Adversarial Laws of Large Numbers and Optimal Regret in Online Classification

Adversarial Laws of Large Numbers and Optimal Regret in Online Classification

Noga Alon
Princeton University, NJ, USA
& Tel Aviv University, Israel
nalon@math.princeton.edu

Omri Ben-Eliezer
Harvard University
Cambridge, MA, USA
omribene@cmsa.fas.harvard.edu

Yuval Dagan
MIT
Cambridge, MA, USA
dagan@mit.edu

Shay Moran
Technion & Google Research
Haifa, Israel
smoran@technion.ac.il

Moni Naor
Weizmann Institute of Science
Rehovot, Israel
moni.naor@weizmann.ac.il

Eylon Yogev
Boston University, MA, USA
& Tel Aviv University, Israel
eylony@gmail.com

ABSTRACT

Laws of large numbers guarantee that given a large enough sample from some population, the measure of any fixed sub-population is well-estimated by its frequency in the sample. We study laws of large numbers in sampling processes that can affect the environment they are acting upon and interact with it. Specifically, we consider the sequential sampling model proposed by Ben-Eliezer and Yogev (2020), and characterize the classes which admit a uniform law of large numbers in this model: these are exactly the classes that are *online learnable*. Our characterization may be interpreted as an online analogue to the equivalence between learnability and uniform convergence in statistical (PAC) learning.

The sample-complexity bounds we obtain are tight for many parameter regimes, and as an application, we determine the optimal regret bounds in online learning, stated in terms of *Littlestone's dimension*, thus resolving the main open question from Ben-David, Pál, and Shalev-Shwartz (2009), which was also posed by Rakhlin, Sridharan, and Tewari (2015).

CCS CONCEPTS

• **Theory of computation → Streaming models**; **Adversary models**; **Online learning theory**; **Sample complexity and generalization bounds**; **Regret bounds**.

KEYWORDS

random sampling, robust sampling, online learning, Littlestone dimension, adversarial robustness

ACM Reference Format:
Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. 2021. Adversarial Laws of Large Numbers and Optimal Regret in Online Classification. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3406325.3451041

STOC '21, June 21–25, 2021, Virtual, Italy

ACM ISBN 978-1-4503-8053-9/21/06…$15.00
https://doi.org/10.1145/3406325.3451041

1 INTRODUCTION

When analyzing an entire population is infeasible, statisticians apply *sampling methods* by selecting a *sample* of elements from a target population as a guide to the entire population. Thus, one of the most fundamental tasks in statistics is to provide bounds on the sample size that is sufficient to soundly represent the population, and probabilistic tools are used to derive such guarantees, under a variety of assumptions. Virtually all of these guarantees are based on classical probabilistic models assuming that *the target population is fixed in advance and does not depend on the sample collected throughout the process.* Such an assumption, that the setting is *offline* (or *oblivious* or *static*), is however not always realistic. In this work we explore an abstract framework which removes this assumption, and prove that natural and efficient sampling processes produce samples which soundly represent the target population.

Situations where the sampling process explicitly or implicitly affects the target population are abundant in modern data analysis. Consider, for instance, navigation apps that optimize traffic by routing drivers to less congested routes: such apps collect statistics from drivers to estimate the traffic-load on the routes, and use these estimates to guide their users through faster routes. Thus, such apps interact with and affect the statistics they estimate. Consequently, the assumption that the measured populations do not depend on the measurements is not realistic.

Similar issues generally arise in settings involving decision-making in the face of an ever-changing (and sometimes even adversarial) environment; a few representative examples include autonomous driving [42], adaptive data analysis [17, 49], security [32], and theoretical analysis of algorithms [12]. Consequently, there has recently been a surge of works exploring such scenarios, a partial list includes [4, 11, 18–20, 22, 23, 31, 32, 48]. In this work, we focus on the sequential sampling model recently proposed by Ben-Eliezer and Yogev [5].

Organization. We next formally describe the sampling setting and the main question we investigate. Then, in Section 2 we state

our main results. Section 3 contains an overview of the proofs and the main techniques. Finally, Section 4 surveys related work in VC theory, online learning, and streaming algorithms. The formal proofs appear in the complete version of this work, available on arXiv [1].

1.1 The Adversarial Sampling Model

Ben-Eliezer and Yogev [5] model sampling processes over a domain X as a sequential game between two players: a sampler and an adversary. The game proceeds in n rounds, where in each round $i = 1, \ldots, n$:

- The adversary picks an item $x_i \in X$ and provides it to the sampler. The choice of x_i might depend on $x_1, \ldots, x_{i-1}$ and on all information sent to the adversary up to this point.
- Then, the sampler decides whether to add x_i to its sample.
- Finally, the adversary is informed of whether x_i was sampled by the sampler.

The number of rounds n is known in advance to both players.[1] We stress that both players can be randomized, in which case their randomness is private (i.e., not known to the other player).

Oblivious Adversaries. In the oblivious (or static) case, the sampling process consists only of the first two bullets. Equivalently, oblivious adversaries decide on the entire stream in advance, without receiving any feedback from the sampler. Unless stated otherwise, the adversary in this paper is assumed to be adaptive (not oblivious).

Uniform Laws of Large Numbers. Uniform laws of large numbers (ULLN) quantify the minimum sample size which is sufficient to *uniformly estimate multiple statistics of the data.* (Rather than just a *single* statistic, as in standard laws of large numbers.) This is relevant, for instance, in the example given above regarding the navigation app: it is desirable to accurately compute the congestion along *all* routes (paths). Otherwise, one congested route may be regarded as entirely non-congested, and it will be selected for navigation.

Given a family $\mathcal{E}$ of subsets of X, we consider ULLNs that estimate the frequencies of each subset $E \in \mathcal{E}$ within the adversarial stream. Formally, let $\overline{x} = \{x_1, \ldots, x_n\}$ denote the input-stream produced by the adversary, and let $\overline{s} = \{x_{i_1}, \ldots, x_{i_k}\}$ denote the sample chosen by the sampler. The sample $\overline{s}$ is called an *ϵ-approximation of the stream $\overline{x}$ with respect to $\mathcal{E}$* if:

$$(\forall E \in \mathcal{E}) : \quad \left| \frac{|\bar{s} \cap E|}{|\bar{s}|} - \frac{|\bar{x} \cap E|}{|\bar{x}|} \right| \leq \epsilon. \tag{1}$$

That is, $\overline{s}$ is an ϵ-approximation of $\overline{x}$ if the *true-frequencies* $|\bar{x} \cap E|/|\bar{x}|$ are uniformly approximated by the *empirical frequencies* $|\bar{s} \cap E|/|\bar{s}|$. The following question is the main focus of this work:

Question (Main Question). *Given a family $\mathcal{E}$, an error-parameter $\epsilon > 0$, and $k \in \mathbb{N}$, is there a sampler that, given any adversarially-produced input stream $\overline{x}$, picks a sample $\overline{s}$ of at most k items which forms an ϵ-approximation of $\bar{x}$, with high probability?*

[1] Though we will also consider samplers which are oblivious to the number of rounds n.

The Story in the Statistical Setting. It is instructive to compare with the statistical setting in which the sample $\overline{s}$ is drawn independently from an unknown distribution over X (the said distribution is assumed not to change with time, so this is a static/offline/oblivious setting). Here, ULLNs are characterized by the Vapnik-Chervonenkis (VC) Theory which asserts that a family $\mathcal{E}$ satisfies a ULLN if and only if its VC dimension, $\mathrm{VC}(\mathcal{E})$, is finite [45].

This fundamental result became a corner-stone in statistical machine learning. In particular, *The Fundamental Theorem of PAC Learning* states that the following properties are equivalent for any family $\mathcal{E}$: (1) $\mathcal{E}$ satisfies a uniform law of large numbers, (2) $\mathcal{E}$ is PAC learnable, and (3) $\mathcal{E}$ has a finite VC dimension. Quantitatively, the sample size required for both ϵ-approximation and for PAC learning with excess-error ϵ is $\Theta((\mathrm{VC}(\mathcal{E}) + \log(1/\delta))/\epsilon^2)$.

Spoiler: Our main result (stated below) can be seen as an online/adversarial analogue of this theorem where the **Littlestone dimension** replaces the VC dimension.

2 MAIN RESULTS

2.1 Adversarial Laws of Large Numbers

The main result in this paper is a characterization of adversarial uniform laws of large numbers in the spirit of VC theory and The Fundamental Theorem of PAC Learning. We begin with the following central definition.

Definition 2.1 (Adversarial ULLN). *We say that a family $\mathcal{E}$ satisfies an* adversarial ULLN *if for any $\epsilon, \delta > 0$, there exist $k = k(\epsilon, \delta) \in \mathbb{N}$ and a sampler $\mathcal{S}$ satisfying the following. For any adversarially-produced input-stream $\overline{x}$ (of any size), $\mathcal{S}$ chooses a sample of at most k items, which form an ϵ-approximation of $\overline{x}$ with probability at least $1 - \delta$. We denote by $k(\mathcal{E}, \epsilon, \delta)$ the minimal such value of k.*

Note that this definition requires the sample complexity $k = k(\epsilon, \delta)$ to be a constant independent of the stream size n. Another reasonable requirement is $k = o(n)$. It turns out that these two requirements are equivalent.

Which families $\mathcal{E}$ satisfy an adversarial law of large numbers? Clearly, $\mathcal{E}$ must have a finite VC-dimension, as otherwise, basic VC-theory implies that any sampler will fail to produce an ϵ-approximation even against oblivious adversaries which draw the input-stream $\overline{x}$ independently from a distribution on X. However, finite VC dimension is not enough in the fully adversarial setting: [5] exhibit a family $\mathcal{E}$ with $\mathrm{VC}(\mathcal{E}) = 1$ that does not satisfy an adversarial ULLN.

Our first result provides a characterization of adversarial ULLN in terms of *Online Learnability*, which is analogous to the Fundamental Theorem of PAC Learning. In this context, the role of VC dimension is played by the *Littlestone dimension*, a combinatorial parameter which captures online learnability similar to how the VC dimension captures PAC learnability. (See the appendix for the formal definition.)

Theorem 2.2 (Adversarial ULLNs – Qualitative Characterization). *Let $\mathcal{E}$ be a family of subsets of X. Then, the following statements are equivalent:*

(1) $\mathcal{E}$ satisfies an adversarial ULLN;

(2) $\mathcal{E}$ is online learnable; and
(3) $\mathcal{E}$ has a finite Littlestone dimension.

Our quantitative upper bound for the sample-complexity $k(\mathcal{E}, \epsilon, \delta)$, which is the main technical contribution of this paper, is stated next.

Theorem 2.3 (Adversarial ULLNs – Quantitative Characterization). *Let $\mathcal{E}$ be a family with Littlestone dimension d. Then, the sample size $k(\mathcal{E}, \epsilon, \delta)$, which suffices to produce an ϵ-approximation satisfies:*

$$k(\mathcal{E}, \epsilon, \delta) \leq O\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right).$$

The above upper bound is realized by natural and efficient samplers; for example it is achieved by: (i) the *Bernoulli sampler* $\mathrm{Ber}(n, p)$ which retains each element with probability $p = k/n$; (ii) the *uniform sampler* $\mathrm{Uni}(n, k)$ that draws a subset $I \subseteq [n]$ uniformly at random from all the subsets of size k and selects the sample $\{x_t : t \in I\}$; and (iii) the *reservoir sampler* $\mathrm{Res}(n, k)$ (see the appendix) that maintains a uniform sample continuously throughout the stream.

2.1.1 Lower Bounds. The upper bound in Theorem 2.3 cannot be improved in general. In particular, it is tight in all parameters for *oblivious samplers*: a sampler is called oblivious if the indices of the chosen subsample are independent of the input-stream. (The Bernoulli, Reservoir, and Uniform samplers are of this type.) A lower bound of $\Omega((d + \log(1/\delta))/\epsilon^2)$ for oblivious samplers directly follows from VC-theory, and applies to any family $\mathcal{E}$ for which the VC dimension and Littlestone dimension are of the same order.[2] For unrestricted samplers we obtain bounds of $\Omega(d/\epsilon^2)$ for ϵ-approximation and $\Omega(d \log(1/\epsilon)/\epsilon)$ for ϵ-nets. We state these results and prove them in the full version of this work [1].

The above lower bound proofs hold for specific "hard" families $\mathcal{E}$. This is in contrast with the statistical or oblivious settings in which a lower bound of $\Omega((\mathrm{VC}(\mathcal{E}) + \log(1/\delta))/\epsilon^2)$ applies to any class. We do not know whether an analogous result holds in the adversarial sampling setting and leave it as an open problem. We do show, however, that the linear dependence in d is necessary for any $\mathcal{E}$, as part of proving Theorem 2.2.

2.2 Online Learning

We continue with our main application to online learning. Consider the setting of online prediction with binary labels; a learning task in this setting can be described as a guessing game between a learner and an adversary. The game proceeds in rounds $t = 1, \ldots, T$, each consisting of the following steps:

- The adversary selects $(x_t, y_t) \in X \times \{0, 1\}$ and reveals x_t to the learner.
- The learner provides a prediction $\hat{y}_t \in \{0, 1\}$ of y_t and announces it to the adversary.
- The adversary announces y_t to the learner.

The goal is to minimize the number of mistakes, $\sum_t \mathbb{1}(y_t \neq \hat{y}_t)$. Given a class $\mathcal{E}$, the *regret* of the learner w.r.t. $\mathcal{E}$ is defined as the difference between the number of mistakes made by the learner and the number of mistakes made by the best $E \in \mathcal{E}$:

$$\sum_t \mathbb{1}(y_t \neq \hat{y}_t) - \min_{E \in \mathcal{E}} \sum_t \mathbb{1}\Big(y_t \neq \mathbb{1}(x_t \in E)\Big).$$

A class $\mathcal{E}$ is *online-learnable* if there exists an online learner whose (expected) regret w.r.t. every adversary is at most $R(T)$, where $R(T) = o(T)$. (The amortized regret $R(T)/T$ vanishes as $T \to \infty$.) Ben-David, Pál, and Shalev-Shwartz [3] proved that for every class $\mathcal{E}$, the optimal regret $R_T(\mathcal{E})$ satisfies

$$\Omega(\sqrt{d \cdot T}) \leq R_T(\mathcal{E}) \leq O(\sqrt{d \cdot T \log T}), \tag{2}$$

where d is the Littlestone dimension of $\mathcal{E}$, and left closing that gap as their main open question. Subsequently, Rakhlin, Sridharan, and Tewari [35–37] defined the notion of *Sequential Rademacher Complexity*, proved that it captures regret bounds in online learning in a general setting, and used it to re-derive (2). They also asked as an open question whether the logarithmic factor in (2) can be removed and pointed on difficulties to achieve this using some known techniques [33, 37].

We show that the sequential Rademacher complexity also captures the sample-complexity of ϵ-approximations and bound it in the proof of Theorem 2.3. This directly implies a tight bound on online learning:

Theorem 2.4 (Tight Regret Bounds in Online Learning). *Let $\mathcal{E}$ be a class with Littlestone dimension d. Then the optimal regret bound in online learning $\mathcal{E}$ is $\Theta(\sqrt{d \cdot T})$.*

The lower bound was shown by [3]. We prove the upper bound in the full version of this work [1].

2.3 Applications and Extensions

We next discuss applications and extensions of our results.

Epsilon Nets. We also provide sample complexity bounds for producing *ϵ-nets*: a subsample $\overline{s}$ of the stream $\overline{x}$ is an ϵ-net if whenever $E \in \mathcal{E}$ satisfies $|E \cap \overline{x}| \geq \epsilon n$, then $\overline{s} \cap E \neq \emptyset$. I.e. the subsample $\overline{s}$ hits every $E \in \mathcal{E}$ which contains at least an ϵ-fraction of the items in the stream.

Epsilon nets are a fundamental primitive in computational geometry and in learning theory. In computational geometry this notion underlies fundamental algorithmic techniques, and in learning theory it is tightly linked to the learnability in the *realizable* setting. In that sense, it is analogous to ϵ-approximations, which correspond to learnability in the *agnostic* setting.

In the full version [1] we show that, as with ϵ-approximations, ϵ-nets are also characterized by the Littlestone dimension; and similarly, our results here provide tight sample-complexity bounds.

Maintaining An ϵ-Approximation Continuously. Some natural applications require that the sampler continuously maintains an ϵ-approximation with respect to the prefix of the stream observed thus-far. To address such scenarios we slightly modify the adversarial sampling setting by allowing the sampler to delete items from its sample. In this modified setting, we prove that the classical *Reservoir sampler* [47], $\mathrm{Res}(n, k)$, enjoys similar guarantees to those of Theorem 2.3 above. Concretely, the exact same bound of Theorem 2.3 is achieved by reservoir sampling if one is only

[2] E.g., projective spaces, Hamming balls, lines in the plane, and others.

interested in ϵ-approximation at the end of the process; for continuous ϵ-approximation, the same bound with an added term of $O(\log\log(n))$ in the numerator suffices. These results are presented and proved in the full version of this work [1].

Notably, allowing deletions does not add significant power to the sampler, and in particular Theorem 2.2 still applies in this setting.

ALLNs for Real-Valued Function Classes. The adversarial sampling setting naturally extends to real-valued function classes $\mathcal{E}$. Moreover, much of the machinery developed in this paper readily applies in this case. In particular, the relationship with the sequential Rademacher complexity is retained. Therefore, since the sequential Rademacher complexity captures regret bounds in online learning, this allows an automatic translation of regret bounds from online learning to sample complexity bounds in adversarial ULLNs w.r.t. real-valued function classes.[3] See [8] for a very recent follow-up work further exploring and extending this connection.

Algorithmic Applications. Part of the reason that the Fundamental Theorem of PAC Learning became a corner-stone in machine learning theory is due to its algorithmic implications. In particular, because it justifies the *Empirical Risk Minimization Principle* (ERM), which asserts that in order to learn a VC class, it suffices to minimize the empirical loss w.r.t. a random sample. This principle reduces the learning problem (of minimizing the loss w.r.t. an unknown distribution) to an optimization problem of minimizing the loss w.r.t. the (known) input sample.

It will be interesting to explore such implications in the adversarial setting. One promising direction is to use these sampling methods to design *lazy streaming/online algorithms*. That is, algorithms that update their internal state only on a small (random) substream. Intuitively, if that substream represents the entire stream appropriately, then the performance of the algorithm will be satisfactory, and the gain in efficiency can be significant. In fact, our proof of Theorem 2.3 identifies and exploits such a phenomenon in online learning: we use a *lazy online learner* that updates its predictor rarely, only in a small random subsample of examples.

3 TECHNICAL OVERVIEW

We next overview the technical parts in this work. We outline the proofs of the main theorems, and try to point out which technical arguments are novel, and which are based on known techniques. A more detailed overview of particular proofs is given in the dedicated sections.

3.1 Upper Bounds

We begin with the sample-complexity upper bound, Theorem 2.3 (which is the longest and most technical derivation in this work).

Reductions Between Samplers. Our goal is to derive an upper bound for the Bernoulli, uniform, and reservoir samplers. In order to abstract out common arguments, we develop a general framework which serves to methodically transform sample-complexity bounds between the different samplers via a type of "online reductions". This framework allows us to bound the sample-complexity with respect to one sampler, and *automatically* deduce them for the other samplers. The reduction relies on transforming one sampling scheme into another in an online fashion, and from a technical perspective, this boils down to coupling arguments, similar to coupling techniques in Markov Chains processes [26]. The full version of this work [1] contains a more detailed overview followed by the formal derivations.

Upper Bounds for The Uniform Sampler. Thus, for the rest of this overview we focus the sampling scheme to be the uniform sampler which uniformly draws a k-index-set $I \subseteq [n]$, and selects the subsample $\overline{x}_I = (x_i : i \in I)$. Our goal is to show that with probability $\geq 1-\delta$,

$$\sup_{E\in\mathcal{E}} \left| \frac{|\overline{x}_I \cap E|}{k} - \frac{|\overline{x} \cap E|}{n} \right| \leq O\Big(\sqrt{\frac{d+\log(1/\delta)}{k}}\Big), \tag{3}$$

where d is the Littlestone dimension of $\mathcal{E}$ and $\overline{x}$ is the *adversarially* produced sequence. The proof consists of two main steps which are detailed below.

3.1.1 Step 1: Reduction to Online Discrepancy via Double Sampling. The first step in the proof consists of an online variant of the celebrated *double-sampling argument* due to [45]. This argument serves to replace the error w.r.t. the entire population by the error w.r.t. a small *test-set* of size k, thus effectively restricting the domain to the $2k$ items in the union of the selected sample and the test-set. In more detail, let $J \subseteq [n]$ be a uniformly drawn *ghost* subset of size k which is disjoint from I, and is not known to the adversary. Consider the maximal deviation between the sample $\overline{x}_I$ and the "test-set" $\overline{x}_J$:

$$\sup_{E\in\mathcal{E}} \left| \frac{|\overline{x}_I \cap E|}{k} - \frac{|\overline{x}_J \cap E|}{k} \right|. \tag{4}$$

The argument proceeds by showing that for a typical J, the deviation w.r.t. the entire population $\overline{x}$ in the LHS of (3) has the same order of magnitude like the deviation w.r.t. the test-set $\overline{x}_J$ in (4) above. Hence, it suffices to bound (4).

In order to bound (4), consider sampling I, J according to the following process: (i) First sample the $2k$ indices in $I \cup J$ uniformly from $[n]$, and reveal these $2k$ indices to both players (in advance). (ii) Then, the sampler draws I from these $2k$ indices in an online fashion (i.e., the adversary does not know in advance the sample I). Intuitively, this modified process only helps the adversary who has the additional information of a superset of size $2k$, which contains I. *What we gain is that the modified process is essentially equivalent to reducing the horizon from n to $2k$.* The case of $n = 2k$ can be interpreted as an online variant of the well-studied *Combinatorial Discrepancy* problem, which is described next.

Online Combinatorial Discrepancy. The online discrepancy game w.r.t. $\mathcal{E}$ is a sequential game played between a painter and an adversary which proceeds as follows: at each round $t = 1, \ldots, 2k$ the adversary places an item x_t on the board, and the painter colors x_t in either red or blue. The goal of the painter is that each set in $\mathcal{E}$ will be colored in a balanced fashion; i.e., if we denote by I the set of

[3] The reduction from bounds on ϵ-approximations to bounds on the sequential Rademacher complexity appear in the full version of this work [1]. They rely on concentration inequalities for $\{0,1\}$ valued random variables that have analogues for $[0,1]$ valued random variables with the same guarantees. This enables a direct extension of this reduction.

indices of items colored red, her goal is to minimize the discrepancy

$$\mathrm{Disc}_{2k}(\mathcal{E}, \overline{x}, I) := \max_{E \in \mathcal{E}} \left| |\overline{x}_I \cap E| - |\overline{x}_{[2k]\setminus I} \cap E| \right|.$$

One can verify that minimizing the discrepancy is equivalent to minimizing (4). Moreover, each of the samplers $\mathrm{Ber}(2k, 1/2)$ and $\mathrm{Uni}(2k, k)$ corresponds to natural coloring strategies of the painter; in particular, $\mathrm{Uni}(2k, k)$ colors a random subset of k of the items in red (and the rest in blue.) Thus, we focus now on analyzing the performance of $\mathrm{Uni}(2k, k)$ in the online discrepancy problem.

3.1.2 Step 2: From Online Discrepancy to Sequential Rademacher. Instead of analyzing the discrepancy of $\mathrm{Uni}(2k, k)$, it will be more convenient to consider the discrepancy of $\mathrm{Ber}(2k, 1/2)$, which colors each item in red/blue uniformly and independently of its previous choices. Towards this end, we show that these two strategies are essentially equivalent, using the reduction framework described at the beginning of this section.

The discrepancy of $\mathrm{Ber}(2k, 1/2)$ connects directly to the *Sequential Rademacher Complexity* [34], defined as the expected discrepancy $\mathrm{Rad}_{2k}(\mathcal{E}) = \mathbb{E}\mathrm{Disc}_{2k}(\mathcal{E}, \overline{x}, I)$, where the expectation is taken according to a uniformly drawn $I \subseteq [2k]$. (Which is precisely the coloring strategy of $\mathrm{Ber}(2k, 1/2)$.)

3.1.3 Step 3.1: Bounding Sequential Rademacher Complexity – Oblivious Case. In what follows, it is convenient to set $n = 2k$. Our goal here is to bound $\mathrm{Rad}_n(\mathcal{E}) \le O(\sqrt{d \cdot n})$. As a prelude, it is instructive to consider the oblivious setting where the items $x_1, \ldots, x_n$ are fixed in advance, before they are presented to the painter. Here, the analysis is exactly as in the standard i.i.d. setting, and the sequential Rademacher complexity becomes the standard Rademacher complexity. Consider the following three approaches, in increasing level of complexity.

First Approach: a Union Bound. Assume $\mathcal{E}$ is finite. Then, for each $E \in \mathcal{E}$ it is possible to show by concentration inequalities that with high probability, the discrepancy $||\overline{x}_I \cap E| - |\overline{x}_{[n]\setminus I} \cap E||$ is small. By applying a union bound over all $E \in \mathcal{E}$, one can derive that $\mathrm{Rad}_n(\mathcal{E}) \le O(\sqrt{n \log |\mathcal{E}|})$.

Second Approach: Sauer-Shelah-Parles Lemma. Since $\mathcal{E}$ can be very large or even infinite, the bound in the previous attempt may not suffice. An improved argument relies on the celebrated Sauer-Shelah-Perles (SSP) Lemma [39], which asserts that the number of distinct intersection-patterns of sets in $\mathcal{E}$ with $\{x_1, \ldots, x_n\}$ is at most $\binom{n}{\le \mathrm{VC}(\mathcal{E})} \le O(n^{\mathrm{VC}(\mathcal{E})})$. The proof then follows by union bounding the discrepancy over $\{\overline{x} \cap E \colon E \in \mathcal{E}\}$, resulting in a bound of

$$O\left(\sqrt{n \log\left(n^{\mathrm{VC}(\mathcal{E})}\right)}\right) \le O\left(\sqrt{\mathrm{VC}(\mathcal{E}) n \log n}\right),$$

which is off only by a factor of $\sqrt{\log n}$.

Third Approach: Using Approximate Covers and Chaining. Shaving the extra logarithmic factor is a non-trivial task which was achieved in the seminal work by Talagrand [43] using a technique called *chaining* [16]. It relies on the notion of *approximate covers*:

Definition 3.1 (Approximate Covers). *A family C is an ϵ-cover of $\mathcal{E}$ with respect to $x_1, \ldots, x_n$ if for every $E \in \mathcal{E}$ there exists $C \in C$ such that E and C agree on all but at most $\epsilon \cdot n$ of the x_i's.*

In a nutshell, the chaining approach starts by finding covers $C_0, C_1, \ldots$ where C_i is a 2^{-i}-cover for $\mathcal{E}$ w.r.t. $\overline{x}$, then writing the telescopic sum

$$\mathrm{Disc}_n(\mathcal{E}, \overline{x}, I) = \mathrm{Disc}_n(C_0, \overline{x}, I) + \sum_{i=1}^{\infty} (\mathrm{Disc}_n(C_i, \overline{x}, I) - \mathrm{Disc}_n(C_{i-1}, \overline{x}, I))$$

and bounding each summand using a union bound.

Note that the SSP Lemma provides a bound of $|C| \le \binom{n}{\le \mathrm{VC}(\mathcal{E})}$ in the case of $\epsilon = 0$, where d is the VC-dimension of $\mathcal{E}$. For $\epsilon > 0$, a classical result by Haussler [24] asserts that every family admits an ϵ-cover of size $(1/\epsilon)^{O(d)}$. The latter bound allows via chaining to remove the redundant logarithmic factor and bound $\mathrm{Rad}_n(\mathcal{E}) \le O(\sqrt{\mathrm{VC}(\mathcal{E}) n})$.

3.1.4 Step 3.2: Bounding Sequential Rademacher Complexity – Adversarial Case. We are now ready to outline the last and most technical step in this proof. Our goal is twofold: first, we discuss how previous work [3, 35] generalized the above arguments to the adversarial (or the online learning) model, culminating in a bound of the form $\mathrm{Rad}_n(\mathcal{E}) = O(\sqrt{dn \log n})$. Then, we describe the proof approach for our improved bound of $O(\sqrt{dn})$.

An $O(\sqrt{dn \log n})$ Bound via Adaptive SSP. First, the union bound approach generalizes directly to the adversarial setting. However, the second approach, via the SSP lemma, does not. The issue is that in the adversarial setting, the stream $\overline{x}$ can depend on the coloring that the painter chooses, and hence $\{E \cap \{x_1, \ldots, x_n\} : E \in \mathcal{E}\}$ depends on the coloring as well. In particular, it is not possible to apply a union bound over a small number of such patterns. Moreover, it is known that a non-trivial bound depending only on the VC dimension and n does not exist [36]. To overcome this difficulty we use an adaptive variant of the SSP Lemma due to [3], which is based on the following notion:

Definition 3.2 (Dynamic Sets). *A dynamic set $\mathbb{B}$ is an online algorithm that operates on a sequence $\overline{x} = (x_1, \ldots, x_n)$. At each time $t = 1, \ldots, n$, the algorithm decides whether to retain x_t as a function of $x_1, \ldots, x_t$. Let $\mathbb{B}(\overline{x})$ denote the set of elements retained by $\mathbb{B}$ on a sequence $\overline{x}$.*[4]

Ben-David, Pál, and Shalev-Shwartz [3] proved that any family $\mathcal{E}$ whose Littlestone dimension is d can be covered by $\binom{n}{\le d}$ dynamic sets. That is, for every n there exists a family C of $\binom{n}{\le d}$ dynamic sets such that for every sequence $\overline{x} = (x_1, \ldots, x_n)$ and for every $E \in \mathcal{E}$ there exists a dynamic set $\mathbb{B} \in C$ which agrees with E on the sequence $\overline{x}$, namely, $\mathbb{B}(\overline{x}) = E \cap \overline{x}$.

Using this adaptive SSP Lemma, one can proceed to bound the discrepancy as in the oblivious case by applying a union bound over the $(2k)^d$ dynamic sets, and bounding the discrepancy with respect to each dynamic set using Martingale concentration bounds. Implementing this reasoning yields a bound of $\mathrm{Rad}_n(\mathcal{E}) \le O(\sqrt{dn \log n})$ which is off by a logarithmic factor.

[4] [3] refers to dynamic-sets as experts, which is compatible with the terminology of online learning.

Removing the Logarithmic Factor. To adapt the chaining argument to the adversarial setting we first need to find small ϵ-covers. This raises the following question:

Can every Littlestone family be ϵ-covered by $\epsilon^{-O(d)}$ dynamic sets?

Unfortunately, we cannot answer this question and leave it for future work. In fact, [37] identified a variant of this question as a challenge towards replicating the chaining proof in the online setting. To circumvent the derivation of dynamic approximate covers, we introduce a fractional variant which we term *fractional-covers*. It turns out that any Littlestone family admits "small" approximate fractional covers and these can be used to complete the chaining argument.

Definition 3.3 (Approximate Fractional-Covers). *A probability measure μ over dynamic sets $\mathbb{B}$ is called an* (ϵ, γ)-fractional cover *for $\mathcal{E}$ if for any $\overline{x} = (x_1, \ldots, x_n)$ and any $E \in \mathcal{E}$,*

$$\mu\left(\{\mathbb{B} \colon E \text{ and } \mathbb{B}(\overline{x}) \text{ agree on all but at most } \epsilon n \text{ of the } x_i s\}\right) \geq 1/\gamma.$$

The parameter γ should be thought of as the size of the cover. Observe that fractional-covers are relaxations of covers: indeed, if C is an ϵ-cover for $\mathcal{E}$ then the uniform distribution over C is an (ϵ, γ)-fractional cover for $\mathcal{E}$ with $\gamma = |C|$.

Small Approximate Fractional-Covers Exist. In the full version of this work [1] we prove that every Littlestone family $\mathcal{E}$ admits an (ϵ, γ)-fractional cover of size

$$\gamma = (O(1)/\epsilon)^d.$$

This fractional cover is essentially a mixture of non-fractional covers for subsets of the sequence $\overline{x}$ of size d/ϵ. In more detail, the distribution over dynamic sets is defined by the following two-step sampling process: (1) draw a uniformly random subset $\overline{s}$ of $\overline{x}$ of size d/ϵ, and let C_s denote the (non-fractional) cover of $\mathcal{E}$ with respect to $\overline{s}$, which is promised by the dynamic variant of the SSP-Lemma. (2) Draw $\mathbb{B}$ from the uniform distribution over C_s.

We outline the proof that this is an (ϵ, γ)-fractional cover with $\gamma = O(1/\epsilon)^d$. Fixing E and $\overline{x}$, our goal is to show that with probability at least $1/\gamma$ over μ, the drawn $\mathbb{B}$ agrees with E on all but at most $\epsilon \cdot n$ elements of $\overline{x}$. This relies of the following two arguments: (1) For every $\overline{s}$ there exists $\mathbb{B}_{\overline{s}} \in C_s$ that agrees with E on $\overline{s}$; and (2) it can be shown that with high probability over the selection of the subset $\overline{s}$, $\mathbb{B}_{\overline{s}}$ agrees with E on all but at most ϵn of the stream $\overline{x}$. We call such values of $\overline{s}$ as *good*, and conclude from the two steps above:

$$\begin{aligned}
&\Pr_{\mathbb{B}\sim\mu}[\mathbb{B} \text{ agrees with } E \text{ on } (1-\epsilon)n \text{ of the } x_i\text{'s}] \\
&\quad \geq \Pr[\overline{s} \text{ is good}] \Pr_{\mathbb{B}\sim\text{uniform}(C_s)}[\mathbb{B} = \mathbb{B}_{\overline{s}}] \\
&\quad \geq \frac{1}{2} \cdot \frac{1}{|C_s|} \geq \frac{1}{2\binom{d/\epsilon}{\leq d}} \geq \Omega(\epsilon)^d \geq \frac{1}{\gamma} \ .
\end{aligned}$$

We further comment on the proof that $\overline{s}$ is *good* with high probability: the proof relies on analyzing a *lazy* online learner that updates its internal state only once encountering elements from $\overline{s}$. We show that if $\overline{s}$ is drawn uniformly, then with high probability such a learner will make $\leq \epsilon \cdot n$ mistakes and this will imply that w.h.p. $\mathbb{B}_{\overline{s}}$ agrees with E on $(1-\epsilon)n$ stream elements.

Chaining with Fractional Covers: Challenges and Subtleties. Here, we discuss how approximate fractional covers are used to bound the sequential Rademacher complexity. We do so by describing how to modify the bound that uses 0-covers to use $(0, \gamma)$-fractional covers instead. Recall that this argument goes by two steps: (1) bounding the discrepancy for each dynamic set in the cover, and (2) arguing by a union bound that, with high probability the discrepancies of *all* dynamic sets in the cover are bounded. In comparison, with fractional covers, the second step is modified to: (2') arguing that with high probability (over the random coloring), the discrepancies of *nearly all* the dynamic sets are bounded. In particular, if more than a $(1-\gamma)$-fraction of the dynamic sets have bounded discrepancies, then the discrepancies of all sets in $\mathcal{E}$ are bounded. Indeed, this follows since every $E \in \mathcal{E}$ is covered by at least a γ-fraction of the dynamic-sets, and therefore, the pigeonhole principle implies that at least one such dynamic set also has bounded discrepancy, and hence E has bounded discrepancy as well.

We note that multiple further technicalities are required to generalize the chaining technique for fractional covers and refer the reader to the full version of this work [1] for a short overview of this method followed by its adaptation to the adversarial setting.

3.2 Lower Bounds

Beyond the $\Omega((d + \log(1/\delta))/\epsilon^2)$ lower bound for oblivious samplers, which follows immediately from the VC literature, we prove several non-trivial lower bounds in other contexts. We distinguish between two types of approaches used to derive our lower bounds, described below. As the proofs are shorter than those of the upper bounds and more self-contained, we omit the exact technical details of the proofs in this overview and refer the reader to the full version [1].

Universal Lower Bound by Adversarial Arguments. The main lower bound in [5] exhibits a separation between the static and adversarial setting by proving an adversarial lower bound for the family of one-dimensional *thresholds*. We identify that their proof implicitly constructs a tree as in the definition of the Littlestone dimension, and generalize their argument to derive an $\Omega(d)$ lower bound for *all* families of Littlestone dimension d.

Lower Bounds on the Minimum Sizes of ϵ-Approximations/Nets. These lower bounds actually exhibit a much stronger phenomenon, showing that small ϵ-approximations/nets *do not exist* for some families $\mathcal{E}$. Thus, obviously, these cannot be captured by a sample of the same size.

It is natural to seek lower bounds of this type in the VC-literature. The main challenge is that many of the known lower bounds apply for geometric VC classes whose Littlestone dimension is unbounded. To overcome this, we present two lower bounds where Ldim can be controlled: one for ϵ-approximation, which carefully analyzes a simple randomized construction, and another for ϵ-nets, which combines intersection properties of lines in the projective plane with probabilistic arguments.

4 RELATED WORK

4.1 VC Theory

As suggested by the title, the results presented by this work are inspired by uniform laws of large numbers in the statistical i.i.d. setting and in particular by VC theory. (A partial list of basic manuscripts on this subject include [15, 44–46].) Moreover, the established equivalence between online learning and adversarial laws of large numbers is analogous to the the equivalence between PAC learning and uniform laws of large numbers in the i.i.d. setting. (See e.g. [9, 10, 40, 41, 45].) From a technical perspective, our approach for deriving sample complexity upper bound is based on the chaining technique [13, 14, 16], which was analogously used to establish optimal sample complexity bounds in the statistical setting [43]. (The initial bounds by [45] are off by a $\log(1/\epsilon)$ factor.)

From the lower bound side, our proofs are based on ideas originated from combinatorial discrepancy and ϵ approximations. (E.g., [30]; see the book by Matoušek [29] for a text-book introduction.)

4.2 Online Learning

The first works in online learning can be traced back to [6, 7, 21, 38]. In terms of learning binary functions, Littlestone's dimension was first proposed in [27] to characterize online learning in the realizable (noiseless) setting. The agnostic (noisy) setting was first proposed by [24] in the statistical model and later extended to the online setting by [28] who studied function-classes of bounded cardinality and then by [3] and [35] who provided both upper and lower bounds with only a logarithmic gap.

We note that Rakhlin, Sridharan, and Tewari [35–37], in the same line of work that proved the equivalence between online learning and sequential Rademacher complexity, analyzed uniform martingales laws of large numbers in the context of online learning. These laws of large numbers are conceptually different from ours: roughly, they assert uniform concentration of certain properties of martingales, where the uniformity is over a given family of martingales. In particular, in contrast with our work, there is no aspect of sub-sampling in these laws. Below, we compare their techniques to those of this paper:

- [35] used a symmetrization argument to reduce from Martingale quantities relating to online learning to the Rademacher complexity. This does not reduce the effective sample size, which is what we achieve using the double sampling argument.
- [35] developed methods suitable for analyzing the sequential Rademacher complexity. In particular, they developed a notion of covering numbers that is generally more powerful than the *non-fractional* cover that uses dynamic sets, which was developed by [3] and was the baseline for our analysis. Yet, obtaining tight bound on the sequential Rademacher of Littlestone classes remained open.
- Reductions between sampling schemes did not appear in the above work as they did not study sampling.

4.3 Streaming Algorithms

The streaming model of computation is useful when analyzing massive datasets [2]. There is a wide variety of algorithms for solving different tasks. One common method that is useful for various approximation tasks in streaming is random sampling. To approximate a function f, each element is sampled with some small probability p, and at the end, the function f is computed on the sample. For tasks such as computing a center point of a high-dimensional dataset, where the objective is (roughly speaking) preserved under taking an ϵ-approximation, this can result in improved space complexity and running time. Motivated by streaming applications, Ben-Eliezer and Yogev [5] proposed the adversarial sampling model that we study in this paper, and proved preliminary bounds on it. Their main result, a weaker quantitative analogue of our Theorem 2.3, is an upper bound of $O((\log(|\mathcal{E}|)+\log(1/\delta))/\epsilon^2)$ for any finite family $\mathcal{E}$.

Streaming algorithms in the adversarial setting is an emerging topic that is not well understood. Hardt and Woodruff [22] showed that linear sketches are inherently *non-robust* and cannot be used to compute the Euclidean norm of its input (where in the static setting they are used mainly for this reason). Naor and Yogev [32] showed that Bloom filters are susceptible to attacks by an adversarial stream of queries. Kaplan et al. [25] constructed a streaming problem naturally inspired by the adaptive data analysis literature, which exhibits a large separation between the space complexities in the adversarial and oblivious regimes. On the positive side, several recent works [4, 23, 48] present generic compilers that transform non-robust randomized streaming algorithms into efficient adversarially robust ones, for various classical problems such as distinct elements counting and F_p-sampling, among others.

ACKNOWLEDGMENTS

Noga Alon is supported in part by National Science Foundation (NSF) grant DMS-1855464, US-Israel Binational Science Foundation (BSF) grant 2018267, and by the Simons Foundation. Research partially conducted while Omri Ben-Eliezer was at Weizmann Institute of Science, supported in part by a Israel Science Foundation (ISF) grant no. 950/15. Shay Moran is a Robert J. Shillman Fellow and was supported in part by the ISF (grant No. 1225/20), by an Azrieli Faculty Fellowship, and by BSF grant 2018385. Moni Naor is Supported in part by ISF grants (no. 950/15 and 2686/20) and by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. Incumbent of the Judith Kleeman Professorial Chair. Eylon Yogev is supported in part by ISF grants 484/18, 1789/19, Len Blavatnik and the Blavatnik Foundation, and The Blavatnik Interdisciplinary Cyber Research Center at Tel Aviv University.

A BASIC DEFINITIONS

For completeness, we formally define the Littlestone dimension and the sampling procedures discussed in this paper (in the upper bound context; the lower bounds are for any sampler). See also Section 5 in the full version [1].

Littlestone Dimension. Let X be a domain and let $\mathcal{E}$ be a family of subsets of X. The definition of the *Littlestone Dimension* [27], denoted $\mathsf{Ldim}(\mathcal{E})$, is given using mistake-trees: these are binary decision trees whose internal nodes are labelled by elements of X. Any root-to-leaf path corresponds to a sequence of pairs $(x_1, y_1), \ldots, (x_d, y_d)$, where x_i is the label of the i'th internal node in the path, and $y_i = 1$ if the $(i+1)$'th node in the path is the right

child of the i'th node, and otherwise $y_i = 0$. We say that a tree T is shattered by $\mathcal{E}$ if for any root-to-leaf path $(x_1, y_1), \ldots, (x_d, y_d)$ in T there is $E \in \mathcal{E}$ such that $x_i \in R \iff y_i = +1$, for all $i \leq d$. $\mathrm{Ldim}(\mathcal{E})$ is the depth of the largest complete tree shattered by $\mathcal{E}$, with the convention that $\mathrm{Ldim}(\emptyset) = -1$.

Sampling Algorithms. Our results are achieved by three simple and commonly used sampling procedures: Bernoulli sampling, uniform sampling, and reservoir sampling.

- *Bernoulli sampling*: $\mathrm{Ber}(n, p)$ samples the element arriving in each round $i \in [n]$ independently with probability p.
- *Uniform sampling*: $\mathrm{Uni}(n, k)$ randomly draws k indices $1 \leq i_1 < \ldots < i_k \leq n$ and samples the elements arriving at rounds $i_1, \ldots, i_k$. (Note that the uniform sampler can be implemented efficiently in an online way: after i rounds, the probability that the next element x_{i+1} will be sampled depends only on i, n, and the number of elements sampled so far.)
- *Reservoir sampling*: $\mathrm{Res}(n, k)$ [47] maintains a sample of size k at all times using insertions and deletions: the first k elements are always added to the sample, and for any $i > k$, with probability k/i the element arriving in round i is added to the sample while one of the existing elements (picked uniformly) is removed from the sample.

REFERENCES

[1] Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. 2021. Adversarial Laws of Large Numbers and Optimal Regret in Online Classification. *arXiv preprint arXiv:2101.09054* (2021). https://arxiv.org/abs/2101.09054

[2] Noga Alon, Yossi Matias, and Mario Szegedy. 1999. The Space Complexity of Approximating the Frequency Moments. *J. Comput. System Sci.* 58, 1 (1999), 137–147. https://doi.org/10.1006/jcss.1997.1545

[3] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. 2009. Agnostic Online Learning. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*.

[4] Omri Ben-Eliezer, Rajesh Jayaram, David P. Woodruff, and Eylon Yogev. 2020. A Framework for Adversarially Robust Streaming Algorithms. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*. 63–80. https://doi.org/10.1145/3375395.3387658

[5] Omri Ben-Eliezer and Eylon Yogev. 2020. The Adversarial Robustness of Sampling. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*. 49–62. https://doi.org/10.1145/3375395.3387643

[6] David Blackwell. 1954. Controlled random walks. In *Proceedings of the International Congress of Mathematicians*, Vol. 3. 336–338.

[7] David Blackwell. 1956. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6, 1 (1956), 1–8. https://doi.org/10.2140/pjm.1956.6.1

[8] Adam Block, Yuval Dagan, and Sasha Rakhlin. 2021. Majorizing Measures, Sequential Complexities, and Online Learning. *arXiv preprint arXiv:2102.01729* (2021).

[9] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM* 36 (1989), 929–965. https://doi.org/10.1145/76359.76371

[10] Olivier Bousquet. 2004. Introduction to Statistical Learning Theory. In *Advanced lectures on machine learning*. Vol. 3176. Springer, 169–207. https://doi.org/10.1007/978-3-540-28650-9_8

[11] Yeshwanth Cherapanamjeri and Jelani Nelson. 2020. On Adaptive Distance Estimation. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

[12] Timothy Chu, Yu Gao, Richard Peng, Sushant Sachdeva, Saurabh Sawlani, and Junxing Wang. 2018. Graph Sparsification, Spectral Sketches, and Faster Resistance Computation, via Short Cycle Decompositions. In *Proceedings of the 59th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. 361–372. https://doi.org/10.1109/FOCS.2018.00042

[13] Richard M. Dudley. 1973. Sample Functions of the Gaussian Process. *The Annals of Probability* 1, 1 (1973), 66–103. https://doi.org/10.1214/aop/1176997026

[14] Richard M. Dudley. 1978. Central Limit Theorems for Empirical Measures. *The Annals of Probability* 6, 6 (1978), 899–929. https://doi.org/10.1214/aop/1176995384

[15] Richard M. Dudley. 1984. A course on empirical processes. In *École d'Été de Probabilités de Saint-Flour XII - 1982*, P. L. Hennequin (Ed.). Springer Berlin Heidelberg, 1–142. https://doi.org/10.1007/BFb0099432

[16] Richard M. Dudley. 1987. Universal Donsker Classes and Metric Entropy. *The Annals of Probability* 15, 4 (1987), 1306–1326. https://doi.org/10.1214/aop/1176991978

[17] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349, 6248 (2015), 636–638. https://doi.org/10.1126/science.aaa9375

[18] Anna C. Gilbert, Brett Hemenway, Atri Rudra, Martin J. Strauss, and Mary Wootters. 2012. Recovering simple signals. In *Information Theory and Applications Workshop (ITA)*. 382–391. https://doi.org/10.1109/ITA.2012.6181772

[19] Anna C. Gilbert, Brett Hemenway, Martin J. Strauss, David P. Woodruff, and Mary Wootters. 2012. Reusable low-error compressive sampling schemes through privacy. In *IEEE Statistical Signal Processing Workshop (SSP)*. 536–539. https://doi.org/10.1109/SSP.2012.6319752

[20] Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. 2020. Smoothed Analysis of Online and Differentially Private Learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

[21] James Hannan. 1957. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games* 3 (1957), 97–139.

[22] Moritz Hardt and David P. Woodruff. 2013. How robust are linear sketches to adaptive inputs?. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*. 121–130. https://doi.org/10.1145/2488608.2488624

[23] Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Yossi Matias, and Uri Stemmer. 2020. Adversarially Robust Streaming Algorithms via Differential Privacy. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

[24] David Haussler. 1992. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation* 100, 1 (1992), 78 – 150. https://doi.org/10.1016/0890-5401(92)90010-D

[25] Haim Kaplan, Yishay Mansour, Kobbi Nissim, and Uri Stemmer. 2021. Separating Adaptive Streaming from Oblivious Streaming. *arXiv preprint arXiv:2101.10836* (2021).

[26] David A. Levin and Yuval Peres. 2017. *Markov chains and mixing times*. Vol. 107. American Mathematical Society. https://doi.org/10.1090/mbk/107

[27] Nick Littlestone. 1988. Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm. *Machine Learning* 2, 4 (1988), 285–318. https://doi.org/10.1023/A:1022869011914

[28] Nick Littlestone and Manfred K. Warmuth. 1994. The weighted majority algorithm. *Information and computation* 108, 2 (1994), 212–261. https://doi.org/10.1006/inco.1994.1009

[29] Jiří Matoušek. 2009. *Geometric Discrepancy: An Illustrated Guide*. Springer-Verlag Berlin Heidelberg. 847 pages. https://doi.org/10.1007/978-3-642-03942-3

[30] Jiří Matoušek, Emo Welzl, and Lorenz Wernisch. 1993. Discrepancy and approximations for bounded VC-dimension. *Combinatorica* 13, 4 (1993), 455–466. https://doi.org/10.1007/BF01303517

[31] Ilya Mironov, Moni Naor, and Gil Segev. 2011. Sketching in Adversarial Environments. *SIAM J. Comput.* 40, 6 (2011), 1845–1870. https://doi.org/10.1137/080733772

[32] Moni Naor and Eylon Yogev. 2019. Bloom Filters in Adversarial Environments. *ACM Transactions on Algorithms* 15, 3 (2019), 35. https://doi.org/10.1145/3306193

[33] Alexander Rakhlin and Karthik Sridharan. 2014. Statistical learning and sequential prediction. *Book Draft* (2014).

[34] Alexander Rakhlin and Karthik Sridharan. 2015. On Martingale Extensions of Vapnik–Chervonenkis Theory with Applications to Online Learning. In *Measures of Complexity*. Springer, 197–215. https://doi.org/10.1007/978-3-319-21852-6_15

[35] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. 2010. Online learning: Random averages, combinatorial parameters, and learnability. In *Advances in Neural Information Processing Systems*. 1984–1992. https://doi.org/10.5555/2997046.2997117

[36] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. 2015. Online learning via sequential complexities. *J. Mach. Learn. Res.* 16, 1 (2015), 155–186.

[37] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. 2015. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields* 161, 1-2 (2015), 111–153. https://doi.org/10.1007/s00440-013-0545-5

[38] Herbert Robbins. 1951. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California.

[39] Norbert Sauer. 1972. On the Density of Families of Sets. *Journal of Combinatorial Theory, Series A* 13, 1 (1972), 145–147.

[40] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning*. Cambridge university press. https://doi.org/10.1017/CBO9781107298019

[41] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. 2010. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research* 11 (2010), 2635–2670. https://doi.org/10.5555/1756006.1953019

[42] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. 2018. DARTS: Deceiving Autonomous Cars with Toxic Signs. *CoRR* abs/1802.06430 (2018).

[43] Michel Talagrand. 1994. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability* 22, 1 (1994), 28–76. https://doi.org/10.1214/aop/1176988847
[44] Vladimir N. Vapnik. 1998. *Statistical Learning Theory.* John Wiley & Sons, Inc.
[45] Vladimir N. Vapnik and Alexey Y. Chervonenkis. 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications* 16, 2 (1971), 264–280. https://doi.org/10.1137/1116025
[46] Vladimir N. Vapnik and Alexey Y. Chervonenkis. 1974. *Theory of Pattern Recognition.* Nauka, Moscow.
[47] Jeffrey S. Vitter. 1985. Random Sampling with a Reservoir. *ACM Trans. Math. Software* 11, 1 (1985), 37–57. https://doi.org/10.1145/3147.3165
[48] David P. Woodruff and Samson Zhou. 2020. Tight Bounds for Adversarially Robust Streams and Sliding Windows via Difference Estimators. *arXiv preprint arXiv:2011.07471* (2020).
[49] Blake E. Woodworth, Vitaly Feldman, Saharon Rosset, and Nati Srebro. 2018. The Everlasting Database: Statistical Validity at a Fair Price. In *Advances in Neural Information Processing Systems 31 (NeurIPS).* 6532–6541. https://doi.org/10.5555/3327757.3327760