Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020)
Main Track

# Reasoning about Measures of Unmeasurable Sets

**Marco Console**[1] , **Matthias Hofer**[1] , **Leonid Libkin**[1,2,3]
[1] University of Edinburgh ,
[2] ENS-Paris, PSL ,
[3] Neo4j

**Abstract**

In a variety of reasoning tasks, one estimates the likelihood of events by means of volumes of sets they define. Such sets need to be measurable, which is usually achieved by putting bounds, sometimes ad hoc, on them. We address the question how unbounded or unmeasurable sets can be measured nonetheless. Intuitively, we want to know how likely a randomly chosen point is to be in a given set, even in the absence of a uniform distribution over the entire space.

To address this, we follow a recently proposed approach of taking intersection of a set with balls of increasing radius, and defining the measure by means of the asymptotic behavior of the proportion of such balls taken by the set. We show that this approach works for every set definable in first-order logic with the usual arithmetic over the reals (addition, multiplication, exponentiation, etc.), and every uniform measure over the space, of which the usual Lebesgue measure (area, volume, etc.) is an example. In fact we establish a correspondence between the good asymptotic behavior and the finiteness of the VC dimension of definable families of sets. Towards computing the measure thus defined, we show how to avoid the asymptotics and characterize it via a specific subset of the unit sphere. Using definability of this set, and known techniques for sampling from the unit sphere, we give two algorithms for estimating our measure of unbounded unmeasurable sets, with deterministic and probabilistic guarantees, the latter being more efficient. Finally we show that a discrete analog of this measure exists and is similarly well-behaved.

## 1 Introduction

In a variety of reasoning tasks in areas including temporal and spatial reasoning or answering queries over missing data one regularly estimates likelihoods of certain events by computing volumes of sets that serve as a mathematical representation of such events. Very often these sets are given by constraints $p(\bar{x}) \leq 0$, where $p$ is a multi-variate polynomial (for instance, $x^2 + y^2 + z^2 - 1 \leq 0$ defines the unit ball in $\mathbb{R}^3$). This of course assumes that the set itself is *measurable* (i.e., its volume is well-defined). In most cases of reasonable constraints one uses in applications, bounded sets, such as the unit ball, will be measurable. Indeed, the construction of unmeasurable bounded sets invariably requires the axiom of choice, and it is not very natural to think of a practical reasoning task that would use such sets.

On the other hand, there are many tasks (with some examples outlined below) where we have *unbounded* sets of infinite area, volume, etc. That is, the reason for non-measurability is their unboundedness. We then want to measure such sets, or rather define the proportion of the entire space occupied by it. Thinking of it in a different way, if we somehow had a uniform distribution on the entire $\mathbb{R}^n$, the measure of a set $X \subseteq \mathbb{R}^n$ would be the probability that a randomly chosen point in $\mathbb{R}^n$ belongs to $X$.

To start with an abstract example, consider the positive quadrant $R^2_+ = \{(x, y) \mid x, y \in \mathbb{R}, x, y > 0\}$. Intuitively it is clear that it occupies a quarter of $\mathbb{R}^2$. If in addition we restrict it to $X_1 = \{(x, y) \in \mathbb{R}^2+ \mid y \geq x\}$, then this set occupies 1/8th (or 0.125) of $\mathbb{R}^2$. But what about $X_2 = \{(x, y) \in \mathbb{R}^2 \mid y \geq 2x\}$? Suddenly the intuition is lost. We shall later argue that the measure is roughly $0.074$.

Before we explain how we arrive at this, we look at typical AI applications where such reasoning about measures of unmeasurable sets could be of use.

**Temporal reasoning** Temporal networks are a formalism that captures relationships between points in time. They can define bounds (event $x$ happens between 1pm and 3pm), relative constraints (event $x$ precedes event $y$), and metric constraints (event $x$ happens three hours before event $y$). As argued in (Dechter, Meiri, and Pearl 1991) and much followup work, temporal networks can be represented by semi-linear sets, i.e., sets definable in the theory of the reals with addition. For example, to model a multi-step chemical reaction one needs a Boolean combination of conditions saying that forming $y$ from $x$ requires at least time $t$, defined by $y - x \geq t$. The standard reasoning task performed with temporal networks is satisfiability, i.e., checking whether there exists a configuration of the events consistent with the constraints. This gives a yes/no answer, but it cannot tell *how likely* a consistent configuration is. To answer the latter, we need to impose fixed (and quite possibly ad hoc) bounds.

Consider, e.g., constraints $(0 \leq y \leq 3) \vee (y \geq x + 12)$ indicating a successful outcome if $y$ happens in the first 3 hours of observation, or 12 hours after $x$ happen. If we restrict our attention to 24 hours only, these constraints are satisfied by points $(x, y) \in \mathbb{R}^2$ whose area is 25% of the area of $[0, 24] \times [0, 24]$. But what if we do not have this upper bound

of 24 hours and want to see how likely these constraints are? Intuitively, without an upper bound the constraint $0 \leq y \leq 3$ is very unlikely, and $y \geq x + 12$ will happen with probability 1/2 for randomly chosen points $x$ and $y$. The question is then how to formalize this intuition.

**Querying missing data** In database applications, or applications involving reasoning about data and querying (e.g., data integration (Lenzerini 2002) or OBDA (Bienvenu and Ortiz 2015)) one often has to deal with missing data represented by null values: essentially markers saying that some piece of data is not known at the moment. In data science, one often uses imputation techniques, i.e., replacing missing data by concrete values and then applying usual querying techniques. This is often suboptimal as it loses any information about the fact that data was not initially there, and in database applications techniques for querying complete and incomplete data are rather different (Imielinski and Lipski 1984; Libkin 2016a). The latter is based on the notion of certain answers assumed to be the correct one, although it tends to be computationally very expensive. Various solutions have been proposed such as approximation schemes (Feng et al. 2019; Greco, Molinaro, and Trubitsyna 2019; Libkin 2016b) or probabilistic guarantees (Libkin 2018). The latter essentially estimates how likely a query is to be true in a randomly chosen database. They work well for simple constraints used in queries – essentially equalities of values. In real-life queries with arithmetic operations are ubiquitous, and one needs to estimate the likelihood of conditions exactly like those specified above: $y \geq x$ or $y \geq 2x$.

If variables have naturally restricted ranges (e.g., professorial salaries, or class sizes in logic), it is easy to estimate the volume of the set defined by the constraints. If however the ranges of variables could be unbounded (e.g., salaries of executives, or class sizes in machine learning) we need to estimate a proportion of the entire space cut by the constraints – the exact problem described above.

**Default reasoning** To understand what a default is, (Koutras et al. 2018) proposed to connect it with topological properties of the set of worlds $[\![\alpha]\!]$ that satisfy a formula $\alpha$. Then $\beta$ is a default consequence of $\alpha$ if $[\![\alpha \wedge \neg\beta]\!]$ is "small" and $[\![\alpha \wedge \beta]\!]$ is "large". When dealing with measurable sets, one can define this via a degree of overlap of $[\![\alpha]\!]$ and $[\![\beta]\!]$, i.e., by $\mathsf{Vol}(\alpha \wedge \beta) \leq \tau\mathsf{Vol}(\alpha)$, for some threshold $\tau$. This notion of default is very natural, but it falls short in the case of unbounded sets of infinite volume. If we could measure such sets, we could extend such a default reasoning approach to more general settings permitting fewer restrictions on sets of possible worlds.

**Spatial reasoning** Reasoning about the spatial configuration of geometrical objects is a prominent AI application. An example of languages for spatial reasoning is RCC8 (Egenhofer and Franzosa 1991; Kontchakov, Pratt-Hartmann, and Zakharyaschev 2010) based on topological primitives such as containment and disjointness relations. Sometimes however a quantitative aspect needs to be added, for instance for reasoning on how big an overlap of two sets is. This was proposed in (Godoy and Rodríguez 2002); the overlap of two measurable sets $A$ and $B$ can be defined as $\text{Overlap}(A, B) = \mathsf{Vol}(A)/\mathsf{Vol}(A \cup B)$. However if sets $A$ and $B$ are unmeasurable this is again undefined. The problem can be "fixed" by imposing some arbitrary restrictions, e.g, looking at the restriction of $A$ and $B$ to points whose Euclidean norm is bounded by some number $r$, but then it depends on an ad hoc choice of $r$. By measuring unmeasurable sets, we would eliminate this ad hoc dependence.

Our goal is to define such measures and study their structural and computational properties. This problem was first studied (Console, Hofer, and Libkin 2019) using the following approach. It first defined the measure of a set $X$ restricted to the ball of radius $r$ as $\mathfrak{m}_r(X) = \mathsf{Vol}(X \cap B^n_r)/\mathsf{Vol}(B^n_r) \in [0, 1]$, and then set $\mathfrak{m}(X) = \lim_{r\to\infty} \mathfrak{m}_r(X)$. It showed the following:

- For sets $X$ definable with the common functions such as $+, \cdot, -, \div, s^x$ for $s > 0$ etc, $\mathfrak{m}(X)$ is well-defined, i.e., the limit exists (but see more on this below);
- Even with $+$ available, $\mathfrak{m}(X)$ is usually an irrational number and needs to be approximated. An approximation scheme with additive error guarantees was given for linear constraints in a specific syntactic shape (essentially a union of convex polytopes). Additive error guarantees are necessary since there are provably no approximation schemes with multiplicative error guarantees, even for order constraints.

This leaves a multitude of questions unanswered that we outline now, together with our contributions.

*Why Lebesgue measure?* It was motivated by the fact that we know how to sample uniformly from the $n$-dimensional ball, and this could be useful in approximation schemes. But could we get different values of $\mathfrak{m}(X)$ with different uniform measures on $\mathbb{R}^n$? We show that this is not the case.

*Existence of the measure.* The proof of the existence of $\mathfrak{m}(X)$ cited above relied on a result from (Karpinski and Macintyre 1997) on approximability of volumes by first-order formulae. It used the result as *stated* in that paper, but upon examining the actual proof of (Karpinski and Macintyre 1997) one discovers that it only works for subsets of $[0, 1]^n$ which, not surprisingly, breaks the argument for *unbounded* sets! We nonetheless find a (rather nontrivial) workaround and prove an even more general existence result for $\mathfrak{m}(X)$. The key point of it is that sets definable with the help of functions that do not exhibit a periodic behavior, such as $+, \cdot, -, \div, s^x, \log_s(x)$, are fine, but those with a periodic behavior such as $\sin, \cos, \tan, \arcsin$ etc, are not.

*Measure via unit spheres.* We find a new characterization of $\mathfrak{m}(X)$ for $X \subseteq \mathbb{R}^n$ in terms of areas of some sets, called $\mathsf{Ult}(X)$, on the unit sphere in $\mathbb{R}^n$. That is, one only has to look at one specific sphere of a fixed radius to determine the asymptotic behavior.

*Approximating $\mathfrak{m}(X)$ by sampling from the sphere.* The advantage of the new characterization is due to well-understood algorithms for generating points uniformly on the unit sphere in $\mathbb{R}^n$, see (Blum, Hopcroft, and Kannan 2020). This leads to two new algorithms for estimating

$\mathfrak{m}(X)$, i.e., returning a number in the interval $\mathfrak{m}(X) \pm \varepsilon$ for a given $\varepsilon > 0$. One is a deterministic algorithm, while the other is randomized (using sampling from the unit sphere in an essential way) and having much better complexity.

Moreover, existing algorithms were very restrictive: they worked only for sets definable by linear inequalities, and only guaranteed good complexity when such sets had a specific representation that was effectively a union of convex polytopes. The new approach gives us an algorithm that works for sets definable with any of the usual arithmetic function (addition, multiplication, exponentiation, logarithms). Furthermore, in its randomized version it remains efficient without any restrictions on the syntactic shape of the formula that defines the set.

*Discrete measure.* Finally, we also study a discretization of the measure, that, instead of using the volume, counts the number of integer points in a set. We show that our approach is robust, as this measure in the limit coincides with the previously defined one.

*Remark* While we use the standard terminology of measure theory, readers unfamiliar with it can simply think of the Lebesgue measure as the extension of length, area, and volume to multi-dimensional spaces. This will be sufficient for understanding the paper.

## 2 Preliminaries

We use $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ to denote natural, integer, rational, and real numbers. Elements of $\mathbb{R}^n$ are $n$-dimensional points, i.e., tuples $\bar{a} = (a_1, \ldots, a_n)$ of real numbers. By $\|\bar{a}\|$ we denote its Euclidean norm $(\sum_{i=1}^n a_i^2)^{1/2}$. We write $B_r^n$ for the $n$-dimensional ball of radius $r$, i.e., $\{\bar{a} \in \mathbb{R}^n \mid \|\bar{a}\| \leq r\}$ and $S_r^{n-1}$ for its boundary, the $(n-1)$-sphere, i.e., $\{\bar{a} \in \mathbb{R}^n \mid \|\bar{a}\| = r\}$.

If $X \subseteq \mathbb{R}^n$, we denote by $\mathsf{Vol}^n(X)$ the $n$-dimensional Lebesgue measure of $X$ (i.e., the $n$-dimensional volume). More generally, $\mathsf{Vol}^k(X)$ denotes the $k$-dimensional Lebesgue measure when the dimension of $X$ is $k$ or smaller. For example, $\mathsf{Vol}^n(B_r^n) = (\sqrt{\pi}r)^n/\Gamma(n/2+1)$ and $\mathsf{Vol}^{n-1}(S_r^{n-1}) = 2\sqrt{\pi}^n r^{n-1}/\Gamma(n/2)$. We omit indices and write just $\mathsf{Vol}$ when the dimension is clear from the context.

For a set $X \subseteq \mathbb{R}^n$, a point $\bar{a} \in \mathbb{R}^n$, and $c \in \mathbb{R}$, we write $\bar{a} + X$ for $\{\bar{a} + \bar{x} \mid \bar{x} \in X\}$ and $c \cdot X$ for $\{c \cdot \bar{x} \mid \bar{x} \in X\}$.

We deal with sets definable in first-order languages over structures on numbers (primarily $\mathbb{R}$). They are determined by the vocabulary $\Omega$ of allowed functions such as $+, \cdot, e^x, \ln x, \sin x$. We assume that the standard comparisons $<$ and $=$ are always available. First-order formulae in the language of $\Omega$ are defined in the usual way. Given a countably infinite set VAR of variables, each number and each variable are terms, and if $t_1, \ldots, t_n$ are terms, and $f$ is a $n$-ary function then $f(t_1, \ldots, t_n)$ is a term. If $t, t'$ are terms, then $t = t'$ and $t < t'$ are atomic formulae. Formulae are closed under the Boolean connectives $\vee, \wedge$, and $\neg$, and under quantifiers $\exists$ and $\forall$. Definitions of free and bound variables are standards, and we write $\varphi(\bar{x})$ to indicate that $\bar{x}$ is the tuple of free variables of $\varphi$.

The semantics of these is standard. Given a formula $\varphi(\bar{x})$ with $\bar{x}$ of length $n$, and a $n$-tuple $\bar{a}$ over $\mathbb{R}$, we say that $\varphi(\bar{a})$ is true if it is satisfied in the structure $\langle \mathbb{R}, \Omega_\varphi \rangle$ where $\Omega_\varphi$ contains all the functions and comparison predicates used in $\varphi$. The set of all $\bar{a}$ such that $\varphi(\bar{a})$ is true is denoted by $[\![\varphi]\!]$; it is a subset of $\mathbb{R}^n$. For example, if $\varphi(x, y) = (x \cdot x + y \cdot y \leq 1)$ then $[\![\varphi]\!] = B_1^2$.

## 3 When can we Measure Unmeasurable Sets?

We start by recalling the measure as defined in (Console, Hofer, and Libkin 2019). Consider an arbitrary set $X \subseteq \mathbb{R}^n$ so that for each $n$-dimensional ball $B_r^n$ of radius $r$, the set $X \cap B_r^n$ is Lebesgue-measurable. Define then

$$\mathfrak{m}_r(X) = \frac{\mathsf{Vol}(X \cap B_r^n)}{\mathsf{Vol}(B_r^n)} \text{ and } \mathfrak{m}(X) = \lim_{r \to \infty} \mathfrak{m}_r(X). \quad (1)$$

In other words, we take the proportion of the $r$-ball occupied by $X$, and see how it behaves when $r$ increases. For sets definable by formulae, we shall write $\mathfrak{m}(\varphi)$ for $\mathfrak{m}([\![\varphi]\!])$.

Of course $B_r^n$ is the set of points $\bar{a}$ with $\|\bar{a}\| \leq r$, and in principle one could have used another norm $\|\bar{a}\|_p = \left(\sum_{i=1}^n |a_i|^p\right)^{1/p}$ for $p$ between 1 and $\infty$, but for all of those except for the case of $p = 2$ the measure $\mathfrak{m}(X)$ defined via them would not be invariant under volume-preserving transformations such as rotations. This is the reason for using the Euclidean norm.

We first show that instead of the Lebesgue measure $\mathsf{Vol}(\cdot)$, we could use any well-behaved measure on $\mathbb{R}^n$. A measure $\mu$ on $\mathbb{R}^n$ is *uniform* if for every two points $\bar{a}_1$ and $\bar{a}_2$ and every $r > 0$ we have $\mu(\bar{a}_1 + B_r^n) = \mu(\bar{a}_2 + B_r^n)$, see (Mattila 1995). Note that this is not the same as a uniform distribution, which does not exist over $\mathbb{R}^n$, while uniform measures exist (e.g., the Lebesgue measure). Given such a measure $\mu$, we normalize it by considering $\mu_r(\cdot) = \mu(\cdot)/\mu(B_r^n)$ so that $\mu_r(B_r^n) = 1$. Then, for a set $X$, we define $\mathfrak{m}_r^\mu(X) = \mu_r(X \cap B_r^n)$ and $\mathfrak{m}^\mu(X) = \lim_{r \to \infty} \mathfrak{m}_r^\mu(X)$.

It turns out that with every reasonable uniform measure $\mu$, we get the same value $\mathfrak{m}^\mu$ as was given by (1) for the Lebesgue measure. The condition we need is that $\mu$ does not assign non-zero values to sets $X$ with $\mathsf{Vol}(X) = 0$. This is known in measure theory as being absolutely continuous.

**Proposition 1.** *If $\mu$ is a uniform absolutely continuous measure, then $\mathfrak{m}^\mu = \mathfrak{m}$.*

With this Proposition and the justified use of the measure $\mathfrak{m}$, we now move to the existence result.

### 3.1 Existence of the Measure

The measure $\mathfrak{m}(X)$ may not exist for two reasons: first, some of the sets $X \cap B_r^n$ may not be Lebesgue-measurable, and second, the limit in (1) may not exist. We now give a general model-theoretic criterion for the existence of $\mathfrak{m}(X)$ for sets $X$ definable in first-order logic over structures $\langle \mathbb{R}, \cdots \rangle$; such structures use various arithmetic functions and comparisons such as $=$ and $<$. We then show that these general conditions apply in many structures of interest and capture the usual arithmetic functions.

A set $X$ is *definable* if there is a first-order formula $\varphi$ so that $[\![\varphi]\!] = X$. A function is definable if its graph is

definable. For example, subtraction is definable in $\langle \mathbb{R}, + \rangle$, division in $\langle \mathbb{R}, +, \cdot \rangle$, and $\log_s(x)$ in $\langle \mathbb{R}, +, \cdot, e^x \rangle$ for each fixed $s > 0$. To see the latter notice that $y = \log_s(x)$ iff $e^{(\ln s) \cdot y} = x$. We assume that constants, i.e., elements of $\mathbb{R}$ such as $\ln s$, can be used in formulae.

We say that $\langle \mathbb{R}, \cdots \rangle$ has *definable Skolemization* if for each definable $X \subseteq \mathbb{R}^2$ there is a definable function $f_X$ so that $(x, f_X(x)) \in X$ whenever $x$ is such that there is some pair $(x, y) \in X$. We say that a definable function $f : \mathbb{R} \to \mathbb{R}$ is *eventually monotone* if there is an $x_0 \in \mathbb{R}$ such that $f$ is monotone on $(x_0, \infty)$. That is to say, $f$ is either monotonically increasing ($x \geq x'$ imply $f(x) \geq f(x')$) or monotonically decreasing ($x \geq x'$ imply $f(x) \leq f(x')$).

Given a formula $\varphi(\bar{x}, \bar{y}, \bar{z})$, with $|\bar{x}| = n$ and $|\bar{y}| = m$, for each tuple $\bar{c}$ of the same length as $\bar{z}$ it defines a family $\Phi_{\bar{c}} = \{\bar{a} \in \mathbb{R}^n \mid \varphi(\bar{a}, \bar{b}, \bar{c}) \text{ holds}\}_{\bar{b} \in \mathbb{R}^m}$ of subsets of $\mathbb{R}^n$. Recall that for a family $\Phi$ of subsets of $\mathbb{R}^n$, its *VC dimension* is the maximum cardinality of a finite $C \subseteq \mathbb{R}^n$ such that $\{C \cap F \mid F \in \Phi\} = 2^C$; if no such maximum exists, then the VC dimension is infinite, cf. (Anthony and Biggs 1992). If for every such first-order formula $\varphi(\bar{x}, \bar{y}, \bar{z})$ their is a number $v$ so that the VC dimension of all $\Phi_{\bar{c}}$ is at most $v$ we say that every *parameterized definable family has finite VC dimension*.

Now we are ready to formulate the existence theorem.

**Theorem 1.** *Suppose we have a structure $\langle \mathbb{R}, +, \cdot, <, \cdots \rangle$ such that:*

- *every parameterized definable family has finite VC dimension;*
- *it has definable Skolemization for every definable subset of $\mathbb{R}^2$;*
- *every definable function $f : \mathbb{R} \to \mathbb{R}$ is eventually monotone;*
- *every bounded definable set is Lebesgue-measurable.*

*Then $\mathfrak{m}(X)$ exists for every definable set $X$.*

**Examples** The key example of structures on reals satisfying conditions of the Theorem are *o-minimal* structures, cf. (Van den Dries 1998), that are structures on $\mathbb{R}$ so that for every first-order formula $\varphi(x)$ with one free variable, $[\![\varphi]\!] \subseteq \mathbb{R}$ is a finite union of intervals. The best known example is the real field $\langle \mathbb{R}, +, \cdot, 0, 1, < \rangle$; we can also assume that each element $c \in \mathbb{R}$ is available as a constant. This follows from Tarski's quantifier elimination: every formula $\varphi(x)$ is equivalent to a Boolean combination of polynomial inequalities $p(x) \leq 0$, and in intervals defined by roots of all such polynomials the truth value of $\varphi$ does not change.

A remarkable example of o-minimality is the exponential field $\langle \mathbb{R}, +, \cdot, e^x \rangle$, see (Wilkie 1996). In this structure one has $+, -, \cdot, \div, s^x, \log_s(x)$ definable (for $s > 0$). What is not definable in o-minimal structures are functions and sets with a periodic behavior, e.g., $\mathbb{N}$, or trigonometric functions $\sin$, $\cos$, $\arcsin$, $\arccos$, etc.

That o-minimal structures satisfy the conditions of the theorem follows from their well-known properties. Finiteness of VC dimension (with parameters) and eventual monotonicity are explicitly stated in (Van den Dries 1998). Cell decomposition (Van den Dries 1998) ensures that each definable set is a finite union of cells which are open sets (in their dimension) and thus measurable if bounded. Finally, for a definable set $X \subseteq \mathbb{R}^2$, by o-minimality each $X_x = \{y \mid (x, y) \in X\}$ is a finite union of intervals, and one can simply take the midpoint of the smallest one (or $a-1$ for the interval $(-\infty, a)$ and $a + 1$ for $(a, \infty)$).

There are generalizations of o-minimality that have finite VC dimension and other properties such as quasi-o-minimality (Belegradek, Peterzil, and Wagner 2000) to which the result would apply as well.

*Remark* Before we sketch the proof, we point out that o-minimality (or its generalization as in the statement of the theorem) is essential. One might be tempted to think that standard techniques from measure theory would deliver the result for arbitrary sets $X$ defined with smooth functions, but this is not the case: with the addition of functions with periodic behavior (e.g., triginometry) or a predicate for natural numbers, the measure actually does *not* exist (see the discussion in the conclusions). And it is known that these essentially deliniate the boundary of o-minimality: if adding nicely behaved functions to $\langle \mathbb{R}, +, \cdot, e^x \rangle$ results in a non-o-minimal structure, then such a structure would define either $\mathbb{N}$ or $\sin(x)$, see (Miller 2011). Hence the techniques from the field of o-minimality and tame topology (Van den Dries 1998) are essential for our proof.

**Proof sketch of the existence theorem** Assume we have a structure on the reals as in the statement of the Theorem. Given a formula $\alpha(\bar{x}, \bar{y})$ with $|\bar{x}| = n$ and $|\bar{y}| = m$, and $\bar{a} \in \mathbb{R}^n$, we shall use the notation $\alpha(\bar{a}, \mathbb{R}^m)$ for the set of $\bar{b} \in \mathbb{R}^m$ so that $\alpha(\bar{a}, \bar{b})$ holds. In the proof below, every reference to $\mathsf{Vol}(\cdot)$ will involve a bounded definable set, and thus by the last assumption the Lebesgue measure is well defined. Suppose we have a formula $\alpha(\bar{y}, \bar{u})$, and fix $\varepsilon > 0$. Then a formula $\beta(\bar{y}, z)$ is called an $\varepsilon$-volume approximation for $\alpha$ if two conditions are satisfied. First, for every interpretation $\bar{a}$ of $\bar{y}$, there exists $v \in \mathbb{R}$ such that $\beta(\bar{a}, v)$ holds. Second, if $\beta(\bar{a}, v)$ holds, then $|v - \mathsf{Vol}(\alpha(\bar{a}, \mathbb{R}^m)| < \varepsilon$, where $m$ is the length of $\bar{u}$. Recall that $\alpha(\bar{a}, \mathbb{R}^m)$ is the set of all $\bar{b} \in \mathbb{R}^m$ such that $\alpha(\bar{a}, \bar{b})$ holds.

It was shown in (Karpinski and Macintyre 1997) (see also (Koiran 1995)) that, under some conditions, as long as $\alpha(\bar{a}, \mathbb{R}^m) \subseteq [0, 1]^m$ for each $\bar{a}$, such $\varepsilon$-volume approximation exists for each $\varepsilon > 0$. The conditions, as one follows specifically the construction in (Karpinski and Macintyre 1997), are the availability of $+$ in the vocabulary of the structure, and the finite VC dimension of parameterized definable families. Essentially, using $+$, it constructs from $\alpha$ another formula with parameters, and based on the VC dimension of the family it defines, constructs yet another formula defining the volume, again referring to $\alpha$ and using $+$. The result is even stronger: produced formulae ensure that $|v - \mathsf{Vol}(\alpha(\bar{a}, \mathbb{R}^m)| < \varepsilon/4$ implies that $\beta(\bar{a}, v)$ holds.

Now suppose we are given a formula $\varphi(\bar{x})$ in $n$ variables. We shall prove that $\mathfrak{m}(\varphi)$ exists. Note that there is an FO formula in $n + 1$ variables defining the condition $\|\bar{x}\| \leq r$, where the length of $\bar{x}$ is $n$, simply by checking $x_1^2 + \cdots x_n^2 \leq$

$r^2$; where $\bar{x} = (x_1, \ldots, x_n)$. Thus, for $\varphi(\bar{x})$ we can define a new formula $\varphi'(r, \bar{x}) = \varphi(\bar{x}) \wedge (\|x\| \leq r)$. The formula $\varphi'(r, \bar{x})$ defines the set $[\![\varphi]\!] \cap B^n_r$, i.e., $\varphi'(r, \bar{a})$ is true for every $\bar{a} \in ([\![\varphi]\!] \cap B^n_r)$. Consider next the formula

$$\psi(r, \bar{x}) \;=\; \exists y_1, \ldots, y_n \left(\varphi'(r, \bar{y}) \wedge \bigwedge_{i=1}^{n} x_i = (y_i + r)/2r\right).$$

Then $\psi(r, \mathbb{R}^n) = \frac{1}{2r}\big(\varphi'(r, \mathbb{R}^n) + (r, \ldots, r)\big)$ and thus $\psi(r, \mathbb{R}^n) \subseteq [0, 1]^n$. Adding a fixed vector $(r, \ldots, r)$ does not change the volume, and dividing each coordinate by $2r$ corresponds to a linear transformation whose matrix has $2r$ in each diagonal position; thus we have $\mathsf{Vol}(\psi(r, \mathbb{R}^n)) = \mathsf{Vol}(\varphi'(r, \mathbb{R}^n))/(2r)^n$, and hence, since $\mathsf{Vol}(B^n_r) = b_n \cdot r^n$ where $b_n = \pi^{n/2}/\Gamma(n/2 + 1)$,

$$\mathfrak{m}_r(\varphi) \;=\; \frac{\mathsf{Vol}(\varphi'(r, \mathbb{R}))}{b_n \cdot r^n} \;=\; \frac{2^n}{b_n}\mathsf{Vol}(\psi(r, \mathbb{R}^n)). \qquad (2)$$

Using the volume approximating formulae described above, we can now define a formula $\alpha'_\varepsilon(z, y)$, for each $\varepsilon > 0$, so that $\alpha'(r, \mathbb{R}) \neq \emptyset$ for each $r > 0$ and $\alpha'_\varepsilon(r, v)$ implies $|v - \mathsf{Vol}(\psi(r, \mathbb{R}^n))| < \varepsilon$. By (2) it means that there are formulae $\alpha_\varepsilon(z, y)$ such that $\alpha(r, \mathbb{R}) \neq \emptyset$ for each $r > 0$ and $\alpha_\varepsilon(r, v)$ implies $|v - \mathfrak{m}_r(\varphi)| < \varepsilon$. Using this, we show

**Lemma 1.** *For each $\varepsilon > 0$, there is a function $f_\varepsilon : \mathbb{R} \to \mathbb{R}$ such that $|\mathfrak{m}_r(\varphi) - f_\varepsilon(r)| < \varepsilon$ for each $r > 0$ and $\lim_{r\to\infty} f_\varepsilon(r)$ exists.*

Indeed, the formula $\alpha_\varepsilon(z, y)$ defines a subset $\mathbb{R}^2$ and thus there exists a definable Skolem function $f_\varepsilon : \mathbb{R} \to \mathbb{R}$ such that $\alpha_\varepsilon(z, f_\varepsilon(z))$ holds and hence $|\mathfrak{m}_r(\varphi) - f_\varepsilon(r)| < \varepsilon$. We can assume this function to be total by assigning it a fixed value for all $r \leq 0$. Thus again by the assumption on the structure this function is eventually monotone. Note that the function is also bounded as $|\mathfrak{m}_r(\varphi) - f_\varepsilon(r)| < \varepsilon$ and $\mathfrak{m}_r(\varphi) \in [0, 1]$. Assume that function $f_\varepsilon$ is monotonically nondecreasing after some $r_0 > 0$ (the case of $f_\varepsilon$ being monotonically nonincreasing is symmetric). Because of boundedness we have $s = \sup_{r>r_0} f_\varepsilon(r)$ exists. By monotonicity, it easily follows that $s = \lim_{r\to\infty} f_\varepsilon(r)$.

Using the lemma define $L_\varepsilon = \lim_{r\to\infty} f_\varepsilon(r)$. Since $|f_\varepsilon(r) - f_\delta(r)| < \varepsilon + \delta$, we have $|L_\varepsilon - L_\delta| < \varepsilon + \delta$. Next we show that $L = \lim_{\varepsilon\to 0+} L_\varepsilon$ exists. Indeed, fix any $\delta > 0$. Then for $\varepsilon_1, \varepsilon_2 < \delta/2$ we have $|L_{\varepsilon_1} - L_{\varepsilon_2}| < \delta$ and thus by Cauchy convergence (viewing $L_\varepsilon$ as a function from $\mathbb{R}^+$ to $\mathbb{R}$) the limit of $L_\varepsilon$ exists.

We finally prove the theorem by showing that $\lim_{r\to\infty} \mathfrak{m}_r(\varphi) = L$. Fix $\delta > 0$. Then there is $\varepsilon > 0$ and $r_0 > 0$ such that for all $r > r_0$ we have that each of $|\mathfrak{m}_r(\varphi) - f_\varepsilon(r)|$, $|f_\varepsilon(r) - L_\varepsilon|$, and $|L - L_\varepsilon|$ is smaller than $\delta/3$. Then $|\mathfrak{m}_r(\varphi) - L| \leq |\mathfrak{m}_r(\varphi) - f_\varepsilon(r)| + |f_\varepsilon(r) - L_\varepsilon| + |L - L_\varepsilon| < \delta$ as required. All of these three statements are routine to prove. □

### 3.2 Basic Properties

The Lebesgue measure is well-behaved, and some of its good properties carry over to our measure $\mathfrak{m}$. We analyze the most important aspects: positivity, monotonicity, translation invariance, the characterization of zero sets and the interaction with Cartesian product. The latter two will require techniques to be developed in the next section.

**Proposition 2.** *For any collection $\{X_i\}_{i\in\mathbb{N}}$ of subsets of $\mathbb{R}^n$ such that $X_i \subseteq X_j$ whenever $i \leq j$, if $\mathfrak{m}(X_i)$ exists for all $i \in \mathbb{N}$ then:*

- $\mathfrak{m}(X_i) \geq 0$;
- $\mathfrak{m}(X_i) \leq \mathfrak{m}(X_j)$;
- $\mathfrak{m}(\bigcup_{i\geq 1} X_i) = \lim_{n\to\infty} \mathfrak{m}(X_n)$;

*Furthermore, if $\mathfrak{m}(X)$ exists for $X \subseteq \mathbb{R}^n$, then $\mathfrak{m}(c \cdot X) = \mathfrak{m}(\bar{a} + X) = \mathfrak{m}(X)$ for every $c > 0$ and $\bar{a} \in \mathbb{R}^n$.*

## 4 An Alternative Characterization

We now provide a characterization of the measure $\mathfrak{m}(X)$ in terms of the usual Lebesgue measure of subsets of a unit sphere (or a sphere of a fixed radius). One immediate advantage is that this characterization eliminates the need to reason about asymptotics. The second advantage is that such a characterization can be easily used in algorithms. Indeed, there are well-known techniques for sampling points from a sphere (Blum, Hopcroft, and Kannan 2020), and these will be used later to provide good estimates for $\mathfrak{m}(X)$. In this section we concentrate on sets definable in o-minimal structures over the reals, i.e., sets definable in first-order logic with operations such as $+, \cdot, s^x$, etc.

Since the likelihood measure $\mathfrak{m}$ is defined as a limit, it is intuitively clear that it does not depend on any finite parts, i.e., intersections of $X$ with balls of small radius. This intuition is right, and the important aspect, when it comes to the computation of $\mathfrak{m}(X)$, is the asymptotic satisfaction of the formula into the individual directions. That means, for any point from the unit sphere, $\bar{z} \in S^{n-1}_1$, we only need to determine whether $X$ ultimately covers the ray $c \cdot \bar{z}$, i.e. whether there is a value $c_0 > 0$ such that for all $c > c_0$ we have $(c \cdot \bar{z}) \in X$. The proportion of points $\bar{z}$ from the unit sphere for which this is the case determines $\mathfrak{m}(X)$. More formally, we first give the following definition.

**Definition 1.** *A vector $\bar{z} \in S^{n-1}_r$* ultimately covers *$X \subseteq \mathbb{R}^n$ if there exists $c_0 \in \mathbb{R}^+$ such that $c \cdot \bar{z} \in X$ for all $c > c_0$. The set of such vectors is denoted by $\mathsf{Ult}_r(X)$.*

If $X = [\![\varphi]\!]$, we write $\mathsf{Ult}(\varphi)$ and refer to $\bar{z} \in \mathsf{Ult}(\varphi)$ as ultimately satisfying $\varphi$. Note that the radius $r$ is irrelevant in this definition, as $\mathsf{Ult}_r(\varphi) = r \cdot \mathsf{Ult}_1(\varphi)$, and thus, when we write $\mathsf{Ult}(\varphi)$, or $\mathsf{Ult}(X)$, we actually mean $\mathsf{Ult}_1(\varphi)$, i.e., the set of all ultimately satisfying vectors on the unit sphere.

The power of o-minimality tells us that for a formula $\varphi(\bar{x})$ with $|\bar{x}| = n$ over such a structure, we have a dichotomy: every point $\bar{z}$ on the unit sphere is either in $\mathsf{Ult}(\varphi)$ or $\mathsf{Ult}(\neg\varphi)$.

**Lemma 2.** *If $\varphi$ is over an o-minimal structure on $\mathbb{R}$, then $\mathsf{Ult}(\varphi) = S^{n-1} \setminus \mathsf{Ult}(\neg\varphi)$.*

*Proof.* Given $\varphi(\bar{x})$, and $\bar{z} \in S^{n-1}$, define a formula $\psi(y)$ in one free variable as $\varphi(y \cdot \bar{z})$. By o-minimality (which remains true if we add constants for all elements of $\mathbb{R}$), this formula defines a finite union of intervals. Thus, there is $c_0$ such that either for all $c > c_0$ we have $\psi(c)$ or for all $c > c_0$ we have $\neg\psi(c)$. In turn, the former implies $\bar{z} \in \mathsf{Ult}(\varphi)$ and the latter that $\bar{z} \in \mathsf{Ult}(\neg\varphi)$. The claim follows straightforwardly. □

Now using this lemma we prove the main characterization result for $\mathfrak{m}(X)$ where $X$ is definable in an o-minimal structure.

**Theorem 2.** *Let $\varphi(\bar{x})$, with $|\bar{x}| = n$, be a formula over an o-minimal structure $\langle \mathbb{R}, <, \cdots \rangle$. Then*

$$\mathfrak{m}(\varphi) = \frac{\mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))}{\mathsf{Vol}^{n-1}(S_1^{n-1})}\,.$$

*Proof Sketch.* Let $Z \subseteq S^{n-1}$. Define $\mathrm{cone}(Z) = \{c \cdot \bar{z} \mid \bar{z} \in Z,\ 0 \le c \le 1\}$. We show that for every $\varepsilon > 0$ there is a value $r_0$ such that for all $r \ge r_0$ we have that:

$$\left| \frac{\mathsf{Vol}^n([\![\varphi]\!] \cap B_r^n)}{\mathsf{Vol}^n(B_r^n)} - \frac{\mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))}{\mathsf{Vol}^{n-1}(S_1^{n-1})} \right| \le \varepsilon \qquad (3)$$

From now on, let $\varepsilon > 0$ be fixed. We start by proving the following

$$\frac{\mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))}{\mathsf{Vol}^{n-1}(S_1^{n-1})} = \frac{\mathsf{Vol}^n(\mathrm{cone}(\mathsf{Ult}_r(\varphi)))}{\mathsf{Vol}^n(B_r^n)}$$

to ensure the same denominator. Next, we have that

$$\mathsf{Vol}^n([\![\varphi]\!] \cap B_r^n) = \int_0^r \mathsf{Vol}^{n-1}([\![\varphi]\!] \cap S_x^{n-1})dx \ \text{ and}$$

$$\mathsf{Vol}^n(\mathrm{cone}(\mathsf{Ult}_r(\varphi))) = \int_0^r x^{n-1} \cdot \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))dx\,.$$

Setting $S_{\varphi,x} = [\![\varphi]\!] \cap S_x^{n-1}$, we get that

$$\begin{aligned}
&\int_0^r \mathsf{Vol}^{n-1}(S_{\varphi,x})dx - \int_0^r x^{n-1} \cdot \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))dx \le \\
&\quad \le \int_0^r |\mathsf{Vol}^{n-1}(S_{\varphi,x}) - x^{n-1} \cdot \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))|dx \\
&\quad \le r^{n-1} \int_0^r |\mathsf{Vol}^{n-1}(\frac{1}{x} \cdot S_{\varphi,x}) - \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))|dx
\end{aligned}$$

We refer to the last value in the above equation as $u$. Next, we note that

$$\lim_{x\to\infty} \mathsf{Vol}^{n-1}(\frac{1}{x} \cdot S_{\varphi,x}) = \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))\,.$$

That is, for all $\varepsilon' > 0$ there is a value $x_0$ such that for all $x \ge x_0$ we have that $|\mathsf{Vol}^{n-1}(\frac{1}{x} \cdot S_{\varphi,x}) - \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))| \le \varepsilon'$.

Since we are interested in proving (3) for $r \to \infty$ we assume that we take $r$ sufficiently large so that $r > x_0$. Fix $\varepsilon' < c_n \cdot \varepsilon$, where $b_n = \pi^{\frac{n}{2}}/\Gamma(\frac{n}{2}+1)$ is the ball constant. Then we get the following estimate on $u$:

$$\begin{aligned}
u = r^{n-1} \cdot \Big(&\int_0^{x_0} |\mathsf{Vol}^{n-1}(\frac{1}{x} \cdot S_{\varphi,x}) - \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))| + \\
&+ \int_{x_0}^r |\mathsf{Vol}^{n-1}(\frac{1}{x} \cdot S_{\varphi,x}) - \mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))|\Big) \\
&\le r^{n-1} \cdot (x_0 \cdot \mathsf{Vol}^{n-1}(S_{x_0}^{n-1}) + \varepsilon' \cdot r)
\end{aligned}$$

Now, we set $c_0 = x_0 \cdot \mathsf{Vol}^{n-1}(S_{x_0}^{n-1})$, and get that

$$\left| \frac{\mathsf{Vol}^n([\![\varphi]\!] \cap B_r^n) - \mathsf{Vol}^n(\mathrm{cone}(\mathsf{Ult}_r(\varphi)))}{\mathsf{Vol}^n(B_r^n)} \right| \le \frac{c_0}{c_n \cdot r} + \frac{\varepsilon'}{c_n}$$

which is smaller than $\varepsilon$ as long as $r > c_0/(c_n \cdot \varepsilon - \varepsilon')$. This proves (3) and the theorem. □

For the example in the introduction of $X = \{(x,y) \mid x, y \ge 0,\ y \ge 2x\}$, the set $\mathsf{Ult}(X)$ is the arc of $S^1$ corresponding to angles between $\arctan(2)$ and $\pi/2$, and hence $\mathsf{Vol}^1(\mathsf{Ult}(X))/2\pi = 1/4 - \arctan(2)/2\pi$, or approximately 0.074, as claimed.

An immediate corollary of Theorem 2 is that we can use uniform sampling from the unit sphere to evaluate $\mathfrak{m}(X)$.

**Corollary 1.** *For a set $X \subseteq \mathbb{R}^n$ definable in an o-minimal structure over $\mathbb{R}$, we have*

$$\mathfrak{m}(X) = \mathsf{Prob}(Z \in \mathsf{Ult}(X))$$

*where $Z \sim \mathbb{U}(S_1^{n-1})$, i.e. $Z$ is a random variable that is uniformly distributed on the unit-sphere.*

**Additional properties of the measure** We can use Theorem 2 to establish additional properties of the measure about its zero sets, and its interaction with the Cartesian product.

**Proposition 3.** *Consider an o-minimal structure over $\mathbb{R}$, and two definable sets $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$. Then*

- $\mathfrak{m}(X) = 0$ *iff* $\mathsf{Vol}^{n-1}(\mathsf{Ult}(X)) = 0$.
- $\mathfrak{m}(X \times Y) = \mathfrak{m}(X) \cdot \mathfrak{m}(Y)$.

*Proof sketch.* The first item is a direct consequence of Theorem 2. For the second, let $n+m$ independent and identically distributed random variables $Z_i$ be given, with $Z_i \sim N(0,1)$, i.e. the $Z_i$ are distributed according to the standard normal distribution. Then, we know that for $\bar{Z} = (Z_1, \ldots, Z_{n+m})$ we have that $\bar{Z}/\|\bar{Z}\| \sim \mathbb{U}(S_1^{n+m-1})$, i.e it is uniformly distributed on the $(n+m-1)$-sphere, see (Blum, Hopcroft, and Kannan 2020). When we define the following two random variables, $\bar{Z}_X = (Z_1, \ldots, Z_n)$ and $\bar{Z}_Y = (Z_{n+1}, \ldots, Z_{n+m})$. It is easy to see that $(\bar{x}, \bar{y}) \in \mathsf{Ult}(X \times Y)$ iff $\bar{x}/\|\bar{x}\| \in \mathsf{Ult}(X)$ and $\bar{y}/\|\bar{y}\| \in \mathsf{Ult}(Y)$. Using this and Corollary 1 we show:

$$\begin{aligned}
\mathfrak{m}(X \times Y) &= \mathsf{Prob}(\bar{Z}/\|\bar{Z}\| \in \mathsf{Ult}(X \times Y)) \\
&= \mathsf{Prob}\Big(\bar{Z}_X/\|\bar{Z}_X\| \in \mathsf{Ult}(X) \text{ and } \bar{Z}_Y/\|\bar{Z}_Y\| \in \mathsf{Ult}(Y)\Big) \\
&= \mathfrak{m}(X) \cdot \mathfrak{m}(Y)\,.
\end{aligned}$$

## 5 Robustness of the Definition

Before using the characterizations of Theorem 2 and Corollary 1, we show that the measure $\mathfrak{m}(X)$ is very robust. We propose two alternatives – one using spheres instead of balls, and the other one using integer lattice points instead of computing the volume – and show that the measures they give rise to coincide with $\mathfrak{m}(X)$.

### 5.1 Using Spheres Rather Than Balls

Instead of using $X \cap B_r^n$ in the definition of $\mathfrak{m}(X)$, we could have used $X \cap S_r^{n-1}$ and $\mathsf{Vol}^{n-1}(\cdot)$. That is, we could define

$$\mathfrak{m}_r^\circ(\varphi) = \frac{\mathsf{Vol}^{n-1}([\![\varphi]\!] \cap S_r^{n-1})}{\mathsf{Vol}^{n-1}(S_r^{n-1})}$$

and $\mathfrak{m}^\circ(\varphi) = \lim_{r\to\infty} \mathfrak{m}_r^\circ(\varphi)$. When $X = [\![\varphi]\!]$ we shall write $\mathfrak{m}^\circ(\varphi)$. In general it is easy to find cases when $\mathfrak{m}(X)$

exists while $\mathfrak{m}^\circ(X)$ does not. Consider, for instance, the set $X$ of all $\bar{z}$ in $\mathbb{R}^n$ such that $\|\bar{z}\|$ is irrational. Then $\mathfrak{m}_r(X) = 1$ for all $r$, and yet $\mathfrak{m}_r^\circ(X)$ is 1 when $r$ is irrational and 0 when $r$ is rational and hence $\mathfrak{m}^\circ(X)$ does not exist. Nonetheless, for sets definable in nice logical theories like $\langle \mathbb{R}, +, \cdot, e^x \rangle$, the measure defined over spheres exists and is the same as the one we used.

**Theorem 3.** *Let $X$ be definable in an o-minimal structure. Then $\mathfrak{m}^\circ(\varphi) = \mathfrak{m}(\varphi)$.*

The proof uses techniques similar to those in the proof of Theorem 2.

### 5.2 Using Integers Rather Than Reals

An alternative approach is not to define the measure of the sets via volumes but rather by the number of points in the integer lattice $\mathbb{Z}^n$. More precisely, for a set $Y \subseteq \mathbb{R}^n$, define $\mathsf{Vol}^{\mathbb{Z}}(Y)$ as the cardinality of the set $|Y \cap \mathbb{Z}^n|$, and then set

$$\mathfrak{m}_r^{\mathbb{Z}}(\varphi) = \frac{\mathsf{Vol}^{\mathbb{Z}}([\![\varphi]\!] \cap B_r^n)}{\mathsf{Vol}^{\mathbb{Z}}(B_r^n)} \text{ and } \mathfrak{m}^{\mathbb{Z}}(\varphi) = \lim_{r\to\infty} \mathfrak{m}_r^{\mathbb{Z}}(\varphi)\,.$$

This is indeed a natural measure when we deal with constraints on integers, which are extremely common. The classical Gauss ball problem tells us that $\mathsf{Vol}^{\mathbb{Z}}(B_r^n)$ is a good approximation of $\mathsf{Vol}^n(B_r^n)$ for large $r$, specifically $|\mathsf{Vol}^{\mathbb{Z}}(B_r^n) - \mathsf{Vol}^n(B_r^n)| = o(r^n)$, see (Krätzel 1988).

**Theorem 4.** *If $X$ is definable in an o-minimal structure over $\mathbb{R}$, then $\mathfrak{m}^{\mathbb{Z}}(X) = \mathfrak{m}(X)$.*

*Proof Sketch.* For $X \subseteq \mathbb{R}^n$, we define $Z = \bigcup_{r>0}\{x \in \mathbb{R}^{1+n} \mid x = (r, y), y \in X \cap B_r^n\} \subseteq \mathbb{R}^{1+n}$. The fibers of $Z$ are given as $Z_r = \{x \in \mathbb{R}^n : (r, x) \in Z\}$, and we have that $Z_r = X \cap B_r^n$. To see that, we first choose $y \in Z_r$, which means $(r, y) \in Z$, which means $y \in X \cap B_r^n$. On the other hand, if $y \in X \cap B_r^n$, then $(r, y) \in Z$, which means $y \in Z_r$.

We next need a bit of terminology. For $j > 0$, the term $V_j(Z_r)$ denotes the sum of the $j$-dimensional volumes of the orthogonal projections of $Z_r$ on every $j$-dimensional coordinate subspace of $\mathbb{R}^n$, while $V_0(Z_r) = 1$. It is well defined, since all these are definable in the o-minimal structure, and bounded, and thus are Lebesgue-measurable.

Then (Barroero and Widmer 2014)[Theorem 1.3] says that there is a constant $c_Z$ depending only on the family $Z$, i.e. it can be chosen uniformly w.r.t. $r$, so that

$$\left|\, \mathsf{Vol}^{\mathbb{Z}}(Z_r) - \mathsf{Vol}^n(Z_r) \,\right| \leq c_Z \cdot \sum_{j=0}^{n-1} V_j(Z_r) = f(r) \quad (4)$$

Since $c_Z$ and $n$ are constant, and $\mathit{diam}(Z_r) \leq 2 \cdot r$, we have:

$$f(r) \leq c_Z + c_Z \cdot \sum_{j=1}^{n-1} \binom{n}{j} \cdot (2 \cdot r)^{n-j} \leq C \cdot r^{n-1} \quad (5)$$

where $C$ is a constant that depends only on $c_Z$ and $n$. Hence, $f(r) \in O(r^{n-1})$. From the existence theorem in Section 3, we know that $\mathfrak{m}(\varphi)$ exists. Therefore, it is enough to show

$$\lim_{r\to\infty} |\mathfrak{m}_r^{\mathbb{Z}}(\varphi) - \mathfrak{m}_r(\varphi)| = 0\,. \quad (6)$$

We just showed that there is a constant $c_1$ for which we have that for all large enough $r > 0$:

$$|\mathsf{Vol}^{\mathbb{Z}}(Z_r) - \mathsf{Vol}^n(Z_r)| \leq c_1 \cdot r^{n-1} \quad (7)$$

Also, from (Krätzel 1988), we know that there is a constant $c_2$ for which we have that for all large enough $r > 0$:

$$|\mathsf{Vol}^{\mathbb{Z}}(B_r^n) - \mathsf{Vol}^n(B_r^n)| \leq c_2 \cdot r^{n-1} \quad (8)$$

Now, we make the following case distinction: $\mathsf{Vol}^{\mathbb{Z}}(Z_r)\theta_1\mathsf{Vol}^n(Z_r)$ and $\mathsf{Vol}^{\mathbb{Z}}(B_r^n)\theta_2\mathsf{Vol}^n(B_r^n)$, where each of $\theta_1$ and $\theta_2$ is either $\leq$ or $\geq$. In each of those it is easy to prove (6). □

## 6 Computational Aspects

Computing $\mathfrak{m}(\varphi)$ is often a challenging task. Even for the very simple language of quantifier-free first order formulae over $\langle \mathbb{R}, < \rangle$, it is complete for #P (the class of problems that ask for the number of accepting paths of an NP machine), see (Console, Hofer, and Libkin 2019). Problems hard for #P are intractable, unless P = NP (Arora and Barak 2009). As we go to slightly more expressive languages, computing $\mathfrak{m}(\varphi)$ becomes even more challenging: for every formula $\varphi$ of the form $\alpha x \leq y$, with $\alpha \notin \{0, \pm 1\}$, $\mathfrak{m}(\varphi)$ is irrational.

Thus we look at *approximating* $\mathfrak{m}(\varphi)$ for $\varphi$ definable in logical theories over $\mathbb{R}$ with familiar arithmetic functions. The only known result of this kind is for a very syntactically restricted subclass of formulae over $\langle \mathbb{R}, +, < \rangle$, i.e., linear constraints on the reals. We now provide two different ways of approximating $\mathfrak{m}(\varphi)$ for much larger classes of functions, without syntactic restrictions on the shape of the formulae.

The first of those shows that, for FO formulae $\varphi$ over $\langle \mathbb{R}, +, \cdot, < \rangle$, an approximation of $\mathfrak{m}(\varphi)$ is definable in the same structure. Thus, existing decision procedures can be used to approximate $\mathfrak{m}(\varphi)$. The size of $M_\varphi^\varepsilon(x)$ grows very fast however. To overcome this, we present an efficient *randomized* approximation algorithm for $\mathfrak{m}(\varphi)$, provided that $\varphi$ satisfies some mild conditions. These conditions capture, e.g., Boolean combinations of polynomial inequalities.

### 6.1 Definability and Deterministic Approximation

We now focus on FO formulae over $\langle \mathbb{R}, +, \cdot, < \rangle$. We assume, as was already mentioned earlier, that the logical language has a constant symbol for each $r \in \mathbb{R}$. This preserves o-minimality and many other properties of $\langle \mathbb{R}, +, \cdot, < \rangle$. Since $\langle \mathbb{R}, +, \cdot, < \rangle$ is o-minimal, every parameterized definable family over it has finite VC dimension, see (Dries 1998) (this is what we used earlier to prove the existence result for $\mathfrak{m}(\varphi)$). Using this property, we will show that, for every FO formula $\varphi$ over $\langle \mathbb{R}, +, \cdot, < \rangle$, written $\varphi \in \mathrm{FO}(\langle \mathbb{R}, +, \cdot, < \rangle)$, and for every $\varepsilon > 0$, there exists an *$\varepsilon$-approximation* formula $M_\varphi^\varepsilon(x)$ that satisfies two conditions:

- If $M_\varphi^\varepsilon(m)$ is true, then $|m - \mathfrak{m}(\varphi)| < \varepsilon$; and
- If $|m - \mathfrak{m}(\varphi)| < \varepsilon/4$, then $M_\varphi^\varepsilon(m)$ is true.

Assume that $\varphi(\bar{x})$ has $n$ free variables. Define the set $\mathsf{Ult}(B_r^n(\bar{c}), \varphi)$ that contains all the points in the $r$-ball around $\bar{c}$ that ultimately satisfy $\varphi$ as defined in Section 4:

$$\mathsf{Ult}(B_r^n(\bar{c}), \varphi) = \{\bar{a} \mid \bar{a} \in B_{r,\bar{c}} \text{ and } \exists r_0 \forall r > r_0\, \varphi(r \cdot \bar{a})\}$$

We can then characterize $\mathfrak{m}(\varphi)$ as follows:

$$\mathfrak{m}(\varphi) = \frac{\mathsf{Vol}(\mathsf{Ult}(B_r^n(\bar{c}), \varphi))}{\mathsf{Vol}(B_r^n)}. \tag{9}$$

Observe also that $\mathsf{Ult}(B_r^n(\bar{c}), \varphi)$ is definable itself by a formula denoted by $U_\varphi(\bar{x})$. Fix $r = 1/2$ and $\bar{c} = (1/2)^n$, then $\mathsf{Ult}(B_r^n(\bar{c}), \varphi) \subseteq [0,1]^n$. Using the results in (Karpinski and Macintyre 1997; Koiran 1995) discussed in Section 3, we can conclude that there exists an $\varepsilon$-volume approximation $V_\varphi^\varepsilon(x) \in \mathrm{FO}(\langle \mathbb{R}, +, \cdot, < \rangle)$ of $\mathsf{Ult}(B_r^n(\bar{c}), \varphi)$.

With $V_\varphi^\varepsilon(x)$ in place, to define an $\varepsilon$-approximation of $\mathfrak{m}(\varphi)$ we need a formula that defines $\mathsf{Vol}(B_{\frac{1}{2}}^n)$. For $n \geq 0$, let $b_n = \mathsf{Vol}(B_{\frac{1}{2}}^n)$. We know $b_n = q \cdot \pi^n$, for some $q \in \mathbb{Q}$, and thus it is a term in $\mathrm{FO}(\langle \mathbb{R}, +, \cdot, < \rangle)$ since we have all elements of $\mathbb{R}$ as constants. Then define

$$M_\varphi^\varepsilon(x) = \exists v \left( V_\varphi^\varepsilon(v) \wedge x = (v/b_n) \right) \tag{10}$$

Next, we prove that $M_\varphi^\varepsilon(x)$ is an $\varepsilon$-approximation of $\mathfrak{m}(\varphi)$.

**Theorem 5.** *For every $\varphi \in \mathrm{FO}(\langle \mathbb{R}, +, \cdot, < \rangle)$ and every $\varepsilon_0 \in [0,1]$, the formula $M_\varphi^\varepsilon(x)$ in (10), with $\varepsilon = \varepsilon_0 \cdot b_n$, is an $\varepsilon_0$-approximation of $\mathfrak{m}(\varphi)$.*

*Proof sketch.* With $r = \frac{1}{2}$ and $\bar{c} = (\frac{1}{2})^n$, $V_\varphi^\varepsilon(v)$ is an $\varepsilon$-volume approximation of $\mathsf{Ult}(B_r^n(\bar{c}), \varphi)$. The result follows from $x = (v/b_n)$ and (9). □

Can we use $M_\varphi^\varepsilon(x)$ to compute $\mathfrak{m}(\varphi)$ approximately? To answer this question, we examine the size of $M_\varphi^\varepsilon(x)$ for quantifier-free formulae over $\langle \mathbb{R}, +, \cdot, < \rangle$, i.e., Boolean combinations of polynomial equalities and inequalities.

**Theorem 6.** *Let $\varphi(\bar{x})$ be the Boolean combination of $s$ polynomial inequalities of degree at most $d$, and $n = |\bar{x}|$. The formula $M_\varphi^\varepsilon(x)$, is of the form $\exists^* \forall^* \exists^* \forall^* \alpha$ where the number of quantifiers is:*

- *logarithmic in $s$ and $d$;*
- *polynomial in $1/\varepsilon$;*
- *exponential in $n$;*

*and the size of the quantifier-free formula $\alpha$ is*

- *logarithmic in $d$*
- *polynomial in $s$ and in $1/\varepsilon$;*
- *exponential in $n$.*

The long and routine proof of Theorem 6 consists of applying the construction of (Karpinski and Macintyre 1997) and VC dimension bounds over the real field (Goldberg and Jerrum 1995).

Theorem 6 tells us that the number of quantifiers and the size of $M_\varphi^\varepsilon(x)$ grow very fast, in fact exponentially, in the size of $\varphi$. Also, to compute $\mathfrak{m}(\varphi)$ via $M_\varphi^\varepsilon(x)$, we need to run a (super-polynomial) quantifier-elimination procedure for $\langle \mathbb{R}, +, \cdot, < \rangle$. Thus, while giving us a deterministic approximation algorithm, the technique cannot be realistically applied except to very simple formulae. Therefore, we turn our attention to randomized procedures.

## 6.2 Absolute Error Approximation Scheme

When a function is hard to compute, a natural way to deal with it is via approximation schemes. Roughly, an approximation scheme for a function $f(\bar{x})$ is an algorithm that can compute $f(\bar{x})$ efficiently, within an input level of precision.

There are many different approaches to approximation schemes (Vazirani 2001), all giving different kinds of guarantees on the output. Here we we deal with *absolute error fully polynomial randomized approximation schemes* (AFPRAS). An AFPRAS for a function $f : A \to B$ is an algorithm that takes as input $a \in A$ and a value $\varepsilon \in (0,1]$ and, in time polynomial in the size of $a$ and the value $\varepsilon^{-1}$, outputs a random variable $A(\varphi, \varepsilon)$ with the following guarantees:

$$\mathsf{Prob}(|A(\varphi,\varepsilon) - f(a)| \leq \varepsilon) \geq \frac{3}{4}$$

AFPRASs are a popular way to deal with functions that are hard to compute, especially when the range of these functions lies in $[0,1]$ (Arora and Barak 2009). Another way to deal with hard functions is via *relative error fully polynomial randomized approximation schemes* (FPRASes). An FPRAS for a function $f : A \to B$ is an algorithm that takes as input $a \in A$ and a value $\varepsilon \in (0,1]$, and outputs, in time polynomial in the size of $a$ and the value $\varepsilon^{-1}$, a random variable $A(\varphi, \varepsilon)$ with the property that $\mathsf{Prob}(|A(\varphi,\varepsilon) - f(a)| \leq f(a) \cdot \varepsilon) \geq \frac{3}{4}$. Under reasonable complexity-theoretic assumptions however it can be shown that there exists no FPRAS for $\mathfrak{m}(\varphi)$, even for quantifier-free $\varphi$ in $\mathrm{FO}(\langle \mathbb{R}, < \rangle)$, see (Console, Hofer, and Libkin 2019).

We now present a technique to obtain an AFPRAS for formulae of o-minimal structures. Assume a set of functions and predicates so that the associated structure $\mathfrak{A}$ over $\mathbb{R}$ is o-minimal (e.g., the usual arithmetic $+, -, \cdot, \div, <$), and let $\mathcal{L}$ be the FO language of $\mathfrak{A}$. We write $\mathfrak{m}_\mathcal{L}$ for the family of functions $\mathfrak{m}(\varphi)$ with $\varphi \in \mathcal{L}$. Assume an algorithm $\mathsf{Sample}(\varphi)$ that, with input an $n$-ary formula $\varphi \in \mathcal{L}$, outputs a Bernoulli random variable $\mathcal{U}$ such that

$$\mathsf{Prob}(\mathcal{U} = 1) = \frac{\mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))}{\mathsf{Vol}^{n-1}(S^{n-1})}$$

From Theorem 2 in Section 4, we know that $\mathsf{Prob}(\mathcal{U} = 1) = \mathfrak{m}(\varphi)$, and therefore the expected value $E[\mathcal{U}]$ is equal to $\mathfrak{m}(\varphi)$. To obtain an approximation for $\mathfrak{m}(\varphi)$ then, we can use the following standard statistical technique. Let $\mathcal{U}_1, \ldots, \mathcal{U}_n$ be $n$ independent Bernulli random variables, and define a new random variable $\mathcal{M}_n = (\sum_{i=1}^n \mathcal{U}_i)/n$. By Hoeffding's inequality (Blum, Hopcroft, and Kannan 2020) we then get

$$\mathsf{Prob}(|\mathcal{M}_n - E[\mathcal{U}]| \leq \varepsilon) \geq 1 - 2^{-2n\varepsilon^2} \tag{11}$$

Consider now the algorithm $\mathsf{Apx}(\varphi, \varepsilon)$ that, for input $\varphi \in \mathcal{L}$ and $\varepsilon \in (0,1]$, computes the mean of $n \geq \varepsilon^{-2}$ executions of $\mathsf{Sample}(\varphi)$, for the algorithm $\mathsf{Sample}$ given below. From (11), we conclude

$$\mathsf{Prob}(|\mathsf{Apx}(\varphi,\varepsilon) - \mathfrak{m}(\varphi)| \leq \varepsilon) \geq \frac{3}{4} \tag{12}$$

Hence, $\mathsf{Apx}$ is an AFPRAS for $\mathfrak{m}(\varphi)$ as long as $\bar{a} \in \mathsf{Ult}(\varphi)$ can be tested efficiently.

**Algorithm 1** Sample

**Input:** An $n$-ary $\varphi(\bar{x}) \in \mathcal{L}$
**Output:** either 0 or 1
Pick $\bar{c}$ uniformly at random from $S^{n-1}$
**if** $\bar{c} \in \mathsf{Ult}(\varphi)$ **then**
  **return** 1
**else**
  **return** 0
**end if**

**Lemma 3.** *Let $\mathfrak{A}$ be an o-minimal structure over $\mathbb{R}$. For every* FO *formula $\varphi$ over $\mathfrak{A}$,* $\mathsf{Sample}(\varphi)$ *outputs a Bernulli random variable such that $E[\mathsf{Sample}(\varphi)] = \mathfrak{m}(\varphi)$. Moreover, if $\bar{a} \in \mathsf{Ult}(\varphi)$ can be tested in time polynomial in $\varphi$, then* $\mathsf{Sample}(\varphi)$ *runs in time polynomial in the size of $\varphi$.*

*Proof Sketch.* Since $\mathsf{Sample}(\varphi)$ outputs a Bernulli random variables, $E[\mathsf{Sample}(\varphi)] = \mathsf{Prob}(\mathsf{Sample}(\varphi) = 1)$. Since $\bar{c}$ has been picked according to a uniform distribution over $S^{n-1}$, the latter is equal to $\frac{\mathsf{Vol}^{n-1}(\mathsf{Ult}(\varphi))}{\mathsf{Vol}^{n-1}(S^{n-1})} = \mathfrak{m}(\varphi)$. For the complexity, observe that we can sample uniformly at random from $S^{n-1}$ in time polynomial in $n-1$ (Blum, Hopcroft, and Kannan 2020). □

Summing up, these results give us the following.

**Theorem 7.** *Let $\mathcal{L}$ be a language giving rise to an o-minimal structure over $\mathbb{R}$. If there exists an algorithm that checks whether $\bar{a} \in \mathsf{Ult}(\varphi)$ in time polynomial in the size of $\varphi$, then there exists an AFPRAS for $\mathfrak{m}_{\mathcal{L}}$.*

### 6.3 Asymptotic Satisfaction Oracles

The missing ingredient in Theorem 7 that gives us an AFPRAS is the ability to test $\bar{a} \in \mathsf{Ult}(\varphi)$ efficiently. This condition is related to the existence of an efficient oracle of the following kind.

**Definition 2.** *An* Asymptotic Satisfaction Oracle *(ASO) for a structure $\mathfrak{A}$ over $\mathbb{R}$ is a procedure* $\mathsf{Asym}$ *defined as follows. For every $n$-ary* FO *atomic formula $\alpha(\bar{x})$ over $\mathfrak{A}$ and every point $\bar{a} \in \mathbb{R}^n$*

- $\mathsf{Asym}(\alpha, \bar{a}) = 1$*, if there exists $r_0 \in \mathbb{R}$ such that for every $r \geq r_0$ the formula $\alpha(r \cdot \bar{a})$ is true;*
- $\mathsf{Asym}(\alpha, \bar{a}) = 0$*, otherwise.*

A structure $\mathfrak{A}$ admits a *polynomial-time* Asymptotic Satisfaction Oracle whenever there exists a procedure $\mathsf{Asym}(\alpha, \bar{a})$ such that $\mathsf{Asym}(\alpha, \bar{a})$ is an ASO for $\mathfrak{A}$ and $\mathsf{Asym}(\varphi, \bar{a})$ runs in time polynomial in the size of the input formula $\alpha$. Whenever $\mathfrak{A}$ is o-minimal and admits a polynomial-time Asymptotic Satisfaction Oracle, we can devise an efficient algorithm to check whether a point $\bar{a} \in \mathbb{R}^n$ lies inside $\mathsf{Ult}(\varphi)$.

**Lemma 4.** *Let $\mathfrak{A}$ be an o-minimal structure over $\mathbb{R}$ that has the operation of multiplication. If there exists an ASO for $\mathfrak{A}$, there exists an algorithm that checks whether $\bar{a} \in \mathsf{Ult}(\varphi)$ in time polynomial in the size of $\varphi$, for every quantifier-free* FO *formula $\varphi$ over $\mathfrak{A}$ and $\bar{a} \in \mathbb{R}^n$.*

This immediately gives us the following:

**Theorem 8.** *Given an o-minimal structures $\mathfrak{A}$ over $\mathbb{R}$ that has the operation of multiplication, if there exists a polynomial-time ASO for it, then there exists an AFPRAS for computing $\mathfrak{m}(\varphi)$ for quantifier-free formulae over $\mathfrak{A}$.*

We conclude by showing that there is an AFPRAS computing $\mathfrak{m}(\varphi)$ for arbitrary linear and polynomial constraints over $\mathbb{R}$. The only previously known approximation result was for linear constraints in disjunctive normal form, which used the syntactic shape in an essential way. We now eliminate syntactic restrictions, and allow multiplication and thus conditions of the form $p(\bar{x}) \leq 0$ where $p$ is a polynomial.

**Theorem 9.** *There exists a polynomial time ASO for $\langle \mathbb{R}, +, \cdot, < \rangle$. Consequently, there is an AFPRAS computing $\mathfrak{m}(\varphi)$ for arbitrary Boolean combinations of linear and polynomial equalities and inequalities over $\mathbb{R}$.*

*Proof idea.* Given constraint $p(\bar{x}) \leq 0$ and direction $\bar{a}$, to check the asymptotic behavior of $p$ in direction $\bar{a}$ we replace $p$ by a univariate polynomial $p'(r) = p(r \cdot \bar{a})$ and check its leading coefficient. □

## 7 Conclusions and Future Work

There are two main directions for further study. One concerns applications of our techniques in areas mentioned in the introduction: answering queries over incomplete databases, temporal reasoning, spatial reasoning, etc. They key is to see how well the sampling algorithm, that has good theoretical guarantees, behaves in real-life scenarios.

The second idea is to extend techniques to cases when $\mathfrak{m}(X)$ does not exist due to the oscillating behavior of $\mathfrak{m}_r(X)$ as a function of $r$. In this case one could define an *interval measure*

$$\mathfrak{m}^*(X) = [\liminf_{r\to\infty}(\mathfrak{m}_r(X)),\ \limsup_{r\to\infty}(\mathfrak{m}_r(X))]\,.$$

This is well defined since sets $\{\mathfrak{m}_r(X) \mid r > 0\}$ are bounded. An example provided in (Console, Hofer, and Libkin 2019) to prove that $\mathfrak{m}(X)$ may not be defined outside o-minimal structures gave a specific set $X$ defined by a quantifier-free formula using $\sin(x)$ and $\ln x$ such that (approximately) $\mathfrak{m}^*(X) = [0.00012, 0.06366]$. While not an exact number, it is a small interval nonetheless. It is however far from clear, and far from trivial, to see how large these intervals can be if the usual arithmetic is supplement with oscillating functions such as $\sin(x)$.

More generally though these intervals can be arbitrarily large, even for structures with nicely behaving theories over integers. Consider for example the structure $\langle \mathbb{N}, +, 2^x \rangle$. By Semenov's theorem it is decidable and admits quantifier-elimination, but it is easy to see that for $X = \bigcup_k [2^{2^{2k}}, 2^{2^{2k+1}}]$ we have $\mathfrak{m}^*(X) = [0, 1]$. Understanding when we can find a bound on the measure of sets which is a small interval, as well as algorithmic consequences of it, is left for further work.

## Acknowledgements

Work partially supported by EPSRC grants M025268 and N023056, a grant from the Foundation Sciences Mathématiques de Paris under the FSMP Chairs program,

and Royal Society through a Wolfson Research Merit Award.

## References

Anthony, M., and Biggs, N. 1992. *Computational Learning Theory*. USA: Cambridge University Press.

Arora, S., and Barak, B. 2009. *Computational Complexity - A Modern Approach*. Cambridge University Press.

Barroero, F., and Widmer, M. 2014. Counting lattice points and o-minimal structures. *International Mathematics Research Notices* 2014(18):4932–4957.

Belegradek, O. V.; Peterzil, Y.; and Wagner, F. 2000. Quasi-o-minimal structures. *J. Symb. Log.* 65(3):1115–1132.

Bienvenu, M., and Ortiz, M. 2015. Ontology-mediated query answering with data-tractable description logics. In *Reasoning Web*, 218–307.

Blum, A.; Hopcroft, J.; and Kannan, R. 2020. *Foundations of data science*. Cambridge University Press.

Console, M.; Hofer, M.; and Libkin, L. 2019. Measuring the likelihood of numerical constraints. In Kraus, S., ed., *IJCAI*, 1654–1660. ijcai.org.

Dechter, R.; Meiri, I.; and Pearl, J. 1991. Temporal constraint networks. *Artificial Intelligence* 49(1):61 – 95.

Dries, L. V. D. 1998. *Tame topology and o-minimal structures*, volume 248. Cambridge university press.

Egenhofer, M. J., and Franzosa, R. D. 1991. Point set topological relations. *International Journal of Geographical Information Systems* 5(2):161–174.

Feng, S.; Huber, A.; Glavic, B.; and Kennedy, O. 2019. Uncertainty annotated databases - A lightweight approach for approximating certain answers. In *SIGMOD*, 1313–1330.

Godoy, F., and Rodríguez, A. 2002. A quantitative description of spatial configurations. In *Advances in Spatial Data Handling*. Springer. 299–311.

Goldberg, P. W., and Jerrum, M. 1995. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning* 18(2-3):131–148.

Greco, S.; Molinaro, C.; and Trubitsyna, I. 2019. Approximation algorithms for querying incomplete databases. *Inf. Syst.* 86:28–45.

Imielinski, T., and Lipski, W. 1984. Incomplete information in relational databases. *Journal of the ACM* 31(4):761–791.

Karpinski, M., and Macintyre, A. 1997. Approximating the volume of general Pfaffian bodies. In *Structures in Logic and Computer Science, A Selection of Essays in Honor of Andrzej Ehrenfeucht*, 162–173.

Koiran, P. 1995. Approximating the volume of definable sets. In *FOCS*, 134–141. IEEE Computer Society.

Kontchakov, R.; Pratt-Hartmann, I.; and Zakharyaschev, M. 2010. Interpreting topological logics over euclidean spaces. In *KR*.

Koutras, C. D.; Liaskos, K.; Moyzes, C.; and Rantsoudis, C. 2018. Default reasoning via topology and mathematical analysis: A preliminary report. In *KR*, 267–276.

Krätzel, E. 1988. *Lattice Points*. Kluwer.

Lenzerini, M. 2002. Data integration: a theoretical perspective. In *ACM Symposium on Principles of Database Systems (PODS)*, 233–246.

Libkin, L. 2016a. Certain answers as objects and knowledge. *Artif. Intell.* 232:1–19.

Libkin, L. 2016b. SQL's three-valued logic and certain answers. *ACM Trans. Database Syst.* 41(1):1:1–1:28.

Libkin, L. 2018. Certain answers meet zero-one laws. In *PODS*, 195–207.

Mattila, P. 1995. *Geometry of Sets and Measures in Euclidean Spaces*. Cambridge University Press.

Miller, C. 2011. Expansions of o-minimal structures on the real field by trajectories of linear vector fields. *Proceedings of the American Mathematical Society* 139(1):319–330.

Van den Dries, L. 1998. *Tame Topology and o-minimal Structures*, volume 248. Cambridge university press.

Vazirani, V. V. 2001. *Approximation algorithms*. Springer.

Wilkie, A. J. 1996. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *Journal of the American Mathematical Society* 9(4):1051–1094.