac{1}{2}, \frac{c_1}{c_2})$. We claim that with a $(c_1, c_2)$-relative approximation of the volume we can define a sentence $\Phi$ such that $n < c' \cdot m$ implies $D \models \Phi$ and $n > c'' \cdot m$ implies $D \models \neg \Phi$. This will suffice, as such a sentence $\Phi$ would be a $(c'', 1/c')$-separating sentence, which cannot be defined in $\text{FO} + \Omega$.

Given a finite set $X = \{x_1 < \ldots < x_p\} \subseteq [0, 1]$ and $\delta > 0$, define

$$X(\delta) = \bigcup_{i=1}^{p-1} [x_i, x_i + \delta] \cup [x_p - \delta, x_p]$$

Note that for a given $\delta$ and $X$, this is $\text{FO} + \text{LIN}$-definable. We now let

$$\delta = \frac{1}{3} \cdot \min_{a_1, a_2 \in \text{dom}(U_1 \cup U_2), a_1 \neq a_2} |a_1 - a_2|$$

Then $U_1(\delta), U_2(\delta) \subseteq [0, 1]$ and $\text{VOL}(U_1(\delta)) = n\delta, \text{VOL}(U_2(\delta)) = m\delta$. Let $\Omega' = \Omega \cup \{+, \cdot\}$. We now have two queries in $\text{FO} + \Omega'$, $\varphi_1(y)$ and $\varphi_2(y)$ defining $U_1(\delta)$ and $U_2(\delta)$; assuming that $(c_1, c_2)$-relative approximation of the volume is definable, we have two queries $\psi_1(z)$ and $\psi_2(z)$ that produce such an approximation for $U_1(\delta)$ and $U_2(\delta)$. We next define

$$\Phi \equiv \exists z_1 \exists z_2. \psi_1(z_1) \land \psi_2(z_2) \land (z_1/z_2 < 1/2).$$

Suppose $\psi_1(v_1) \land \psi_2(v_2)$ holds. It follows then that

$$\frac{c_1n}{c_2m} < \frac{v_1}{v_2} < \frac{c_2n}{c_1m}$$

Thus $n < c' \cdot m$ implies $v_1/v_2 < 1/2$ for any $v_1$ and $v_2$ that satisfy $\psi_1$ and $\psi_2$, and hence $D \models \Phi$. Conversely, $n > c'' \cdot m$ implies $v_1/v_2 > 1/2$ for any $v_1$ and $v_2$ that satisfy $\psi_1$ and $\psi_2$, and thus in this case $D \models \neg \Phi$. This completes the proof. 

□
4.3 Absolute approximation

We shall now prove the strongest of the inexpressibility results: that $\text{Vol}_\epsilon$, for $\epsilon < 1/2$, cannot be defined in languages like FO + LIN and FO + POLY. First note:

**Proposition 5** FO + LIN defines $\text{Vol}_\epsilon$ for $\epsilon \geq 1/2$.

*Proof sketch.* If the volume is not 0 or 1, then 1/2 is the $\epsilon$-approximation. □

It turns out that this trivial approximation is the best one can hope for in languages like FO + LIN and FO + POLY.

**Theorem 3** Let $\mathcal{M} = (\mathbb{R}, \Omega)$, and let $(\mathbb{R}, \Omega, +, *)$ be o-minimal. Assume that $\epsilon < 1/2$. Then $\text{Vol}_\epsilon$ is not definable in FO + $\Omega$.

*Proof.* Let SC consist of two unary relations $A$ and $B$. Call a finite instance good if two properties are satisfied: $A$ is an initial fragment of natural numbers (that is, $\{0, 1, 2, \ldots, k\}$) and $B$ is a nonempty proper subset of $A$. Let

$$c_1 = \frac{1 - 2\epsilon}{3} \quad \text{and} \quad c_2 = \frac{2 + 2\epsilon}{3}$$

We have $0 < c_1 < c_2 < 1$ and $c_1 + c_2 = 1$.

Consider a sentence $\Phi$ in the language of SC and $\Omega$. We call it a $(c_1, c_2)$-good sentence is the following two conditions hold, whenever $(A, B)$ is a good instance:

1. If $\text{card}(B) < c_1 \cdot \text{card}(A)$, then $D \models \neg \Phi$;

2. If $\text{card}(B) > c_2 \cdot \text{card}(A)$, then $D \models \Phi$.

Note that this is the same as having a separating sentence for $B$ and $A - B$; however, here we only require that the above conditions hold for a good instance. The result now follows two lemmas.

**Lemma 1** Assume $\text{Vol}_\epsilon$ is definable in FO + $\Omega$. Then for $c_1, c_2$ as above there exists a signature $\Omega'$ extending $\Omega$ and a $(c_1, c_2)$-good sentence in FO_{act}(SC, $\Omega'$).

*Proof of Lemma 1.* Assume that an instance $(A, B)$ with $B \subset A$ is given. Let $n = \text{card}(B)$ and $m = \text{card}(A - B)$; $n, m > 0$. We now construct $A'$ and $B'$ by translating $A$ and $B$ into $[0, 1]$. That is, each element $x$ of $A$ is replaced by $x/x_M$ where $x_M$ is the maximal element of $A$. Note that $A', B'$ are FO + POLY-definable.

$$P = \bigcup_{b \in B', a \in A', (a, b) \cap A' = \emptyset} [b, a] \quad \text{and} \quad R = \bigcup_{b \in A' - B', a \in A', (a, b) \cap A' = \emptyset} [b, a]$$

Note that both $P$ and $R$ are definable in FO + POLY. We now have the following: if the instance $(A, B)$ is good, then

$$\frac{n - 1}{n + m - 1} \leq \text{Vol}(P) \leq \frac{n}{n + m - 1} \quad \text{and} \quad \frac{m - 1}{n + m - 1} \leq \text{Vol}(R) \leq \frac{m}{n + m - 1}$$
If \( \text{Vol}^\mu \) is definable in \( \text{FO} + \Omega \), we have a \( \text{FO}(\text{SC}, \Omega \cup \{+, \ast\}) \) queries \( \psi_P(z) \) and \( \psi_R(z) \) such that
\[ D \models \psi_P(v) \text{ implies } |v - \text{Vol}(P)| < \epsilon, \text{ and likewise for } R. \]
We now define \( \Psi \) as
\[ \exists z_1 \exists z_2. \psi_P(z_1) \land \psi_R(z_2) \land z_1 > z_2 \]
Let \( c' = \frac{\epsilon}{n+1} \). Assume \( \text{card}(B) > c_2 \cdot \text{card}(A) \); then \( n > c'm \). Then simple calculations show that
\[ \frac{n-1}{n+1} > 1/2 + \epsilon \text{ and } \frac{n}{n+1} < 1/2 - \epsilon \]
which implies that no matter which \( \epsilon \)-approximations \( v_1 \) and \( v_2 \) for \( \text{Vol}^\mu(P) \) and \( \text{Vol}^\mu(R) \) we have, it is the case that \( v_1 - v_2 > 0 \). Since \( \psi_P \) and \( \psi_R \) are satisfiable, we conclude that, under the assumption that the instance is good and \( \text{card}(B) > c_2 \cdot \text{card}(A) \), \( D \models \Psi \).

Next we assume that \( c_1 \cdot \text{card}(A) \). Then we get \( m > c'n \). Again, with simple calculations we obtain \( \frac{m}{n+1} > 1/2 + \epsilon \text{ and } \frac{m}{n+1} < 1/2 - \epsilon \) hence, for every \( \epsilon \)-approximations \( v_1 \) and \( v_2 \) for \( \text{Vol}^\mu(P) \) and \( \text{Vol}^\mu(R) \), it is the case that \( v_1 - v_2 < 0 \), and thus \( D \models \neg \Psi \).

Now the lemma follows from Fact 2, b).

\[ \square \]

**Lemma 2** Let \( \Theta \) be an arbitrary signature on \( \mathbb{R} \). Then \( \text{FO}_{\text{act}}(\text{SC}, \Theta) \) cannot define a \((c_1, c_2)\)-good sentence.

**Proof of Lemma 2.** Suppose for \( 0 < c_1 < c_2 < 1 \) and for some signature \( \Theta \), there is a \( \text{FO}_{\text{act}}(\text{SC}, \Theta) \) sentence \( \Phi \) that is \((c_1, c_2)\)-good according to the definition above. We may assume without loss of generality (just by adding existential quantifiers over the active domain) that all atomic formulae are either \( A(x) \) or \( B(x) \), where \( x \) is a variable, or \( \Theta \)-atomic formulae. Next, make a signature \( \Theta_\Phi \) by putting a \( k_\gamma \)-ary symbol \( P_\gamma \) for each \( \Theta \)-atomic subformula \( \gamma(x_1, \ldots, x_{k_\gamma}) \) of \( \Phi \) into it. We then define a sentence \( \Psi \) in the language of \( \Theta_\Phi \) and \( B \) by replacing, in \( \Phi \), each atomic \( \Theta \)-formula \( \gamma \) by the corresponding symbol \( P_\gamma \), and \( A(\cdot) \) by true.

Next, with each \( n > 1 \) and each \( B \subseteq \{0, \ldots, n-1\} \) associate a \( \Theta_\Phi \cup \{U\} \) structure \( S(B, n) \) whose universe is \( \{0, \ldots, n-1\} \), the unary symbol \( U \) is interpreted as \( B \), and \( \Theta_\Phi \) predicates inherit their interpretation from \( \langle \mathbb{R}, \Theta \rangle \) (this is possible since \( \Theta_\Phi \) does not contain any function symbols). We then have, by a straightforward induction on the structure of a formula
\[ S(B, n) \models \Psi \text{ iff } \langle 0, \ldots, n-1 \rangle \models \Phi \]
where \( \langle 0, \ldots, n-1 \rangle \) is the good instance with \( A \) interpreted as \( \{0, \ldots, n-1\} \). Thus, for \( \text{card}(B) < c_1 n \) we have \( S(B, n) \models \neg \Psi \) and for \( \text{card}(B) > c_2 n \) we have \( S(B, n) \models \Psi \).

It follows from [12] that \( \Psi \) is definable by a family of non-uniform \( AC^0 \) circuits, with size bounded by some polynomial \( p(n) \), and depth \( d \). This is because \( \Psi \) can be transformed into a Boolean formula by replacing each \( \exists x \in \text{dom} \) by a disjunction over \( \{0, 1, \ldots, n-1\} \) and each \( \forall x \in \text{dom} \) by a conjunction over \( \{0, 1, \ldots, n-1\} \). Once quantifiers are replaced, each occurrence of a \( \Theta_\Phi \) predicate only mentions constants and is replaced by its truth value (this is why the circuit may be non-uniform). It now follows from [12] that such a family of formulae is definable by a polynomial-size constant depth family of \( AC^0 \) circuits.

According to Lemma 5 from [12], for large enough inputs, constant-depth circuits cannot distinguish cardinalities in \( \sqrt{n}, n - \sqrt{n} \). Thus, there is a number \( N_1 \in \mathbb{N} \) such that for all \( n > N_1 \) it is the case that for any \( p, q \in \sqrt{n}, n - \sqrt{n} \), \( p \neq q \), there exist sets \( B_1 \) and \( B_2 \) of cardinalities \( p \) and \( q \) respectively such that \( S(B_1, n) \) and \( S(B_2, n) \) agree on \( \Psi \). We now let \( N \) be an integer that exceeds both \( N_1 \) and \( \frac{1}{c_1} \). Let \( n \) be an arbitrary integer bigger than \( N \). Then there are integers \( n_1, n_2 \) such
that $n_1, n_2 \in [\sqrt{n}, n - \sqrt{n}]$ and $n_1 < c_1 n$, $n_2 > c_2 n$. In particular, for any two $B_1$ and $B_2$ such that $n_1 = \text{card}(B_1)$ and $n_2 = \text{card}(B_2)$, we have $S(B_1, n) = \neg \Psi$ and $S(B_2, n) = \Psi$ (since $\Psi$ is equivalent to $\Phi$, which is a $(c_1, c_2)$-good sentence). However, this contradicts the above observation that for some $B_1$ and $B_2$ as above, $S(B_1, n)$ and $S(B_2, n)$ must agree on $\Psi$. This contradiction concludes the proof of the lemma and the theorem. □

**Corollary 1** FO + LIN, FO + POLY and FO + EXP cannot express $\text{Vol}_f$ for any $\epsilon < 1/2$. □

Theorem 3 shows that one cannot possibly adjust the method of [23, 24, 25] to get the approximation operators uniformly definable. This is somewhat surprising, for the following reasons. It is possible that there exists an o-minimal structure which is closed under taking integrals. That is, for every $\varphi(\vec{x}, \vec{y})$ in the language of the structure, there is a formula $\psi(\vec{x}, z)$ such that $\models \varphi(\vec{a}, v)$ iff $v = \int \cdots \int_{\varrho(\vec{a}, \mathbb{R}^n) \cap I^n} \varrho(\vec{x}, z)$. The existence of such a structure is conjectured in [24]. By Theorem 3, even if such a structure $\mathcal{M} = (\mathbb{R}, \Omega)$ existed, the volume of outputs of very simple queries on finite instances could not be approximated in FO + \Omega!

Is it possible that one can express the approximate volume computation over outputs of some particularly simple queries? We now show that for two very simple classes, this remains impossible in FO + POLY and similar languages.

**Corollary 2** In languages FO + LIN, FO + POLY, FO + EXP, it is impossible to express $\text{Vol}_f$ even restricted to a) outputs of conjunctive $<$-queries over finite instances, or b) schema predicates, interpreted as for instances definable with dense-order constraints.

**Proof.** Let the scheme consist of three unary symbols $A$, $B$, $C$, and one binary symbol $E$. A finite instance $D$ is called good if $B, C$ form a partition of $A$, the distance between any two consecutive elements of $A$ is the same, and $E$ is the successor relation on $A \subset [0, 1]$. With this, we follow the proof of Theorem 3. We define $P$ and $R$ as before, and note that with $C$ and $E$ in the signature, they can be defined by conjunctive queries. For example, for $P$: $\varphi_p(z) \equiv \exists b \in \text{adom}(x) \in \text{adom}. B(b) \land C(c) \land E(b, c) \land b < z \land z < c$. Now, assuming $\text{Vol}_f$ is definable in FO(SC, $\Omega$), we obtain, as in Lemma 1, that a $(c_1, c_2)$-good sentence is definable in FO(SC, $\Omega$), for a good instance as defined above. This easily leads to contradiction: if a $(c_1, c_2)$-good sentence is definable in FO(SC, $\Omega$) for instances with $A \subset [0, 1]$, it is definable in FO(SC, $\Omega, +, \cdot$) for instances with $A$ being an initial fragment of natural numbers. Then the proof of Lemma 2 applies, as in the translation into a family of Boolean formulae the symbols $C$ and $E$ can be eliminated: $C(x)$ is replaced by $\neg B(x)$, and $E(x, y)$ by $y - x = 1$. This completes the proof. □

**Remarks** One may ask where the procedure of [23, 24, 25] fails if we try to apply it, in a uniform way, to, say, FO + POLY queries. Note that the method of [23, 24, 25] produces a formula whose quantifier prefix is proportional to the VC dimension of the family of sets defined by the input formula. However, for relational calculus queries, this may depend on the size of the database, thus making it impossible to quantify uniformly over random samples. For a query $\varphi(\vec{x}, \vec{y})$ with and a database $\mathcal{D}$, the *definable family given by* $\varphi$ and $\mathcal{D}$ is $F_\varphi(\mathcal{D}) = \{ \varphi(\vec{a}, D) \mid \vec{a} \in \mathcal{U}\}$ where $\varphi(\vec{a}, D) = \{ \vec{b} \mid D \models \varphi(\vec{a}, \vec{b}) \}$. The size of a finite database $D$, $|D|$, is defined to be $\text{card}(\text{adom}(D))$.

**Proposition 6** There exists a (quantifier-free) relational calculus query $\varphi(x, y)$, and a sequence of databases $\mathcal{D}_1, \mathcal{D}_2, \ldots$ of increasing size such that $\text{VCdim}(F_\varphi(\mathcal{D}_n)) \geq \log |\mathcal{D}_n|$. 16
\textbf{Proof.} Let $SC$ contain a single binary symbol $P$. Let $D_n$ be an instance with the second projection being an $n$-element set $A_n$, and the first projection coding the powerset of $A_n$ (as in \cite[page 462]{[1]}). That is, for each $B \subseteq A_n$, there is $a_B$ such that $(a_B, b) \in P$ iff $b \in B$. Let $\varphi(x,y) \equiv P(x,y)$. We now consider the family $\mathcal{F}_n = \{ \varphi(a,D) \mid a \in U \}$. It follows immediately from the construction that $\mathcal{F}_n$ shatters $A_n$; thus, $\text{VCdim}(\mathcal{F}_n(D_n)) \geq n$. Since one needs $2^n$ elements to code the powerset of $A_n$, one can choose $D_n$ to have the active domain of $2^n$ elements. This proves the proposition. \hfill \qedsymbol

We also remark that under some special assumptions on the outputs of the queries, their volumes can be approximated. One can show, using Löwner-John ellipsoids \cite{[1]}, that for a $\text{FO} + \text{POLY}$ query $\varphi(\vec{x}, \vec{y})$ with $|\vec{y}| = k$, under the assumption that $\varphi(\vec{a}, D)$ is convex, a relative $(c_1, c_2)$ approximation of its volume can be found with $c_1 = \frac{k^k + 1}{2^{k+1}} - \epsilon$ and $c_2 = \frac{k^k + 1}{2} + \epsilon$ for an arbitrarily small $\epsilon > 0$.

5 \textbf{FO + POLY + SUM: An aggregate language for constraint databases}

We now introduce a language for extending $\text{FO} + \text{POLY}$ with a summation operator. The main difficulty is to make sure that when summation is done over all elements in some query output, we are guaranteed that the query output is finite. To do this, we use techniques from \cite{[6]} for guaranteeing that a query is safe (that is, that a query yields finite output).

Let $Q$ be a non-boolean query over a database schema $SC$. We say that $Q$ is a semi-algebraic query if it gives semi-algebraic output on semi-algebraic inputs. We say $Q$ is semi-algebraic-to-finite and write $Q \in \text{SAF}$ if $Q$ produces finite output on semi-algebraic input databases. If $Q$ is expressed as $\varphi(y, \vec{x})$, we say that $Q$ is $\vec{x}$-semi-algebraic-to-finite if for every $\vec{a}$ the query $\varphi(y, \vec{a})$, with one free variable $y$, is in SAF. In the language $\text{FO} + \text{POLY} + \text{SUM}$, all queries are semi-algebraic queries, but in the construction we will have to ensure that certain subqueries are in the smaller class SAF.

A first-order formula $\gamma(x, \vec{w})$ with distinguished variable $x$ in the language of the real field is said to be deterministic if it produces at most one output $x$ for every vector of real numbers $\vec{w}$. Deterministic formulas are the building blocks from which safe queries can be formed. Given a deterministic formula $\gamma(x, \vec{w})$ and a finite set of tuples of reals $A$ (having the same length as $\vec{w}$), we let $\gamma(A)$ refer to the bag $\{\vec{a} \in A \mid f(\vec{a}) \}$, where $f$ is the corresponding partial function taking $\vec{w}$ to the unique $x$ such that $\gamma(x, \vec{w})$ holds. Note that it is decidable if a formula is deterministic.

\textbf{Definition of FO + POLY + SUM} The query language $\text{FO} + \text{POLY} + \text{SUM}$ is defined inductively as follows. Atomic queries are the same as for $\text{FO} + \text{POLY}$. The formulae of $\text{FO} + \text{POLY} + \text{SUM}$ are closed under boolean connectives and quantification $\forall$ and $\exists$ (over the reals).

Next, we define the summation term-former. Given any $\text{FO} + \text{POLY} + \text{SUM}$ formula $\varphi(y, \vec{z})$, we let $\text{END}[y, \varphi(y, \vec{z})](u, \vec{z})$ be the query that holds for a tuple $(b, \vec{a})$ on an input database $D$ iff $b$ is an endpoint of the intervals that compose $\varphi(D, \vec{a}) = \{ c \in \mathbb{R} \mid D \models \varphi(c, \vec{a}) \}$. Note that if $\varphi$ is a semi-algebraic query (which is guaranteed by Theorem 4 below), then $\text{END}[y, \varphi(y, \vec{z})]$ is $\vec{z}$-SAF.

A range-restricted $\text{FO} + \text{POLY} + \text{SUM}$ expression is an expression of the form $\rho(\vec{w}, \vec{z}) \equiv (\varphi_1(\vec{w}, \vec{z})[\text{END}[y, \varphi_2(y, \vec{z})]])$ where $\varphi_1(\vec{w}, \vec{z})$ and $\varphi_2(y, \vec{z})$ are $\text{FO} + \text{POLY} + \text{SUM}$ queries. It binds $y$, that is, the free variables are $\vec{z}, \vec{w}$. We have $D = \rho(\vec{a}, \vec{b})$ for $\vec{a} = (a_1, \ldots, a_n)$ iff $D \models \varphi_1(\vec{a}, \vec{b})$ and $D \models (\text{END}[y, \varphi_2(y, \vec{z})])(a_i, \vec{b})$, $i = 1, \ldots, n$. 

17
It then follows from the closure property (Theorem 4) that for any $D$ and any $\vec{b}$, the set $\rho(D, \vec{b}) = \{\vec{a} \mid D = \rho(\vec{a}, \vec{b}) \}$ is finite.

For any deterministic formula $\gamma(x, \vec{w})$ and any range-restricted expression $\rho(\vec{w}, \vec{z})$ as above we now define a term $t(\vec{z})$ by

$$t(\vec{z}) = \left[ \sum_{\rho(\vec{w}, \vec{z})} \gamma(\vec{z}) \right]$$

Given $D$ and $\vec{b}$, the value of $t(\vec{b})$ in $D$ is the sum of all the members of the finite bag $\gamma(A)$, where $A = \rho(\vec{a}, \vec{b})$.

Finally, new terms in $\text{FO + POLY + SUM}$ can be built by applying composition with the real functions $+, \cdot$. If $t, \varphi$ are terms and $\varphi$ is a formula, then $t_1 = t_2, t_1 < t_2$ and $\gamma(t_1, \ldots, t_k)$ are $\text{FO + POLY + SUM}$ formulæ.

**Examples of FO + POLY + SUM queries** Let $\varphi(w)$ be an $\text{FO + POLY}$ query. Let $\gamma(x, w) \equiv (x = w)$ and $\rho(w) = (w = w) \in \text{END}[w, \varphi(w)]$. Then the $\text{FO + POLY + SUM}$ term (without free variables) $\sum_{\rho(w)} \gamma$ gives the sum of all the endpoints of the intervals that compose $\varphi(D)$.

The area of a convex polygon in $\mathbb{R}^2$ can be defined in $\text{FO + POLY + SUM}$. The idea of the query is illustrated in Figure 1. Suppose we triangulate the polygon as shown. Then the area of the polygon is the sum of the areas of triangles. We thus have to define the triangulation and then apply the summation term of $\text{FO + POLY + SUM}$ to calculate the area.

This is done as follows. Assume that the polygon is given by a predicate $P(x, y)$ (it could be an input relation or the output of a query). There is a $\text{FO + POLY}$ query $\varphi_P(x, y)$ that computes all the vertices of $P$ – this is because $\vec{a}$ is vertex iff $\vec{a} \not\in \text{conv}(P - \{\vec{a}\})$. Since one can compute the boundary of $P$ by a $\text{FO + POLY}$ query, it follows that there is a $\text{FO + POLY}$ query $\nu_P(\vec{x}, \vec{y})$ that tests if $\vec{x}, \vec{y}$ are two adjacent vertices of $P$.

We now form two $\text{FO + POLY}$ queries. The query $\psi_2(u)$ tests if $u$ is a coordinate of a vertex of $P$. The query $\psi_1(\vec{x}, \vec{y}, \vec{z})$ tests the following conditions: (1) $\varphi_P(\vec{x}) \land \varphi_P(\vec{y}) \land \varphi_P(\vec{z})$ holds; (2) $\vec{x}$ is a lexicographically minimal vertex of $P$; (3) either $\nu_P(\vec{y}, \vec{z})$ holds and $\vec{y}$ is lexicographically less than $\vec{z}$ and $-\nu_P(\vec{x}, \vec{y}) \land -\nu_P(\vec{x}, \vec{z})$, or $\nu_P(\vec{x}, \vec{y}) \land \nu_P(\vec{y}, \vec{z}) \land -\nu_P(\vec{x}, \vec{z})$.

We now let $\rho(\vec{x}, \vec{y}, \vec{z})$ be the range-restricted expression $(\psi_1(\vec{x}, \vec{y}, \vec{z}) | \text{END}[u, \psi_2(u)])$. It can be easily seen that for $P$ convex, the output of $\rho$ is finite and produces a triangulation of $P$. That is, $\rho(\vec{a}, \vec{b}, \vec{c})$
holds iff $\tilde{a}, \tilde{b}, \tilde{c}$ are the vertices of one of the triangles such as those shown in Figure 1.

Since for each triangle with vertices $(a_1, a_2), (b_1, b_2), (c_1, c_2)$, its area is computable as $|a_1 b_2 - a_2 b_1 + a_2 c_1 - a_1 c_2 + b_1 c_2 - c_1 b_2|/2$, we obtain a deterministic formula $\gamma(v, \bar{x}, \bar{y}, \bar{z})$ saying that $v$ is the area of the triangle with vertices $\bar{x}, \bar{y}, \bar{z}$. We then conclude that the term $\sum_{\rho(\bar{x}, \bar{y}, \bar{z})} \gamma$ defines the area of $P$.

Note that the above method codes a standard computation of area used in computational geometry [33] which generalizes to nonconvex polygons, and is in fact used in GISs for area computation [38].

Properties of $\text{FO} + \text{POLY} + \text{SUM}$ The language $\text{FO} + \text{POLY} + \text{SUM}$ has a number of attractive features. It extends both $\text{FO} + \text{POLY}$ and the relational calculus with summation and other standard aggregates. It is also related to aggregate languages for statistical databases studied recently in [17]. Furthermore, we have the following property.

Theorem 4 $\text{FO} + \text{POLY} + \text{SUM}$ is closed. That is, every $\text{FO} + \text{POLY} + \text{SUM}$ query returns semi-algebraic output on a semi-algebraic input.

Proof. We show this by structural induction on the construction of the query. Suppose we know inductively that $\varphi(v, \bar{z})$ is a semi-algebraic query, and fix a semi-algebraic database $D$. There is an integer $n$ such that for any $\tilde{a}, D \models \text{END}[y, \varphi(y, \bar{z})](c, \tilde{a})$ for at most $n$ distinct values of $c$ (by $\alpha$-minimality of the real field and the uniform bounds result of [32]). Moreover, this integer can be effectively computed given $\varphi$ and $D$. Hence, for every $\bar{z}$,

$$\rho(\bar{w}, \bar{z}) \equiv \varphi_1(\bar{w}, \bar{z}) \land \text{END}[y, \varphi(y, \bar{z})]$$

holds for at most $n^m$ tuples $\bar{w}$, where $m$ is the length of $\bar{w}$. We then see that the set $\{(v, \bar{a}) \mid D \models v = \sum_{\rho(\bar{w}, \bar{a})} \gamma(x, \bar{w})\}$ is semi-algebraic, since it is definable by the disjunction of $v = 0 \land \forall \bar{w} \neg \rho(\bar{w}, \bar{a})$ with

$$\bigvee_{1 \leq k \leq n^m} \exists \bar{w}_1 \cdots \exists \bar{w}_k \left( \bigwedge_{i=1}^k \rho(\bar{w}_i, \bar{a}^i) \land (\forall \bar{w} \rho(\bar{w}, \bar{a}) \rightarrow \bigvee_{i} (\bar{w} = \bar{w}_i) \land \bigwedge_{i \neq j} (\bar{w}_i \neq \bar{w}_j) \right)$$

$$\land \exists u_1 \cdots \exists u_k \left( (\bigwedge_{i=1}^k \gamma(u_i, \bar{u}_i) \vee (\forall z \neg \gamma(z, \bar{u}_i) \land u_i = 0) \right) \land (v = u_1 + \cdots + u_k) \right).$$

The language is also closed under the standard relational aggregation.

Proposition 7 • For any SAF $\text{FO} + \text{POLY} + \text{SUM}$ query $\varphi(\bar{z})$, we can express in $\text{FO} + \text{POLY} + \text{SUM}$ the cardinality of the output of $\varphi$.

• For any SAF query $\text{FO} + \text{POLY} \varphi(\bar{z})$ and any deterministic formula $\chi(x, \bar{w})$ we can express in $\text{FO} + \text{POLY} + \text{SUM}$ the sum of the $x$ values of $\chi$ for $\bar{w}$ ranging over the output of $\varphi$ and the average of the $x$ values of $\chi$ over the output of $\varphi$.

Proof. To see the first item, consider an arbitrary SAF $\text{FO} + \text{POLY} + \text{SUM}$ query $\varphi(\bar{w})$. Let $\varphi'(w)$ be the query returning the active domain of the output of $\varphi$. Then $\varphi'$ is clearly SAF as well, and
END[w, \varphi'(w)] is the same as \varphi'(w). Let \rho(\bar{w}) = \varphi(\bar{w})|\text{END}[w, \varphi'(w)] and \gamma(x, \bar{w}) be \gamma = 1. Then 
\sum_{\rho(\bar{w})} \gamma \text{ is an FO + POLY + SUM query returning the number of items in the output of } \varphi.

To see the second item, let \rho be as in the previous paragraph. For any deterministic formula \chi(x, \bar{w}) we have that \sum_{\rho(\bar{w})} \chi \text{ is an FO + POLY + SUM query returning the sum of the x-values of } \chi \text{ over the output of } \varphi. \text{ The average of } \varphi \text{ is simply the quotient of the sum of } \varphi \text{ and the cardinality of } \varphi. \text{ Since the FO + POLY definable functions are closed under division, we can define average.} \quad \Box

6 Computing the volume of Semi-linear sets in FO + POLY + SUM

In this section we show how to use the aggregate language FO + POLY + SUM for volume computation and approximation. Our goal is to prove that FO + POLY + SUM can compute the volume of semi-linear sets. We start by noting that taking volumes of semi-linear sets does not take us out of the semi-algebraic setting. This fact is easily derived from known results in the literature (and may have been published before, see, for example, [8] for a closely related result).

Lemma 3 For any formula \varphi(\bar{x}, \bar{y}) over the real ordered group \mathbb{R}^n, the volume of \varphi is semi-algebraic. That is, \{\bar{r}, s \mid [\text{VOL } \bar{y}, \varphi(\bar{x}, \bar{y})]((\bar{r}, s))\} is a semi-algebraic set.

Proof. By Fubini's Theorem, [\text{VOL } \bar{y}, \varphi(\bar{x}, \bar{y})]((\bar{x}, z)) holds exactly when \( z = \int \cdots \int \chi(\bar{x}, \bar{y})d y_n \cdots d y_1 \), where \( \chi_{\varphi} \) is the characteristic function of the set defined by \( \varphi \).

Let \( F_1(y_1, \ldots, y_{n-1}, \bar{x}) \) be the innermost integral \( \int \chi_{\varphi}(\bar{y}, \bar{x})d y_n \). We first show that \( F_1(y_1, \ldots, y_{n-1}, \bar{x}) \) is semi-algebraic. Let \( l_i(y_1, \ldots, y_{n-1}, \bar{x}) \) and \( u_i(y_1, \ldots, y_{n-1}, \bar{x}) \) be the ith lower and upper endpoint of the set \( \varphi_{\bar{x}, y_1, \ldots, y_{n-1}} = \{ y_n \mid \varphi(\bar{x}, y_1, \ldots, y_{n-1}) \} \). We know that \( u_i \) and \( l_i \) are semi-linear definable partial functions. We now note that any such function is piecewise linear with the coefficients in the linear polynomial being rational, cf. [35]. That is, for each function, its domain can be partitioned into finitely many semi-linear sets on which it is linear. To see this, note that on its domain \( U_i \), \( u_i(y_1, \ldots, y_{n-1}, \bar{x}) \) is the unique solution to a disjunction of conjunctions of linear inequalities in \( y_1, \ldots, y_{n-1}, \bar{x} \). Each disjunct must then have at most one solution. Let a disjunct be a conjunction \( \bigwedge_{l \in T_1} C_l(y_1, \ldots, y_{n-1}, \bar{x}) \theta \), where \( \theta \in \{<, >, \leq, \geq \} \). We know that this must have at most one solution \( r \) for each \( r_1, \ldots, r_{n-1}, s \in U_i \). But this solution must then be the solution to the conjunction of some subset of the corresponding equalities \( C_l(y_1, \ldots, y_{n-1}, \bar{x}) = 0 \) where \( l \in T_2 \subset T_1 \). (Otherwise fix a counterexample \( r_1, \ldots, r_{n-1}, s \) and let \( T_2 \) be the set of \( l \in T_1 \) such that the solution \( r \) satisfies \( C_l(r_1, \ldots, r_{n-1}, s) = 0 \). If the corresponding solution space is not 0-dimensional, then the set of proper inequalities of the form \( C_l(y_1, \ldots, y_{n-1}, \bar{x}) \{<, >\}0 \) with \( l \in T_1 - T_2 \) satisfies \( r, s \) defines an open subset of this space, which would then have to be infinite or empty, giving a contradiction.) But by linear algebra, we know that when a set of linear equalities \( C_l(y_1, \ldots, y_{n-1}, \bar{x}) \) has a unique solution \( y_1, \ldots, y_{n-1}, \bar{x} \), this solution is given by a linear function with coefficients in the field generated by \( y_1, \ldots, y_{n-1}, \bar{x} \). Hence piecewise \( u_i \) is linear, and similarly for \( l_i \).

Hence we can find a decomposition of \( \mathbb{R}^{n+1} \) into semilinear sets \( A_1, \ldots, A_k \), and find a function \( b : k \to N \) and linear functions \( f_{ij}(y_1, \ldots, y_{n-1}, \bar{x}) : i \leq k, j \leq b(i) \) such that

\[ \forall r_1, \ldots, r_{n-1}, s_1, \ldots, s_m \in \bigwedge_{i \leq b(i)} (F_1(r_1, \ldots, r_{n-1}, s_1, \ldots, s_m) = \sum_{k \leq b(i)} f_{ij}(r_1, \ldots, r_{n-1}, s_1, \ldots, s_m). \]

But now we have that \( [\text{VOL } \bar{y}, \varphi(\bar{x}, \bar{y})]((\bar{x}, z)) \) holds exactly when \( z = \int \cdots \int F_1(x_1, \ldots, x_{n-1}, y)dx_1 \cdots dx_{n-1} \), so we can partition \( \mathbb{R}^n \) into finitely many pieces, on each one of which \( [\text{VOL } \bar{y}, \varphi(\bar{x}, \bar{y})]((\bar{x}, z)) \) is given by the graph of a polynomial in \( \bar{x} \). Hence \( \text{VOL } \bar{y}, \varphi(\bar{x}, \bar{y}) \) is semi-algebraic. \quad \Box
We now prove that the language FO + POLY + SUM can express volumes of semi-linear sets.

**Theorem 5**  
• For every schema predicate $S \in SC$ there is an FO + POLY + SUM term $\tau$ which, for any semi-linear database $D$, computes the volume of $S$ in $D$.

• For every FO + LIN query $\varphi$ there is an FO + POLY + SUM term $\tau_\varphi$ such that for any semi-linear database $D$, $\tau_\varphi(D)$ returns the volume of $\varphi(D)$.

**Proof:** Note that the first item clearly implies the second, because, given such a term $\tau$ we can compose it with the query $\varphi$ to get the necessary term in the second item. Hence we only prove the first item here.

For any semi-linear $S$ we have $\text{Vol}(S) = \int \cdots \int \chi_S(\bar{x})dx_n \cdots dx_1$, where $\chi_S$ is the characteristic function. The innermost integral is $[\sum_{\rho_1(w,x_1,\ldots,x_{n-1})} \gamma(x_1,\ldots,x_{n-1})$, where $\rho_1(w,x_1,\ldots,x_{n-1})$ is the query saying $w$ is the sum of difference of consecutive endpoints of the set $\{x_n \mid S(x_1,\ldots,x_{n-1},x_n)\}$, and $\gamma(w) \equiv (w = w)$. Note that by $\alpha$-minimality, $\rho_1$ is an FO + POLY + SUM query mapping semi-algebraic sets to finite sets. The proof of Proposition 7 shows that any such query can be written as a range-restricted expression in FO + POLY + SUM.

Let $f^1_{r_1,\ldots,r_{n-2}} = [\sum_{\rho_1(w,x_1,\ldots,x_{n-1})} \gamma](x_1,\ldots,x_{n-1})$. We know from the proof of Lemma 3 that for each fixed $r_1,\ldots,r_{n-2}$, the function $g^1_{r_1,\ldots,r_{n-2}}(x_{n-1}) = f^1(r_1,\ldots,r_{n-2},x_{n-1})$ is piecewise a linear function of $x_{n-1}$. Since $f^1$ is an FO + POLY + SUM definable function, we can also define in FO + POLY + SUM the set of points $(r_1,\ldots,r_{n-2},x_{n-1})$ such that the function $g^1_{r_1,\ldots,r_{n-2}}$ is not smooth at $r_{n-1}$. We can do this because a piecewise linear function is smooth whenever it is differentiable, and the latter property can be tested by an FO + POLY query.

Let $f^2(x_1,\ldots,x_{n-2})$ be the sum of all values of the function $(mu^2 - ml^2)/2 + b(u-l)$, where the quadruples $(u,l,m,b)$ vary over all quadruples of points such that $(l,u)$ are consecutive points of nonsmoothness of $g^1_{x_1,\ldots,x_{n-2}}$, and $g^2_{x_1,\ldots,x_{n-2}} = mx + b$ on the interval $(l,u)$.

Note that since $g^1_{x_1,\ldots,x_{n-2}}$ is piecewise linear, there are only finitely many points where $f^1$ is not smooth, hence only finitely many pairs of consecutive points of nonsmoothness. Therefore there are only finitely many quadruples $(u,l,m,b)$ as above. Also note that the formula $\gamma(w,l,u,m,b)$ given by $w = (mu^2 - ml^2)/2 + b(u-l)$ is a deterministic formula. Hence, by Proposition 7, there is an FO + POLY + SUM query returning the sum of all $\gamma$ output values $w$ as $(l,u,m,b)$ vary. Hence $f^2(x_1,\ldots,x_{n-2})$ is an FO + POLY + SUM definable function.

**Claim 1** $f^2(x_1,\ldots,x_{n-2})$ is exactly the volume of the fiber of $S$ based on $x_1,\ldots,x_{n-2}$. That is, $f^2(x_1,\ldots,x_{n-2}) = \text{Vol}(\{x_n \mid x_n \in S \land (x_1,\ldots,x_{n-2},x_n) \in D\})$.

**Proof:** By Fubini’s theorem, the volume is the integral of the one variable function $g^1_{x_1,\ldots,x_{n-2}}(x_{n-1})$. Since this function is piecewise linear, for each fixed $r_1,\ldots,r_{n-2}$ there are finitely many $a_1,\ldots,a_k \in \mathbb{R} \cup \{-\infty,\infty\}$ with $a_1 < \ldots < a_k$ such that $g^1$ is linear on $(a_j,a_{j+1})$. Hence the integral of $g^1_{r_1,\ldots,r_{n-2}}$ is just the sum of the integral of $g^1$ over the intervals $(a_j,a_{j+1})$. But the integral of a linear function $h(x) = mx + b$ over an interval $l,u$ is just $mu^2/2 + bu^2$, and hence the result follows.

Continuing this inductively, we have the function $f^{k-1}(x_1,\ldots,x_{n-k+1})$ giving the volume of the fiber of $S$ defined by $x_1,\ldots,x_{n-k+1}$. If we fix the first $n-k$ parameters in this function, we obtain a function $g^{k-1}_{x_1,\ldots,x_{n-k}}(x_{n-k+1})$ which is piecewise polynomial of degree at most $k-1$. That is, $\mathbb{R}$ is
partitioned into finitely many intervals, and on each of them \( g_{x_1, \ldots, x_{n-1}}^{k-1}(y) \) is given by \( b_{k-1} y^k + \ldots + b_0 \). One can now determine all the points of non-smoothness (since this is the same as not being \( k-1 \)-times differentiable) of \( g_{x_1, \ldots, x_{n-1}}^{k-1} \) by a FO + POLY + SUM query. Furthermore, one can write a query, using polynomial constraints, that on every point in every interval between the points of non-smoothness finds the coefficients of the polynomial of degree \( k-1 \) that gives \( g_{x_1, \ldots, x_{n-1}}^{k-1} \) on every such interval (e.g., by computing the derivatives and applying Taylor’s theorem). Thus, we have a range-restricted FO + POLY + SUM query \( \rho_k(b_{k-1}, b_{k-2}, \ldots, b_0, u, l, x_1, \ldots, x_{n-k}) \) that for every \( x_1, \ldots, x_{n-k} \) produces the tuples \( (b_{k-1}, b_{k-2}, \ldots, b_0, u, l) \) such that on \( (u, l) \), \( g_{x_1, \ldots, x_{n-1}}^{k-1} \) is given by the polynomial \( b_{k-1} y^{k-1} + \ldots + b_0 \), and furthermore \( (u, l) \) list all such intervals, which cover all \( \mathbb{R} \) except finitely many points of non-smoothness.

Now let \( \gamma_k(b_{k-1}, b_{k-2}, \ldots, b_0, u, l) \) be defined by

\[
\frac{b_{k-1}(u^k - l^k)}{k} + \frac{b_{k-2}(u^{k-1} - l^{k-1})}{k-1} + \ldots + b_0(u - l)
\]

Hence, \( f^k(x_1, \ldots, x_{n-k}) \) given by

\[
\left[ \sum_{\rho_k(b_{k-1}, b_{k-2}, \ldots, b_0, u, l, x_1, \ldots, x_{n-k})} \gamma_k(b_{k-1}, b_{k-2}, \ldots, b_0, u, l) \right] (x_1, \ldots, x_{n-k})
\]

defines, for each \( (x_1, \ldots, x_{n-k}) \),

\[
\int g_{x_1, \ldots, x_{n-1}}^{k-1}(x_{n-k+1}) dx_{n-k+1},
\]

and thus by Fubini’s theorem it is the volume of the fiber of \( S \) over \( x_1 \ldots x_{n-k} \).

Now it immediately follows that \( f^n \) is a FO + POLY + SUM function giving the volume of \( S \). Theorem 5 is proved. \( \Box \)

7 Conclusions

This paper has dealt with the key question of how to add aggregation to constraint query languages. The first fundamental question is whether there can be a language that is closed under the natural spatial aggregation operators, and which also retains the basic closure property that is fundamental to a constraint-based approach: namely, that every query output can be again represented as a constraint solution set. Our results give indication that this is impossible: these two closure properties are fundamentally incompatible. Perhaps more surprisingly, we show that the problem is not particular to the polynomial or linear constraint model: even going to a larger well-behaved constraint set does not remedy the problem.

The results above motivated us to look for languages that are not closed under volume operators, but which are closed under natural discrete aggregations and which permit the computation of volumes for semi-linear sets. The language FO + POLY + SUM defined here gives a natural approach to the addition of discrete aggregation operators to a constraint language. The key idea is the notion of range-restricted querying: allowing aggregation to be formed only on sets that are guaranteed to be finite. We show not only that FO + POLY + SUM has some attractive closure properties analogous to classical aggregate languages, but it allows one to do a significant amount of spatial aggregation — e.g., volumes of semi-linear sets, averages over semi-linear sets — as well.
The approach given here based on classical summation over range-restricted sets is natural, and allows one to re-use many of the evaluation strategies for classical aggregation operators; it is clear, however, that the syntax given here for FO + POLY + SUM is quite awkward. We hope to find more streamlined and natural syntax for FO + POLY + SUM, and we are looking at subsets of FO + POLY + SUM that can be more efficiently evaluated than the full language. It remains to discover how one could best provide support for directly expressing volumes in some language built 'on top of' FO + POLY + SUM, and how to add grouping constructs to the language.

A challenging issue on the theoretical side is how to prove expressive bounds on aggregate constraint database languages like FO + POLY + SUM. For example, the results of this paper give strong evidence that FO + POLY + SUM does not suffice to calculate volumes of semi-algebraic sets, but this is at this point only a conjecture.

Acknowledgements Part of this work was done while the second author was visiting INRIA. We thank Serge Abiteboul, Stéphane Grumbach, Michel Scholl and Luc Segoufin for helpful discussions. Libkin thanks all the members of the Verso team for their hospitality.

References


