

Stability and List-Replicability for Agnostic Learners

Ari Blondal

McGill University

ARI.BLONDAL@MAIL.MCGILL.CA

Shan Gao

McGill University

SHAN.GAO5@MAIL.MCGILL.CA

Hamed Hatami

McGill University, Supported by an NSERC grant

HATAMI@CS.MCGILL.CA

Pooya Hatami

Ohio State University

POOYAHAT@GMAIL.COM

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Two seminal papers—Alon, Livni, Malliaris, Moran (STOC 2019) and Bun, Livni, and Moran (FOCS 2020)—established the equivalence between online learnability and globally stable PAC learnability in binary classification. However, Chase, Chornomaz, Moran, and Yehudayoff (STOC 2024) recently showed that this equivalence does not hold in the agnostic setting. Specifically, they proved that in the agnostic setting, only finite hypothesis classes are globally stable learnable. Therefore, agnostic global stability is too restrictive to capture interesting hypothesis classes.

To address this limitation, Chase *et al.* introduced two relaxations of agnostic global stability. In this paper, we characterize the classes that are learnable under their proposed relaxed conditions, resolving the two open problems raised in their work.

First, we prove that in the setting where the stability parameter can depend on the excess error (the gap between the learner’s error and the best achievable error by the hypothesis class), agnostic stability is fully characterized by the Littlestone dimension. Consequently, as in the realizable case, this form of learnability is equivalent to online learnability.

As part of the proof of this theorem, we strengthen the celebrated result of Bun *et al.* by showing that classes with infinite Littlestone dimension are not stably PAC learnable, even if we allow the stability parameter to depend on the excess error.

For the second relaxation proposed by Chase *et al.*, we prove that only finite hypothesis classes are globally stable learnable even if we restrict the agnostic setting to distributions with small population loss.

Keywords: Replicability, PAC learning, Stability, Littlestone dimension.

1. Introduction

We follow the standard PAC learning framework for *binary classification* as, for example, described by Shalev-Shwartz and Ben-David (2014). In this model, a learner receives a sample of i.i.d. *examples* from an unknown distribution \mathcal{D} over $X \times \{0, 1\}$, where X is the domain set, and $\{0, 1\}$ represents the two possible *labels* in binary classification. The learner’s goal is to produce a *hypothesis* $h : X \rightarrow \{0, 1\}$ that minimizes the *population loss*

$$\mathcal{L}_{\mathcal{D}}(h) := \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [h(\mathbf{x}) \neq y].$$

Here, and throughout the paper, we use boldface letters to denote random variables and use the notation $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ to express that (\mathbf{x}, \mathbf{y}) is a random variable distributed according to \mathcal{D} .

Formally, a *learning rule* is a (randomized) function \mathcal{A} that maps any sample $S \in (X \times \{0, 1\})^* := \bigcup_{n=0}^{\infty} (X \times \{0, 1\})^n$ to a hypothesis $\mathcal{A}(S) \in \{0, 1\}^X$. Thus, for any given sample S , $\mathcal{A}(S)$ is a random variable taking values in $\{0, 1\}^X$.

Throughout this paper, all learning rules are assumed to be randomized. We consistently use X to denote the domain, $\{0, 1\}$ to represent the two possible labels, and \mathcal{D} always refers to a distribution over $X \times \{0, 1\}$. For an integer $n > 0$, we use $[n]$ to denote the set $\{1, \dots, n\}$.

Given a *hypothesis class* $\mathcal{H} \subseteq \{0, 1\}^X$, the goal of PAC (Probably Approximately Correct) learning is for the learner to produce, with high probability, a hypothesis whose population loss is close to the best achievable within \mathcal{H} , defined as

$$\mathcal{L}_{\mathcal{D}}(\mathcal{H}) := \inf_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h).$$

A class \mathcal{H} is *PAC learnable* if there is a learning rule \mathcal{A} and a function $n(\epsilon, \delta)$ such that for any $\epsilon, \delta > 0$,

$$\Pr_{S \sim \mathcal{D}^n} [\mathcal{L}_{\mathcal{D}}(\mathcal{A}(S)) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon] \geq 1 - \delta \text{ where } n = n(\epsilon, \delta). \quad (1)$$

PAC learning is studied in the *realizable case*, where we assume $\mathcal{L}_{\mathcal{D}}(\mathcal{H}) = 0$, and the *agnostic case*, where $\mathcal{L}_{\mathcal{D}}(\mathcal{H}) > 0$.

Replicability and Global Stability. Replicability is a fundamental principle of the scientific method. A study is replicable if it consistently yields the same results when repeated with new data drawn from the same distribution or source. In recent years, machine learning has seen a growing need to address the replication crisis (Ball, 2023; Baker, 2016). Impagliazzo et al. (2022) initiated a formal theoretical framework for studying replicability in machine learning. Since their work, a rapidly growing body of research has emerged that introduced various notions of replicability. These works and subsequent research showed that many of these notions of replicability are essentially equivalent. Furthermore, they established deep connections to other foundational concepts in learning theory, such as differential privacy (Chase et al., 2023; Bun et al., 2023; Kalavasis et al., 2023; Ghazi et al., 2021; Chase et al., 2024). Additionally, a growing body of work has explored replicability in many data analysis and learning settings (Impagliazzo et al., 2022; Bun et al., 2023; Karbasi et al., 2023; Esfandiari et al., 2023a,b; Eaton et al., 2023; Kalavasis et al., 2024, 2023).

In this paper, we focus on the notion of replicability where the learning algorithm is expected to often produce the same predictor when applied to two independent and identically distributed inputs. This concept was first introduced under the term *global stability* by Bun et al. (2020) and has since been refined and explored in subsequent works (Ghazi et al., 2021; Kalavasis et al., 2023; Chase et al., 2023, 2024). We start by defining global stability.

Definition 1 (ρ -Global Stability, Chase et al. (2024)) Given a function $\rho : (0, 1) \rightarrow (0, 1)$, a learning rule \mathcal{A} is a ρ -global stable learner for a hypothesis class \mathcal{H} if the following holds. For every $\epsilon > 0$, there exists $n = n(\epsilon)$ such that for every realizable distribution \mathcal{D} , there exists a hypothesis h satisfying

$$\mathcal{L}_{\mathcal{D}}(h) \leq \epsilon$$

and

$$\Pr_{S \sim \mathcal{D}^n} [\mathcal{A}(S) = h] \geq \rho(\epsilon). \quad (2)$$

Similarly, we call \mathcal{A} a ρ -global stable agnostic learner for \mathcal{H} , if there exists $n = n(\epsilon)$ such that for every distribution \mathcal{D} on $X \times \{0, 1\}$, there exists a hypothesis $h \in \{0, 1\}^X$ that satisfies (2) and

$$\mathcal{L}_{\mathcal{D}}(h) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon. \quad (3)$$

To simplify terminology, we use the term ρ -global stable to describe a hypothesis class \mathcal{H} with a ρ -global stable learner. Likewise, we call \mathcal{H} agnostically ρ -global stable if it has a ρ -global stable agnostic learner.

Definition 2 (Global Stability, Bun et al. (2020)) We say that a hypothesis class \mathcal{H} is globally stable if it is ρ -global stable for a fixed constant $\rho \in (0, 1)$. Similarly, a hypothesis class \mathcal{H} is agnostically globally stable if it is agnostically ρ -stable for such a constant.

In short, global stability requires the stability parameter in Equation (2) to be uniform, meaning it must not depend on ϵ .

Bun et al. (2020) showed that in the realizable setting, global stability is fully characterized by bounded Littlestone dimension (Theorem 12). This result, combined with the seminal works of Littlestone (1988) and Alon et al. (2022), shows that global stability is equivalent to online learnability and approximately private learnability, as well as some other notions of replicability (Kalavasis et al., 2023; Ghazi et al., 2021; Bun et al., 2023).

In contrast, the agnostic setting reveals a different picture. Chase et al. (2024) proved the following characterization using a topological approach.

Theorem 3 (Chase et al. (2024)) A hypothesis class \mathcal{H} is agnostically globally stable if and only if \mathcal{H} is finite.

This striking result shows that agnostic global stability is far more restrictive than its realizable counterpart. Since finite classes are trivially global stable, Theorem 3 shows that agnostic global stability is too restrictive to lead to interesting learnability phenomena. To remedy this, Chase et al. (2024) introduced two relaxations of agnostic global stability and proposed a study of which hypothesis classes can be learned under these relaxed notions of stability.

Excess-error dependent stability. The first suggested relaxation, coincides with our definition of ρ -global stability in Theorem 1. A hypothesis class \mathcal{H} is called *excess-error dependent stable* if it is agnostically ρ -global stable for some $\rho : (0, 1) \rightarrow (0, 1)$. Here, the excess-error refers to the parameter ϵ in Eq. (3).

Our main theorem provides a complete characterization of such classes. We show that a hypothesis class is agnostically ρ -global stable learnable for some ρ if and only if it has a bounded Littlestone dimension. We denote the Littlestone dimension of \mathcal{H} as $\text{Ldim}(\mathcal{H})$.

Theorem 4 (Main Theorem) Let \mathcal{H} be a binary concept class.

- (i) If $\text{Ldim}(\mathcal{H}) = \infty$, then \mathcal{H} is not ρ -global stable for any $\rho : (0, 1) \rightarrow (0, 1)$.
- (ii) If $\text{Ldim}(\mathcal{H}) < \infty$, then \mathcal{H} is agnostically ρ -global stable for some $\rho : (0, 1) \rightarrow (0, 1)$.

We will prove Theorem 4 (i) in Section 2 and defer the proof of Theorem 4 (ii) to Appendix B.

Note that Theorem 4 (i) states that if $\text{Ldim}(\mathcal{H}) = \infty$, then even in the realizable case, we cannot achieve ρ -global stability for any $\rho : (0, 1) \rightarrow (0, 1)$. This strengthens the result of Bun et al. (2020), which only overrules ρ -global stability when $\rho > 0$ is a fixed constant.

Shortly after a draft of this paper was posted online, Hopkins and Moran (2025) communicated to us that in an independent work, they have proved an equivalent statement to Theorem 4 by utilizing the known relation between stability and differential privacy. They use a ρ -global stable learner to achieve weak DP learning, which in turn is boosted to a strong DP learner. It is well known that strong DP learning is achievable if and only if the Littlestone dimension is finite (Alon et al., 2022). In contrast, our proof is direct and relies solely on notions of stability and list-replicability.

Combined with the work of Alon et al. (2022), Theorem 4 implies that agnostic ρ -global stability is equivalent to global stability, as well as to approximate private learnability and online learnability.

Class-error dependent stability. In many practical learning scenarios, while we cannot assume realizability, we may have prior knowledge that the hypothesis class performs reasonably well. This corresponds to a more restricted version of agnostic learning, where the learning task is limited to distributions \mathcal{D} that satisfy $\mathcal{L}_{\mathcal{D}}(\mathcal{H}) \leq \gamma$ for some small $\gamma > 0$.

Definition 5 (Class-error Dependent Stability, Chase et al. (2024)) *Let $\gamma \in [0, 1]$ be a fixed constant. We say \mathcal{H} is γ -agnostically globally stable if there exists a constant $\rho > 0$ and a learning rule \mathcal{A} such that the following holds. For every $\epsilon > 0$, there exists $n = n(\epsilon)$ such that for every distribution \mathcal{D} with $\mathcal{L}_{\mathcal{D}}(\mathcal{H}) \leq \gamma$, there exists a hypothesis h satisfying*

$$\mathcal{L}_{\mathcal{D}}(h) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon,$$

and

$$\Pr_{\mathcal{S} \sim \mathcal{D}^n} [\mathcal{A}(\mathcal{S}) = h] \geq \rho.$$

The case $\gamma = 0$ corresponds to the realizable case, where Bun et al. (2020) show that global stability is fully characterized by bounded Littlestone dimension. On the other hand, $\gamma = 1$ corresponds to the agnostic case, where Theorem 3 shows that only finite classes are agnostically globally stable.

Chase et al. (2024) ask which hypothesis classes are γ -agnostically globally stable for all sufficiently small γ . Our next theorem shows that the realizable case, $\gamma = 0$, is the only scenario in which infinite hypothesis classes can be γ -agnostically globally stable. Therefore, the relaxation of agnostic global stability to γ -agnostic global stability does not lead to any generalization, as only finite hypothesis classes can be γ -agnostically globally stable if $\gamma > 0$.

To prove our theorem, we show that agnostic global stability reduces to γ -agnostic global stability, for any arbitrary $\gamma > 0$.

Theorem 6 *If a class $\mathcal{H} \subseteq \{0, 1\}^X$ is γ -agnostically globally stable for some $\gamma > 0$, then \mathcal{H} is finite.*

We defer the proof of Theorem 6 to Appendix A.

Relation to list replicability. In learning theory, global stability is more useful when paired with a guarantee that the learner typically outputs a hypothesis with a low population loss. The definition of global stability (Theorem 2) only requires that the learner outputs a low-error hypothesis with some probability $\rho > 0$, and the learner can output hypotheses with large population loss with probability $1 - \rho$. However, it is known that global stability implies bounded VC dimension Alon et al. (2022), and assuming bounded VC dimension, this can be easily remedied. The learner can estimate the population loss of its output by comparing it to that of the hypothesis produced by the empirical risk minimization (ERM) rule. If the population loss is unsatisfactory, the learner can fall back on the ERM hypothesis instead.

Next, we introduce a seemingly stronger notion of replicability, originally proposed by Chase et al. (2023), known as *list replicability*. This notion strengthens the standard guarantee by requiring that, with high probability, the output hypothesis lies within a small list of hypotheses, each with low population loss.

Definition 7 (List Replicability, Chase et al. (2023)) *Given a function $L : (0, 1) \rightarrow \mathbb{N}$, we say that a learner \mathcal{A} is an L -list-replicable learner for a hypothesis class \mathcal{H} if the following holds. For every $\epsilon, \delta > 0$, there exists $n = n(\epsilon, \delta)$ such that for every realizable distribution \mathcal{D} , there exists a list of $L = L(\epsilon)$ hypotheses h_1, \dots, h_L satisfying*

$$\mathcal{L}_{\mathcal{D}}(h_i) \leq \epsilon \text{ for all } 1 \leq i \leq L$$

and

$$\Pr_{S \sim \mathcal{D}^n}[\mathcal{A}(S) \in \{h_1, \dots, h_L\}] \geq 1 - \delta. \quad (4)$$

Similarly, \mathcal{A} is an agnostic L -list-replicable learner for \mathcal{H} if there exists $n = n(\epsilon, \delta)$ such that for every distribution \mathcal{D} on $X \times Y$, there exists a list of $L = L(\epsilon)$ hypotheses $h_1, \dots, h_L \in \mathcal{H}$ satisfying (4) and

$$\mathcal{L}_{\mathcal{D}}(h_i) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon \text{ for all } 1 \leq i \leq L.$$

Similar to stability, we define the notions of *global list-replicability* and *agnostic global list-replicability* to describe the uniform case where the learner in Theorem 7 exists for a fixed constant $L > 0$ independent of ϵ .¹

It is worth noting that Equation (4) easily implies stability, as there must exist some $i \in [L]$ with

$$\Pr_{S \sim \mathcal{D}^n}[\mathcal{A}(S) = h_i] \geq \frac{1 - \delta}{L}.$$

Chase et al. (2023) showed that the converse is also true: global stability implies global list-replicability.

Theorem 8 (Chase et al. (2023)) *For any fixed constant $L > 0$, a hypothesis class is L -list-replicable if and only if it is ρ -global stable for all $\rho < \frac{1}{L}$.*

Analogous to Theorem 5, given a parameter $\gamma \in [0, 1]$, we refer to a class \mathcal{H} as γ -agnostically list-replicable if we relax the requirement of the agnostic global list-replicability to only consider distributions \mathcal{D} with $\mathcal{L}_{\mathcal{D}}(\mathcal{H}) \leq \gamma$.

1. In the literature, what we refer to as global list-replicability is simply called list-replicability.

The foregoing relaxations of global list-replicability were proposed by [Chase et al. \(2024\)](#), where they asked for a characterization of hypothesis classes that can be agnostically learned under these notions, termed excess-error dependent and class-error dependent list-replicability.

Our next theorem extends Theorem 8 to show that these new notions coincide with their stability counterparts. Consequently, Theorem 4 and Theorem 6 completely resolve the questions posed by [Chase et al. \(2024\)](#).

Theorem 9 *Consider a parameter $\gamma \in [0, 1]$.*

- (i) *Given $L : (0, 1) \rightarrow \mathbb{N}$, if a class \mathcal{H} is γ -agnostically L -list replicable, then it is γ -agnostically ρ -global stable for any $\rho : (0, 1) \rightarrow [0, 1]$ satisfying $\rho(\epsilon) < \frac{1}{L(\epsilon)}$ for all $\epsilon \in (0, 1)$.*
- (ii) *Given $\rho : (0, 1) \rightarrow (0, 1]$, if a class \mathcal{H} is γ -agnostically ρ -global stable, then it is γ -agnostically L -list replicable, for $L(\epsilon) := \left\lfloor \frac{1}{\rho(\epsilon/4)} \right\rfloor$.*

We defer the proof of Theorem 9 to Appendix C.

1.1. Preliminaries: VC Dimension, Uniform Convergence, and Littlestone Dimension

This section outlines a few key concepts and results from learning theory. More specifically, we state the connections between PAC learnability, VC dimension, and uniform convergence, and we state the definition of the Littlestone dimension. For a detailed exposition, see [Shalev-Shwartz and Ben-David \(2014\)](#).

A fundamental result of learning theory is that a class \mathcal{H} is PAC-learnable if and only if it satisfies the *Uniform Convergence* property. For a sample of m examples $S \in (X \times \{0, 1\})^m$, and a hypothesis $h : X \rightarrow \{0, 1\}$, let

$$\mathcal{L}_S(h) := \Pr_{(x,y) \sim S} [h(x) \neq y],$$

denote the *empirical population loss* of h with respect to S .

Definition 10 (Uniform Convergence) *A binary hypothesis class \mathcal{H} has the Uniform Convergence property if, for any $\epsilon, \delta \in (0, 1)$, there exists $n(\epsilon, \delta)$ such that for any distribution \mathcal{D} , we have*

$$\Pr_{S \sim \mathcal{D}^n} [|\mathcal{L}_S(h) - \mathcal{L}_{\mathcal{D}}(h)| < \epsilon \text{ for all } h \in \mathcal{H}] \geq 1 - \delta.$$

The fundamental theory of PAC learning states that the Uniform Convergence property and, consequently, PAC-learnability are characterized by having a finite Vapnik-Chervonenkis (VC) dimension.

Definition 11 (VC dimension) *The VC dimension of a binary hypothesis class \mathcal{H} is the size of the largest subset X' of X such that, for every binary labelling of X' , there is a hypothesis $h \in \mathcal{H}$ consistent with that labelling. Such a set X' is said to be shattered by \mathcal{H} . If arbitrarily large sets can be shattered, the VC dimension is defined to be ∞ .*

The Littlestone dimension relaxes the VC dimension by shattering decision trees instead of sets. A *mistake tree* of depth d over a domain X is a complete binary tree of depth d with the following properties:

- Each internal node in the tree is labelled by an element $x \in X$.
- Each edge is labeled by a binary value $b \in \{0, 1\}$ where $b = 0$ indicates a left child and $b = 1$ indicates a right child.

Every *root-to-leaf path* in the tree is described by a sequence $(x_1, b_1), \dots, (x_d, b_d)$ where $x_i \in X$ is the label of the i th internal node on the path and b_i specifies whether the path moves to the left or right child at each level.

We say that a mistake tree is *shattered* by a hypothesis class $\mathcal{H} \subseteq \{0, 1\}^X$ if for every root-to-leaf path $(x_1, b_1), \dots, (x_d, b_d)$ where $x_i \in X$ and $b_i \in \{0, 1\}$, there exists a hypothesis $h \in \mathcal{H}$ with $h(x_i) = b_i$ for all $i \in [d]$.

Definition 12 (Littlestone Dimension) *The Littlestone dimension of a hypothesis class \mathcal{H} , denoted $\text{Ldim}(\mathcal{H})$, is the largest integer d such that there exists a mistake tree of depth d shattered by \mathcal{H} .*

We always have $\text{VCdim}(\mathcal{H}) \leq \text{Ldim}(\mathcal{H})$, since every shattered set $X' = \{x_1, \dots, x_d\}$ gives rise to a mistake tree of depth d where all nodes at level i are labelled with x_i . This tree is clearly shattered by \mathcal{H} .

2. Theorem 4 (i): Stability implies finite Littlestone dimension

By the equivalence of global stability and list-replicability established in Theorem 9, Theorem 4 (i) is equivalent to the following theorem.

Theorem 13 *If $\text{Ldim}(\mathcal{H}) = \infty$, then \mathcal{H} is not L -list replicable for any $L : (0, 1) \rightarrow \mathbb{N}$.*

The rest of this section is devoted to the proof of Theorem 13, which uses a classical result of Shelah (1990) connecting the Littlestone dimension to the threshold dimension.

Definition 14 (Threshold dimension) *The threshold dimension of $\mathcal{H} \subseteq \{0, 1\}^X$ is the largest k such that there exists a set of inputs $\{x_1, \dots, x_k\} \subseteq X$ and classifiers $\{h_1, \dots, h_k\} \subseteq \mathcal{H}$ satisfying*

$$h_t(x_i) = 1 \iff i \geq t \quad \text{for all } i, t \in [k].$$

We refer the reader to Alon et al. (2022) for an accessible proof of the following result of Hodges (1997), which provides effective bounds for a qualitative result of Shelah (1990). Shelah proved that any class \mathcal{H} with infinite Littlestone dimension also has an infinite threshold dimension.

Proposition 15 (Hodges (1997)) *If $\mathcal{H} \subseteq \{0, 1\}^X$ has $\text{Ldim}(\mathcal{H}) = d$, its threshold dimension is at least $\lfloor \log d \rfloor$.*

Throughout the proof, we will use the following observation stating that without loss of generality we may ignore the order of examples in a sample.

Remark 16 *Since the population loss does not depend on the order of the examples in S , and the examples are drawn independently from \mathcal{D} , in the context of PAC learning and stability, we may assume that the learning rule disregards the order of the examples in any sample $S \in (X \times \{0, 1\})^n$. In other words, the learning rule is invariant under the permutations of the examples in any given sample. As a result, we often treat a sample S as a multiset rather than a sequence.*

Next, we prove a lemma to decrease the probability of failure (i.e., δ) in the definition of list replicability to a small function of the sample size n and the population regret ϵ .

Lemma 17 (Boosting success probability) *Suppose \mathcal{H} is L -list-replicable for some $L : (0, 1) \rightarrow \mathbb{N}$. For every $C > 1$, there exists a learning rule \mathcal{A} and a sample complexity $n_C : (0, 1) \rightarrow \mathbb{N}$ such that the following holds. For every $\epsilon > 0$ and every realizable distribution \mathcal{D} , there exists a list of $L = L(\epsilon)$ hypotheses h_1, \dots, h_L satisfying*

$$\mathcal{L}_{\mathcal{D}}(h_i) \leq \epsilon \text{ for all } 1 \leq i \leq L$$

and

$$\Pr_{S \sim \mathcal{D}^n}[\mathcal{A}(S) \in \{h_1, \dots, h_L\}] \geq 1 - \frac{\epsilon}{n^C} \text{ where } n = n_C(\epsilon).$$

Proof Define $\delta_0 := \frac{1}{16L}$. By our assumption, there exists $n_0 = n_0(\epsilon)$ and a learning rule \mathcal{A}' such that for any realizable distribution \mathcal{D} , there exists a list $h_1, \dots, h_{L(\epsilon)}$ of hypotheses satisfying $\mathcal{L}_{\mathcal{D}}(h_i) \leq \epsilon$ for all i and

$$\Pr_{S \sim \mathcal{D}^{n_0}}[\mathcal{A}'(S) \in \{h_1, \dots, h_L\}] \geq 1 - \delta_0 = 1 - \frac{1}{16L}. \quad (5)$$

Since δ_0 is fixed, n_0 depends only ϵ .

Let $k > 0$ be an integer to be determined later. We define a new learning rule \mathcal{A} that uses samples of size kn_0 . Given a sample $S = (S_1, \dots, S_k) \in ((X \times \{0, 1\})^{n_0})^k$, the learner \mathcal{A} outputs the most frequent hypothesis produced by the k independent runs $\mathcal{A}'(S_1), \dots, \mathcal{A}'(S_k)$.

Let \mathcal{D} be any realizable distribution, and let $h_1, \dots, h_{L(\epsilon)}$ be as above. By Equation (5), there exists some $j^* \in [L]$ such that

$$\Pr_{S \sim \mathcal{D}^{n_0}}[\mathcal{A}'(S) = h_{j^*}] \geq \frac{1}{2L}.$$

Consider $S = (S_1, \dots, S_k) \sim (\mathcal{D}^{n_0})^k$. For every $i \in [k]$, define the indicator variable E_i and B_i as

- $E_i = 1$ iff $\mathcal{A}'(S_i) = h_{j^*}$;
- $B_i = 1$ iff $\mathcal{A}'(S_i) \notin \{h_1, \dots, h_L\}$.

The variables E_1, \dots, E_k are independent Bernoulli variables with $\mathbb{E}[E_i] \geq \frac{1}{2L}$. Similarly, B_1, \dots, B_k are independent Bernoulli variables with $\mathbb{E}[B_i] \leq \delta_0 \leq \frac{1}{16L}$. Define $E := \sum E_i$ and $B := \sum B_i$. Applying Hoeffding's inequality, we have

$$\Pr[E \geq \frac{k}{4L}] \geq 1 - \Pr[|E - \mathbb{E}[E]| \geq \frac{k}{4L}] \geq 1 - e^{-\Omega(k/L^2)}$$

and

$$\Pr[B \leq \frac{k}{8L}] \geq 1 - \Pr[|B - \mathbb{E}[B]| \geq \frac{k}{16L}] \geq 1 - e^{-\Omega(k/L^2)}.$$

When both events occur, the output of \mathcal{A} , the most frequent hypothesis, must come from the list $\{h_1, \dots, h_L\}$. We may now choose k such that $\delta := 2 \cdot e^{-\Omega(k/L^2)} \leq \frac{\epsilon}{(n_0 k)^C} = \frac{\epsilon}{n^C}$, concluding the proof. \blacksquare

Let us give an overview of the proof of Theorem 13. By Theorem 15, it is sufficient to prove that hypothesis classes of infinite threshold dimension are not L -list replicable. Assume towards contradiction that there exists an L -list-replicable learner for a hypothesis class of infinite threshold dimension.

Similarly to Alon et al. (2019), we use a hypergraph Ramsey argument to restrict the learning problem to an arbitrarily large subset X of the domain, on which the learner's prediction is essentially determined by the ordered sign pattern of the sample. In particular, the Ramsey argument ensures that to label an element $x \in X$, the learner essentially looks at the ordered sign pattern of the labelled points in its given sample, and also at the “order” of x with respect to these points.

We consider the distribution that is uniform over the threshold about the median of X . Since the learner is a PAC learner, we show that the probability that its output classifies a point x as a 1 ranges from close to 0 to close to 1, as the order of x ranges from the smallest to the largest element compared to the points in the samples. Consequently, we detect a probability jump from some order to the next.

The most significant difference between our proof and the proof of Alon et al. (2019) lies in handling this probability jump. Alon et al. (2019) exploit the jump to create a “privacy leak,” whereas, without a privacy guarantee, we take a different approach based on an “approximate rank” argument. More specifically, we use our boosting lemma (Theorem 17) to show that for many samples, the function corresponding to the output probabilities can be well-approximated (in the L_∞ norm) by a convex combination of a fixed short list of hypotheses. We then apply a “volume-based” argument to derive a contradiction by finding two samples that are well-approximated by the same convex combination, but are supposed to label a point $x \in X$ differently according to the aforementioned probability jump.

We are ready to present the proof of Theorem 13.

Proof [Theorem 13] Fix some $\epsilon \in (0, 1)$, and towards a contradiction, assume that $\text{Ldim}(\mathcal{H}) = \infty$ and \mathcal{H} is L -list replicable for some $L : (0, 1) \rightarrow \mathbb{N}$.

By Theorem 17, there exists a constant n and a learning rule \mathcal{A} such that for every realizable distribution \mathcal{D} , there exists a list of hypotheses h_1, \dots, h_L satisfying

$$\mathcal{L}_{\mathcal{D}}(h_i) \leq \epsilon \text{ for all } 1 \leq i \leq L \quad (6)$$

and

$$\Pr_{S \sim \mathcal{D}^n} [\mathcal{A}(S) \in \{h_1, \dots, h_L\}] \geq 1 - \delta, \quad (7)$$

where $\delta := n^{-10}$. Since the learning rule can always ignore the extra examples in a sample, we may assume that n is arbitrarily large. In particular, we assume $n > \frac{1}{\epsilon}$. Furthermore, by Theorem 16, we assume that $\mathcal{A}(S)$ and $\mathcal{A}(S')$ are identically distributed if S' is a reordering of the same examples as in S .

Let N be a large integer that will be determined later. Since $\text{Ldim}(\mathcal{H}) = \infty$, by Theorem 15, the threshold dimension of \mathcal{H} is infinite. Thus, since the threshold dimension is at least $N + 1$, we may assume (by renaming elements if needed) that $\{1, \dots, N\} \subseteq X$ and that there exist classifiers $h_1, \dots, h_{N+1} \in \mathcal{H}$ with

$$h_t(i) = 1 \iff i \geq t \quad \text{for all } i \in [N] \text{ and } t \in [N + 1].$$

For the remainder of this proof, we focus on the elements in $[N] \subseteq X$ and the classifiers $h_1, \dots, h_{N+1} \in \mathcal{H}$, disregarding the others.

Consider a set $R = \{x_1, \dots, x_n\} \subseteq [N]$ with $x_1 < x_2 < \dots < x_n$. For every $x \in [N]$, define $\text{ord}_R(x) \in [n+1]$ as

$$\text{ord}_R(x) := 1 + |\{x_i \in R : x_i \leq x\}|,$$

which corresponds to the position of x if it were inserted in the increasing sequence (x_1, \dots, x_n) .

For each $t \in [n+1]$, consider the output of the learning rule \mathcal{A} on the labelling of R according to the threshold t :

$$\mathbf{h}_t^R := \mathcal{A}(\{(x_1, 0), \dots, (x_{t-1}, 0), (x_t, 1), \dots, (x_n, 1)\}).$$

Claim 1 *Let M be a positive integer. Provided that N is sufficiently large, there exists a set $X' \subseteq [N]$ of size M , and real numbers $p_{t,k} \in [0, 1]$ for $t, k \in [n+1]$ such that the following holds.*

For every subset $T = \{x_1, \dots, x_n\} \subseteq X'$ and every $x \in X' \setminus T$, we have

$$p_{t,k} - \delta \leq \Pr[\mathbf{h}_t^T(x) = 1] \leq p_{t,k}, \quad \text{where } k := \text{ord}_T(x)$$

for all $t \in [n+1]$.

Proof The claim is a consequence of the hypergraph Ramsey theorem. Given any subset $T = \{x_1, \dots, x_{n+1}\} \subseteq [N]$ and $t, k \in [n+1]$, let

$$q_{t,k}^T := \Pr[\mathbf{h}_t^{T \setminus \{x_k\}}(x_k) = 1],$$

and let $p_{t,k}^T$ be $q_{t,k}^T$ rounded up to an integer multiple of δ , namely

$$p_{t,k}^T := \left\lceil \frac{q_{t,k}^T}{\delta} \right\rceil \delta.$$

Define the “colour” of the set T as the matrix

$$c(T) := [p_{t,k}^T]_{t,k \in [n+1]},$$

and note that there are at most $\lceil \frac{2}{\delta} \rceil^{(n+1) \times (n+1)}$ possible colours. By the hypergraph Ramsey theorem (Ramsey, 1930), for sufficiently large N , there exists $X' \subseteq [N]$ with $|X'| = M$ such that all subsets T of X' of size $n+1$ share the same colour $[p_{t,k}]_{t,k \in [n+1]}$. The set X' and the values $p_{t,k}$ satisfy the claim. \blacksquare

Let X' be as in Claim 1 for a sufficiently large M . By renaming the elements if necessary, without loss of generality, we assume $X' = [M]$. Let $m^* := \lfloor M/2 \rfloor$ be the median of X' . Let \mathcal{D} be the uniform probability distribution over the set

$$\text{supp}(\mathcal{D}) := \{(x, \mathbf{1}_{[x \geq m^*]}) : x \in [M]\}.$$

In other words, we sample x uniformly at random from $[M]$ and label it according to the hypothesis $\mathbf{1}_{[x \geq m^*]}$. We will show that the learner \mathcal{A} cannot satisfy Equations (6) and (7) for this distribution \mathcal{D} , resulting in a contradiction.

Given $S \in ([M] \times \{0, 1\})^n$, let $S_X \in [M]^n$ be the sequence obtained by removing the labels from the examples in S . Note that if $S \sim \mathcal{D}^n$, then S_X is uniformly distributed over $[M]^n$.

Given a sample $S \in ([M] \times \{0, 1\})^n$, let $t(S)$ denote the number of examples in S with label 0. Let Π denote the set of $S \in \mathcal{D}^n$ with the following desired well-spread-ness properties:

1. S involves n *distinct* elements in $[M]$. In this case, we identify the sequence S_X with the corresponding n -element subset of $[M]$.
2. $t(S) \in [n/4, 3n/4]$.
3. For every interval $I \subseteq [M]$ of size $\frac{M}{8}$, we have

$$\left| |S_X \cap I| - \frac{n}{8} \right| \leq \frac{n}{100}. \quad (8)$$

4. Denoting the elements of S_X by $a_1 < a_2 < \dots < a_n$, we have $a_1 > \frac{M}{2^n}$, $a_n < M - \frac{M}{2^n}$, and $a_{i+1} > a_i + \frac{M}{2^n}$ for all $i = 1, \dots, n-1$.

By taking M to be sufficiently large as a function of n and applying Chernoff and union bounds, we have

$$\Pr_{\mathbf{S} \sim \mathcal{D}^n} [\mathbf{S} \in \Pi] \geq 1 - 2^{-\Omega(n)}.$$

Therefore, we may only focus on the uniformly chosen samples from Π . Note that the uniform distribution over Π corresponds to sampling $\mathbf{S} \sim \mathcal{D}^n$ conditioned on $\mathbf{S} \in \Pi$.

For every $t \in [\frac{n}{4}, \frac{3n}{4}]$, define

$$\Pi_t := \{S \in \Pi \mid t(S) = t\}.$$

Note that for every $a \in [M]$, we have

$$\begin{aligned} \Pr_{\mathbf{S} \sim \Pi_t} [a \in \mathbf{S}_X] &= \frac{\Pr_{\mathbf{S} \sim \mathcal{D}^n} [(a \in \mathbf{S}_X) \wedge (t(\mathbf{S}) = t) \wedge (\mathbf{S} \in \Pi)]}{\Pr_{\mathbf{S} \sim \mathcal{D}^n} [(t(\mathbf{S}) = t) \wedge (\mathbf{S} \in \Pi)]} \\ &\leq \frac{\Pr_{\mathbf{S} \sim \mathcal{D}^n} [a \in \mathbf{S}_X]}{\Pr_{\mathbf{S} \sim \mathcal{D}^n} [(t(\mathbf{S}) = t) \wedge (\mathbf{S} \in \Pi)]} = O_{n,\epsilon} \left(\frac{1}{M} \right), \end{aligned}$$

where $O_{n,\epsilon}(\cdot)$ indicates that the hidden constants in the bound may depend on n and thus also ϵ . By the above discussion and Equation (7), there exists an integer $t_0 \in [n/4, 3n/4]$ such that

$$\Pr_{\mathbf{S} \sim \Pi_{t_0}} [\mathcal{A}(\mathbf{S}) \in \{h_1, \dots, h_L\}] \geq 1 - \delta - 2^{-\Omega(n)}, \quad (9)$$

and for every $a \in [M]$,

$$\Pr_{\mathbf{S} \sim \Pi_{t_0}} [a \in \mathbf{S}_X] \leq O_{n,\epsilon} \left(\frac{1}{M} \right). \quad (10)$$

Fix such a t_0 for the rest of the proof.

For every $S \in \Pi_{t_0}$, define the function $f_S : [M] \rightarrow [0, 1]$ as

$$f_S(x) := \Pr[\mathcal{A}(S)(x) = 1].$$

Since $X' = [M]$ satisfies the assertion of Claim 1, there exists values $p_k := p_{t_0,k} \in [0, 1]$ for $k \in [n+1]$ such that the following holds. For all $S \in \Pi_{t_0}$ and every $x \in X' \setminus S_X$, we have

$$f_S(x) = \Pr[\mathcal{A}(S)(x) = 1] \in [p_k - \delta, p_k] \text{ where } k = \text{ord}_{S_X}(x). \quad (11)$$

Since h_1, \dots, h_L all have low population losses, intuitively, for small x , $f_S(x)$ should be close to 0 and for large x , $f_S(x)$ should be close to 1. Thus, we expect to find a, b such that $|p_b - p_a|$ is large, and consequently there must exist $1 \leq c < n+1$ such that $|p_{c+1} - p_c|$ is not too small.

Claim 2 *There exists $1 \leq c < n + 1$ such that $|p_{c+1} - p_c| \geq \frac{1}{2n}$.*

Proof Let \hat{x} be uniformly sampled from $[M/8]$. Since h_1, \dots, h_L have loss at most ϵ , and since labeling \hat{x} with 1 is incorrect, we have

$$\Pr_{\hat{x} \sim [M/8]} [h_i(\hat{x}) = 1] \leq 8 \Pr_{x \sim [M]} [h_i(x) = 1] \leq 8\epsilon \quad \text{for all } i = 1, \dots, L.$$

Therefore, using Equation (9), we have

$$\begin{aligned} & \Pr_{\substack{S \sim \Pi_{t_0} \\ \hat{x} \sim [M/8]}} [\mathcal{A}(S)(\hat{x}) = 1] \\ & \leq \Pr_{S \sim \Pi_{t_0}} [\mathcal{A}(S) \notin \{h_1, \dots, h_L\}] + \Pr_{\substack{S \sim \Pi_{t_0} \\ \hat{x} \sim [M/8]}} [\mathcal{A}(S)(\hat{x}) = 1 \mid \mathcal{A}(S) \in \{h_1, \dots, h_L\}] \\ & \leq \delta + 2^{-\Omega(n)} + 8\epsilon = O(\epsilon). \end{aligned} \tag{12}$$

Consider $S \in \Pi$. By Equation (8), every $\hat{x} \in [M/8] \setminus S_X$ satisfies $\text{ord}_{S_X}(\hat{x}) < n/4$. Therefore, using Equation (11),

$$\begin{aligned} \Pr_{\substack{S \sim \Pi_{t_0} \\ \hat{x} \sim [M/8]}} [\mathcal{A}(S)(\hat{x}) = 1] & \geq \Pr_{\substack{S \sim \Pi_{t_0} \\ \hat{x} \sim [M/8]}} [\mathcal{A}(S)(\hat{x}) = 1 \mid \hat{x} \notin S_X] - \Pr_{\substack{S \sim \Pi_{t_0} \\ \hat{x} \sim [M/8]}} [\hat{x} \in S_X] \\ & \geq \min_{k \leq n/4} p_k - \delta - \frac{n}{M/8} = \min_{k \leq n/4} p_k - O(\delta). \end{aligned}$$

Combining with Equation (12), we get

$$\min_{k \leq n/4} p_k = O(\epsilon).$$

Using a similar argument, by considering $\hat{x} \sim [\frac{7M}{8}, M]$, we obtain

$$\max_{k \geq 3n/4} p_k = 1 - O(\epsilon).$$

It follows that

$$\left| \max_{k \geq 3n/4} p_k - \min_{k \leq n/4} p_k \right| \geq \frac{1}{2},$$

and therefore there exists some $c \in [n]$ such that $|p_{c+1} - p_c| \geq \frac{1}{2n}$. ■

Let c be as in Claim 2, and suppose that $c \leq t_0$ without loss of generality. Call a sample $S \in \Pi_{t_0}$ *good* if

$$\Pr[\mathcal{A}(S) \in \{h_1, \dots, h_L\}] \geq 1 - \sqrt{\delta}.$$

By Equation (9), we have

$$\delta + 2^{-\Omega(n)} \geq \Pr_{S \sim \Pi_{t_0}} [\mathcal{A}(S) \notin \{h_1, \dots, h_L\}] \geq \Pr_{S \sim \Pi_{t_0}} [S \text{ is not good}] \times \sqrt{\delta}.$$

Consequently,

$$\Pr_{S \sim \Pi_{t_0}} [S \text{ is good}] \geq 1 - \sqrt{\delta} - \frac{2^{-\Omega(n)}}{\sqrt{\delta}} \geq 1 - 2\sqrt{\delta}. \tag{13}$$

Given any good $S \in \Pi_{t_0}$, define

$$\bar{h}_S = \sum_{i=1}^L \Pr[\mathcal{A}(S) = h_i | \mathcal{A}(S) \in \{h_1, \dots, h_L\}] \times h_i,$$

and note that by the definition of goodness, we have

$$|\bar{h}_S(x) - f_S(x)| \leq \sqrt{\delta} \quad \text{for all } x \in [M].$$

The function \bar{h}_S is a convex combination of h_1, \dots, h_L that pointwise $\sqrt{\delta}$ -approximates f_S . Let G be a maximal set of functions $X' \rightarrow [0, 1]$ such that

1. Every function $g \in G$ is a convex combination of h_1, \dots, h_L , and
2. For every pair of distinct functions $g_1, g_2 \in G$, there exists $x \in X'$ such that $|g_1(x) - g_2(x)| \geq \delta$.

By the above two conditions and the above discussion, for any good S , there exists $g \in G$ with

$$\|f_S - g\|_\infty \leq \delta + \sqrt{\delta} \leq 2\sqrt{\delta}.$$

Claim 3 $|G| \leq O(1/\delta)^L$.

Proof Denote by V the set of linear combinations of h_1, \dots, h_L , and let $B = V \cap [0, 1]^{X'}$. For any $\lambda \leq 1$, define $\lambda B := \{\lambda g \mid g \in B\}$. Suppose $m = |G|$, and name the functions in G as g_1, \dots, g_m . Define

$$B_i = g_i + \frac{\delta}{2} B.$$

Now note that B_1, \dots, B_m are disjoint subsets of $(1 + \frac{\delta}{2})B$. Thus the volume of $\cup_i B_i$ is bounded by that of $(1 + \delta/2)B$, and we get that $m \leq \left(\frac{1+\delta/2}{\delta/2}\right)^L = O(1/\delta)^L$. \blacksquare

Given $S \in \Pi_{t_0}$ where the elements of S_X are ordered as $a_1 < a_2 < \dots < a_n$, we define the i -th interval of S , for $i \in [n+1]$, as (a_{i-1}, a_i) when $i > 1$, and $(1, a_1)$ when $i = 1$.

Claim 4 *There exist $g \in G$ and two good samples S_1 and S_2 such that*

$$\|f_{S_1} - g\|_\infty \leq 2\sqrt{\delta} \quad \text{and} \quad \|f_{S_2} - g\|_\infty \leq 2\sqrt{\delta},$$

and moreover, there is an element $x \in [M]$ that belongs to the c -th interval of S_1 and the $c+1$ -th interval of S_2 .

Proof Let $A \subseteq [M/2]$ be the set of $a \in [M/2]$ for which there exists a good sample $S \in \Pi_{t_0}$ such that the c -th smallest element of S_X equals a . For every $a \in A$, let S^a represent an arbitrary choice of such a good sample. Combining Equations (10) and (13), we have $|A| = \Omega_{n,\epsilon}(M)$.

Given any $g \in G$, let A_g denote the set of all $a \in A$ with $\|f_{S^a} - g\|_\infty \leq 2\sqrt{\delta}$. Recall that for every good S , there exists $g \in G$ with $\|f_S - g\|_\infty \leq 2\sqrt{\delta}$. Therefore, by Claim 3, there exists a fixed $g^* \in G$ with

$$|A_{g^*}| \geq |A| \cdot O(\delta)^L = \Omega_{n,\epsilon}(M).$$

Finally, choosing M sufficiently large guarantees that $|A_{g^*}| \times M/2^{n+1} \gg M$ and thus there exist $a, b \in A_{g^*}$ such that $a > b$ and

$$2 \leq |a - b| \leq M/2^{n+1}.$$

Recalling that the gap between every two elements of S_X^a and similarly for S_X^b is at least $M/2^n$, the above guarantees the existence of an element x in the intersection of the $(c + 1)$ -th interval of $S_1 := S^a$ and the c -th interval of $S_2 := S^b$. ■

Let S_1, S_2, g , and x be as guaranteed by Claim 4. In this case,

$$|f_{S_1}(x) - f_{S_2}(x)| \leq \|f_{S_1} - g\|_\infty + \|f_{S_2} - g\|_\infty \leq 4\sqrt{\delta}.$$

Since x belongs to the c -th interval of S_1 and $c + 1$ -th interval of S_2 , by Equation (11), we have $|p_c - f_{S_1}(x)| \leq \delta$ and $|p_{c+1} - f_{S_2}(x)| \leq \delta$. Therefore,

$$|p_c - p_{c+1}| \leq 2\delta + 4\sqrt{\delta}.$$

For sufficiently large n , this inequality contradicts our choice of c from Claim 2 which satisfies $|p_{c+1} - p_c| \geq \frac{1}{2n}$. ■

References

- Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019.
- Noga Alon, Mark Bun, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private and online learnability are equivalent. *J. ACM*, 69(4), August 2022. ISSN 0004-5411. doi: 10.1145/3526074. URL <https://doi.org/10.1145/3526074>.
- Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016. doi: 10.1038/533452a. URL <https://doi.org/10.1038/533452a>.
- Philip Ball. Is AI leading to a reproducibility crisis in science? *Nature*, 624(7990):22–25, 2023.
- Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402, 2020. doi: 10.1109/FOCS46700.2020.00044.
- Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 520–527, 2023.
- Zachary Chase, Shay Moran, and Amir Yehudayoff. Stability and Replicability in Learning . In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2430–2439. IEEE Computer Society, 2023. doi: 10.1109/FOCS57990.2023.00148. URL <https://doi.ieeecomputersociety.org/10.1109/FOCS57990.2023.00148>.

- Zachary Chase, Bogdan Chornomaz, Shay Moran, and Amir Yehudayoff. Local borsuk-ulam, stability, and replicability. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, page 1769–1780, 2024. ISBN 9798400703836. doi: 10.1145/3618260.3649632. URL <https://doi.org/10.1145/3618260.3649632>.
- Eric Eaton, Marcel Husing, Michael Kearns, and Jessica Sorrell. Replicable reinforcement learning. *Advances in Neural Information Processing Systems*, 36:15172–15185, 2023.
- Hossein Esfandiari, Alkis Kalavasis, Amin Karbasi, Andreas Krause, Vahab Mirrokni, and Grigoris Velegkas. Replicable bandits. In *The Eleventh International Conference on Learning Representations*, 2023a. URL <https://openreview.net/forum?id=gcD2UtCGMc2>.
- Hossein Esfandiari, Amin Karbasi, Vahab Mirrokni, Grigoris Velegkas, and Felix Zhou. Replicable clustering. *Advances in Neural Information Processing Systems*, 36:39277–39320, 2023b.
- Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. *Advances in Neural Information Processing Systems*, 34:20172–20184, 2021.
- Wilfrid Hodges. *A shorter model theory*. Cambridge university press, 1997.
- Max Hopkins and Shay Moran. The role of randomness in stability. *arXiv preprint arXiv:2502.08007*, 2025.
- Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 818–831, 2022. ISBN 9781450392648. doi: 10.1145/3519935.3519973. URL <https://doi.org/10.1145/3519935.3519973>.
- Alkis Kalavasis, Amin Karbasi, Shay Moran, and Grigoris Velegkas. Statistical indistinguishability of learning algorithms. In *International Conference on Machine Learning*, pages 15586–15622. PMLR, 2023.
- Alkis Kalavasis, Amin Karbasi, Kasper Green Larsen, Grigoris Velegkas, and Felix Zhou. Replicable learning of large-margin halfspaces. In *International Conference on Machine Learning*, pages 22861–22878. PMLR, 2024.
- Amin Karbasi, Grigoris Velegkas, Lin Yang, and Felix Zhou. Replicability in reinforcement learning. *Advances in Neural Information Processing Systems*, 36:74702–74735, 2023.
- Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.
- F. P. Ramsey. On a problem of formal logic. *Proceedings of the London Mathematical Society*, s2-30(1):264–286, 1930. doi: <https://doi.org/10.1112/plms/s2-30.1.264>. URL <https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-30.1.264>.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

S. Shelah. *Classification theory and the number of nonisomorphic models*, volume 92 of *Studies in Logic and the Foundations of Mathematics*. North-Holland Publishing Co., Amsterdam, second edition, 1990. ISBN 0-444-70260-1.

Appendix A. Proof of Theorem 6

For the reader's convenience, we recall the statement of the theorem.

Theorem 6 *If a class $\mathcal{H} \subseteq \{0, 1\}^X$ is γ -agnostically globally stable for some $\gamma > 0$, then \mathcal{H} is finite.*

Proof Assume towards a contradiction that an infinite $\mathcal{H} \subseteq \{0, 1\}^X$ is γ -agnostically globally stable for some $\gamma > 0$, and let $\rho > 0$, \mathcal{A} , and $n(\cdot)$ be as in Theorem 5.

Pick any $x^* \in X$, and let $b^* \in \{0, 1\}$ be such that the subclass $\mathcal{H}^* := \{h \in \mathcal{H} : h(x^*) = b^*\}$ is infinite. Let $\gamma' := \min\{\gamma, \frac{1}{10}\}$. We obtain a contradiction with Theorem 3 by showing that \mathcal{H}^* is agnostically globally stable despite being infinite.

Given $\epsilon > 0$ and access to a distribution \mathcal{D} on $X \times \{0, 1\}$, let $n := n(\epsilon\gamma')$, and define the distribution

$$\mathcal{D}' := \gamma'\mathcal{D} + (1 - \gamma')\mathbf{1}_{(x^*, b^*)},$$

which corresponds to sampling from \mathcal{D} with probability γ' , and sampling (x^*, b^*) with probability $1 - \gamma'$. Consider the learning rule \mathcal{A}' described as follows:

1. Given a sample $S \sim \mathcal{D}^n$, independently replace each example in S with (x^*, b^*) with probability $1 - \gamma'$. Let T denote the resulting modified sample.
2. Output $\mathcal{A}(T)$.

Note that $\mathcal{A}'(S)$ with $S \sim \mathcal{D}^n$ has the same distribution as $\mathcal{A}(T)$ with $T \sim (\mathcal{D}')^n$.

Since every $h \in \{0, 1\}^X$ with $h(x^*) = b^*$, satisfies $\mathcal{L}_{\mathcal{D}'}(h) = \gamma'\mathcal{L}_{\mathcal{D}}(h) \leq \gamma'$, we have $\mathcal{L}_{\mathcal{D}'}(\mathcal{H}) \leq \gamma'$. Therefore, by our choice of $n := n(\epsilon\gamma')$ and our assumption of the γ -agnostic global stability of \mathcal{A} on \mathcal{H} , there exists $h^* \in \{0, 1\}^X$ with

$$\mathcal{L}_{\mathcal{D}'}(h^*) \leq \mathcal{L}_{\mathcal{D}'}(\mathcal{H}) + \epsilon\gamma' \quad \text{and} \quad \Pr_{S \sim \mathcal{D}^n} [\mathcal{A}'(S) = h^*] = \Pr_{T \sim (\mathcal{D}')^n} [\mathcal{A}(T) = h^*] \geq \rho. \quad (14)$$

If $h^*(x^*) \neq b^*$, then since $\gamma' < \frac{1}{10}$, we have

$$\mathcal{L}_{\mathcal{D}'}(h^*) \geq 1 - \gamma' > \gamma' + \epsilon\gamma' \geq \mathcal{L}_{\mathcal{D}'}(\mathcal{H}) + \epsilon\gamma',$$

which contradicts the first inequality in Equation (14). Therefore, $h^*(x^*) = b^*$, and consequently, we have $\mathcal{L}_{\mathcal{D}'}(h^*) = \gamma'\mathcal{L}_{\mathcal{D}}(h^*)$. Furthermore, $\mathcal{L}_{\mathcal{D}'}(\mathcal{H}) = \gamma'\mathcal{L}_{\mathcal{D}}(\mathcal{H}^*)$. Replacing these in Equation (14) shows

$$\mathcal{L}_{\mathcal{D}}(h^*) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}^*) + \epsilon \quad \text{and} \quad \Pr_{S \sim \mathcal{D}^n} [\mathcal{A}'(S) = h^*] \geq \rho.$$

Therefore, the infinite class \mathcal{H}^* is agnostically globally stable, contradicting Theorem 3. ■

Appendix B. Theorem 4 (ii): Stability from finite Littlestone dimension

Bun et al. (2020) showed that every class with finite Littlestone dimension has a *globally* stable learner in the *realizable* case.

Theorem 18 (Global Stable Learning from Finite Littlestone Dimension, Bun et al. (2020)) *Suppose $\mathcal{H} \subseteq \{0, 1\}^X$ satisfy $\text{Ldim}(\mathcal{H}) \leq d$. Then there exists a sample complexity $n : (0, 1) \rightarrow \mathbb{N}$ and a learning rule \mathcal{A} such that for every $\epsilon > 0$, and every realizable \mathcal{D} , there exists a hypothesis $h \in \mathcal{H}$ with*

$$\mathcal{L}_{\mathcal{D}}(h) \leq \epsilon \quad \text{and} \quad \Pr_{S \sim \mathcal{D}^n} [\mathcal{A}(S) = h] \geq \frac{1}{(d+1)2^{2d+1}} \quad \text{where } n = n(\epsilon).$$

We show that the algorithm of Bun et al. (2020) is already essentially an agnostic ρ -global stable learner for these classes for some $\rho : (0, 1) \rightarrow (0, 1)$. For a minor technical reason, in the following lemma, we require that the population loss of the distribution \mathcal{D} is bounded away from 1. The constant $2/3$ in the statement of the lemma is quite arbitrary and can be replaced by any larger constant strictly less than 1.

Lemma 19 (Agnostic ρ -Global Stable Learning of Classes with Finite Littlestone Dimension)

Suppose $\mathcal{H} \subseteq \{0, 1\}^X$ satisfy $\text{Ldim}(\mathcal{H}) \leq d$. There exists a learning rule \mathcal{A} , a stability parameter $\rho : (0, 1) \rightarrow (0, 1)$, a sample complexity $n : (0, 1) \rightarrow \mathbb{N}$, such that for every $\epsilon > 0$ and every distribution \mathcal{D} with $\mathcal{L}_{\mathcal{D}}(\mathcal{H}) < \frac{2}{3}$, there exists a hypothesis $h \in \mathcal{H}$ with

$$\mathcal{L}_{\mathcal{D}}(h) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon \quad \text{and} \quad \Pr_{S \sim \mathcal{D}^n} [\mathcal{A}(S) = h] \geq \rho(\epsilon).$$

Moreover, ρ is given explicitly by $\rho(\epsilon) = \frac{1}{(d+1)2^{2d+1}4^{n(\epsilon)}}$.

Proof Let \mathcal{A} be a globally stable learner for \mathcal{H} in the realizable case as guaranteed by Theorem 18, and let $n := n_{18}(\epsilon/2)$ where $n_{18}(\cdot)$ is the sample complexity in Theorem 18. Note that \mathcal{A} is intended for realizable distributions, but our samples come from a non-realizable distribution \mathcal{D} .

Fix a hypothesis $h^* \in \mathcal{H}$ with $\gamma := \mathcal{L}_{\mathcal{D}}(h^*) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \frac{\epsilon}{2} \leq \frac{3}{4}$. Note

$$\Pr_{(x,y) \sim \mathcal{D}} [\mathbf{y} = h^*(x)] = 1 - \gamma,$$

and let \mathcal{D}' be the distribution obtained by conditioning \mathcal{D} on the event that the example is consistent with h^* . We have $\mathcal{L}_{\mathcal{D}'}(h^*) = 0$ and thus \mathcal{D}' is realizable by \mathcal{H} . By our choice of n and Theorem 18, there is a hypothesis $h \in \mathcal{H}$ with

$$\mathcal{L}_{\mathcal{D}'}(h) \leq \frac{\epsilon}{2} \quad \text{and} \quad \Pr_{S \sim (\mathcal{D}')^n} [\mathcal{A}(S) = h] \geq \frac{1}{(d+1)2^{2d+1}}.$$

Since $\gamma \leq \frac{3}{4}$, we have

$$\Pr_{S \sim \mathcal{D}^n} [\mathcal{A}(S) = h] \geq (1 - \gamma)^n \Pr_{S \sim (\mathcal{D}')^n} [\mathcal{A}(S) = h] \geq \frac{(1 - \gamma)^n}{(d+1)2^{2d+1}} \geq \frac{1}{(d+1)2^{2d+1}4^n}.$$

Moreover,

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(h) &= \Pr_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}}[h(\mathbf{x}) \neq \mathbf{y}] \\
&\leq \Pr_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}}[h(\mathbf{x}) \neq \mathbf{y} \mid h^*(\mathbf{x}) = \mathbf{y}] + \Pr_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}}[h^*(\mathbf{x}) \neq \mathbf{y}] \\
&= \mathcal{L}_{\mathcal{D}'}(h) + \mathcal{L}_{\mathcal{D}}(h^*) \leq \frac{\varepsilon}{2} + \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \frac{\varepsilon}{2} \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon.
\end{aligned}$$

■

Proof [Proof of Theorem 4 (ii)] Consider the following learning rule: with equal probability $\frac{1}{3}$, output one of the following hypotheses.

- The hypothesis that always predicts label 1 (the all-1 hypothesis);
- The hypothesis that always predicts label 0 (the all-0 hypothesis);
- The output of the learning rule \mathcal{A} from Theorem 19.

If $\mathcal{L}_{\mathcal{D}}(\mathcal{H}) \geq \frac{2}{3}$, then we have global stability, since at least one of the all-1 or the all-0 hypothesis has population loss at most $\frac{1}{2}$ which is smaller than $\mathcal{L}_{\mathcal{D}}(\mathcal{H})$. Otherwise, the guarantee of Theorem 19 ensures stability. ■

Appendix C. Theorem 9: Stability and List-Replicability are equivalent

In this section, we prove Theorem 9, which establishes the equivalence between stability and list replicability. This result generalizes the equivalence between global stability and global list replicability of Chase et al. (2023).

For the reader's convenience, we recall the statement of the theorem and provide a summary of the proof.

Theorem 9 Consider a parameter $\gamma \in [0, 1]$.

- (i) Given $L : (0, 1) \rightarrow \mathbb{N}$, if a class \mathcal{H} is γ -agnostically L -list replicable, then it is γ -agnostically ρ -global stable for any $\rho : (0, 1) \rightarrow [0, 1]$ satisfying $\rho(\epsilon) < \frac{1}{L(\epsilon)}$ for all $\epsilon \in (0, 1)$.
- (ii) Given $\rho : (0, 1) \rightarrow (0, 1]$, if a class \mathcal{H} is γ -agnostically ρ -global stable, then it is γ -agnostically L -list replicable, for $L(\epsilon) := \left\lfloor \frac{1}{\rho(\epsilon/4)} \right\rfloor$.

The proof of (i) is straightforward. Given $\varepsilon > 0$, let $\delta > 0$ be arbitrary and let $n = n(\epsilon, \delta)$ be the sample complexity of a γ -agnostic L -list-replicable learner for \mathcal{H} . Let \mathcal{D} be a distribution with population loss at most γ , and let $h_1, \dots, h_{L(\epsilon)}$ be the list of hypotheses satisfying Equation (4). At least one of these hypotheses h_i satisfies

$$\Pr_{\mathbf{S} \sim \mathcal{D}^n}[\mathcal{A}(\mathbf{S}) = h_i] \geq \frac{1 - \delta}{L(\epsilon)} \geq \frac{1}{L(\epsilon)} - \delta.$$

Since this statement holds for every $\delta > 0$, \mathcal{H} is γ -agnostically ρ -global stable for all $\rho(\epsilon) < \frac{1}{L(\epsilon)}$.

To prove (ii), we construct an agnostic L -list replicable learner by running the ρ -global stable learning algorithm multiple times. We return any output hypothesis whose empirical loss is close to that of the best output hypothesis, and whose empirical frequency is not much smaller than ρ . This guarantees that we typically output a hypothesis with low population loss and high likelihood of being an output of the globally stable learner. The latter ensures that our output is typically confined to a small list.

Consider an $\epsilon > 0$, and let $\delta > 0$ be any confidence parameter. For the sake of brevity, denote $\rho := \rho(\epsilon/4)$ and $L := L(\epsilon) = \left\lfloor \frac{1}{\rho(\epsilon/4)} \right\rfloor$. Thus, we have $\rho \in \left(\frac{1}{L+1}, \frac{1}{L} \right]$. Let

$$\alpha := \rho - \frac{1}{L+1} > 0.$$

Let $n_0 = n_0(\rho, \epsilon)$ be sufficiently large such that the global stability property holds, namely, for every \mathcal{D} with population loss at most γ , there exists $h^* : X \rightarrow \{0, 1\}$ satisfying

$$\mathcal{L}_{\mathcal{D}}(h^*) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \frac{\epsilon}{4} \text{ and } \Pr_{\mathbf{S} \sim \mathcal{D}^{n_0}}[\mathcal{A}(\mathbf{S}) = h^*] \geq \rho. \quad (15)$$

For any $h \in \{0, 1\}^X$, define

$$p(h) := \Pr_{\mathbf{S} \sim \mathcal{D}^{n_0}}[\mathcal{A}(\mathbf{S}) = h],$$

and consider

$$\Lambda := \left\{ h \in \{0, 1\}^X : p(h) > \frac{1}{L+1} \text{ and } \mathcal{L}_{\mathcal{D}}(h) \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon \right\}.$$

Note that $|\Lambda| \leq L$, and Λ is nonempty, as it contains h^* . It suffices to design a learning rule \mathcal{A}' such that with probability at least $1 - \delta$, it outputs a hypothesis from Λ .

Since $\text{VCdim}(\mathcal{H}) < \infty$, by the uniform convergence property of \mathcal{H} , there exists $n_1 \in \mathbb{N}$ such that for any distribution \mathcal{D} ,

$$\Pr_{\mathbf{Q} \sim \mathcal{D}^{n_1}} \left[\sup_{h \in \mathcal{H}} |\mathcal{L}_{\mathbf{Q}}(h) - \mathcal{L}_{\mathcal{D}}(h)| \leq \frac{\epsilon}{4} \right] \geq 1 - \frac{\delta}{2}. \quad (16)$$

Let $t := t(\alpha, \delta)$ be a sufficiently large integer to be determined later. We propose the following learning rule \mathcal{A}' with sample complexity $tn_0 + n_1$:

1. Sample $\mathbf{S} = (\mathbf{P}, \mathbf{Q}) \sim \mathcal{D}^{tn_0+n_1}$, where $\mathbf{P} = (\mathbf{P}_1, \dots, \mathbf{P}_t) \sim (\mathcal{D}^{n_0})^t = \mathcal{D}^{tn_0}$ and $\mathbf{Q} \sim \mathcal{D}^{n_1}$.
2. For every $i \in [t]$, let $\mathbf{h}_i = \mathcal{A}(\mathbf{P}_i)$. Define the empirical estimate of $p(h)$ as

$$\hat{p}_{\mathbf{S}}(h) := \frac{|\{i \in [t] \mid \mathbf{h}_i = h\}|}{t}.$$

3. Output any hypothesis $h \in \{0, 1\}^X$ that satisfies the following two conditions.

- (a) $\hat{p}_{\mathbf{S}}(h) \geq \rho - \frac{\alpha}{2}$;
- (b) $\mathcal{L}_{\mathbf{Q}}(h) \leq \inf_{h' \in \mathcal{H}} \mathcal{L}_{\mathbf{Q}}(h') + \frac{3\epsilon}{4}$;

If no such h exists, output an arbitrary h corresponding to “failure”.

We show that \mathcal{A}' is a γ -agnostic L -list replicable learner with error at most ϵ . Let \mathcal{D} be any distribution with population loss at most γ .

Claim 5 *We have*

$$\Pr_{\mathbf{S} \sim \mathcal{D}^{tn_0+n_1}} \left[|p(h) - \widehat{p}_{\mathbf{S}}(h)| < \frac{\alpha}{2} \text{ for all } h \in \{0, 1\}^X \right] \geq 1 - \frac{\delta}{2}. \quad (17)$$

Proof We use the uniform convergence property of the family of indicator functions on \mathcal{H} . More precisely, for $f \in \{0, 1\}^X$, define $\mathbf{1}_f : \{0, 1\}^X \rightarrow \{0, 1\}$ as

$$\mathbf{1}_f(f') := \begin{cases} 1 & f' = f \\ 0 & \text{otherwise} \end{cases}.$$

The class

$$\mathcal{I} := \left\{ \mathbf{1}_f : f \in \{0, 1\}^X \right\}$$

has VC dimension 1, and therefore, it satisfies the uniform convergence property. For $\mathbf{S} \sim \mathcal{D}^{tn_0+n_1}$, $\mathcal{A}(\mathbf{S})$ induces a probability distribution μ on $\{0, 1\}^X$, and we have

$$1 - p(h) = \Pr_{\mathbf{S} \sim \mathcal{D}^{n_0}} [\mathcal{A}(\mathbf{S}) \neq h] = \mathcal{L}_{\mu}(\mathbf{1}_h),$$

while $1 - \widehat{p}_{\mathbf{S}}(h)$ corresponds to the empirical loss of $(\mathbf{1}_{h_1}, \dots, \mathbf{1}_{h_t}) \sim \mu^t$. By the uniform convergence property for \mathcal{I} , for sufficiently large $t = t(\alpha, \delta)$, Equation (17) holds. \blacksquare

The following claim completes the proof.

Claim 6 *Consider $\mathbf{S} = (\mathbf{P}, \mathbf{Q}) \sim \mathcal{D}^{tn_0+n_1}$ and let $\mathbf{h} = \mathcal{A}'(\mathbf{S})$.*

$$\Pr[\mathbf{h} \in \Lambda] \geq 1 - \delta.$$

Proof By Equations (16) and (17) and the union bound, with probability at least $1 - \delta$, we have

$$|\mathcal{L}_{\mathbf{Q}}(h) - \mathcal{L}_{\mathcal{D}}(h)| \leq \frac{\epsilon}{4} \text{ for all } h \in \mathcal{H},$$

and

$$|p(h) - \widehat{p}_{\mathbf{S}}(h)| < \frac{\alpha}{2} \text{ for all } h \in \{0, 1\}^X.$$

Let \mathcal{E} denote the event that \mathbf{S} satisfies both these statements. Conditioning on \mathcal{E} , we have

$$\left| \inf_{h' \in \mathcal{H}} \mathcal{L}_{\mathbf{Q}}(h') - \mathcal{L}_{\mathcal{D}}(\mathcal{H}) \right| \leq \frac{\epsilon}{4}, \quad (18)$$

and any $h \in \{0, 1\}^X$ satisfying Conditions 3(a) and 3(b) satisfies

$$p(h) \geq \rho - \frac{\alpha}{2} - \frac{\alpha}{2} > \frac{1}{L+1}.$$

and

$$\mathcal{L}_{\mathcal{D}}(h) \leq \inf_{h' \in \mathcal{H}} \mathcal{L}_{\mathcal{Q}}(h') + \frac{3\epsilon}{4} \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \epsilon.$$

Therefore, all such h belong to Λ .

Finally, let h^* be the hypothesis from Equation (15). We have $\widehat{p}_{\mathcal{S}}(h^*) > \rho - \frac{\alpha}{2}$ and

$$\mathcal{L}_{\mathcal{Q}}(h^*) \leq \mathcal{L}_{\mathcal{D}}(h^*) + \frac{\epsilon}{4} \leq \mathcal{L}_{\mathcal{D}}(\mathcal{H}) + \frac{\epsilon}{4} + \frac{\epsilon}{4} \leq \inf_{h' \in \mathcal{H}} \mathcal{L}_{\mathcal{Q}}(h') + \frac{3\epsilon}{4}.$$

Therefore, h^* satisfies Conditions 3(a) and (b), and the output of \mathcal{A}' will not correspond to “failure”.
 ■