# Strategic PAC Learnability via Geometric Definability

Yuval Filmus[1,2] Shay Moran[1,2,3,4] Elizaveta Nesterova[1] Nir Rosenfeld[2] Alexander Shlimovich[1]

[1]Faculty of Mathematics, Technion [2]Faculty of Computer Science, Technion [3]Faculty of Data and Decision Sciences, Technion [4]Google Research

## Extended abstract

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. This creates a gap between the hypothesis chosen by the learner and the effective classifier induced after individuals strategically adapt. A basic learning-theoretic question is therefore whether strategic behavior preserves learnability: if the original hypothesis class is learnable, and the cost structure governing feasible manipulations is simple, must the induced strategic class also be learnable?

A natural intuition suggests a positive answer. If agents can only make simple changes, then one might expect the strategic transformation to preserve the statistical complexity of the original class, at least up to a controlled increase. This intuition is supported by several existing positive results, particularly for linear classifiers and norm-based costs. However, these results leave open whether learnability is preserved in any general sense.

We first show that, without additional structure, the answer is negative. We construct a hypothesis class over the real line with VC dimension 1 such that, even when each point can move only within a fixed-radius interval, the induced strategic class has infinite VC dimension. Thus, strategic behavior can turn a class with minimal combinatorial complexity into a non-learnable one. This shows that low VC dimension of the base class, even together with a simple and uniform neighborhood structure, is not enough to guarantee strategic learnability.

The negative result suggests that the relevant notion of simplicity is not purely combinatorial (e.g. based on restricting the VC dimension). We therefore introduce a geometric definability framework. In this framework, both the hypothesis class and the neighborhood relation induced by the cost are described by first-order formulas over the real field with exponentiation, $\mathbb{R}_{\exp}$. Informally, this means that hypotheses and feasible manipulations can be defined using arithmetic operations, comparisons, exponentiation, logarithms, and logical quantifiers. This captures a broad family of natural examples, including linear and polynomial threshold classes, intersections of halfspaces, decision-tree-like classes with polynomial tests, neural-network-type classes with common activations, and neighborhood systems induced by $\ell_p$ distances, Wasserstein-type distances, and information-theoretic divergences.

Our main positive result shows that this geometric structure restores learnability. If both the hypothesis class and the neighborhood relation are definable in $\mathbb{R}_{\exp}$, then the induced strategic class is PAC-learnable. The key observation is that strategic classification naturally introduces existential quantification: the induced strategic hypothesis labels an input $x$ positively precisely when some point $x' \in N_x$ is labeled positively by the original hypothesis. Definability provides a language in which this existential transformation can be controlled.

We complement this qualitative guarantee with quantitative bounds in more restricted settings. For semialgebraic definitions, we use quantifier elimination and real-algebraic sign-pattern bounds to obtain explicit VC-dimension and ERM sample-complexity estimates. We further extend the quantitative theory to existential formulas involving exponentiation, using tools from Pfaffian and tame geometry. Together, these results give a unified geometric framework for understanding when strategic behavior preserves learnability beyond previously studied linear and norm-based models.

Full paper: https://arxiv.org/abs/2605.13426