Strategic PAC Learnability via Geometric Definability

Yuval Filmus
Faculty of Computer Science and
Faculty of Mathematics
Technion – Israel Institute of Technology
yuvalfi@cs.technion.ac.il

Shay Moran
Faculties of Mathematics, Computer Science,
and Data and Decision Sciences
Technion – Israel Institute of Technology
Google Research
smoran@technion.ac.il

Elizaveta Nesterova
Faculty of Mathematics
Technion – Israel Institute of Technology
elizavetan@campus.technion.ac.il

Nir Rosenfeld
Faculty of Computer Science
Technion – Israel Institute of Technology
nirr@cs.technion.ac.il

Alexander Shlimovich
Faculty of Mathematics
Technion – Israel Institute of Technology
ashlimovich@campus.technion.ac.il

Abstract

Strategic classification studies learning settings in which individuals can modify their features, at a cost, in order to influence the classifier's decision. A central question is how the sample complexity of the induced (strategic) hypothesis class depends on the complexities of the underlying hypothesis class and the cost structure governing feasible manipulations. Prior work has shown that in several natural settings, such as linear classifiers with norm costs, the induced complexity can be controlled. We begin by showing that such guarantees fail in general — even in simple cases: there exist hypothesis classes of VC dimension 1 on the real line such that, even under the simplest interval neighborhoods, the induced class has infinite VC dimension. Thus, strategic behavior can turn an easy learning problem into a non-learnable one. To overcome this, we introduce structure via a geometric definability assumption: both the hypothesis class and the cost-induced neighborhood relation can be defined by first-order formulas over $\mathbb{R}_{\exp}$. Intuitively, this means that hypotheses and costs can be described using arithmetic operations, exponentiation, logarithms, and comparisons. This captures a broad range of natural classes and cost functions, including ℓ_p distances, Wasserstein distance, and information-theoretic divergences. Under this assumption, we prove that learnability is preserved, with sample complexity controlled by the complexity of the defining formulas.

1 Introduction

The field of strategic classification studies learning in settings where individuals can modify their features in response to a classifier, at some cost, in order to obtain a favorable outcome [BKS12, HMPW16]. For example, consider a university that designs a rule for admissions based on applicants' features such as grades, test scores, or extracurricular activities. Once this rule is made public, prospective students may adapt their behavior to meet the requirements — by studying harder, retaking exams, or investing in private tutoring. While such adaptations may improve

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

their chances of admission, they also change the distribution of inputs seen by the classifier. The university's true goal, however, is to admit students who will successfully graduate and contribute to society, rather than those who merely optimize for the admission rule.

This illustrates an inherent tension between systems that aim to make accurate predictions and individuals who benefit from being classified as positive. Applications of this kind arise in many domains, including loan approval, hiring, university admissions, scholarships, social benefit programs, and medical eligibility. The central objective in strategic classification is therefore to learn a classifier that is accurate under such strategic behavior, i.e., achieves strategic robustness.

A fundamental question in learning theory is which hypothesis classes are learnable, and how many samples are needed. Despite over a decade of sustained research, the implications of strategic behavior on learnability and sample complexity remain poorly understood. This gap persists even in simple and well-studied settings, where the cost function is uniform across individuals, feature manipulation is unconstrained by causal relationships, and the classifier is fully known to the agents.

A natural intuition is that if a hypothesis class is learnable and the cost function is "simple," then the induced strategic class should also be learnable, since local input perturbations are unlikely to cause significant statistical degradation. Indeed, the literature provides several positive results supporting this view: linear classes with semi-norm costs [ZC21, SVXY23]; polytope classes, piecewise-linear classes, and classes closed under input scaling with ℓ_p costs [TR25]; and general classes with separable costs [HMPW16] or discrete manipulation graphs [CMMS24].

However, these results stop short of resolving the basic general question: *does learnability of a hypothesis class, together with a well-behaved cost structure, imply learnability under strategic behavior?*

Combinatorial simplicity is not enough. Our first result shows that, in full generality, the answer to the above question is negative. We consider perhaps the simplest possible setting: a hypothesis class with VC dimension 1, defined over the real line, together with a uniform and highly regular cost structure in which each point x can move within distance at most 1 (i.e., its neighborhood is an interval of radius 1 around x). Despite this extreme simplicity, we construct a class for which the induced strategic class is not learnable.

This result demonstrates that learnability is not preserved under strategic behavior, even under minimal combinatorial complexity and very simple cost structures.

Geometric structure restores learnability. The negative result above suggests that strategic learnability requires structure beyond combinatorial simplicity. We therefore turn to a broad and well-studied family of learning problems with *geometric structure*.

By geometric, we mean that the instance space is $\mathbb{R}^d$, and that both the hypothesis class $\mathcal{H}$ and the cost function are described using a fixed collection of basic operations, such as the arithmetic operations, comparison operators, and functions such as $\exp$ and $\log$. This setting captures many standard classes in machine learning, including linear classifiers, polynomial threshold functions, intersections of halfspaces, decision trees with polynomial decision nodes, neural networks with common activation functions, as well as cost functions based on distances such as ℓ_p norms or KL divergence.

Within this framework, we focus on a basic and canonical strategic setting: each point x is associated with a cost-induced neighborhood N_x to which it can move freely, but cannot move outside of it—i.e., N_x is the set of points to which moving is potentially cost-effective. Our goal is to understand when the induced strategic class is learnable, and to characterize its sample complexity. Despite its simplicity, this setting is already non-trivial, and serves as a natural platform for developing our approach. Moreover, we believe the techniques we introduce extend beyond this setting to more general models of strategic behavior.

A key feature of strategic classification is that it naturally introduces *existential quantification*. Indeed, for a hypothesis h, let h' denote the induced strategic hypothesis. An input x is labeled positively after adaptation if and only if there exists a point x' in its neighborhood that is labeled positively by h:

$$h'(x) = +1 \iff \exists x' \text{ s.t. } (x' \in N_x) \wedge (h(x') = +1). \tag{1}$$

Thus, the induced strategic hypotheses are defined by formulas that include existential quantifiers. This observation plays a central role in our approach.

To formalize our geometric setting, we model hypotheses and neighborhoods as being *definable* by first-order formulas over the reals. As a first intuition, this can be viewed as specifying a procedure that, given $h \in \mathcal{H}$ and points x, x', determines whether $h(x) = 1$ and whether x' lies in the neighborhood of x using only the allowed operations (i.e., arithmetic operations, comparisons, $\exp$, and $\log$). However, this intuition is not entirely accurate: allowing quantifiers extends this framework beyond purely algorithmic descriptions. For example, a formula of the form $\psi(x) = \exists x' \, \varphi(x, x')$ can express properties that require searching for a witness in an infinite set, and are therefore not *a priori* constructive—i.e., it is not clear how to compute them.

We note that a closely related line of work in learning theory has studied (non-strategic) hypothesis classes defined using arithmetic operations alone, without $\exp$ and $\log$. In this more restrictive setting, terms correspond to polynomials, and tools from real algebraic geometry have been used to analyze their complexity; a partial list of papers includes [BDL93, GJ95, BMM98, BES02, AMY16, BHLM19, AAE+23]. Much of this work focuses on quantifier-free definitions, a simplification closely tied to the availability of quantifier elimination in those settings [Tar51]. Extending this framework to include exponentiation substantially increases expressiveness, and requires different tools, which we draw from geometry and model theory.

1.1 Overview of Main Results

We now summarize our main contributions.

- **Combinatorial simplicity does not suffice.** Our first result shows that strategic learnability can fail even in the simplest possible settings. We construct a hypothesis class over $\mathbb{R}$ with VC dimension 1, together with a very simple cost structure in which each point x can move within a fixed radius such that the induced strategic class is not PAC-learnable. In Appendix C we present stronger variants of this phenomenon: one in which non-learnability holds for all radii simultaneously, and another in which the neighborhood system itself has VC dimension 1, and in Appendix D we give a different example whose logical complexity is low. These demonstrate that low combinatorial complexity and simple costs are not sufficient to guarantee strategic learnability.
- **Qualitative learnability via definability in $\mathbb{R}_{\exp}$.** Our first positive result shows that learnability is restored under a geometric definability assumption. Specifically, if both the hypothesis class and the neighborhood relation are definable using arithmetic operations together with $\exp$ and $\log$ (i.e., within the structure $\mathbb{R}_{\exp}$), then the induced strategic class is PAC-learnable. Moreover, the sample complexity depends on the number of parameters used in the defining formulas. Our analysis yields refined bounds for *Empirical Risk Minimization* (ERM) learners that can be sharper than standard VC-based guarantees, as they are derived directly from growth function bounds rather than via Sauer–Shelah–Perles Lemma.
- **From qualitative to quantitative bounds.** The above result is inherently qualitative: it guarantees the existence of finite sample complexity bounds, and even yields asymptotic ERM learning rates that improve over VC-based guarantees. However, it does not provide explicit bounds on the constants appearing in the bounds on the sample complexity and VC dimension. This limitation stems from the inherent non-constructive nature of the underlying model-theoretic tools, and motivates the search for settings in which explicit, quantitative bounds can be obtained.
- **Quantitative bounds via restricted definability.** Our next results address this limitation by providing explicit and constructive sample complexity bounds for large families of strategically defined geometric classes. We focus on settings in which the defining formulas lie in Σ_1, i.e., use only existential quantifiers. The form in Equation (1) shows that existential formulas naturally define the strategic transformation. Existential formulas can model many natural examples, including neighborhoods based on ℓ_p norms, earth mover distance, KL divergence, and related distances. We provide detailed examples in Appendix A.

  As a warm-up, we first consider the case without exponentiation, corresponding to semialgebraic definitions. In this setting, definable sets reduce to Boolean combinations of polynomial inequalities, and a single application of quantifier elimination suffices to eliminate the existential quantifier in Equation (1). Combining this with classical tools from real algebraic geometry, we obtain explicit bounds on the sample complexity and VC dimension.

  We then extend this approach to the setting with exponentiation. Here, polynomials are replaced by a richer class of real-valued functions for which analogous complexity bounds are known.

Using these tools, we derive explicit upper bounds on the VC dimension and sample complexity of the induced strategic class.

1.2 Organization

In Section 2 we provide basic preliminaries from PAC learning and strategic classification. In Section 3 we state and explain our main results in detail. We do not assume prior expertise in logic, and aim to present the relevant concepts in a way that is accessible to the broad learning theory community. In Section 4 we survey related work. In Appendix A we give detailed examples of natural hypothesis classes and neighborhood systems, together with their first-order definitions, and illustrate how our theorems apply to these settings. The remainder of the paper is devoted to proofs.

2 Preliminaries

This section collects a few notions needed for stating our main results. We focus on sample complexity in the PAC model, properties of empirical risk minimization (ERM), and basic definitions from strategic PAC learning.

PAC learning and ERM. We work in the standard PAC model of binary classification. A hypothesis class is a set $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. In the realizable setting, examples of input-label pairs are drawn i.i.d. from a distribution D over $\mathcal{X} \times \{0,1\}$ for which there exists $h^\star \in \mathcal{H}$ with zero error. In the agnostic setting no such assumption is made, and labels can admit arbitrary conditional distributions.

For a hypothesis $h \in \mathcal{H}$, write $L_D(h) = \Pr_{(x,y)\sim D}[h(x) \neq y]$ for its population error, and for a sample set $S = ((x_1,y_1),\dots,(x_m,y_m))$, write $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\}$ for its empirical error.

The central combinatorial parameter governing PAC learnability is the VC dimension. In particular, a class is PAC learnable if and only if it has finite VC dimension, and the optimal sample complexity is determined, up to constant factors, by $\mathbb{VC}(\mathcal{H})$. The prevailing algorithmic principle in PAC learning is empirical risk minimization: given a sample, output a hypothesis in $\mathcal{H}$ minimizing the empirical error.

We distinguish between optimal PAC sample complexity and ERM sample complexity. Let $\varepsilon, \delta > 0$ be the error and confidence parameters. The realizable PAC sample complexity of $\mathcal{H}$ is the smallest number $m^{\mathrm{real}}_{\mathrm{PAC}}(\varepsilon,\delta)$ such that there exists a learning algorithm which, given at least this many realizable examples, outputs with probability at least $1-\delta$ over the sample set a hypothesis of error at most ε. The agnostic PAC sample complexity $m^{\mathrm{agn}}_{\mathrm{PAC}}(\varepsilon,\delta)$ is defined similarly, with the guarantee $L_D(A(S)) \leq \inf_{h\in\mathcal{H}} L_D(h) + \varepsilon$.

Our quantitative main theorem statements give explicit realizable ERM sample-complexity bounds together with bounds on $\mathbb{VC}(\mathcal{H})$, which by classical VC theory immediately imply optimal agnostic PAC and agnostic ERM guarantees up to universal constants.

The realizable ERM sample complexity of $\mathcal{H}$, denoted $m^{\mathrm{real}}_{\mathrm{ERM}}(\varepsilon,\delta)$, is the smallest number such that *every* empirical risk minimizer over $\mathcal{H}$, when run on at least m realizable examples, returns with probability at least $1-\delta$ a hypothesis of error at most ε.

The classical VC bounds imply that for $d = \mathbb{VC}(\mathcal{H})$,

$$m^{\mathrm{real}}_{\mathrm{PAC}}(\varepsilon,\delta) = \Theta\left(\frac{d + \log(1/\delta)}{\varepsilon}\right),$$

and this rate is tight up to constant factors for every class [VC74, BEHW89, Han16a]. Moreover, every ERM satisfies

$$m^{\mathrm{real}}_{\mathrm{ERM}}(\varepsilon,\delta) = O\left(\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right),$$

and this dependence on $\log(1/\varepsilon)$ is unavoidable for some classes [VC74, Han16b, BHMZ20].

In the agnostic setting, ERM is optimal up to universal constants and is equal to $\Theta\left(\frac{d+\log(1/\delta)}{\varepsilon^2}\right)$, and this rate is tight for every class ([Tal94]; see [vdVW96, Theorems 2.14.1 and 2.6.7]).

Growth function. In contrast to the VC-based bounds above, our results provide refined upper bounds on the realizable ERM sample complexity that can be substantially smaller than the worst-case VC guarantee. These bounds are expressed in terms of the growth function.

Definition 2.1 (Growth function)**.** Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. The growth function of $\mathcal{H}$ is

$$\Pi_{\mathcal{H}}(m) = \max_{x_1,\ldots,x_m \in \mathcal{X}} \left|\{(h(x_1),\ldots,h(x_m)) : h \in \mathcal{H}\}\right|.$$

Equivalently, $\Pi_{\mathcal{H}}(m)$ is the maximum number of distinct labelings induced by $\mathcal{H}$ on a set of m points.

Lemma 2.2 (ERM bound from the growth function)**.** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, and suppose that $\Pi_{\mathcal{H}}(m) \leq Cm^k$ for every $m \geq 1$, where $C \geq 1$ and $k \geq 1$. Then,*

$$m^{\text{real}}_{\text{ERM}}(\varepsilon,\delta) = O\left(\frac{k\log(k/\varepsilon) + \log C + \log(1/\delta)}{\varepsilon}\right).$$

The proof of this lemma is standard. For completeness, we provide a complete proof in Appendix I.

The usual VC-based ERM bound follows by bounding the growth function using Sauer's lemma: if $\mathbb{VC}(\mathcal{H}) = d$, then $\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d$ for all $m \geq d$, which yields the standard dependence $O\left(\frac{d\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon}\right)$. However, in several classes considered in this work, Sauer's lemma gives a loose upper bound on the growth function. The direct growth-based bound above can therefore yield significantly sharper guarantees.

In particular, when $\Pi_{\mathcal{H}}(m) \leq Cm^k$, the dominant term as $\varepsilon \to 0$ is

$$\frac{k\log(1/\varepsilon)}{\varepsilon},$$

so the exponent k of the growth function governs the leading asymptotic dependence of the ERM sample complexity.

Strategic classification. We briefly recall some basic definitions from strategic classification that we will use throughout the paper.

Let $h\colon \mathcal{X} \to \{0,1\}$ be a hypothesis, and let $\mathcal{N} = \{N_x \subseteq \mathcal{X} : x \in \mathcal{X}\}$ be a neighborhood system, where N_x represents the set of points to which an individual located at x may move. Note we permit non-uniform neighborhoods, i.e., N_x and $N_{x'}$ can be structurally distinct for different x and x'. We assume by default that $x \in N_x$ for all x, although our results apply even without this assumption.

The induced strategic hypothesis $h^{\mathcal{N}}$ is defined by

$$h^{\mathcal{N}}(x) = 1 \quad \text{iff} \quad \exists x' \in N_x \text{ such that } h(x') = 1.$$

For a hypothesis class $\mathcal{H}$, the induced strategic class is $\mathcal{H}^{\mathcal{N}} = \{\, h^{\mathcal{N}} : h \in \mathcal{H} \,\}$.

3 Main results

3.1 Strategic non-learnability

Our first result shows that the VC dimension of a strategically induced hypothesis class $\mathcal{H}^{\mathcal{N}}$ can be arbitrarily larger than the VC dimension of $\mathcal{H}$ – in fact, unbounded – even if $\mathcal{N}$ is natural and simple.

Main Theorem 1 (Strategic response can destroy learnability)**.** *For $r > 0$, let $\mathcal{N}_r$ be the neighborhood system given by $x' \in N_x \iff |x - x'| \leq r$.*

There exists a hypothesis class $\mathcal{H}$ over $\mathbb{R}$ such that $\mathbb{VC}(\mathcal{H}) = 1$ but for every $r > 0$, $\mathbb{VC}(\mathcal{H}^{\mathcal{N}_r}) = \infty$.

This parallels earlier results showing increased VC dimension due to strategic behavior in related settings, including discrete neighborhoods [ZC21], non-uniform neighborhoods [SVXY23], alternative loss functions [ZC21], and label improving responses [ABN+25]. We discuss these connections in more detail in Section 4.

The proof uses the same coding idea as [MHS19, Lemma 2]. In fact, Lemma C.1 essentially recovers their construction in the setting of interval neighborhoods on $\mathbb{R}$, while Theorem 1 adapts it to obtain a single class whose strategic VC dimension is infinite for every radius $r > 0$.

We illustrate the main idea of the construction by describing a simpler variant: for any *fixed* $n \in \mathbb{N}$ we construct a hypothesis class $\mathcal{H}$ over $\mathbb{R}$ such that $\mathbb{VC}(\mathcal{H}) = 1$ while $\mathbb{VC}(\mathcal{H}^{\mathcal{N}_{1/3}}) = n$.

For each $S \subseteq [n]$, the hypothesis class includes the hypothesis $f_S = \mathbf{1}[x \in \{i + x_S : i \in S\}]$, where $x_S = \sum_{j \in S} 10^{-j}$. By construction, the sets $\{x : f_S(x) = 1\}$ are disjoint, and so $\mathbb{VC}(\mathcal{H}) = 1$. Conversely, $\mathcal{H}^{\mathcal{N}_{1/3}}$ shatters the points $1 + \frac{1}{3}, \ldots, n + \frac{1}{3}$. The choice of radius $1/3$ is only for illustrative purposes; an arbitrary fixed radius $r > 0$ is obtained by rescaling the real line. See Appendix C and Appendix D for the full proof and more sophisticated constructions.

3.2 Qualitative bounds

The phenomenon described in Main Theorem 1 disappears if we restrict both $\mathcal{H}$ and $\mathcal{N}$ to be *definable in* $\mathbb{R}_{\exp}$. Loosely, this means that hypotheses and neighborhood relations can be described using (parametric) first-order logic formulas in the *language of* $\mathbb{R}_{\exp}$.

A formula in the language of $\mathbb{R}_{\exp}$ is a formula in first-order logic which can make use of arbitrary real constants as well as the functions $+, \cdot, \exp$ and the relations $<, =$, in addition to the usual logical connectives and quantifiers. If the formula uses no quantifiers, we say that it is *quantifier-free*. If it uses only existential quantifiers, and these appear only in the beginning, we say that it is *existential* (in logic, such formulas are called Σ_1 formulas). If the formula doesn't use $\exp$, then it is also a formula in the language of $\mathbb{R}$.

Definition 3.1 (Definable hypothesis class)**.** A hypothesis class $\mathcal{H}$ over $\mathbb{R}^l$ is *definable in* $\mathbb{R}_{\exp}$ *with* k *parameters* if there is a formula $\Phi_{\mathcal{H}}(x, a)$ in the language of $\mathbb{R}_{\exp}$, where $x = (x_1, \ldots, x_l)$ and $a = (a_1, \ldots, a_k)$, such that

$$\mathcal{H} = \{h_a : a \in \mathbb{R}^k\}, \text{ where } h_a = \mathbf{1}\big\{x \in \mathbb{R}^l : \Phi_{\mathcal{H}}(x, a)\big\}.$$

That is, $\mathcal{H}$ includes all hypotheses h_a corresponding to some parameterization a, where $h_a(x) = 1$ iff the formula $\Phi_{\mathcal{H}}(x, a)$ is satisfied.

As a simple concrete example, consider the hypothesis class of all indicators of halfspaces in $\mathbb{R}^l$. This hypothesis class is definable in $\mathbb{R}_{\exp}$ with $k = l + 1$ parameters, using the following formula:

$$\Phi_{\text{Halfspace}}(x, a) = a_1 x_1 + \cdots + a_l x_l \geq a_{l+1}.$$

For any choice of the parameters $a_1, \ldots, a_{l+1}$, the corresponding hypothesis $h_a = \mathbf{1}\big\{x \in \mathbb{R}^l : \Phi_{\text{Halfspace}}(x, a)\big\}$ is the indicator function of a closed halfspace. As we vary a over all of $\mathbb{R}^{l+1}$, we obtain all such hypotheses.

Definition 3.2 (Definable neighborhood system)**.** A neighborhood system $\mathcal{N} = \{N_x : x \in \mathbb{R}^l\}$ over $\mathbb{R}^l$ is *definable in* $\mathbb{R}_{\exp}$ if there is a *neighborhood relation* formula $\Phi_{\mathcal{N}}(x, x')$ (where each of x, x' is a block of l variables) in the language of $\mathbb{R}_{\exp}$ such that

$$N_x = \{x' \in \mathbb{R}^l : \Phi_{\mathcal{N}}(x, x')\}.$$

For example, let $\mathcal{N}_p$ (for $p > 0$) be the neighborhood system in which N_x consists of all points in the ℓ_p unit ball around x. This neighborhood system is definable in $\mathbb{R}_{\exp}$ using the formula

$$\Phi_{\mathcal{N}_p}(x, x') \equiv \exists y \in \mathbb{R}^l, z \in \mathbb{R}^l \left[y_1 + \cdots + y_l \leq 1 \wedge \right.$$
$$\left. \bigwedge_{i=1}^{l} \Big((x_i = x'_i \wedge y_i = 0) \ \vee \ \big((\exp z_i = x_i - x'_i \vee \exp z_i = x'_i - x_i) \wedge y_i = \exp(p z_i) \big) \Big) \right]. \quad (2)$$

Here $z_i = \log |x_i - x'_i|$ and $y_i = |x_i - x'_i|^p = \exp(p \log |x_i - x'_i|)$. When $p = 2$, Equation (2) reduces to a simple quadratic formula: $\Phi_{\mathcal{N}_2}(x, x') \equiv \sum_{i=1}^{l} (x_i - x'_i)^2 \leq 1$.

We can now state our second main theorem, which shows that strategic classes $\mathcal{H}^{\mathcal{N}}$ are learnable if both $\mathcal{H}$ and $\mathcal{N}$ are definable in $\mathbb{R}_{\exp}$.

Main Theorem 2 (Strategic classes definable in $\mathbb{R}_{\exp}$ are learnable)**.** *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}^l$ definable in $\mathbb{R}_{\exp}$ with k parameters, and let $\mathcal{N}$ be a neighborhood system definable in $\mathbb{R}_{\exp}$. Then the strategic class $\mathcal{H}^{\mathcal{N}}$ is PAC learnable.*

Moreover, the ERM sample complexity of $\mathcal{H}^{\mathcal{N}}$ satisfies

$$m^{\text{real}}_{\text{ERM}}(\varepsilon,\delta) = O\Big(\frac{k\log(1/\varepsilon)+\log(1/\delta)+K_{\mathcal{H},\mathcal{N}}}{\varepsilon}\Big),$$

where $K_{\mathcal{H},\mathcal{N}}$ is a constant depending on $\mathcal{H}$ and $\mathcal{N}$.

Main Theorem 2 applies to a rich collection of natural strategic settings, including:

- Polynomial classifiers with various neighborhood systems such as ℓ_p balls for $1 \le p \le \infty$, where the radius of the ball can depend on the inpu x.
- Hypothesis classes for inputs that are categorical distributions (i.e., over a finite set), with neighborhood systems such as ℓ_p balls, KL divergence, and earth mover distance.

We elaborate on these examples and others in Appendix A.

The proof relies on the observation that the strategic class $\mathcal{H}^{\mathcal{N}}$ is defined by the formula

$$\Phi_{\mathcal{H}^{\mathcal{N}}}(x,a) = \exists y\ \Phi_{\mathcal{N}}(x,y) \wedge \Phi_{\mathcal{H}}(y,a),$$

where $\Phi_{\mathcal{H}}, \Phi_{\mathcal{N}}$ are the formulas defining $\mathcal{H}, \mathcal{N}$, respectively. Given this observation, Main Theorem 2 follows by applying a result of Johnson and Laskowski [JL10] which bounds the growth function of hypothesis classes defined in an o-minimal structure such as $\mathbb{R}_{\exp}$ (see Appendix B for the definition of o-minimality), and applying Lemma 2.2. The details are given in Appendix E.

3.3 Quantitative bounds

Notice that as $\varepsilon \to 0$, the constant $K_{\mathcal{H},\mathcal{N}}$ in Main Theorem 2 becomes a lower-order term in the sample complexity. However, for constant ε, it becomes relevant, though it remains unspecified. This limitation of Main Theorem 2 originates in Wilkie's proof of the o-minimality of $\mathbb{R}_{\exp}$ [Wil96], a deep and inherently nonconstructive result.

If instead of $\mathbb{R}_{\exp}$ we restrict ourselves to quantifier-free formulas in the language of $\mathbb{R}$ (that is, not using $\exp$), then classical results in real algebraic geometry, which have been used in several prior works in learning theory (for example [BDL93, GJ95, BES02, AMY16, AAE+23]), enable us to obtain completely explicit bounds.

Main Theorem 3 (Quantitative learnability of strategic classes definable in quantifier-free $\mathbb{R}$)**.** *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}^l$ definable with k parameters in the language of $\mathbb{R}$, and let $\mathcal{N}$ be a neighborhood system definable in the same language. Suppose that both $\mathcal{H}$ and $\mathcal{N}$ are defined by quantifier-free formulas involving polynomial equalities and inequalities whose* total description complexity *is at most D (that is, the formulas involve at most s polynomials, each of degree at most D', with $D = sD'$). Then*

$$m^{\text{real}}_{\text{ERM}}(\varepsilon,\delta) = O\Big(\frac{k\log(1/\varepsilon)+k^2 l\log D+\log(1/\delta)}{\varepsilon}\Big),$$

and

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O\big(k^2 l\log D\big).$$

If a hypothesis class is defined using a quantifier-free formula, then Warren-type sign-pattern bounds in the style of [GJ95] give the desired bounds on the growth function, and so on the sample complexity, via Lemma 2.2, as in the preceding main theorem. In our case, $\mathcal{H}^{\mathcal{N}}$ is given by the existential formula

$$\Phi_{\mathcal{H}^{\mathcal{N}}}(x,a) = \exists y\ \big(\Phi_{\mathcal{N}}(x,y) \wedge \Phi_{\mathcal{H}}(y,a)\big).$$

Fortunately, in $\mathbb{R}$ we have *quantifier elimination*, a deep result of Tarski [Tar51] which states that every first-order formula over $\mathbb{R}$ is equivalent to one which is quantifier-free. We complete the proof of Main Theorem 3 using modern quantitative quantifier elimination results [BPR06, Chapter 14]. The procedure is constructive, enabling to derive explicit bounds. Details are given in Appendix F.

While Main Theorem 3 covers many situations of interest, it falls short of capturing important cases such as standard metric and similarity measures used in practice, including earth mover distance

(which requires existential formulas), KL divergence, and ℓ_p-balls for irrational p. Addressing these requires the richer language of $\mathbb{R}_{\exp}$. This setting is covered by our final main theorem. The theorem involves two complexity parameters, *format* and *degree*, which we define after stating the theorem.

Main Theorem 4 (Quantitative learnability of strategic classes definable in existential $\mathbb{R}_{\exp}$)**.** *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}^l$ definable with k parameters in the language of $\mathbb{R}_{\exp}$, and let $\mathcal{N}$ be a neighborhood system definable in the language of $\mathbb{R}_{\exp}$. Suppose, moreover, that both $\mathcal{H}$ and $\mathcal{N}$ are defined using existential formulas of format F and degree D. Then*

$$m^{\text{real}}_{\text{ERM}}(\varepsilon,\delta) = O\Big(\frac{k\log(1/\varepsilon)+\gamma(F)\log(D)+\log(1/\delta)}{\varepsilon}\Big)$$

and

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O_F(\log D),$$

where $\gamma\colon \mathbb{N}\to\mathbb{N}$ is an explicitly computable function.

The format and degree are defined for formulas in the language of $\mathbb{R}_{\exp}$ in the following normal form, for blocks of variables X, Y:

$$\Phi(X) = \exists Y\, \theta(X,Y),$$

where θ is a quantifier-free formula which is a logical combination of atomic predicates of one of the forms

$$P(X,Y)\geq 0 \quad \text{or} \quad z = \exp w,$$

where P is an arbitrary polynomial in the variables X, Y, and z, w are variables, each belonging to one of the blocks X, Y.

As an example, the definition above in Equation (2) of the neighborhood system $\mathcal{N}_p$ is almost of this form (we need to replace $x_i - x_i'$, $x_i' - x_i$, and $p\cdot z_i$ with new variables s_i, t_i, u_i belonging to the Y block, and add the constraints $s_i = x_i - x_i'$, $t_i = x_i' - x_i$, and $u_i = p\cdot z_i$).

The *format* of Φ is the sum of three terms: (i) the number of variables in the block X, (ii) the number of variables in the block Y, (iii) the number of atomic predicates of the form $z = \exp w$.

The *degree* of Φ is the sum of two terms: (i) the sum of the degrees of all polynomials $P(X,Y)$ occurring as atomic formulas, (ii) the number of atomic predicates of the form $z = \exp w$.

Compared to Main Theorem 2, here we are restricted only to existential formulas. This is not a major restriction in practice: e.g. all examples in Appendix A.1 are of this form.

The proof of Main Theorem 4 uses recent quantitative results in tame topology [BV20, BNZ22].[1] The basic idea is to use the approach of [JL10], which is also used in the proof of Main Theorem 2. In order to obtain explicit bounds, we rely on the effective cell decomposition theorem for $\mathbb{R}_{\text{rexp}}$, where rexp denotes restricted exponentiation, as established in [BV20, BNZ22]. The language of $\mathbb{R}_{\text{rexp}}$ extends $\mathbb{R}$ with the function $\exp|_{[-M,M]}$, which equals $\exp$ for inputs in $[-M,M]$ and is zero otherwise. The proof involves converting the defining formulas of $\mathcal{H}$ and $\mathcal{N}$ from $\mathbb{R}_{\exp}$ to $\mathbb{R}_{\text{rexp}}$, a step that we only know how to carry out for existential formulas. See Appendix G for the complete proof.

The results of [BV20, BNZ22] extend more generally to sharply o-minimal structures, a notion defined in these papers. Consequently, Main Theorem 4 extends beyond $\mathbb{R}_{\exp}$, allowing the usage of functions like erf; see Appendix H for more details.

4 Related work

Strategic classification. Since its introduction in [BKS12, HMPW16], the field of strategic classification has drawn much attention. Many works have extended the basic setting to support causal effects [MMH20, SEA20, HR23], varied sources of uncertainty [GNE+21, BPWZ22, SBM23, LUBD23, RR24], and flexible user incentives [LR22, SVXY23]. Nonetheless, due to the challenges strategic behavior introduces, most works restrict their focus to linear classifiers and ℓ_2 costs. Some

[1] Grothendieck [Gro97] defined tame topology ("topologie modérée") as a topology which avoids pathological situations such as space-filling curves. The modern usage refers more specifically to the study of o-minimal structures [vdD98]. These are structures such as $\mathbb{R}_{\exp}$ whose definable sets exhibit many of the properties of semialgebraic sets; see Appendix B.

advances have been made in establishing strategic learnability and sample complexity – first for linear classifiers [SVXY23], and then for broader structured hypothesis classes [CMMS24, TR25].

A standard way to formalize the complexity of strategic classifiers is through the induced *effective* class: for each hypothesis, one considers the classifier obtained after individuals best-respond to it, equivalently by composing the original hypothesis with its induced strategic response map [ZC21, Definition 2]. Building on this perspective, we express the induced strategic hypotheses as definable existential formulas. This provides a geometric model-theoretic way to study their complexity, and may be useful for future work on strategic learning.

Finite VC dimension of the original class is not enough. Several related works show that finite VC dimension of the original class need not control the complexity or learnability of the effective model class induced by user behavior. Zhang and Conitzer [ZC21] study incentive-aware PAC learning through manipulation graphs, and construct an example using non-uniform graphs (i.e., that can differ across users) in which the effective VC can blow up arbitrarily. Lechner and Urner [LU22] extend this framework and show a similar result for an ad-hoc 'strategic' loss which differs from the conventional 0-1 loss. Sundaram et al. [SVXY23] study a richer response model in which users seek arbitrary (yet observable) predictions, rather than positive only, and demonstrate VC blow-up using non-uniform (referred to as 'instance-wise') neighborhoods. While bearing similarities to our construction, Main Theorem 1 remains within the scope of the original strategic classification setting of Hardt et al. [HMPW16]: which neighborhood are continuous, simple, and uniform across users, and learning aims to minimize the standard 0-1 loss.

Relatedly, Attias et al. [ABN+25] study a setting where manipulating an instance x also changes its true label y, which they refer to as 'PAC learning with improvements'. This differs from the conventional strategic setting capturing gaming behavior in which y remains unchanged. Interestingly, their non-learnability example is similar in spirit to our negative result: [ABN+25, Example 2] gives a class of finite VC dimension that is not learnable with improvements, while our Main Theorem 1 gives a class whose ordinary VC dimension is 1 but whose induced strategic class has infinite VC dimension for every radius $r > 0$. A closely related phenomenon also appears in adversarially robust learning; from a strategic modeling perspective, this is akin to assuming users seek predictions which are opposite to their true label—a special case of Sundaram et al. [SVXY23]. Here, Montasser, Hanneke, and Srebro [MHS19] construct a class of VC dimension at most 1 for which proper robust PAC learning fails under metric-ball perturbations.

Geometric hypothesis classes. There is a long line of work bounding the VC dimension of hypothesis classes described by polynomial inequalities and related analytic operations. Goldberg and Jerrum [GJ95] gave general VC-dimension bounds for hypothesis classes parameterized by real numbers and defined by polynomial conditions. Related real-algebraic and sign-pattern methods have been used in learning theory to analyze semialgebraic classes, a partial list includes [BDL93, BES02, AMY16, AAE+23]. Closely related counting questions have also been studied in combinatorics: Sauermann [Sau21] proved essentially tight bounds for graph families whose edge relations are determined by polynomial sign conditions. Another important line concerns neural networks, where VC-dimension bounds have been obtained for networks with piecewise polynomial and piecewise linear structure [BMM98, BHLM19].

Our semialgebraic bounds follow this tradition, but the strategic transformation introduces an existential quantifier, requiring quantitative quantifier elimination before sign-pattern bounds can be applied.

Model-theoretic and tame-geometric tools. Quantifier elimination over real closed fields goes back to Tarski [Tar51] and Seidenberg [Sei54]. Effective algorithmic versions were later developed via cylindrical algebraic decomposition by Collins [Col75], and refined by Renegar [Ren92a, Ren92b, Ren92c] and by Basu–Pollack–Roy [BPR96, Bas99].

Beyond the semialgebraic setting, o-minimality provides a general framework for tame geometry. Wilkie proved that the real exponential field $\mathbb{R}_{\exp}$ is o-minimal [Wil96]. A closely related line, rooted in Khovanskii's theory of fewnomials and Pfaffian functions [Kho91], studies restricted Pfaffian structures: Wilkie's theorem of the complement established their o-minimality [Wil99]. Effective bounds for Pfaffian and sub-Pfaffian sets were developed by Gabrielov and Vorobjov [GV04, GV09]. More recently, Binyamini and Vorobjov obtained quantitative cell decom-

position results for restricted sub-Pfaffian sets [BV20, BNZ22], which we use in our quantitative bounds.

Recent work of [BNS25] studies data-driven algorithm design for parameterized algorithm families whose performance has Pfaffian or piecewise-Pfaffian structure, obtaining learning guarantees in both statistical and online settings. In their online-learning analysis, [BNS25] use the sub-Pfaffian cell-decomposition bounds of [GV01] to control the number of regions induced by Pfaffian transition boundaries. Our application requires a closely related but slightly different kind of control: in order to derive ERM sample-complexity bounds via growth-function estimates, we need bounds not only on the number of cells in a decomposition, but also bounds on the complexity. For this purpose, the effective cell-decomposition theorem of [BV20] is particularly well suited.

These tools have also appeared in learning theory. In particular, Karpinski and Macintyre [KM97] used Pfaffian methods to derive VC-dimension bounds for sigmoidal networks and more general Pfaffian neural networks, providing an early connection between Pfaffian geometry and learning theory. In a more general model-theoretic direction, Johnson and Laskowski [JL10] showed that families definable with k parameters in an o-minimal structure have polynomially bounded growth function with exponent k. Relatedly, Livni and Simon [LS13] used model-theoretic tools to derive sample compression bounds for definable families.

Our work brings this model-theoretic and tame-geometric perspective to strategic PAC learnability.

5 Discussion

Summary of results. Our results show that strategic behavior can fundamentally change the learnability of a hypothesis class: even a VC-dimension-one class on the real line can become non-learnable after interval-neighborhood expansion. On the positive side, we show that this pathology disappears under tame geometric definability assumptions: $\mathbb{R}_{\exp}$-definability gives qualitative PAC learnability, while semialgebraic and existential $\mathbb{R}_{\exp}$ descriptions yield explicit sample-complexity and VC-dimension bounds in terms of their description complexity.

Limitations. These positive results should be read within the limits of the tame-definable framework. The approach excludes neighborhood relations with unbounded oscillatory or periodic structure, does not give quantitative bounds for arbitrary quantified $\mathbb{R}_{\exp}$-formulas, and does not automatically apply to distributional neighborhoods defined through integration or infinite-dimensional optimization. Such distributional models are covered only when the resulting neighborhood relation admits a controlled finite-dimensional definable description. We discuss these limitations in Appendix A.3.

Open directions. Our framework captures the reachability aspect of strategic behavior: a point is positive if it can reach some positive point. A natural next step is to enrich this model by incorporating how agents choose among reachable points, and how these choices reshape the data distribution. It would be interesting to understand whether model-theoretic tools can also be used to analyze such richer response models.

Another direction is to sharpen the quantitative bounds. The dependence on the format in Main Theorem 4 is inherited from general effective cell decomposition theorems and is unlikely to be optimal for learning-theoretic applications. Improving these bounds, or obtaining sharper estimates for special classes of strategic models, would make the theory more informative.

Acknowledgements

We thank Steve Hanneke for insightful discussions.

Nir Rosenfeld is supported by Israel Science Foundation (ISF) Grant No. 278/22. Yuval Filmus is supported by ISF Grant No. 507/24. Shay Moran is supported by ISF Grant No. 2362/26, the Technion Center for Machine Learning and Intelligent Systems (MLIS), and the European Union (ERC, GENERALIZATION, 101039692).

Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[AAE+23] Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. *CoRR*, abs/2312.00379, 2023.

[ABN+25] Idan Attias, Avrim Blum, Keziah Naggita, Donya Saless, Dravyansh Sharma, and Matthew Walter. PAC learning with improvements. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 1950–1981. PMLR, 13–19 Jul 2025.

[Alo95] Noga Alon. Tools from higher algebra. *Handbook of combinatorics*, 2:1749–1783, 1995.

[AMY16] Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank versus VC dimension. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, JMLR Workshop and Conference Proceedings, pages 47–80. JMLR.org, 2016.

[Bas99] Saugata Basu. New results on quantifier elimination over real closed fields and applications to constraint databases. *J. ACM*, 46(4):537–555, July 1999.

[BDL93] Shai Ben-David and Michael Lindenbaum. Localization vs. identification of semi-algebraic sets. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 327–336, 1993.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[BES02] Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. Limitations of learning via embeddings in euclidean half spaces. *J. Mach. Learn. Res.*, 3:441–461, 2002.

[BHLM19] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[BHMZ20] Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal SVM bound. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, Proceedings of Machine Learning Research, pages 582–609. PMLR, 2020.

[BKS12] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

[BMM98] Peter Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc dimension bounds for piecewise polynomial networks. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998.

[BNS25] Maria-Florina Balcan, Anh Tuan Nguyen, and Dravyansh Sharma. Algorithm configuration for structured pfaffian settings. *Trans. Mach. Learn. Res.*, 2025, 2025.

[BNZ22] Gal Binyamini, Dmitri Novikov, and Benny Zak. Sharply o-minimal structures and sharp cellular decomposition. *arXiv preprint arXiv:2209.10972*, 2022.

[BNZ24] Gal Binyamini, Dmitry Novikov, and Benny Zak. Wilkie's conjecture for Pfaffian structures. *Annals of Mathematics*, 199(2):795 – 821, 2024.

[BPR96] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. On the combinatorial and algebraic complexity of quantifier elimination. *Journal of the ACM*, 43(6):1002–1045, 1996.

[BPR06] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in real algebraic geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer, 2006. 2nd ed.

[BPWZ22] Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1691–1715. PMLR, 17–23 Jul 2022.

[BV20] Gal Binyamini and Nicolai Vorobjov. Effective cylindrical cell decompositions for restricted sub-pfaffian sets. *International Mathematics Research Notices*, 2022, 11 2020.

[CMMS24] Lee Cohen, Yishay Mansour, Shay Moran, and Han Shao. Learnability gaps of strategic classification. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1223–1259. PMLR, 2024.

[Col75] George E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In *Automata Theory and Formal Languages*, volume 33 of *Lecture Notes in Computer Science*, pages 134–183, Berlin, Heidelberg, 1975. Springer.

[GJ95] Paul W. Goldberg and Mark R. Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2):131–148, 1995.

[GNE+21] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.

[Gro97] Alexandre Grothendieck. Esquisse d'un programme. In *Geometric Galois actions, 1*, volume 242 of *London Math. Soc. Lecture Note Ser.*, pages 5–48. Cambridge Univ. Press, Cambridge, 1997. With an English translation on pp. 243–283.

[GV01] Andrei Gabrielov and Nicolai Vorobjov. Complexity of cylindrical decompositions of sub-pfaffian sets. *Journal of Pure and Applied Algebra*, 164(1):179–197, 2001. Effective Methods in Algebraic Geometry.

[GV04] Andrei Gabrielov and Nicolai Vorobjov. Complexity of computations with pfaffian and noetherian functions. In Yuri Ilyashenko and Christiane Rousseau, editors, *Normal Forms, Bifurcations and Finiteness Problems in Differential Equations*, volume 137 of *NATO Science Series II: Mathematics, Physics and Chemistry*, pages 211–250. Kluwer Academic Publishers, Dordrecht, 2004.

[GV09] Andrei Gabrielov and Nicolai Vorobjov. Approximation of definable sets by compact families, and upper bounds on homotopy and homology. *Journal of the London Mathematical Society*, 80(1):35–54, 2009.

[Han16a] Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

[Han16b] Steve Hanneke. Refined error bounds for several learning algorithms. *J. Mach. Learn. Res.*, 17:135:1–135:55, 2016.

[HMPW16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[HR23] Guy Horowitz and Nir Rosenfeld. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, pages 13233–13253. PMLR, 2023.

[JL10] H. R. Johnson and M. C. Laskowski. Compression schemes, stable definable families, and o-minimal structures. *Discrete & Computational Geometry*, 43(4):914–926, 2010.

[Kho91] Askold G. Khovanskii. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1991.

[KM97] Marek Karpinski and Angus Macintyre. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Sciences*, 54(1):169–176, 1997.

[LR22] Sagi Levanon and Nir Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*, pages 12593–12618. PMLR, 2022.

[LS13] Roi Livni and Pierre Simon. Honest compressions and their application to compression schemes. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, JMLR Workshop and Conference Proceedings, pages 77–92. JMLR.org, 2013.

[LU22] Tosca Lechner and Ruth Urner. Learning losses for strategic classification. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 7337–7344, 2022.

[LUBD23] Tosca Lechner, Ruth Urner, and Shai Ben-David. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*. PMLR, 2023.

[MHS19] Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 25–28 Jun 2019.

[MMH20] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

[Pia19] Nico Philipp Piatkowski. *Exponential families on resource-constrained systems*. Gesellschaft für Informatik eV, 2019.

[Ren92a] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part i: Introduction. preliminaries. the geometry of semi-algebraic sets. the decision problem for the existential theory of the reals. *Journal of Symbolic Computation*, 13(3):255–299, 1992.

[Ren92b] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part ii: The general decision problem. preliminaries for quantifier elimination. *Journal of Symbolic Computation*, 13(3):301–327, 1992.

[Ren92c] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part iii: Quantifier elimination. *Journal of Symbolic Computation*, 13(3):329–352, 1992.

[RR24] Elan Rosenfeld and Nir Rosenfeld. One-shot strategic classification under unknown costs. In *Forty-first International Conference on Machine Learning*, 2024.

[Sau21] Lisa Sauermann. On the speed of algebraically defined graph classes. *Advances in Mathematics*, 380:107593, 2021.

[SBM23] Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[SEA20] Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 8676–8686, 2020.

[Sei54] Abraham Seidenberg. A new decision method for elementary algebra. *Annals of Mathematics*, 60(2):365–374, 1954.

[SVXY23] Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. PAC-learning for strategic classification. *Journal of Machine Learning Research*, 24(192):1–38, 2023.

[Tal94] M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22:28–76, 1994.

[Tar51] Alfred Tarski. *A Decision Method for Elementary Algebra and Geometry*. University of California Press, Berkeley and Los Angeles, second edition, revised edition, 1951. Prepared for publication with the assistance of J. C. C. McKinsey.

[TR25] Benyamin Trachtenberg and Nir Rosenfeld. Strategic classification with non-linear classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[VC74] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

[vdD98] Lou van den Dries. *Tame topology and o-minimal structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1998.

[vdVW96] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[War68] Hugh E Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

[Wil96] Alex J. Wilkie. 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, 1996.

[Wil99] Alex J. Wilkie. A theorem of the complement and some new o-minimal structures. *Selecta Mathematica*, 5(4):397–421, 1999.

[ZC21] Hanrui Zhang and Vincent Conitzer. Incentive-aware PAC learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5797–5804, 2021.

A Examples

The theorems in Section 3 apply to a broad collection of natural strategic settings. In particular, many commonly used hypothesis classes and neighborhood systems $\mathcal{N} = \{N_x : x \in \mathbb{R}^d\}$ are definable either in $\mathbb{R}$ or in $\mathbb{R}_{\exp}$. We now describe several representative examples to which our main theorems apply.

A.1 Representative examples

Halfspaces with neighborhoods depending on the representative point. Consider the base class $\mathcal{H}$ of halfspaces in $\mathbb{R}^2$, defined by

$$\Phi_{\text{Halfspace}}(x, a) := a_1 x_1 + a_2 x_2 + a_3 > 0,$$

where $x \in \mathbb{R}^2$ and $a \in \mathbb{R}^3$.

Define the neighborhood system $\mathcal{N} = \{N_x : x \in \mathbb{R}^2\}$ by

$$y \in N_x \iff \|x - y\|_2 \le r(x), \qquad r(x_1, x_2) = \max\{x_2, 0\}.$$

Here x_2 can be interpreted as the amount of strategic resources available to the point: points with larger x_2 can move farther in order to obtain a positive classification, while points with $x_2 < 0$ cannot move at all.

Equivalently, the neighborhood relation is defined by the semialgebraic formula

$$\Phi_{\mathcal{N}}(x, y) := \Big(x_2 \ge 0 \wedge \|x - y\|_2^2 \le x_2^2\Big) \ \vee \ \Big(x_2 < 0 \wedge x = y\Big).$$

Thus $\mathcal{N}$ is semialgebraic. Therefore Theorem 3 applies to the strategically transformed halfspace class induced by $\mathcal{N}$, and gives

$$m^{\text{real}}_{\text{ERM}}(\varepsilon, \delta) = O\left(\frac{\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

Polynomial classifiers with ℓ_∞ neighborhoods. Fix a degree bound $D \ge 1$. Consider the base class $\mathcal{H}$ of polynomial threshold functions in $\mathbb{R}^l$, defined by

$$\Phi_{\mathcal{H}}(x, \theta) := P_\theta(x) > 0,$$

where P_θ ranges over all real polynomials in l variables of degree at most D, parametrized by their coefficients θ.

Fix $r > 0$, and define the neighborhood system $\mathcal{N}_\infty = \{N_x : x \in \mathbb{R}^l\}$ by

$$y \in N_x \iff \|x - y\|_\infty \le r.$$

Equivalently, the neighborhood relation is defined by

$$\Phi_{\mathcal{N}_\infty}(x, y) \iff \bigwedge_{i=1}^{l} (-r \le x_i - y_i \le r).$$

Both $\Phi_{\mathcal{H}}$ and $\Phi_{\mathcal{N}_\infty}$ are quantifier-free semialgebraic formulas: $\Phi_{\mathcal{H}}$ involves one polynomial inequality of degree at most D, while $\Phi_{\mathcal{N}_\infty}$ is given by $2l$ linear inequalities.

Therefore Theorem 3 applies to the strategically transformed class $\mathcal{H}^{\mathcal{N}_\infty}$, whose defining formula is

$$\exists y \ \Big(\Phi_{\mathcal{N}_\infty}(x, y) \wedge P_\theta(y) > 0\Big).$$

Here the number of parameters is

$$k = \binom{l + D}{D},$$

corresponding to the coefficients of a degree-D polynomial in l variables. The defining formula uses $2l$ linear inequalities for the ℓ_∞-neighborhood, together with one polynomial inequality $P_\theta(y) > 0$. Thus, in the notation of Theorem 3, one may take $s = 2l + 1$.

Decision trees with polynomial decision rules. Fix integers $D_{tree}, q \geq 1$. A *decision tree with polynomial decision rules* is a binary decision tree of depth at most D_{tree} in which every internal node v is labeled by a polynomial inequality

$$P_v(x) \geq 0,$$

where P_v is a real polynomial in l variables of degree at most q, and every leaf is labeled by either 0 or 1. An input $x \in \mathbb{R}^l$ is classified by starting at the root and moving left or right according to the sign of the corresponding polynomial inequality until reaching a leaf.

Our framework applies to the class of all such trees. Indeed, there are only finitely many tree topologies and leaf labelings of depth at most D_{tree}, so it suffices to encode each fixed topology by a single semialgebraic formula.

A complete binary tree of depth D_{tree} has at most

$$T = 2^{D_{tree}} - 1$$

internal nodes. A polynomial of degree at most q in l variables is determined by

$$M = \binom{l+q}{q}$$

coefficients. Thus the entire tree can be parametrized by at most TM real parameters.

For every leaf λ, there is a unique path from the root of the tree to λ. Along this path, each internal node contributes one polynomial condition: either $P_v(x) \geq 0$ if the path follows the right branch, or $P_v(x) < 0$ if it follows the left branch.

For example, suppose a depth-2 tree first tests

$$P_1(x) \geq 0,$$

and then, on the right branch, tests

$$P_2(x) \geq 0.$$

Then the inputs reaching the leaf corresponding to "right then left" are exactly those satisfying

$$P_1(x) \geq 0 \quad \wedge \quad P_2(x) < 0.$$

More generally, for every leaf labeled by 1, the set of inputs reaching that leaf is described by a conjunction of polynomial inequalities. Since the tree outputs 1 whenever the input reaches one of the positively labeled leaves, the set classified positively by the tree is obtained by taking the disjunction over all such leaves. Hence the decision region of the tree is defined by a formula of the form

$$\bigvee_{\lambda \in L_1} \bigwedge_{v \in \mathrm{path}(\lambda)} \big(P_v(x)\ \sigma_{v,\lambda}\ 0\big),$$

where L_1 is the set of leaves labeled by 1, and each $\sigma_{v,\lambda}$ denotes either $\geq$ or $<$, according to the branch chosen at node v.

This is a quantifier-free semialgebraic formula. Since there are only finitely many tree topologies and leaf labelings of depth at most D_{tree}, taking a finite disjunction over all of them gives a single formula defining the entire class of depth-D_{tree} decision trees with polynomial tests of degree at most q. Consequently, this class is definable in $\mathbb{R}$ with at most

$$k = (2^{D_{tree}} - 1)\binom{l+q}{q}$$

real parameters, where l is the input dimension. These parameters correspond to the coefficients of one degree-q polynomial at each possible internal node of the complete depth-D_{tree} binary tree.

Remark. Bounded-complexity semialgebraic threshold classes are handled in the same way. After fixing the Boolean structure of a formula with s polynomial equalities and inequalities of degree at most q, the only parameters are the coefficients of these polynomials. The resulting positive region is already a quantifier-free semialgebraic formula, so the same uniform-definability argument as above applies.

Neural networks. Feedforward neural networks also fit naturally into our framework. Fix layer widths

$$d_0, d_1, \ldots, d_L,$$

where $d_0 = l$ is the input dimension, and let

$$\sigma \colon \mathbb{R} \to \mathbb{R}$$

be an activation function. A feedforward network with activation σ is a function of the form

$$f_\theta(x) = A_L \sigma\Big(A_{L-1} \sigma\big(\cdots \sigma(A_1 x + b_1) \cdots \big) + b_{L-1} \Big) + b_L,$$

where the matrices A_i and vectors b_i are the trainable parameters, collectively denoted by θ.

If σ were omitted, the composition above would collapse to a single affine map, and the resulting class would already be definable in $\mathbb{R}$. Thus the definability of the network class is determined entirely by the definability of the activation function.

Many standard activations are definable in tame structures. For example, the ReLU activation $\sigma(t) = \max\{t, 0\}$ is semialgebraic, and therefore definable in $\mathbb{R}$. The logistic sigmoid $\sigma(t) = \frac{1}{1+\exp(-t)}$ is definable in $\mathbb{R}_{\exp}$. More generally, activations constructed from arithmetic operations together with $\exp$, such as $\tanh$, softplus, and ELU, are also definable in $\mathbb{R}_{\exp}$.

Consequently, for every fixed architecture and every definable activation function, the corresponding thresholded network class

$$\mathcal{H} = \{x \mapsto \mathbf{1}[f_\theta(x) \geq 0] : \theta\}$$

is uniformly definable in the same structure. An explicit logical formula encoding the computation of sigmoid neural networks is given in Example A.4.

Remark A.1 (Composition and definability). The neural-network example illustrates a general closure principle: definable maps may be combined by finite algebraic operations and finite composition without leaving the ambient structure. Thus, once the elementary building blocks of a model are definable, for instance affine maps and the activation function, any fixed finite architecture obtained by composing these blocks is again definable. In this sense, definability is well suited to compositional model classes such as neural networks.

ℓ_p-distance neighborhoods. ℓ_p-distance neighborhoods are among the standard ways to model bounded perturbations in feature space. They arise when an individual may modify its features only within a limited budget, measured in a fixed norm. Different choices of p capture different types of constraints: $p = 2$ gives Euclidean perturbations, $p = 1$ bounds the total coordinate-wise change, and $p = \infty$ gives coordinate-wise bounded deviations. The same definability argument also covers fixed $0 < p < 1$, where the ℓ_p quasi-norm increasingly favors changes supported on few coordinates; in the limit $p \to 0$, this resembles a Hamming-type constraint on the number of modified features.

Formally, for fixed $p > 0$, define the neighborhood system $\mathcal{N}_p = \{N_x : x \in \mathbb{R}^l\}$ by

$$y \in N_x \iff \|x - y\|_p \leq 1,$$

equivalently,

$$\sum_{i=1}^{l} |x_i - y_i|^p \leq 1.$$

The corresponding neighborhood relation is definable in $\mathbb{R}_{\exp}$. Indeed, absolute values can be expressed using algebraic constraints, and for $t > 0$ the map $t \mapsto t^p$ can be written as

$$t^p = \exp(p \log t),$$

with the case $t = 0$ handled separately. For completeness, we give an explicit formula $\Phi_{\mathcal{N}_p}(x, y)$ defining ℓ_p-balls in Example A.1. In particular, the formula $\Phi_{\mathcal{N}_p}(x, y)$ has format and degree $O(l)$.

Consequently, whenever one has a complexity bound for a formula defining the base hypothesis class $\mathcal{H}$, Theorem 4 gives an effective ERM sample-complexity bound for the strategically transformed class $\mathcal{H}^{\mathcal{N}_p}$.

For example, let $A = \{\alpha_i\}_{i\in I} \subseteq \mathbb{N}^l$ be a set of multiindices with $|A| = O(l)$, and consider sparse polynomial threshold functions of the form

$$P_\theta(x) = \sum_{i\in I} \theta_i x^{\alpha_i}.$$

Such polynomials are commonly called *fewnomials*.

This class has only $O(l)$ parameters and format $O(l)$, while its degree

$$q := \max_{i\in I} |\alpha_i|$$

may be arbitrarily large independently of l. Thus Theorem 4 applies to the strategically transformed class $\mathcal{H}^{\mathcal{N}_p}$, giving non-trivial bounds.

In some cases, there may be no effective way to control the format and degree of a formula defining the hypothesis class $\mathcal{H}$. In such cases, one can instead use the asymptotic bounds provided by Main Theorem 2.

KL-divergence neighborhoods on finite domains. KL-divergence neighborhoods arise naturally in settings where one models multiplicative or information-theoretic perturbations of distributions. They are also standard in robust optimization and distribution-shift models, where one allows the underlying distribution to vary within a KL ball around a nominal distribution.

Let $X = \{a_1, \dots, a_l\}$ be a finite set. We consider a strategic classification problem whose domain is the simplex

$$\Delta_l = \left\{ x \in \mathbb{R}^l_{\geq 0} : \sum_{i=1}^{l} x_i = 1 \right\}.$$

A point $x = (x_1, \dots, x_l)$ represents a probability distribution on X, where x_i is the mass assigned to a_i.

Define the neighborhood system $\mathcal{N}_{\mathrm{KL}} = \{N_x : x \in \Delta_l\}$ as follows. On the interior of the simplex,

$$y \in N_x \iff \mathrm{KL}(x\|y) \leq 1,$$

where

$$\mathrm{KL}(x\|y) = \sum_{i=1}^{l} x_i \log\left(\frac{x_i}{y_i}\right).$$

Equivalently, the neighborhood relation is defined by the formula

$$\Phi_{\mathcal{N}_{\mathrm{KL}}}(x, y) \iff \sum_{i=1}^{l} x_i \log\left(\frac{x_i}{y_i}\right) \leq 1,$$

together with the domain restrictions

$$x_i > 0, \qquad y_i > 0, \qquad \sum_{i=1}^{l} x_i = \sum_{i=1}^{l} y_i = 1.$$

This relation is definable in $\mathbb{R}_{\exp}$, since the logarithm is definable from the exponential function and the remaining operations are algebraic. For completeness, we give an explicit defining formula in Example A.2.

Earth mover distance (Wasserstein-1) neighborhoods. The earth mover distance (EMD), also known as the Wasserstein-1 distance, measures the cost of transforming one distribution into another by transporting mass across a ground metric. Unlike ℓ_p or KL distances, it captures the *geometry* of the domain, penalizing movement of mass proportionally to the distance it travels. This makes it a natural model in applications involving structured domains, such as histograms, spatial data, or ordered categories.

Formally, let $X = \{a_1, \dots, a_l\}$ be a finite domain equipped with a ground metric

$$\rho \colon X \times X \to \mathbb{R}_{\geq 0},$$

assumed to be definable in the underlying structure.[2]

As in the previous example, the domain is the family of probability distributions over X, identified with the simplex

$$\Delta_l = \left\{ x \in \mathbb{R}^l_{\geq 0} : \sum_{i=1}^{l} x_i = 1 \right\}.$$

For two distributions $x, y \in \Delta_l$, the earth mover distance is

$$\mathrm{EMD}(x, y) := \min_{\gamma \in \Gamma(x,y)} \sum_{i,j=1}^{l} \gamma_{ij} \rho(i, j),$$

where $\Gamma(x, y)$ is the set of couplings between x and y, i.e., nonnegative matrices $\gamma = (\gamma_{ij})$ such that

$$\sum_{j=1}^{l} \gamma_{ij} = x_i \quad \text{and} \quad \sum_{i=1}^{l} \gamma_{ij} = y_j.$$

Define the neighborhood system $\mathcal{N}_{\mathrm{EMD}} = \{N_x : x \in \Delta_l\}$ by

$$y \in N_x \iff \mathrm{EMD}(x, y) \leq 1.$$

Equivalently, the neighborhood relation is defined by

$$\Phi_{\mathcal{N}_{\mathrm{EMD}}}(x, y) \iff \exists \gamma \in \mathbb{R}^{l \times l}_{\geq 0}$$

such that

$$\sum_{j=1}^{l} \gamma_{ij} = x_i \ (1 \leq i \leq l), \qquad \sum_{i=1}^{l} \gamma_{ij} = y_j \ (1 \leq j \leq l), \qquad \sum_{i,j=1}^{l} \gamma_{ij} \rho(i, j) \leq 1.$$

This relation is definable in $\mathbb{R}$, since the constraints defining feasible transport plans are linear, and the transportation cost is a linear function of γ. Therefore this neighborhood system can be used in any of the three main theorems, depending on the hypothesis class being transformed. For completeness, we provide an explicit defining formula in Example A.3.

KL-divergence for Gaussian families. As in the discrete case above, KL-divergence is defined via an integral,

$$\mathrm{KL}(p \| q) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} \, dx.$$

At first sight, this falls outside the definable framework, since our setting is not closed under integration.

However, for Gaussian families, the integral admits a closed-form expression. For instance, for the location family

$$p_\mu = \mathcal{N}(\mu, 1),$$

one has

$$\mathrm{KL}(p_\mu \| p_{\mu'}) = \frac{(\mu - \mu')^2}{2}.$$

More generally, for Gaussians with arbitrary means and variances, the KL-divergence is given by an explicit expression involving polynomials, logarithms, and division. Consequently, the neighborhood system $\mathcal{N}_{\mathrm{KL}}$ defined by

$$p_{\mu'} \in N_{p_\mu} \iff \mathrm{KL}(p_\mu \| p_{\mu'}) \leq 1$$

is definable in $\mathbb{R}_{\exp}$.

Thus, despite being defined via an integral, KL-divergence neighborhoods for Gaussian families fall within the scope of Theorem 4.

[2] In the finite case, this simply means that each value $\rho(i, j)$ is a fixed real constant appearing in the formula. For example, if $X = \{1, 2, 3\}$, one may specify

$$\rho(1, 2) = 1, \quad \rho(1, 3) = 2, \quad \rho(2, 3) = 1,$$

and these constants are treated as part of the description of the neighborhood. More generally, one may allow ρ to be given by a definable function when working over continuous domains.

Total-variation neighborhoods for Gaussian families. In contrast to KL-divergence, total-variation (TV) distance typically does not admit a closed-form algebraic expression, and therefore provides a natural example where integration leads outside the definable $\mathbb{R}_{\exp}$ setting.

Given a parametric family $\{p_\theta\}$, define the neighborhood system $\mathcal{N}_{\mathrm{TV}} = \{N_{p_\theta}\}$ by

$$p_{\theta'} \in N_{p_\theta} \iff d_{\mathrm{TV}}(p_\theta, p_{\theta'}) \le \varepsilon,$$

where

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2} \int_{\mathbb{R}} |p(x) - q(x)|\, dx.$$

Unlike the KL case, the integral defining TV does not, in general, simplify to a finite expression in terms of elementary operations.

For instance, for Gaussian location families,

$$p_\mu = \mathcal{N}(\mu, 1),$$

one has the explicit formula

$$d_{\mathrm{TV}}(p_\mu, p_{\mu'}) = 2\Phi\left(\frac{|\mu - \mu'|}{2}\right) - 1,$$

where Φ is the Gaussian cumulative distribution function. This function is expressed via the error function

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt,$$

which is known not to be definable in $\mathbb{R}_{\exp}$ (see, e.g., [Pia19, Chapter 4.4]).

However, this does not imply that such neighborhoods are intractable. The key observation is that the error function is a Pfaffian function (see Definition G.1). Indeed, setting

$$f_1(x) = e^{-x^2}, \qquad f_2(x) = \mathrm{erf}(x),$$

we obtain a Pfaffian chain:

$$f_1'(x) = -2x f_1(x), \qquad f_2'(x) = \frac{2}{\sqrt{\pi}} f_1(x).$$

Thus erf is definable in the restricted Pfaffian structure $\mathbb{R}_{\mathrm{rPfaff}}$.

Consequently, after restricting the parameters to a bounded domain, the neighborhood relation induced by Gaussian TV distance becomes definable in $\mathbb{R}_{\mathrm{rPfaff}}$, and the quantitative bounds from Appendix H apply to this setting. This example illustrates that while definability is not preserved under integration in general, many natural statistical distances remain tame after passing to the Pfaffian setting.

Remark A.2 (Definable dependence of neighborhoods). The examples above can be further generalized by allowing the neighborhoods N_x to depend on x in a definable way. If this dependence can be controlled, and if one can bound the format and degree of the resulting formula $\Phi_{\mathcal{N}}(X, Y)$ defining the neighborhood system, then the same quantitative bounds apply.

More precisely, the parameters defining the neighborhoods may themselves be given by definable functions of x. For example, one may consider KL-balls with radius $\varepsilon = \varepsilon(x)$, ℓ_p-balls where $p = p(x)$ varies with x, or earth mover distances with a ground metric $\rho = \rho(x)$. As long as these dependencies are definable in the underlying structure, the resulting neighborhood system $\mathcal{N} = \{N_x : x \in \mathbb{R}^d\}$ remains definable. The corresponding format and degree can then be bounded by a straightforward semantic argument.

This includes, as a special case, piecewise-defined neighborhood systems obtained by partitioning the domain into finitely many definable regions and assigning a different defining formula $\Phi_{\mathcal{N}}(X, Y)$ on each region.

A.2 Explicit formulas

Example A.1. *If $p < \infty$ then we can define the neighborhood relation corresponding to $\|x-x'\|_p \leq 1$ in $\mathbb{R}^l$ as*

$$\Phi_{\mathcal{N}_p}(x,x') \iff \exists y_1,\dots,y_l,z_1,\dots,z_l \left[\sum_{i=1}^{l} y_i \leq 1 \wedge \bigwedge_{i=1}^{l} \big((x_i = x_i' \wedge y_i = 0) \vee (y_i = \exp(pz_i) \wedge (\exp z_i = x_i - x_i' \vee \exp z_i = x_i' - x_i))\big)\right].$$

If $p = \infty$ then the neighborhood relation corresponding to $\|x - x'\|_\infty \leq 1$ in $\mathbb{R}^l$ is given by

$$\Phi_{\mathcal{N}_\infty}(x,x') \iff \bigwedge_{i=1}^{l} \big(x_i - x_i' \leq 1 \wedge x_i' - x_i \leq 1\big).$$

Example A.2. *The neighborhood relation corresponding to $\mathrm{KL}(x\|x') \leq 1$ for distributions on a finite set of size l, represented as points in $\mathbb{R}^l$ is*

$$\Phi_{\mathcal{N}_{\mathrm{KL}}}(x,x') \iff \exists z_1,\dots,z_l \left[\bigwedge_{i=1}^{l} (x_i \geq 0 \wedge x_i' \geq 0) \wedge \sum_{i=1}^{l} x_i = 1 \wedge \sum_{i=1}^{l} x_i' = 1 \wedge \sum_{i=1}^{l} x_i z_i \leq 1 \wedge \bigwedge_{i=1}^{l} \big((x_i > 0 \wedge x_i' > 0 \wedge x_i' \exp z_i = x_i) \vee (x_i = 0 \wedge z_i = 0)\big)\right].$$

Example A.3. *The neighborhood relation corresponding to $\mathrm{EMD}(x,x')$ for distributions on a finite set of size l, represented as points in $\mathbb{R}^l$ with respect to a metric $\rho\colon [l] \times [l] \to \mathbb{R}_+$ (which we think of as constants) is*

$$\Phi_{\mathcal{N}_{\mathrm{EMD}}}(x,x') \iff \exists \gamma_{ij}|_{1\leq i,j\leq l} \left[\bigwedge_{i=1}^{l} (x_i \geq 0 \wedge x_i' \geq 0) \wedge \sum_{i=1}^{l} x_i = 1 \wedge \sum_{i=1}^{l} x_i' = 1 \wedge \bigwedge_{i,j=1}^{l} \gamma_{ij} \geq 0 \wedge \bigwedge_{i=1}^{l} \sum_{j=1}^{l} \gamma_{ij} = x_i \wedge \bigwedge_{j=1}^{l} \sum_{i=1}^{l} \gamma_{ij} = x_j' \wedge \sum_{i,j=1}^{l} \gamma_{ij}\rho(i,j) \leq 1\right].$$

Example A.4. *Fix a feedforward architecture with layer widths*

$$d_0, d_1, \dots, d_L,$$

where $d_0 = l$ is the input dimension and $d_L = 1$, and logistic sigmoid activation

$$\sigma(t) = \frac{1}{1+\exp(-t)}.$$

A network in this architecture is determined by weight matrices $A^{(j)} \in \mathbb{R}^{d_j \times d_{j-1}}$ and bias vectors $b^{(j)} \in \mathbb{R}^{d_j}$ for $1 \leq j \leq L$. Let

$$\theta = (A^{(1)}, b^{(1)}, \dots, A^{(L)}, b^{(L)})$$

denote the trainable parameters. Writing $z^{(0)} := x$, the network is computed recursively by

$$z^{(j)} = \sigma\big(A^{(j)} z^{(j-1)} + b^{(j)}\big), \qquad 1 \leq j \leq L,$$

where σ is applied coordinate-wise. The associated thresholded network class is

$$\mathcal{H} = \Big\{ x \mapsto \mathbf{1}[z_1^{(L)} \geq 0] : \theta \Big\}.$$

This class is definable in $\mathbb{R}_{\exp}$. Indeed, introduce auxiliary variables $r_i^{(j)}, q_i^{(j)}, z_i^{(j)}$, where $r_i^{(j)}$ is the affine input to the neuron, $q_i^{(j)} = \exp(-r_i^{(j)})$, and $z_i^{(j)}$ is the neuron output. Denote $z^{(0)}$ by x. The network computation is encoded by

$$\Phi_{\mathrm{NN}}(x,\theta) \iff \exists r_i^{(j)}, q_i^{(j)}, z_i^{(j)} \,|_{1\le j\le L,\ 1\le i\le d_j} \left[\bigwedge_{j=1}^{L} \bigwedge_{i=1}^{d_j} \left(r_i^{(j)} = \sum_{s=1}^{d_{j-1}} A_{is}^{(j)} z_s^{(j-1)} + b_i^{(j)} \ \wedge \right.\right.$$
$$\left.\left. q_i^{(j)} = \exp(-r_i^{(j)}) \ \wedge \ z_i^{(j)}(1+q_i^{(j)}) = 1 \right) \ \wedge \ r_1^{(L)} \ge 0 \right].$$

Thus sigmoid neural networks of fixed architecture are uniformly definable in $\mathbb{R}_{\exp}$. Matrix-vector notation is used only as shorthand: all equations expand into coordinate-wise polynomial equations together with applications of exp.

A.3 Limitations of the definability framework

Outside the tame-definable regime. Our positive results rely on tameness. Although we focus primarily on hypothesis classes and neighborhood systems definable in $\mathbb{R}_{\exp}$, the same approach also applies beyond the specific examples emphasized in the main text; see the Gaussian examples in Appendix A and the extensions discussed in Appendix H. However, not every natural neighborhood system can be accommodated by this model-theoretic framework.

For example, define the neighborhood system $\mathcal{N} = \{N_x : x \in \mathbb{R}\}$ by

$$y \in N_x \iff |\sin(y-x)| \le \varepsilon.$$

One cannot obtain analogues of Main Theorem 2 simply by adding sin to the language of the real field. The obstruction is not merely technical. The neighborhood

$$N_0 = \{y : |\sin y| \le \varepsilon\}$$

is an infinite periodic union of disjoint intervals, repeating the same local pattern across infinitely many regions. Such behavior is incompatible with the tameness underlying $\mathbb{R}_{\exp}$, formalized by o-minimality; see Appendix B for the formal definition of o-minimality.

More importantly, in structures that permit this kind of unbounded periodic repetition, one cannot expect general learnability guarantees for strategic expansions of learnable hypothesis classes. A closely related mechanism is exploited in Lemma D.1, which yields infinite VC dimension after neighborhood expansion; see also Remark D.2. Thus the exclusion of oscillatory or periodic structure is an essential feature of the framework, not just a limitation of the proof.

Qualitative versus quantitative guarantees in $\mathbb{R}_{\exp}$. The quantitative bounds in Theorem 4 apply to existentially definable formulas, and therefore do not directly cover formulas with universal quantifiers. This limitation is quantitative rather than qualitative: Main Theorem 2 still applies to arbitrary first-order definable hypothesis classes and neighborhood systems in $\mathbb{R}_{\exp}$.

For example, robustly defined hypothesis classes of the form

$$\Phi_{\mathcal{H}}(x,a) \iff \forall y \ \big(|y-x| \le r \ \Rightarrow \ f_a(y) \ge 0\big)$$

are first-order definable whenever f_a is definable, and hence fall under the qualitative guarantee. However, they are not covered by our quantitative existential bounds unless the universal condition can be converted, with controlled complexity, into an equivalent existential or quantifier-free form.

Distributional and optimization-based neighborhoods. Our framework is not closed under operations such as integration or optimization over infinite-dimensional spaces. Consequently, some natural neighborhood systems modeling strategic behavior fall outside the definable regime. This does not mean that all distributional neighborhood systems are excluded: in some special cases, integral-defined quantities admit explicit finite-dimensional descriptions and therefore remain definable; see Appendix A for Gaussian examples. However, such reductions are not available in general.

For example, consider the neighborhood system $\mathcal{N}_{\mathrm{EMD}}$ defined by

$$p_{\theta'} \in N_{p_\theta} \iff \mathrm{EMD}(p_\theta, p_{\theta'}) \leq \varepsilon,$$

where

$$\mathrm{EMD}(p, q) = \inf_{\gamma \in \Gamma(p,q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| \, d\gamma(x, y),$$

and $\Gamma(p, q)$ denotes the set of couplings of p and q.

Even when the distributions p_θ are definable, the induced neighborhood relation need not be first-order definable in $\mathbb{R}_{\exp}$, since the definition involves optimization over an infinite-dimensional family of transport plans. This should be contrasted with the finite-domain Earth-Mover example in Appendix A, where the transport problem is finite-dimensional and therefore admits a definable formulation.

B Technical Background

B.1 VC dimension and growth functions

Let $\mathcal{H} \subseteq \{0, 1\}^X$ be a hypothesis class on a domain X.

We freely identify subsets $A \subseteq X$ with their indicator functions $\mathbf{1}_A \colon X \to \{0, 1\}$. Thus, when convenient, we will use set-theoretic notation interchangeably with functional notation for hypothesis classes.

Definition B.1 (Trace and growth function)**.** Let $A \subseteq X$ be finite. The *trace* of $\mathcal{H}$ on A is

$$\mathcal{H}|_A := \{(h(x))_{x \in A} : h \in \mathcal{H}\}.$$

The *growth function* of $\mathcal{H}$ is

$$\Pi_{\mathcal{H}}(m) := \max_{A \subseteq X,\ |A| = m} |\mathcal{H}|_A|.$$

Definition B.2 (VC dimension)**.** A finite set $A \subseteq X$ is *shattered* by $\mathcal{H}$ if

$$\mathcal{H}|_A = \{0, 1\}^A.$$

The *VC dimension* of $\mathcal{H}$ is

$$\mathbb{VC}(\mathcal{H}) := \sup\{|A| : A \subseteq X \text{ is shattered by } \mathcal{H}\}.$$

The following classical bound relates the growth function to the VC dimension.

Theorem B.3 (Sauer–Shelah–Perles)**.** *If $\mathbb{VC}(\mathcal{H}) = d < \infty$, then for all $m \geq d$,*

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d.$$

B.2 O-minimality and tame topology

The positive results in this paper rely on a form of geometric tameness. Historically, Grothendieck [Gro97] used the term *tame topology* (*topologie modérée*) to describe a setting in which pathological topological phenomena are excluded. In modern model theory, one precise and widely used formalization of this idea is *o-minimality*. Informally, o-minimal structures are expansions of the real field whose definable sets behave like semialgebraic sets: they admit finite decompositions into simple pieces and cannot exhibit infinite oscillation or repeated discrete patterns.

Definition B.4 (o-minimality)**.** An expansion $\mathcal{R}$ of the ordered real field is called *o-minimal* if every definable subset of $\mathbb{R}$ is a finite union of points and intervals.

The structures used in this paper are o-minimal. The real field $\mathbb{R}$ is o-minimal by Tarski's quantifier-elimination theorem. The exponential real field $\mathbb{R}_{\exp}$ is o-minimal by Wilkie's theorem [Wil96]. The restricted Pfaffian structure $\mathbb{R}_{\mathrm{rPfaff}}$ is o-minimal by Wilkie's theorem of the complement [Wil99].

Although the definition only refers to subsets of the real line, o-minimality has strong higher-dimensional consequences. In particular, definable subsets of $\mathbb{R}^n$ admit finite cell decompositions, and definable families have controlled combinatorial complexity. See [vdD98] for the general theory.

Definition B.5 (Cylindrical cell decomposition)**.** A *cell* is defined inductively on the ambient dimension. In $\mathbb{R}$, cells are points and open intervals. Suppose cells in $\mathbb{R}^{n-1}$ have already been defined. A cell in $\mathbb{R}^n$ is either

1. the graph of a continuous definable function
$$f : C \to \mathbb{R}$$
over a cell $C \subseteq \mathbb{R}^{n-1}$, or
2. a band of the form
$$\{(x,t) \in C \times \mathbb{R} : f(x) < t < g(x)\},$$
where $C \subseteq \mathbb{R}^{n-1}$ is a cell and $f, g : C \to \mathbb{R}$ are continuous definable functions with $f < g$. We also allow $f = -\infty$ or $g = +\infty$.

A *cylindrical cell decomposition* of $\mathbb{R}^n$ is a finite partition of $\mathbb{R}^n$ into cells. It is called *compatible* with a set $X \subseteq \mathbb{R}^n$ if every cell is either contained in X or disjoint from X. More generally, it is compatible with a finite family $X_1, \dots, X_s \subseteq \mathbb{R}^n$ if it is compatible with each X_i.

The cell decomposition theorem for o-minimal structures states that if $X_1, \dots, X_s \subseteq \mathbb{R}^n$ are definable sets, then there exists a finite cell decomposition of $\mathbb{R}^n$ compatible with all of them.

B.3 VC bounds for polynomially definable classes

Theorem B.6 (Goldberg–Jerrum bound [GJ95, Theorem 2.2])**.** *Let $\mathcal{H} = \{h_a : a \in \mathbb{R}^k\}$ be a hypothesis class on $\mathbb{R}^l$. Assume that membership $x \in h_a$ can be expressed by a quantifier-free Boolean formula*
$$\Phi(x, a)$$
involving at most s distinct polynomial (in)equalities in the variables $(x, a) \in \mathbb{R}^{l+k}$, each of degree at most D.

Then
$$\mathbb{VC}(\mathcal{H}) \;\le\; O(k \log(sD)) .$$

B.4 Quantifier elimination

We use the following quantitative form of one-block quantifier elimination (see Basu–Pollack–Roy [BPR06, Theorem 14.16]).

Theorem B.7 (One-block quantifier elimination)**.** *Let $P = \{P_1, \dots, P_s\} \subset \mathbb{R}[X, Y]$ be a family of s polynomials of degree at most d in variables $X \in \mathbb{R}^k$ and $Y \in \mathbb{R}^\ell$. Let $\Phi(X, Y)$ be a Boolean combination of atomic predicates of the form $P_i(X, Y)\ \sigma\ 0$, where $\sigma \in \{<, =, >\}$. Consider the existential formula*
$$\psi(Y) := \exists X \in \mathbb{R}^k \;\; \Phi(X, Y).$$
Then $\psi(Y)$ is equivalent to a quantifier-free formula of the form
$$\Psi(Y) = \bigvee_{i=1}^{I} \bigwedge_{j=1}^{J_i} \left(\bigvee_{n=1}^{N_{ij}} \operatorname{sign}(Q_{ijn}(Y)) = \sigma_{ijn} \right),$$
where each $Q_{ijn} \in \mathbb{R}[Y]$ is a polynomial and $\sigma_{ijn} \in \{-1, 0, 1\}$. Moreover,
$$I \le s^{(k+1)(\ell+1)} d^{O(k\ell)}, \quad J_i \le s^{(k+1)} d^{O(k)}, \quad N_{ij} \le d^{O(k)}, \quad \deg(Q_{ijn}) \le d^{O(k)}.$$

Corollary B.8 (One-block quantifier elimination: polynomial form)**.** *Let $X \in \mathbb{R}^k$ and $Y \in \mathbb{R}^\ell$, and let*
$$\psi(Y) := \exists X \in \mathbb{R}^k \; \Phi(X, Y),$$
where $\Phi(X, Y)$ is a quantifier-free Boolean formula involving at most s distinct polynomial sign conditions
$$P(X, Y)\ \sigma\ 0, \qquad \sigma \in \{<, =, >\},$$
with each polynomial P having degree at most d.

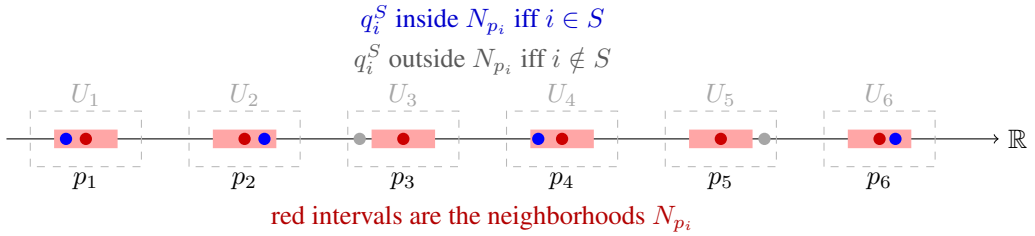


Figure 1: Each anchor p_i has its own private cell U_i. For the set $F_S = \{q_i^S : i \in [n]\}$, the point q_i^S is placed inside N_{p_i} exactly when $i \in S$, and outside N_{p_i} otherwise. Thus $(\mathbf{1}_{F_S})^{\mathcal{N}_s} \cap \{p_1, \dots, p_n\} = \{p_i : i \in S\}$.

Then $\psi(Y)$ is equivalent to a quantifier-free formula $\Psi(Y)$ whose truth value is determined by the sign pattern of a family of polynomials

$$\mathcal{P} \subseteq \mathbb{R}[Y]$$

such that:

- $|\mathcal{P}| \le s^{O(k\ell)} d^{O(k\ell)}$,
- *every $P \in \mathcal{P}$ has degree at most $d^{O(k)}$.*

Proof. Apply Theorem B.7 to the quantifier-free formula $\Phi(X, Y)$, which involves at most s polynomial sign conditions of degree at most d. This yields an equivalent quantifier-free formula $\Psi(Y)$ whose truth value is determined by a Boolean combination of sign conditions of polynomials in Y.

The length of Ψ is bounded (up to a constant factor) by the size of its Boolean representation, which is at most

$$I \cdot \max_i J_i \cdot \max_{i,j} N_{ij}.$$

Substituting the bounds from Theorem B.7 gives

$$L \le s^{O(k\ell)} d^{O(k\ell)}.$$

The degree bound follows directly from the theorem. □

C Strategic expansion breaks learnability

The aim of this section is to prove Main Theorem 1.

Theorem 1. *For $r > 0$, let $\mathcal{N}_r$ be the neighborhood system given by $x' \in N_x \iff |x - x'| \le r$.*

There exists a hypothesis class $\mathcal{H}$ over $\mathbb{R}$ such that $\mathbb{VC}(\mathcal{H}) = 1$ but for every $r > 0$, $\mathbb{VC}(\mathcal{H}^{\mathcal{N}_r}) = \infty$.

We split the proof into two steps. First, in Lemma C.1, we construct, for every n, a hypothesis class on an interval whose VC dimension is 1, but whose thickening has VC dimension n. Then, in Corollary C.2, we extend this construction to obtain Main Theorem 1.

We also prove a related pathology in Lemma C.3: there exists a hypothesis class of VC dimension 1 and a neighborhood system whose associated neighborhood class has VC dimension 1, while the induced strategic class is not PAC learnable.

In this section, we identify binary hypotheses with their positive regions.

Lemma C.1 (VC blows up after interval neighborhoods)**.** *For every $n \ge 1$ and every $r \ge r' > 0$, there exist a finite hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathbb{R}}$ with $|\mathcal{H}| = 2^n$ and a neighborhood system*

$$\mathcal{N}_s = \{N_x : x \in \mathbb{R}\}, \qquad y \in N_x \iff |x - y| \le s,$$

such that for every $s \in [r', r]$,

$$\mathbb{VC}(\mathcal{H}) = 1 \qquad \textit{but} \qquad \mathbb{VC}(\mathcal{H}^{\mathcal{N}_s}) = n.$$

Proof. Fix $r > 0$.

Well-separated anchor points. Let $p_1, \dots, p_n \in \mathbb{R}$ be such that

$$|p_i - p_j| > 4r \quad (i \neq j),$$

for example $p_i = 10ri$.

For each i, define a small interval

$$U_i := (p_i - 2r,\, p_i + 2r),$$

so the intervals U_i are pairwise disjoint.

Encoding subsets. For each subset $S \subseteq [n]$, define a set F_S by choosing one point $q_i^S \in U_i$ for every $i \in [n]$ such that

$$|q_i^S - p_i| < r' \quad \text{if } i \in S, \qquad |q_i^S - p_i| > r \quad \text{if } i \notin S.$$

Set

$$F_S := \{q_1^S, \dots, q_n^S\}.$$

Since each U_i contains infinitely many points, we may choose all q_i^S so that the sets F_S are pairwise disjoint.

VC dimension of $\mathcal{H}$. Let

$$\mathcal{H} := \{\mathbf{1}_{F_S} : S \subseteq [n]\}.$$

The sets F_S are nonempty and pairwise disjoint, hence $\mathcal{H}$ shatters a singleton but cannot shatter any two-point set. Therefore

$$\mathbb{VC}(\mathcal{H}) = 1.$$

Effect of the transformation. We claim that $\mathcal{H}^{\mathcal{N}_s}$ shatters $\{p_1, \dots, p_n\}$ whenever $r' \leq s \leq r$.

Fix $i \in [n]$:

- If $i \in S$, then $|q_i^S - p_i| < r' \leq s$, hence
$$q_i^S \in N_{p_i},$$
and therefore
$$p_i \in (\mathbf{1}_{F_S})^{\mathcal{N}_s}.$$
- If $i \notin S$, then
$$|q_i^S - p_i| > r \geq s, \quad \text{and for } j \neq i, \quad |q_j^S - p_i| > r \geq s$$
(by separation of the U_j). Hence
$$F_S \cap N_{p_i} = \emptyset,$$
so
$$p_i \notin (\mathbf{1}_{F_S})^{\mathcal{N}_s}.$$

Thus

$$(\mathbf{1}_{F_S})^{\mathcal{N}_s} \cap \{p_1, \dots, p_n\} = \{p_i : i \in S\},$$

so all subsets are realized. Hence $\mathcal{H}^{\mathcal{N}_s}$ shatters an n-point set, and therefore

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}_s}) = n.$$

□

Corollary C.2 (A single class that blows up for every radius)**.** *There exists an infinite hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathbb{R}}$ such that*

$$\mathbb{VC}(\mathcal{H}) = 1, \qquad \textit{but for every } s > 0, \qquad \mathbb{VC}(\mathcal{H}^{\mathcal{N}_s}) = \infty,$$

where $\mathcal{N}_s = \{N_x : x \in \mathbb{R}\}$ is given by

$$y \in N_x \iff |x - y| \leq s.$$

Proof. For $n \geq 1$ and $m \in \mathbb{Z}$, let $\mathcal{H}_{n,m}$ be the hypothesis class given by Lemma C.1 with

$$r = 2^{-m}, \qquad r' = 2^{-m-1}.$$

Equivalently, write

$$\mathcal{H}_{n,m} = \{\mathbf{1}_F : F \in \mathcal{A}_{n,m}\},$$

where $\mathcal{A}_{n,m}$ is the corresponding family of subsets of $\mathbb{R}$. From the construction, we may assume that all sets in $\mathcal{A}_{n,m}$ are contained in an interval of length

$$L_{n,m} := 10 \cdot 2^{-m}(n+1).$$

Enumerate the pairs $(n, m) \in \mathbb{N} \times \mathbb{Z}$ as

$$(n_1, m_1), (n_2, m_2), \ldots.$$

Define shifts inductively by

$$a_1 := 0, \qquad a_{t+1} := a_t + L_{n_t,m_t} + 1,$$

and set

$$\mathcal{H} := \bigcup_{t=1}^{\infty} \{\mathbf{1}_{F+a_t} : F \in \mathcal{A}_{n_t,m_t}\}.$$

By construction, the supports of the shifted families are pairwise disjoint. Since within each block the sets are nonempty and pairwise disjoint, it follows that all supports of hypotheses in $\mathcal{H}$ are nonempty and pairwise disjoint. Hence

$$\mathbb{VC}(\mathcal{H}) = 1.$$

Now fix $s > 0$, and choose $m \in \mathbb{Z}$ such that

$$2^{-m-1} \leq s \leq 2^{-m}.$$

For every $n \geq 1$, there exists some block corresponding to the pair (n, m). By Lemma C.1, the strategic transformation of that block by $\mathcal{N}_s$ shatters an n-point set. Therefore

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}_s}) \geq n$$

for all n, which implies

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}_s}) = \infty.$$

□

Lemma C.3 (Pathology with interval neighborhoods)**.** *There exist a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathbb{R}}$ and a neighborhood system $\mathcal{N} = \{N_x : x \in \mathbb{R}\}$ such that*

$$\mathbb{VC}(\mathcal{H}) = 1, \qquad \mathbb{VC}(\{N_x : x \in \mathbb{R}\}) = 1,$$

but

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = \infty.$$

Proof. Partition $\mathbb{R}$ into disjoint intervals

$$I_j := [j, j+1), \qquad j \in \mathbb{Z},$$

and define

$$N_x := I_j \quad \text{whenever } x \in I_j.$$

In other words, $N_x = [\lfloor x \rfloor, \lfloor x \rfloor + 1)$. Thus every point has a neighborhood which is an interval containing it, and the family $\{N_x : x \in \mathbb{R}\}$ consists of pairwise disjoint sets, hence

$$\mathbb{VC}(\{N_x : x \in \mathbb{R}\}) = 1.$$

Now fix points

$$p_i := i + \tfrac{1}{2} \in I_i, \qquad i \geq 1.$$

Choose an injective map

$$\alpha\colon 2^{\mathbb{N}} \to (0, 1/10).$$

For example, identifying $2^{\mathbb{N}}$ with $\{0,1\}^{\mathbb{N}}$, we can take $\alpha(x_1x_2\ldots) = 0.01x_1x_2\ldots$, where the right-hand side is interpreted as a decimal expansion.

For each subset $S \subseteq \mathbb{N}$, define

$$F_S := \{\, i + \tfrac{1}{2} + \alpha(S) : i \in S \,\} \cup \{\, -i - \alpha(S) : i \notin S \,\}.$$

Let

$$h_S := \mathbf{1}_{F_S}, \qquad \mathcal{H} := \{h_S : S \subseteq \mathbb{N}\}.$$

VC dimension of $\mathcal{H}$. The sets F_S are nonempty and pairwise disjoint: different $S \neq T$ have different shifts $\alpha(S) \neq \alpha(T)$, so no point can belong to both F_S and F_T. Hence the supports of the hypotheses in $\mathcal{H}$ are pairwise disjoint. Therefore $\mathcal{H}$ shatters a singleton but no two-point set, and thus

$$\mathbb{VC}(\mathcal{H}) = 1.$$

VC dimension of $\{N_x : x \in \mathbb{R}\}$. By definition, $\{N_x : x \in \mathbb{R}\} = \{I_j : j \in \mathbb{Z}\}$. Since the sets I_j are disjoint,

$$\mathbb{VC}(\{N_x : x \in \mathbb{R}\}) = 1.$$

Effect of the transformation. For $i \geq 1$,

$$h_S^{\mathcal{N}}(p_i) = 1 \iff F_S \cap N_{p_i} \neq \emptyset.$$

Since $N_{p_i} = I_i = [i, i+1)$, the only points of F_S that can lie in this interval are of the form

$$i + \tfrac{1}{2} + \alpha(S),$$

and such a point belongs to F_S if and only if $i \in S$. Thus

$$h_S^{\mathcal{N}}(p_i) = 1 \iff i \in S.$$

Therefore

$$\{\, p_i : h_S^{\mathcal{N}}(p_i) = 1,\ 1 \leq i \leq n \,\} = \{p_i : i \in S \cap [n]\}.$$

Since $S \subseteq \mathbb{N}$ was arbitrary, the class $\mathcal{H}^{\mathcal{N}}$ shatters $\{p_1, \ldots, p_n\}$ for every n, and hence

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = \infty. \qquad \square$$

D The role of definability

The examples in the previous section are constructive, but they may look somewhat artificial: in each case, the hypotheses are chosen in a very specific way so that, after thickening by a neighborhood relation, the transformed class shatters arbitrarily large finite sets.

In light of Main Theorems 3 and 4, one might suspect that such behavior is ruled out whenever the original class and the neighborhood relation are definable in a reasonably tame expansion of the real field. The following example shows that this is not true in general. Already in the expansion

$$(\mathbb{R}, +, \cdot, >, =, \{\cdot\}),$$

where $\{x\} := x - \lfloor x \rfloor$ denotes the fractional part, a definable class of VC dimension 1, together with a definable neighborhood system of VC dimension 1, can give rise to a strategic class of infinite VC dimension.

Lemma D.1. *Fix $0 < r < 1/2$. Consider the definable family $\mathcal{H}$ given by*

$$\Phi_{\mathcal{H}}(x, a) := \exists y \left(\{y\} = 0 \wedge \{a\} = 0 \wedge y \neq 0 \wedge x = y + \{\sqrt{2}\, ay\} \right),$$

and let $\mathcal{N}_r$ be the neighborhood relation defined by

$$\Phi_{\mathcal{N}_r}(x, y) := |y - x| \leq r.$$

Then the strategically transformed family $\mathcal{H}^{\mathcal{N}_r}$ has infinite VC dimension. The same construction works with $\sqrt{2}$ replaced by any fixed irrational number.

Proof. For each parameter a, let $H_a := \{x \in \mathbb{R} : \Phi_{\mathcal{H}}(x, a)\}$ denote the positive region of the corresponding hypothesis. Write $\alpha = \sqrt{2}$. The idea is simple: the set H_a contains one point in each integer cell, namely $y + \{\alpha a y\}$. By choosing the integer parameter a, we can force these fractional parts to fall either close to $1/2$, so that the midpoint $y + 1/2$ is hit after r-thickening, or far from $1/2$, so that it is not hit.

For integer a, the set defined by $\Phi_{\mathcal{H}}(x, a)$ is

$$H_a = \{\, y + \{\alpha a y\} : y \in \mathbb{Z} \setminus \{0\} \,\},$$

and for non-integer a it is empty.

First note that the nonempty sets H_a are pairwise disjoint. Indeed, if $x \in H_{a_1} \cap H_{a_2}$, then

$$x = y_1 + \{\alpha a_1 y_1\} = y_2 + \{\alpha a_2 y_2\}$$

for some $y_1, y_2 \in \mathbb{Z} \setminus \{0\}$. Since the two sides lie in the half-open intervals $[y_1, y_1 + 1)$ and $[y_2, y_2 + 1)$, we have $y_1 = y_2 =: y$. Hence

$$\{\alpha a_1 y\} = \{\alpha a_2 y\},$$

so $\alpha y(a_1 - a_2) \in \mathbb{Z}$. Since α is irrational and $y \neq 0$, this implies $a_1 = a_2$. Thus the original family has $\mathbb{VC}(\mathcal{H}) = 1$.

We now show that the thickened family has infinite VC dimension. Let

$$P := (1/2 - r, 1/2 + r), \qquad Q := (0, 1/2 - r).$$

Both intervals are nonempty. We claim that for every n there are positive integers $b_1, \ldots, b_n$ such that for every $A \subseteq [n]$ there is a nonempty interval $I_A \subseteq [0, 1)$ satisfying

$$\{b_i t\} \in P \quad \text{if } i \in A, \qquad \{b_i t\} \in Q \quad \text{if } i \notin A$$

for all $t \in I_A$.

This is an elementary induction. Suppose $b_1, \ldots, b_k$ and the intervals I_A, $A \subseteq [k]$, have already been constructed, and put

$$v := \min_{A \subseteq [k]} |I_A| > 0.$$

Choose b_{k+1} so large that $b_{k+1} > 2/v$. Then each interval I_A contains a full interval of the form

$$[j/b_{k+1}, (j + 1)/b_{k+1}).$$

Inside such a full interval, the map $t \mapsto \{b_{k+1} t\}$ runs once through $[0, 1)$. Therefore both intersections

$$I_A \cap \{t : \{b_{k+1} t\} \in P\}, \qquad I_A \cap \{t : \{b_{k+1} t\} \in Q\}$$

contain nonempty intervals. Use the first one to define $I_{A \cup \{k+1\}}$, and the second one to redefine I_A. This completes the induction.

Now fix $A \subseteq [n]$. Since α is irrational, the sequence $\{\alpha m\}$, $m \in \mathbb{Z}$, is dense in $[0, 1)$. Choose $m_A \in \mathbb{Z} \setminus \{0\}$ such that $\{\alpha m_A\} \in I_A$. Then, since each b_i is an integer, $\{\alpha m_A b_i\} = \{b_i \{\alpha m_A\}\}$. Hence

$$\{\alpha m_A b_i\} \in P \quad \text{if } i \in A, \qquad \{\alpha m_A b_i\} \in Q \quad \text{if } i \notin A.$$

We claim that the points

$$x_i := b_i + \frac{1}{2}, \qquad i = 1, \ldots, n,$$

are shattered. For the parameter m_A, the corresponding point of H_{m_A} in the b_i-th integer cell is

$$z_i := b_i + \{\alpha m_A b_i\}.$$

If $i \in A$, then $\{\alpha m_A b_i\} \in P$, so $|x_i - z_i| < r$, and therefore $x_i \in H_{m_A}^{\mathcal{N}_r}$.

If $i \notin A$, then $\{\alpha m_A b_i\} \in Q$, so the point z_i in the same integer cell is at distance $> r$ from x_i. Points of H_{m_A} coming from other integer cells are at distance $> 1/2 > r$ from x_i. Hence $x_i \notin H_{m_A}^{\mathcal{N}_r}$. Thus

$$x_i \in H_{m_A}^{\mathcal{N}_r} \quad \iff \quad i \in A.$$

Since every $A \subseteq [n]$ is realized, the set $\{x_1, \ldots, x_n\}$ is shattered. As n was arbitrary,

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}_r}) = +\infty.$$

The proof used only the irrationality of $\sqrt{2}$, so the same argument works with any fixed irrational number. □

Remark D.2 (Defining $\mathbb{Z}$ is essentially sufficient). The construction above relies only on two ingredients:

1. the ability to restrict variables to integers, and
2. the presence of a fixed irrational constant α.

Whenever the set $\mathbb{Z}$ is definable, the fractional-part function $\{x\}$ is also definable in the structure, since

$$\{x\} = x' \iff \exists y\,(\theta(y) \wedge x' = x - y \wedge 0 \le x' < 1),$$

and the proof goes through unchanged.

In particular, the role of the predicates $\{y\} = 0$ and $\{a\} = 0$ in the definition of $\Phi_{\mathcal{H}}$ is only to enforce $y, a \in \mathbb{Z}$.

Thus, the phenomenon of VC-dimension blow-up after thickening already appears in any expansion of the real field in which $\mathbb{Z}$ is definable.

E Proof of Main Theorem 2

Theorem 2. *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}^l$ definable in $\mathbb{R}_{\exp}$ with k parameters, and let $\mathcal{N}$ be a neighborhood system definable in $\mathbb{R}_{\exp}$. Then the strategic class $\mathcal{H}^{\mathcal{N}}$ is PAC learnable.*

Moreover, the ERM sample complexity of $\mathcal{H}^{\mathcal{N}}$ satisfies

$$m^{\mathrm{real}}_{\mathrm{ERM}}(\varepsilon, \delta) = O\Big(\frac{k\log(1/\varepsilon) + \log(1/\delta) + K_{\mathcal{H},\mathcal{N}}}{\varepsilon}\Big),$$

where $K_{\mathcal{H},\mathcal{N}}$ is a constant depending on $\mathcal{H}$ and $\mathcal{N}$.

Proof. The strategic class $\mathcal{H}^{\mathcal{N}}$ is again definable in $\mathbb{R}_{\exp}$. Indeed, suppose $\mathcal{H}$ is defined by $\Phi_{\mathcal{H}}(y, a)$, with parameter $a \in \mathbb{R}^k$, so that

$$h_a = \mathbf{1}\big\{y \in \mathbb{R}^l : \Phi_{\mathcal{H}}(y, a)\big\}.$$

Then

$$h^{\mathcal{N}}_a(x) = 1 \iff \exists y\big(\Phi_{\mathcal{N}}(x, y) \wedge \Phi_{\mathcal{H}}(y, a)\big).$$

Thus $\mathcal{H}^{\mathcal{N}}$ is defined by the single formula

$$\psi(x, a) := \exists y\big(\Phi_{\mathcal{N}}(x, y) \wedge \Phi_{\mathcal{H}}(y, a)\big),$$

with the same parameter vector $a \in \mathbb{R}^k$.

Since $\mathbb{R}_{\exp}$ is o-minimal [Wil96], the polynomial growth theorem for definable families in o-minimal structures [JL10, Corollary 4.6] implies that there exists a constant $C_\psi \ge 1$, depending only on the defining formula ψ, such that

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \le C_\psi m^k$$

for every $m \ge 1$. Write

$$K := \log C_\psi.$$

Applying Lemma 2.2 to $\mathcal{H}^{\mathcal{N}}$, we obtain

$$m^{\mathrm{real}}_{\mathrm{ERM}}(\varepsilon, \delta) = O\left(\frac{k\log(k/\varepsilon) + \log C_\psi + \log(1/\delta)}{\varepsilon}\right).$$

Since k is fixed for the class, the term $k\log k$ is absorbed into the class-dependent constant. Hence

$$m^{\mathrm{real}}_{\mathrm{ERM}}(\varepsilon, \delta) = O\left(\frac{k\log(1/\varepsilon) + K + \log(1/\delta)}{\varepsilon}\right).$$

Finally, since the growth function of $\mathcal{H}^{\mathcal{N}}$ is polynomially bounded, $\mathcal{H}^{\mathcal{N}}$ has finite VC dimension. Therefore $\mathcal{H}^{\mathcal{N}}$ is PAC learnable. □

F Proof of Theorem 3

Proposition F.1 (Growth bound for definable strategic classes over $\mathbb{R}$)**.** *Let $\mathcal{H} = \{h_a : a \in \mathbb{R}^k\}$ be a hypothesis class on $\mathbb{R}^l$, and let $\mathcal{N}$ be a neighborhood system on $\mathbb{R}^l$. Suppose that $h_a(x) = 1$ and the predicate $\Phi_{\mathcal{N}}(x, y)$ defining $y \in N_x$ are definable in the real field $\mathbb{R}$ by quantifier-free formulas of* total description complexity *at most D (that is, the formulas involve at most s polynomials, each of degree at most D', with $D = sD'$).*

Then, for every $m \geq 1$,

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \leq (Cm)^k D^{O(k^2 l)}.$$

Proof. Fix points $y_1, \ldots, y_m \in \mathbb{R}^l$. For $a \in \mathbb{R}^k$, define

$$\varphi(y, a) \iff \exists x \in \mathbb{R}^l \big(h_a(x) = 1 \wedge \Phi_{\mathcal{N}}(y, x)\big).$$

Then

$$h_a^{\mathcal{N}}(y) = 1 \iff \varphi(y, a).$$

For each fixed $i \in [m]$, the condition $h_a^{\mathcal{N}}(y_i) = 1$ is equivalent to

$$\exists x \in \mathbb{R}^l \big(h_a(x) = 1 \wedge \Phi_{\mathcal{N}}(y_i, x)\big).$$

This is an existential formula with quantified variables $x \in \mathbb{R}^l$ and free variables $a \in \mathbb{R}^k$. By assumption, its quantifier-free part involves at most s polynomial sign conditions of degree at most D'.

By Corollary B.8, for each fixed y_i this formula is equivalent to a quantifier-free formula in a, whose truth value is determined by the sign pattern of at most

$$L \leq s^{O(lk)} D'^{O(lk)} = D^{O(lk)}$$

polynomials in a, each of degree at most

$$\Delta \leq D'^{O(l)} \leq D^{O(l)}.$$

Thus, for each i, there is a family

$$\mathcal{P}_i \subseteq \mathbb{R}[a], \qquad |\mathcal{P}_i| \leq D^{O(lk)},$$

such that the truth value of $h_a^{\mathcal{N}}(y_i) = 1$ is determined by the sign pattern of $\mathcal{P}_i$ at a.

Let

$$\mathcal{P} = \bigcup_{i=1}^{m} \mathcal{P}_i.$$

Then

$$M := |\mathcal{P}| \leq m\, D^{O(lk)},$$

and every polynomial in $\mathcal{P}$ has degree at most

$$\Delta \leq D^{O(l)}.$$

For every $a \in \mathbb{R}^k$, the labeling of $\{y_1, \ldots, y_m\}$ by $h_a^{\mathcal{N}}$ is determined by the sign pattern of the polynomials in $\mathcal{P}$ at a. Hence the number of traces is at most the number of sign patterns of M polynomials of degree at most Δ in k variables.

By the standard bound on sign patterns of real polynomials (see, e.g., Warren [War68] or [Alo95, Proposition 5.5]), the number of such sign patterns is at most

$$\left(\frac{C\, M \Delta}{k}\right)^k.$$

Substituting the bounds on M and Δ gives

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \leq \left(C\, D^{O(l)}\, m\, D^{O(lk)}\right)^k.$$

Therefore

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \leq (Cm)^k D^{O(k^2 l)}.$$

Since the sample was arbitrary, the proof is complete. □

Remark F.2 (Relation to Goldberg–Jerrum). The proof above follows the proof of Theorem B.6 of Goldberg and Jerrum: once membership is determined by the sign pattern of finitely many polynomials in the parameter variables, one obtains bounds on the growth function and VC dimension via bounds on the number of realizable sign patterns.

In the quantifier-free setting, if membership $x \in F_a$ is determined by the sign pattern of at most s polynomials in $a \in \mathbb{R}^k$, each of degree at most D, then the Goldberg–Jerrum bound yields

$$\mathbb{VC}(\mathcal{F}) = O\big(k \log(sD)\big).$$

In our setting, the strategic transformation introduces an existential quantifier. Applying quantitative one-block quantifier elimination reduces the problem to the quantifier-free regime, but increases the number of defining polynomials: the condition $y \in C_a^{\mathcal{N}}$ becomes determined by the sign pattern of at most

$$s^{O(kl)} D^{O(kl)}$$

polynomials in the parameter variables a.

Theorem 3. *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}^l$ definable with k parameters in the language of $\mathbb{R}$, and let $\mathcal{N}$ be a neighborhood system definable in the same language. Suppose that both $\mathcal{H}$ and $\mathcal{N}$ are defined by quantifier-free formulas involving polynomial equalities and inequalities whose* total description complexity *is at most D (that is, the formulas involve at most s polynomials, each of degree at most D', with $D = sD'$). Then*

$$m_{\mathrm{ERM}}^{\mathrm{real}}(\varepsilon, \delta) = O\Big(\frac{k \log(1/\varepsilon) + k^2 l \log D + \log(1/\delta)}{\varepsilon}\Big),$$

and

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O\big(k^2 l \log D\big).$$

Proof. By Proposition F.1, for every $m \geq 1$,

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \leq (C_0 m)^k D^{O(k^2 l)}.$$

Equivalently, for some constant $A > 0$,

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \leq C m^k, \qquad C := C_0^k D^{Ak^2 l}.$$

Hence

$$\log C = O(k^2 l \log D).$$

Applying Lemma 2.2 to $\mathcal{H}^{\mathcal{N}}$, we obtain

$$m_{\mathrm{ERM}}^{\mathrm{real}}(\varepsilon, \delta) = O\left(\frac{k \log(k/\varepsilon) + k^2 l \log D + \log(1/\delta)}{\varepsilon}\right).$$

Absorbing the lower-order $k \log k$ term into the displayed bound gives

$$m_{\mathrm{ERM}}^{\mathrm{real}}(\varepsilon, \delta) = O\left(\frac{k \log(1/\varepsilon) + k^2 l \log D + \log(1/\delta)}{\varepsilon}\right).$$

For the VC-dimension bound, applying Lemma I.1 to $\mathcal{H}^{\mathcal{N}}$ gives

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) \leq 2 \log C + 4k \log(4k).$$

Since $\log C = O(k^2 l \log D)$, this yields

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O(k^2 l \log D).$$

□

G Proof of Main Theorem 4

The aim of this section is to prove the following theorem.

Theorem 4. *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}^l$ definable with k parameters in the language of $\mathbb{R}_{\exp}$, and let $\mathcal{N}$ be a neighborhood system definable in the language of $\mathbb{R}_{\exp}$. Suppose, moreover, that both $\mathcal{H}$ and $\mathcal{N}$ are defined using existential formulas of format F and degree D. Then*

$$m_{\mathrm{ERM}}^{\mathrm{real}}(\varepsilon,\delta) = O\Big(\frac{k\log(1/\varepsilon)+\gamma(F)\log(D)+\log(1/\delta)}{\varepsilon}\Big)$$

and

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O_F(\log D),$$

where $\gamma\colon \mathbb{N}\to\mathbb{N}$ is an explicitly computable function.

The proof of Main Theorem 4 uses several notions from the theory of restricted Pfaffian and sub-Pfaffian sets. Since these notions are somewhat abstract and not directly part of the statement of the theorem, we first describe the structure of the appendix.

Organization. In Section G.1, we recall the restricted Pfaffian structure $\mathbb{R}_{\mathrm{rPfaff}}$ and the basic notions of Pfaffian functions, semi-Pfaffian sets, sub-Pfaffian sets, format, and degree. In Section G.2, we explain the refined notions of $*$-format and $*$-degree used in the quantitative cell decomposition results. In Section G.3, we use sharp cell decomposition in $\mathbb{R}_{\mathrm{rPfaff}}$ to prove a growth-function bound for definable hypothesis classes in this structure. In Section G.4, we explain how formulas in $\mathbb{R}_{\exp|_{[-M,M]}}$ can be interpreted in $\mathbb{R}_{\mathrm{rPfaff}}$, with controlled complexity. Finally, in Section G.5, we combine these ingredients to prove Main Theorem 4.

G.1 Restricted Pfaffian structure

Definition G.1 (Pfaffian function)**.** Let $U\subseteq\mathbb{R}^n$ be open. A *Pfaffian chain* on U is a sequence of analytic functions

$$f_1,\dots,f_r\colon U\to\mathbb{R}$$

such that, for every $1\le i\le r$ and $1\le j\le n$, there is a polynomial

$$P_{ij}\in\mathbb{R}[x_1,\dots,x_n,y_1,\dots,y_i]$$

satisfying

$$\frac{\partial f_i}{\partial x_j}(x) = P_{ij}\big(x,f_1(x),\dots,f_i(x)\big).$$

A function $f\colon U\to\mathbb{R}$ is called a *Pfaffian function (with respect to this chain)* if there exists a polynomial Q such that

$$f(x) = Q\big(x,f_1(x),\dots,f_r(x)\big).$$

We say that the Pfaffian function f has *format* $n+r$ and *degree* $\sum_{i,j}\deg P_{ij}+\deg Q$.

In analogy with semialgebraic sets, a set $X\subseteq U$ is called *semi-Pfaffian* if it can be defined by a Boolean combination of atomic formulas of the form

$$g(x)\,\sigma\,0,\qquad \sigma\in\{=,>,<,\ge,\le\},$$

where each $g\colon U\to\mathbb{R}$ is a Pfaffian function.

Suppose that X is defined using Pfaffian functions $g_1,\dots,g_m$, all taken with respect to a common Pfaffian chain $f_1,\dots,f_r$. Then we say that this representation has *format* $n+r$, and its *degree* is the sum of the degrees of the functions g_i appearing in the atomic formulas.

If the functions $g_1,\dots,g_m$ are originally given with respect to different Pfaffian chains, we may pass to a single Pfaffian chain by concatenation. In this way, the notion of format and degree is well-defined up to an additive increase in the format, while the degree remains bounded by the total degree of the defining data.

A *restricted Pfaffian function* on the unit cube I^n, where $I = [0,1]$, is the restriction to I^n of a Pfaffian function defined on an open neighborhood of I^n. We denote by $\mathbb{R}_{\mathrm{rPfaff}}$ the expansion of the real field by all such restricted Pfaffian functions, in all arities, viewed as total functions on $\mathbb{R}^n$ by extending them by zero outside I^n.

Throughout this section, we call a set *sub-Pfaffian* if it is definable in $\mathbb{R}_{\mathrm{rPfaff}}$. By Wilkie's theorem of the complement, $\mathbb{R}_{\mathrm{rPfaff}}$ is o-minimal [Wil99]. Moreover, every sub-Pfaffian set $X \subseteq \mathbb{R}^n$ admits a representation of the form

$$X = \pi(Y),$$

where $Y \subseteq \mathbb{R}^{n+d}$ is a semi-Pfaffian set and $\pi\colon \mathbb{R}^{n+d} \to \mathbb{R}^n$ is the projection onto the first n coordinates. The *format* and *degree* of this representation are defined to be the format and degree of Y.

Equivalently, every formula $\Phi(x)$ in $\mathbb{R}_{\mathrm{rPfaff}}$ is equivalent to one of the form

$$\Phi(x) \;\equiv\; \exists y \in \mathbb{R}^d \; \theta(x,y),$$

where $\theta(x,y)$ is quantifier-free and is a Boolean combination of atomic predicates

$$g(x,y)\,\sigma\,0, \qquad \sigma \in \{=,>,<,\geq,\leq\},$$

with g restricted Pfaffian functions. Thus, every definable set is Σ_1-definable.

In contrast, in the real field $\mathbb{R}$, every formula is equivalent to a quantifier-free formula (quantifier elimination).

Remark G.2. When referring to the format and degree of a sub-Pfaffian set, we implicitly fix a particular representation as a projection of a semi-Pfaffian set. A given set may admit multiple such representations with different complexities. Since our results only require upper bounds, this choice is immaterial: when we say that a sub-Pfaffian set has format at most F and degree at most D, we mean that it admits some representation with these bounds.

G.2 Sharp cell decomposition

In this section we record a quantitative cell decomposition result for $\mathbb{R}_{\mathrm{rPfaff}}$, which we will use as a black box.

Cylindrical cell decomposition. We use the notion of cylindrical cell and cylindrical cell decomposition from Definition B.5.

We refer to the standard literature for further details; see, e.g., [BPR06, Chapter 5] in the semialgebraic setting and [BV20, Definitions 5–6] for the Pfaffian setting.

Example (Cylindrical decomposition of the unit disk). Consider the unit disk

$$X = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$

A cylindrical cell decomposition of $\mathbb{R}^2$ compatible with X is obtained as follows.

Decomposition of the x-axis.

We first decompose $\mathbb{R}$ into the cells

$$S_1 = (-\infty,-1), \quad S_2 = \{-1\}, \quad S_3 = (-1,1), \quad S_4 = \{1\}, \quad S_5 = (1,\infty).$$

Lifting over each base cell.

(i) Over S_1 and S_5.

If $x \in S_1 \cup S_5$, then $x^2 > 1$, so the inequality

$$x^2 + y^2 \leq 1$$

has no solutions in y. Thus, no defining functions arise, and we obtain a single cylindrical cell:

$$S_i \times \mathbb{R}, \quad i = 1,5.$$

(ii) Over S_2 and S_4.

At $x = \pm 1$, the equation becomes

$$y^2 \leq 0,$$

so the only solution is $y = 0$. This corresponds to a single defining function (constant zero), and we obtain three cells:

$$S_i \times (-\infty, 0), \quad S_i \times \{0\}, \quad S_i \times (0, \infty), \quad i = 2, 4.$$

(iii) Over $S_3 = (-1, 1)$.

For $x \in (-1, 1)$, the boundary of X is given by the two functions

$$\xi_1(x) = -\sqrt{1 - x^2}, \qquad \xi_2(x) = \sqrt{1 - x^2}.$$

These are continuous semialgebraic functions with $\xi_1(x) < \xi_2(x)$. They partition the fiber into five cells:

$$\begin{aligned}
S_{3,1} &= \{(x, y) : -1 < x < 1,\ y < \xi_1(x)\},\\
S_{3,2} &= \{(x, y) : -1 < x < 1,\ y = \xi_1(x)\},\\
S_{3,3} &= \{(x, y) : -1 < x < 1,\ \xi_1(x) < y < \xi_2(x)\},\\
S_{3,4} &= \{(x, y) : -1 < x < 1,\ y = \xi_2(x)\},\\
S_{3,5} &= \{(x, y) : -1 < x < 1,\ \xi_2(x) < y\}.
\end{aligned}$$

The example above illustrates the key feature of cylindrical cell decomposition: a definable set is partitioned into finitely many simple regions, each described by continuous functions over lower-dimensional cells.

In general, for definable sets in o-minimal structures (and in particular in the restricted Pfaffian structure), such decompositions always exist. Moreover, for our purposes, it is crucial that one can control the *number* and *complexity* of the resulting cells in terms of the complexity of the defining formulas.

The quantitative theory of restricted Pfaffian structures provides such control via the notions of $*$-format and $*$-degree for sub-Pfaffian sets, introduced in [BV20]. These notions measure definitional complexity in a way that is stable under the operations appearing in cell decomposition (such as Boolean combinations, projections, and lifting).

We will not recall the full definitions here, and instead rely only on the following comparison with the usual notion of format and degree, together with three key structural results.

Remark G.3 (Comparison with usual format and degree). If a sub-Pfaffian set has format F and degree D, then it has $*$-format F and $*$-degree $\mathrm{Poly}_F(D)$ [BNZ24, Remark 10].

Theorem G.4 (Theorem 2 of [BV20])**.** *Let φ be a formula in $\mathbb{R}_{\mathrm{rPfaff}}$, with $*$-format F and $*$-degree D. Then the set defined by φ has $*$-format $O_F(1)$ and $*$-degree $\mathrm{Poly}_F(D)$.*

Theorem G.5 (Sharp cell decomposition for $\mathbb{R}_{\mathrm{rPfaff}}$)**.** *Let $X_1, \ldots, X_k \subseteq \mathbb{R}^\ell$ be definable sets, each of $*$-format at most F and $*$-degree at most D. Then there exists a cylindrical cell decomposition of $\mathbb{R}^\ell$ compatible with all X_i, consisting of at most $\mathrm{Poly}_F(D, k)$ cells.*

Moreover, each cell is definable with $*$-format $O_F(1)$ and $*$-degree $\mathrm{Poly}_F(D)$.

This result follows from [BV20, Theorem 1] together with [BNZ24, Remark 10].

G.3 Growth function of definable sets in $\mathbb{R}_{\mathrm{rPfaff}}$

Theorem G.6. *Let $\mathcal{H} = \{h_a : a \in \mathbb{R}^k\}$ be a hypothesis class on $\mathbb{R}^l$ definable in k parameters by a formula $\Phi_{\mathcal{H}}(X, Y)$ in $\mathbb{R}_{\mathrm{rPfaff}}$ of format F and degree D, where $X \in \mathbb{R}^l$ and $Y \in \mathbb{R}^k$. Then*

$$\Pi_{\mathcal{H}}(m) \ \leq \ \mathrm{Poly}_F(D) \cdot m^k.$$

Proof. We adapt the argument of Johnson–Laskowski [JL10, Lemma 4.2 and Proposition 4.3] to $\mathbb{R}_{\mathrm{rPfaff}}$. The key difference is that, in place of the i-th boundary point of a fiber $\theta(\bar{a}, \mathbb{R}, \bar{e})$ (which exists by o-minimality but with no control on its complexity), we use the threshold functions produced by sharp cell decomposition. This gives explicit bounds on the format and degree of every formula in the construction.

We write $\lg(\bar{X})$ for the length of a tuple $\bar{X}$, i.e., the number of its coordinates.

From now on, we call $*$-format and $*$-degree just format and degree.

Step 1: shaving one variable. Let $\theta(\bar{X}, Y, \bar{Z})$ be a formula with $\bar{X} \in \mathbb{R}^l$, of format $O_F(1)$ and degree $\mathrm{Poly}_F(D)$. By Theorem G.5, applied to the singleton family defined by $\{(\bar{x}, y, \bar{z}) \mid \theta(\bar{x}, y, \bar{z})\}$, we obtain a cylindrical cell decomposition compatible with θ, with respect to the variable order $\bar{Z}, \bar{X}, Y$, so that Y is the last variable. Let $\mathcal{S}$ denote the resulting decomposition of the $(\lg(\bar{Z}) + l)$-dimensional space obtained by projecting away Y. Then each cell $C \in \mathcal{S}$ has format $O_F(1)$ and degree $\mathrm{Poly}_F(D)$, and $|\mathcal{S}| \leq \mathrm{Poly}_F(D)$.

For each cell $C \in \mathcal{S}$, the cells lying above C in the full decomposition are described by continuous definable threshold functions

$$-\infty = \xi_0^C < \xi_1^C(\bar{Z}, \bar{X}) < \cdots < \xi_{N_C}^C(\bar{Z}, \bar{X}) < \xi_{N_C+1}^C = +\infty$$

on C with $N_C \leq \mathrm{Poly}_F(D)$. Moreover, the graph of each function ξ_j^C is definable with format $O_F(1)$ and degree $\mathrm{Poly}_F(D)$. Equivalently, the formula expressing $y = \xi_j^C(\bar{Z}, \bar{X})$ has format $O_F(1)$ and degree $\mathrm{Poly}_F(D)$. In the formulas below, we use ξ_j^C by a slight abuse of notation, with this graph representation understood.

The cells over C are the graphs $\{Y = \xi_j^C\}$ $(1 \leq j \leq N_C)$ and the open strips $\{\xi_j^C < Y < \xi_{j+1}^C\}$ $(0 \leq j \leq N_C)$. Since the decomposition is compatible with θ, for any $(\bar{z}, \bar{x}) \in C$, the truth value of $\theta(\bar{x}, Y, \bar{z})$ is invariant on each open strip, and is also constant on each graph $Y = \xi_j^C(\bar{z}, \bar{x})$.

In analogy with Johnson–Laskowski, define the following formulas (in free variables $\bar{X}, \bar{W}, \bar{Z}$, where $\bar{W}$ plays the role of a parameter from A):

$$\psi_{j,1}^C(\bar{X}, \bar{W}, \bar{Z}) := (\bar{Z}, \bar{W}) \in C \wedge \theta(\bar{X}, \xi_j^C(\bar{Z}, \bar{W}), \bar{Z}),$$
$$\psi_{j,2}^C(\bar{X}, \bar{W}, \bar{Z}) := (\bar{Z}, \bar{W}) \in C \wedge \exists \varepsilon > 0 \, \forall U \left(\xi_j^C(\bar{Z}, \bar{W}) - \varepsilon < U < \xi_j^C(\bar{Z}, \bar{W}) \to \theta(\bar{X}, U, \bar{Z})\right),$$
$$\psi_{j,3}^C(\bar{X}, \bar{W}, \bar{Z}) := (\bar{Z}, \bar{W}) \in C \wedge \exists \varepsilon > 0 \, \forall U \left(\xi_j^C(\bar{Z}, \bar{W}) < U < \xi_j^C(\bar{Z}, \bar{W}) + \varepsilon \to \theta(\bar{X}, U, \bar{Z})\right),$$
$$\psi^*(\bar{X}, \bar{W}, \bar{Z}) := \forall U \, \theta(\bar{X}, U, \bar{Z}).$$

Let $\mathcal{F}_\theta$ be the resulting finite collection of formulas. Each formula in $\mathcal{F}_\theta$ has format $O_F(1)$ and degree $\mathrm{Poly}_F(D)$, and $|\mathcal{F}_\theta| \leq \mathrm{Poly}_F(D)$.

Claim. For every finite $A \subseteq \mathbb{R}^l$, every $c \in \mathbb{R}$, and every $\bar{e} \in \mathbb{R}^{\lg(\bar{Z})}$, there exist $\bar{a} \in A$ and $\psi \in \mathcal{F}_\theta$ such that

$$\{\bar{x} \in A : \theta(\bar{x}, c, \bar{e})\} = \{\bar{x} \in A : \psi(\bar{x}, \bar{a}, \bar{e})\}.$$

For each $\bar{a} \in A$, the pair $(\bar{e}, \bar{a})$ lies in a unique cell $C(\bar{e}, \bar{a}) \in \mathcal{S}$. Let

$$\Delta = \bigcup_{\bar{a} \in A} \{\xi_j^{C(\bar{e}, \bar{a})}(\bar{e}, \bar{a}) : 1 \leq j \leq N_{C(\bar{e}, \bar{a})}\},$$

a finite subset of $\mathbb{R}$. There are four cases:

Case 1: $c \in \Delta$. Then $c = \xi_j^C(\bar{e}, \bar{a})$ for some $\bar{a} \in A, C \in \mathcal{S}$ and some j, and

$$\{\bar{x} \in A : \theta(\bar{x}, c, \bar{e})\} = \{\bar{x} \in A : \psi_{j,1}^C(\bar{x}, \bar{a}, \bar{e})\}.$$

Case 2: $c \notin \Delta$, but $c < d$ for some $d \in \Delta$. Let $d^* \in \Delta$ be the least element greater than c, say $d^* = \xi_j^C(\bar{e}, \bar{a})$ for some $\bar{a} \in A, C \in \mathcal{S}$ and j.

Since d^* is the least boundary value of Δ above c, there exists $\varepsilon > 0$ such that

$$(d^* - \varepsilon, d^*) \subseteq (c, d^*)$$

and this interval contains no element of Δ.

Therefore, for every $\bar{x} \in A$, the truth of $\theta(\bar{x}, y, \bar{e})$ is invariant for $y \in (d^* - \varepsilon, d^*)$, and agrees with its value at c. Hence

$$\{\bar{x} \in A : \theta(\bar{x}, c, \bar{e})\} = \{\bar{x} \in A : \psi^C_{j,2}(\bar{x}, \bar{a}, \bar{e})\}.$$

Case 3: $\Delta \neq \emptyset$ but $c > d$ for every $d \in \Delta$. Let $d^* = \max \Delta$, say $d^* = \xi^C_j(\bar{e}, \bar{a})$ for some $\bar{a} \in A, C \in \mathcal{S}$.

Since there are no elements of Δ greater than d^*, there exists $\varepsilon > 0$ such that

$$(d^*, d^* + \varepsilon)$$

contains no element of Δ.

Therefore, for every $\bar{x} \in A$, the truth of $\theta(\bar{x}, y, \bar{e})$ is invariant for $y \in (d^*, d^* + \varepsilon)$, and agrees with its value at c. Hence

$$\{\bar{x} \in A : \theta(\bar{x}, c, \bar{e})\} = \{\bar{x} \in A : \psi^C_{j,3}(\bar{x}, \bar{a}, \bar{e})\}.$$

Case 4: $\Delta = \emptyset$. The truth of $\theta(\bar{x}, y, \bar{e})$ is invariant on all of $\mathbb{R}$ for every $\bar{x} \in A$, and

$$\{\bar{x} \in A : \theta(\bar{x}, c, \bar{e})\} = \{\bar{x} \in A : \psi^*(\bar{x}, \bar{a}, \bar{e})\}$$

for any choice of $\bar{a} \in A$.

Step 2: iterating. Let $\Phi_{\mathcal{H}}(\bar{X}, \bar{Y})$ with $\bar{X} \in \mathbb{R}^l$ and $\bar{Y} \in \mathbb{R}^k$ be the defining formula of $\mathcal{H}$. Applying Step 1 to $\Phi_{\mathcal{H}}$ shaves off the last coordinate Y_k, producing $\mathrm{Poly}_F(D)$ formulas in variables $\bar{X}, \bar{W}_k, Y_1, \ldots, Y_{k-1}$, each of format $O_F(1)$ and degree $\mathrm{Poly}_F(D)$. Iterating this construction k times, peeling off one coordinate of $\bar{Y}$ at each step, we obtain a finite family $\mathcal{F}_{\Phi_{\mathcal{H}}}$ of formulas $\psi(\bar{X}, \bar{W}_1, \ldots, \bar{W}_k)$ such that for every finite $A \subseteq \mathbb{R}^l$ and every $\bar{c} \in \mathbb{R}^k$, there exist $(\bar{a}_1, \ldots, \bar{a}_k) \in A^k$ and $\psi \in \mathcal{F}_{\Phi_{\mathcal{H}}}$ with

$$\{\bar{x} \in A : \Phi_{\mathcal{H}}(\bar{x}, \bar{c})\} = \{\bar{x} \in A : \psi(\bar{x}, \bar{a}_1, \ldots, \bar{a}_k)\}.$$

At each step the format remains $O_F(1)$ and the number of formulas is multiplied by $\mathrm{Poly}_F(D)$, so $|\mathcal{F}_{\Phi_{\mathcal{H}}}| \leq \mathrm{Poly}_F(D)^k = \mathrm{Poly}_F(D)$ (absorbing $k \leq F$ into the polynomial).

Step 3: counting traces. Every trace of $\mathcal{H}$ on A is represented by some tuple

$$(\psi, \bar{a}_1, \ldots, \bar{a}_k) \in \mathcal{F}_{\Phi_{\mathcal{H}}} \times A^k.$$

Therefore the number of traces is at most

$$|\mathcal{F}_{\Phi_{\mathcal{H}}}| \cdot |A|^k \leq \mathrm{Poly}_F(D) \cdot |A|^k. \qquad \square$$

G.4 The embedding of $\mathbb{R}_{\exp|_{[-M,M]}}$ into $\mathbb{R}_{\mathrm{rPfaff}}$

The choice of the unit cube is only a normalization. If a restricted Pfaffian function is given on another compact box $B \subseteq \mathbb{R}^n$, then after an affine change of variables $B \simeq I^n$, it can be viewed as a restricted Pfaffian function on I^n. Thus allowing compact boxes instead of I^n does not change the resulting structure.

For example, the restricted exponential structure $\mathbb{R}_{\exp|_{[-M,M]}}$ is embedded in $\mathbb{R}_{\mathrm{rPfaff}}$. Indeed, under the affine change of variables

$$x = 2Mt - M, \qquad t \in [0, 1],$$

we replace $\exp(x)$ by the normalized function

$$f_M(t) := \frac{\exp(2Mt - M)}{\exp M},$$

which takes values in $[0, 1]$. Then $f_M \colon [0, 1] \to [0, 1]$ extends to a Pfaffian function on a neighborhood of $[0, 1]$, since it satisfies the differential equation

$$f'_M(t) = 2M\, f_M(t).$$

Thus f_M is a Pfaffian function of order 1, defined by a polynomial relation of degree 1, and in particular contributes $O(1)$ to the $*$-format and $*$-degree (uniformly in M).

Moreover, the original restricted exponential can be recovered via

$$\exp(x) = \exp(M)\, f_M(t), \qquad x = 2Mt - M.$$

Therefore every formula in $\mathbb{R}_{\exp|_{[-M,M]}}$ can be interpreted as a formula in $\mathbb{R}_{\mathrm{rPfaff}}$, after applying the affine rescaling to the input variables and the corresponding rescaling to the function values.

Syntactic complexity of formulas. For our purposes, it will be more convenient to work with a notion of complexity defined directly at the level of formulas, rather than via Pfaffian chains. We therefore introduce a syntactic definition of format and degree for existential formulas in $\mathbb{R}_{\exp|_{[-M,M]}}$, which will serve as the primary measure of

complexity in what follows.

Definition G.7 (Format and degree in $\mathbb{R}_{\exp|_{[-M,M]}}$)**.** Let $\Phi(\bar{Y})$ be a formula in $\mathbb{R}_{\exp|_{[-M,M]}}$. We write Φ in *existential normal form*

$$\Phi(\bar{Y}) \equiv \exists \bar{W}\, \theta(\bar{Y}, \bar{W}),$$

where θ is a quantifier-free formula given as a Boolean combination of atomic predicates of the following types:

$$P(\bar{Y}, \bar{W})\ \sigma\ 0, \qquad \sigma \in \{=, >, <, \geq, \leq\},$$

and

$$U = \exp|_{[-M,M]}(V),$$

where P is a polynomial.

Let $n = |\bar{Y}|$, let $f = |\bar{W}|$, let $P_1, \ldots, P_s$ be the polynomials appearing in θ, and let r be the number of constraints of the form $U = \exp|_{[-M,M]}(V)$. The *format* and *degree* are defined by

$$F := n + f + r, \qquad D := \sum_{i=1}^{s} \deg(P_i) + r.$$

The pair (F, D) is called the *complexity* of Φ.

Comparison of complexities. We compare the syntactic complexity (F, D) of formulas in $\mathbb{R}_{\exp|_{[-M,M]}}$ (Definition G.7) with the $*$-format and $*$-degree used in the restricted Pfaffian framework. These notions are not identical: the former is a syntactic measure for existential formulas, while the latter records the structure of Pfaffian chains and the degrees of the associated polynomial relations. We only need a one-sided comparison.

Let Φ be a formula in $\mathbb{R}_{\exp|_{[-M,M]}}$ of complexity (F, D). After the normalization described above, each restricted exponential constraint

$$U = \exp|_{[-M,M]}(V)$$

is replaced by a restricted Pfaffian function. If several such constraints appear, their Pfaffian chains can be merged into a single chain, increasing the Pfaffian format by at most $r \leq F$. The polynomial atomic predicates remain unchanged, with total degree bounded by D.

Therefore, after interpreting Φ as a formula in $\mathbb{R}_{\mathrm{rPfaff}}$, the formula has $*$-format $O_F(1)$ and $*$-degree $\mathrm{Poly}_F(D)$, uniformly in M. By Theorem G.4, the set defined by Φ has $*$-format $O_F(1)$ and $*$-degree $\mathrm{Poly}_F(D)$.

Equivalently, every set definable by an existential $\mathbb{R}_{\exp|_{[-M,M]}}$-formula of complexity (F, D) admits a restricted sub-Pfaffian presentation of $*$-format $O_F(1)$ and $*$-degree $\mathrm{Poly}_F(D)$.

G.5 Proof of Main Theorem 4

Recall the notion of format and degree for existential formulas (Definition G.7).

In the case of the full exponential structure $\mathbb{R}_{\exp}$, the definition is identical, except that the exponential constraints are of the form

$$U = \exp(V)$$

instead of $U = \exp|_{[-M,M]}(V)$.

Once again for a tuple X, we write $\lg(X)$ for its length (number of coordinates).

Theorem G.8. *Let $\mathcal{H} = \{h_a : a \in \mathbb{R}^k\}$ be a hypothesis class on $\mathbb{R}^l$, definable in k parameters by an existential formula $\Phi(X, A)$ in $\mathbb{R}_{\exp}$ of format F and degree D, where $\lg(X) = l$ and $\lg(A) = k$. Then*

$$\Pi_{\mathcal{H}}(m) \le \mathrm{Poly}_F(D) \cdot m^k.$$

Proof. Let $\mathcal{H}$ be defined by an existential formula

$$\Phi(X, A) \equiv \exists W\, \theta(X, W, A),$$

where

$$X = (X_1, \dots, X_l), \qquad A = (A_1, \dots, A_k), \qquad W = (W_1, \dots, W_f).$$

Assume that this formula has format and degree (F, D) in the sense of Section 3. We also assume that θ is written in graph form, namely as a Boolean combination of atomic predicates of the following two types:

1. Polynomial equalities and inequalities
$$P(X, W, A)\ \sigma\ 0, \qquad \sigma \in \{=, >, <, \ge, \le\},$$
where P is a real polynomial.
2. Exponential graph constraints
$$U = \exp(V),$$
where U, V are variables among X, W, A.

For $M > 0$, let $\Phi_M(X, A)$ be the formula in $\mathbb{R}_{\exp|_{[-M,M]}}$ obtained from $\Phi(X, A)$ by replacing each occurrence of $\exp$ by $\exp|_{[-M,M]}$. Define

$$\Psi_M(X, A) := \exists W \left[\theta_M(X, W, A) \wedge \bigwedge_{j=1}^{l} |X_j| \le M \wedge \bigwedge_{j=1}^{k} |A_j| \le M \wedge \bigwedge_{j=1}^{f} |W_j| \le M\right],$$

where θ_M is obtained from θ by replacing each occurrence of $\exp$ by $\exp|_{[-M,M]}$.

Each formula $\Phi_M(X, A)$ has $*$-format $O_F(1)$ and $*$-degree $\mathrm{Poly}_F(D)$. Passing from Φ_M to Ψ_M adds only boundedness constraints and does not introduce new Pfaffian structure. Hence, by the inductive definition of $*$-format and $*$-degree, the $*$-format increases by at most a constant, and the $*$-degree increases by at most $O(l + k + f) \le O(F)$.

Therefore, by Theorem G.6,

$$\sup_{M \in \mathbb{N}} \Pi_{\Psi_M}(m) \le \mathrm{Poly}_F(D) \cdot m^k.$$

We now show that this uniform bound implies the desired bound for Φ. Fix arbitrary points

$$x_1, \dots, x_m \in \mathbb{R}^l.$$

For every sign pattern on these points realized by Φ, choose one parameter vector $a \in \mathbb{R}^k$ realizing it. Let $A_0 \subseteq \mathbb{R}^k$ be the finite set of all chosen parameter vectors.

Choose M large enough so that all coordinates of all points x_i and all parameters $a \in A_0$ are bounded in absolute value by M. We now enlarge M if necessary as follows. For every pair (x_i, a) with $a \in A_0$, if $\Phi(x_i, a)$ holds, choose a witness $w \in \mathbb{R}^f$ such that

$$\theta(x_i, w, a)$$

holds, and enlarge M so that all coordinates of this witness are also bounded in absolute value by M. Since there are only finitely many such pairs, this gives a finite value M_m.

We claim that Ψ_{M_m} realizes on $x_1, \dots, x_m$ every sign pattern realized by Φ. Indeed, fix $a \in A_0$ and x_i.

If $\Phi(x_i, a)$ holds, then by construction there is a witness $w \in [-M_m, M_m]^f$ such that $\theta(x_i, w, a)$ holds. Since all variables appearing in the exponential graph constraints are bounded by M_m, the restricted exponential agrees with the ordinary exponential on this witness. Thus $\Psi_{M_m}(x_i, a)$ also holds.

Conversely, if $\Psi_{M_m}(x_i, a)$ holds, then there is a witness $w \in [-M_m, M_m]^f$ such that the restricted formula holds. Again, since all variables appearing in the exponential graph constraints are bounded by M_m, the restricted exponential agrees with the ordinary exponential. Hence $\Phi(x_i, a)$ holds.

Therefore Φ and Ψ_{M_m} realize the same labels on $x_1, \dots, x_m$ for every $a \in A_0$. In particular,

$$\Pi_\Phi(m) \le \Pi_{\Psi_{M_m}}(m).$$

Using the uniform bound on $\Pi_{\Psi_M}(m)$, we obtain

$$\Pi_\Phi(m) \le \mathrm{Poly}_F(D) \cdot m^k.$$

This proves the theorem. □

We are now ready to finish the proof of Main Theorem 4.

Theorem 4. *Let $\mathcal{H}$ be a hypothesis class over $\mathbb{R}^l$ definable with k parameters in the language of $\mathbb{R}_{\exp}$, and let $\mathcal{N}$ be a neighborhood system definable in the language of $\mathbb{R}_{\exp}$. Suppose, moreover, that both $\mathcal{H}$ and $\mathcal{N}$ are defined using existential formulas of format F and degree D. Then there is a universal function $\gamma\colon \mathbb{N} \to \mathbb{N}$, computable explicitly, such that*

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \le D^{\gamma(F)} m^k.$$

Consequently,

$$m^{\mathrm{real}}_{\mathrm{ERM}}(\varepsilon, \delta) = O\Big(\frac{k \log(1/\varepsilon) + \gamma(F) \log(D) + \log(1/\delta)}{\varepsilon}\Big)$$

and

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O_F(\log D).$$

Proof. Let $\Phi_{\mathcal{H}}(Y, A)$ be an existential formula defining the hypothesis class $\mathcal{H} = \{h_a : a \in \mathbb{R}^k\}$, so that

$$h_a(Y) = 1 \quad \Longleftrightarrow \quad \Phi_{\mathcal{H}}(Y, a),$$

and let $\Phi_{\mathcal{N}}(X, Y)$ be an existential formula defining the neighborhood relation $\mathcal{N}$. Then the transformed class $\mathcal{H}^{\mathcal{N}}$ is defined by

$$\Phi_{\mathcal{H}^{\mathcal{N}}}(X, A) := \exists Y \, \big(\Phi_{\mathcal{N}}(X, Y) \wedge \Phi_{\mathcal{H}}(Y, A)\big).$$

Indeed, $h^{\mathcal{N}}_a(X) = 1$ holds precisely when there is some Y reachable from X through $\mathcal{N}$ such that $h_a(Y) = 1$.

Since both $\Phi_{\mathcal{H}}$ and $\Phi_{\mathcal{N}}$ have complexity (F, D), the formula $\Phi_{\mathcal{H}^{\mathcal{N}}}$ is again existential, with format at most $2F$ and degree at most $2D$. Here we only add one extra existential block and conjoin the two quantifier-free parts; the total number of variables and the total polynomial degree are bounded by the sum of the two original bounds.

Applying Theorem G.8 to the hypothesis class $\mathcal{H}^{\mathcal{N}}$, whose parameter space still has dimension k, gives

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \le \mathrm{Poly}_F(D)\, m^k.$$

Since the polynomial $\mathrm{Poly}_F(D)$ has degree and coefficients depending only on F, after increasing $\gamma(F)$ we may write

$$\mathrm{Poly}_F(D) \le D^{\gamma(F)}$$

with the convention that $D \geq 2$. Thus

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \leq D^{\gamma(F)} m^k.$$

The sample-complexity bound now follows directly from Lemma 2.2:

$$m_{\mathrm{ERM}}^{\mathrm{real}}(\varepsilon, \delta) = O\Big(\frac{k \log(1/\varepsilon) + \gamma(F) \log D + \log(1/\delta)}{\varepsilon}\Big).$$

Finally, the VC-dimension bound follows from Lemma I.1. Indeed, we have shown that

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \leq D^{\gamma(F)} m^k$$

for every $m \geq 1$. Applying Lemma I.1 with

$$C = D^{\gamma(F)}$$

gives

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) \leq 2\gamma(F) \log D + 4k \log(4k).$$

Since $k \leq F$, the second term is bounded by a constant depending only on F. Hence

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O_F(\log D).$$ □

H Beyond $\mathbb{R}_{\exp}$

The proof of Main Theorem 4, given in Appendix G, does not rely on any specific properties of the exponential function. The only property used is that, after restricting all variables to a bounded domain and applying an affine rescaling, the functions appearing in the defining formulas become *restricted Pfaffian functions* with controlled complexity. In particular, it suffices that every function in the language is *Pfaffian on the whole real line*.

Let

$$\mathbb{R}_g := (\mathbb{R}, +, \cdot, 0, 1, <, g),$$

where $g \colon \mathbb{R} \to \mathbb{R}$ is a Pfaffian function on the whole real line. Assume that g has Pfaffian format $F_\circ$ and Pfaffian degree $D_\circ$, in the sense of Definition G.1.

We assign format and degree to existential formulas in the language of $\mathbb{R}_g$ as follows. This definition is very similar to the corresponding definition for $\mathbb{R}_{\exp}$, given in Definition G.7; the only difference is that we must also account for the format and degree of the function g itself. Let

$$\Phi(X) \equiv \exists Y \, \theta(X, Y)$$

be an existential formula, where θ is quantifier-free and is written as a Boolean combination of atomic predicates of the forms $P(X, Y) \, \sigma \, 0$, with $\sigma \in \{=, >, <, \geq, \leq\}$, and $z = g(w)$. Here P is a polynomial in the variables X, Y, and in the atomic formulas of the form $z = g(w)$, both z and w are variables drawn from $X \cup Y$.

The *format* of Φ is defined as

$$|X| + |Y| + rF_\circ,$$

where r is the number of atomic predicates of the form $z = g(w)$. The *degree* of Φ is defined as

$$\sum_i \deg(P_i) + rD_\circ,$$

where the sum is over all polynomials P_i appearing in polynomial atomic predicates of θ.

Theorem H.1. *Let $\mathcal{H} = \{h_a : a \in \mathbb{R}^k\}$ be a hypothesis class on $\mathbb{R}^\ell$, and let $\mathcal{N}$ be a neighborhood system on $\mathbb{R}^\ell$. Suppose that both $\mathcal{H}$ and $\mathcal{N}$ are definable in $\mathbb{R}_g$ by existential formulas of format at most F and degree at most D. Then*

$$m_{\mathrm{ERM}}^{\mathrm{real}}(\varepsilon, \delta) = O\Big(\frac{k \log(1/\varepsilon) + \gamma(F) \log D + \log(1/\delta)}{\varepsilon}\Big),$$

and

$$\mathbb{VC}(\mathcal{H}^{\mathcal{N}}) = O_F(\log D).$$

Here $\gamma \colon \mathbb{N} \to \mathbb{N}$ is the same explicitly computable function as in Main Theorem 4.

Proof. The proof is the same as the proof of Main Theorem 4. The only additional point is the comparison between formulas in $\mathbb{R}_g$ and restricted Pfaffian formulas.

Fix $M > 0$. On the interval $[-M, M]$, write

$$x = 2Mt - M, \qquad t \in [0, 1].$$

The restriction of g to $[-M, M]$ is represented on the unit interval by the function

$$\widetilde{g}_M(t) := g(2Mt - M).$$

We claim that $\widetilde{g}_M$ has Pfaffian format and degree bounded in terms of $F_\circ$ and $D_\circ$, uniformly in M.

Indeed, suppose that g is defined with respect to a Pfaffian chain

$$f_1, \dots, f_r$$

on $\mathbb{R}$, with

$$\frac{df_i}{dx}(x) = P_i\big(x, f_1(x), \dots, f_i(x)\big).$$

Define

$$\widetilde{f}_i(t) := f_i(2Mt - M).$$

Then

$$\frac{d\widetilde{f}_i}{dt}(t) = 2M\, P_i\big(2Mt - M, \widetilde{f}_1(t), \dots, \widetilde{f}_i(t)\big).$$

Thus $\widetilde{f}_1, \dots, \widetilde{f}_r$ is again a Pfaffian chain. The affine change of variables changes only the coefficients of the defining polynomials, not their degrees, and therefore does not change the relevant Pfaffian format or degree bounds. The same applies to $\widetilde{g}_M$.

Consequently, every formula in $\mathbb{R}_g$, restricted to a box $[-M, M]^n$, can be interpreted as a formula in $\mathbb{R}_{\mathrm{rPfaff}}$ with $*$-format $O_F(1)$ and $*$-degree $\mathrm{Poly}_F(D)$, uniformly in M. The argument of Appendix G.5 then applies with only minor modifications: use the growth-function bound for the restricted formulas via Theorem G.6, and then pass to the unrestricted formula by choosing M large enough for any fixed finite sample and the finitely many witnesses needed on that sample.

Thus

$$\Pi_{\mathcal{H}^{\mathcal{N}}}(m) \le D^{\gamma(F)} m^k.$$

The stated ERM sample-complexity and VC-dimension bounds follow from Lemma 2.2 and Lemma I.1, exactly as in the proof of Main Theorem 4. □

Remark H.2. The same argument applies if the language is expanded by finitely many globally defined Pfaffian functions, possibly of several variables. One only has to add the corresponding Pfaffian formats and degrees to the definitions of format and degree above. After restricting to a compact box and applying an affine rescaling to the unit cube, the resulting functions are again restricted Pfaffian functions with the same degree bounds, up to changes in the coefficients. We omit this straightforward generalization.

I Technical lemmas

Throughout this appendix, logarithms are base 2.

Lemma 2.2. *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, and suppose that $\Pi_{\mathcal{H}}(m) \le Cm^k$ for every $m \ge 1$, where $C \ge 1$ and $k \ge 1$. Then*

$$m^{\mathrm{real}}_{\mathrm{ERM}}(\varepsilon, \delta) = O\left(\frac{k \log(k/\varepsilon) + \log C + \log(1/\delta)}{\varepsilon}\right).$$

Proof. We use the standard realizable ERM bound

$$\Pr\big[L_D(h_S) > \varepsilon\big] \le \Pi_{\mathcal{H}}(2m) e^{-\varepsilon m/2}.$$

By the growth assumption, $\Pi_{\mathcal{H}}(2m) \le C(2m)^k$. Thus it is enough to choose m so that

$$C(2m)^k e^{-\varepsilon m/2} \le \delta, \qquad \text{equivalently} \qquad \frac{\varepsilon m}{2} \ge \log C + k \log(2m) + \log(1/\delta).$$

Set

$$x := \frac{\varepsilon m}{2}, \qquad L := \log C + \log(1/\delta).$$

Since $2m = 4x/\varepsilon$, the sufficient condition becomes

$$x \geq L + k \log\left(\frac{4x}{\varepsilon}\right) = L + k \log(4/\varepsilon) + k \log x.$$

Equivalently, it is enough to solve an inequality of the form $x \geq a + k \log x$, where

$$a := L + k \log(4/\varepsilon).$$

By Lemma I.2, it is enough to take

$$x = O\big(a + k \log k\big) = O\big(\log C + \log(1/\delta) + k \log(1/\varepsilon) + k \log k\big).$$

Since $x = \varepsilon m/2$, this gives

$$m = O\left(\frac{k \log(k/\varepsilon) + \log C + \log(1/\delta)}{\varepsilon}\right).$$

□

Lemma I.1 (VC dimension from polynomial growth)**.** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, and suppose that $\Pi_{\mathcal{H}}(m) \leq C m^k$ for every $m \geq 1$, where $C \geq 1$ and $k \geq 1$. Then*

$$\mathbb{VC}(\mathcal{H}) \leq 2 \log C + 4k \log(4k).$$

Proof. Let $d = \mathbb{VC}(\mathcal{H})$. If $d = 0$, there is nothing to prove. Otherwise, by the definition of VC dimension and by the growth assumption,

$$2^d \leq \Pi_{\mathcal{H}}(d) \leq C d^k.$$

Taking logarithms gives

$$d \leq \log C + k \log d.$$

Set $a := \log C$ and $b := k$. Then $d \leq a + b \log d$, so Lemma I.2 gives

$$d \leq 2a + 4b \log(4b) = 2 \log C + 4k \log(4k).$$

Thus $\mathbb{VC}(\mathcal{H}) \leq 2 \log C + 4k \log(4k)$. □

Lemma I.2 (Logarithmic self-bound)**.** *Let $x \geq 1$, $a \geq 0$, and $b \geq 1$. If*

$$x \leq a + b \log x,$$

then

$$x \leq 2a + 4b \log(4b).$$

Proof. If $x \leq 2a$, then immediately

$$x \leq 2a \leq 2a + 4b \log(4b).$$

It remains to consider the case $x > 2a$. Then $a < x/2$, and therefore

$$x \leq a + b \log x < \frac{x}{2} + b \log x.$$

Hence $x < 2b \log x$. Set $B := 2b$. Since $b \geq 1$, we have $B \geq 2$, and the last inequality becomes

$$x < B \log x.$$

We now claim that $x \leq 2B \log(2B)$. Suppose, toward a contradiction, that $x > 2B \log(2B)$. Since $B \geq 2$, we have $2B \log(2B) \geq 4 \log 4 > e$, and hence $x > e$. The function $t \mapsto \log t/t$ is decreasing for all $t \geq e$, so

$$\frac{\log x}{x} < \frac{\log(2B \log(2B))}{2B \log(2B)}.$$

Also,

$$\log(2B\log(2B)) \le 2\log(2B),$$

because this is equivalent to

$$2B\log(2B) \le (2B)^2,$$

or equivalently $\log(2B) \le 2B$, which follows from the elementary inequality $\log u \le u$ for $u > 0$. Therefore

$$\frac{\log x}{x} < \frac{2\log(2B)}{2B\log(2B)} = \frac{1}{B}.$$

Multiplying by Bx, we get $B\log x < x$, contradicting $x < B\log x$. Hence $x \le 2B\log(2B)$.

Substituting $B = 2b$, we obtain

$$x \le 4b\log(4b).$$

Together with the first case $x \le 2a$, this gives

$$x \le 2a + 4b\log(4b). \qquad \square$$

Lemma I.3. *Let $k \ge 1$, $d \ge 1$, and $A \ge 2$. Suppose that*

$$d \le k\log(Ad/k),$$

where logarithms are to base 2. Then

$$d \le 4k\log A.$$

Proof. We consider two cases.

First, suppose that

$$A \ge d/k.$$

Then

$$d \le k\log(Ad/k) \le k\log(A^2) = 2k\log A.$$

Otherwise,

$$A < d/k.$$

Then

$$d \le k\log(Ad/k) < k\log\big((d/k)^2\big) = 2k\log(d/k).$$

Equivalently, with $t = d/k$,

$$t < 2\log t.$$

This implies $t < 4$, and hence $d < 4k$. Since $A \ge 2$, we have

$$4k \le 4k\log A.$$

$\square$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's theoretical contributions and scope: the negative VC-dimension blow-up result, the positive learnability result under definability in $\mathbb{R}_{\exp}$, and the quantitative bounds for semialgebraic and existential definable settings. The assumptions and scope are stated in the main results and further discussed in Section 5 and Appendix A.3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses its main limitations in a dedicated Limitations section; see Section 5 and Appendix A.3.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are stated with clearly listed assumptions, and complete proofs are provided in the full version of the paper. Sketches also appear in the main text to build intuition.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: The paper does not contain experiments. The results are purely theoretical and proven analytically.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not involve data or code. All contributions are analytical.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: There are no experiments in this work. The paper is entirely theoretical.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not report experiments, so statistical significance does not apply.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: No compute resources were used, as the paper contains no experiments or computational simulations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It does not involve human subjects, data scraping, or dual-use risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper is primarily theoretical and does not introduce a deployed system. The introduction motivates the work through applications of strategic classification, including lending, hiring, admissions, public benefits, and medical eligibility. These applications point to potential positive impacts, through better understanding of strategic learnability, as well as potential negative impacts, since errors or unequal access to strategic improvement may affect individuals in consequential settings.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce models or datasets with high risk of misuse. No safeguards are necessary.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use any external datasets or code.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets, codebases, or pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The study does not involve human subjects and does not require IRB review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as a core component of the methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.