Adversarial Laws of Large Numbers and Optimal Regret in Online Classification

Noga Alon* Omri Ben-Eliezer† Yuval Dagan‡
Shay Moran§ Moni Naor¶ Eylon Yogev||

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).

*Department of Mathematics, Princeton University, Princeton, New Jersey, USA and Schools of Mathematics and Computer Science, Tel Aviv University, Tel Aviv, Israel. Research supported in part by NSF grant DMS-1855464, BSF grant 2018267 and the Simons Foundation. Email: nalon@math.princeton.edu.

†Center for Mathematical Sciences and Applications, Harvard University, Massachusetts, USA. Research partially conducted while the author was at Weizmann Institute of Science, supported in part by a grant from the Israel Science Foundation (no. 950/15). Email: omribene@cmsa.fas.harvard.edu.

‡Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Email: dagan@mit.edu.

§Department of Mathematics, Technion, Israel. Email: smoran@technion.ac.il. Research supported in part by the Israel Science Foundation (grant No. 1225/20), by an Azrieli Faculty Fellowship, and by a grant from the United States - Israel Binational Science Foundation (BSF).

¶Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Supported in part by grants from the Israel Science Foundation (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. Email: moni.naor@weizmann.ac.il.

||Department of Computer Science, Boston University and Department of Computer Science, Tel Aviv University. Email: eylony@gmail.com. Research 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.

Contents

1 Introduction **1**
1.1 The Adversarial Sampling Model 1

2 Main Results **2**
2.1 Adversarial Laws of Large Numbers 2
2.1.1 Lower Bounds 3
2.2 Online Learning 4
2.3 Applications and Extensions 4

3 Technical Overview **6**
3.1 Upper Bounds 6
3.1.1 Step 1: Reduction to Online Discrepancy via Double Sampling 6
3.1.2 Step 2: From Online Discrepancy to Sequential Rademacher 7
3.1.3 Step 3.1: Bounding Sequential Rademacher Complexity – Oblivious Case . . 7
3.1.4 Step 3.2: Bounding Sequential Rademacher Complexity – Adversarial Case . 8
3.2 Lower Bounds 10

4 Related Work **11**
4.1 VC Theory 11
4.2 Online Learning 11
4.3 Streaming Algorithms 12

5 Preliminaries **12**
5.1 Basic Definitions: Littlestone Dimension and Sampling Schemes 12
5.2 Notation 13
5.3 Additional Central Definitions 14
5.4 Sampling Without Replacement 14

6 Epsilon Approximations **15**

7 Epsilon Nets **16**

8 Double Sampling **18**
8.1 Proof of Lemma 6.2 20
8.2 Proof of Lemma 7.2 21

9 Covering Numbers **21**
9.1 Overview 21
9.1.1 Epsilon Approximation and Sequential Rademacher 22
9.1.2 Epsilon Nets 23
9.1.3 Organization 23
9.2 Covering for Littlestone Classes 24
9.2.1 Proof of Lemma 9.3 24
9.2.2 Proof of Lemma 9.5 27
9.3 Deriving Bounds on ϵ-Approximation via Fractional Covering Numbers 30

9.3.1 Basic Lemmas for Deterministic Covers and Proof of Lemma 9.6 30
9.3.2 Basic Lemmas for Fractional Covers 31
9.3.3 Chaining for Non-Fractional Covers 32
9.3.4 Proof of Lemma 9.7 33
9.4 Bounds on ϵ-Nets via Fractional Covering Numbers 37

10 Reductions Between Different Sampling Schemes **38**
10.1 Intuition for the Reduction Method 38
10.2 Abstract Reduction Method 39
10.3 Bounds for Reservoir Sampling via Uniform Sampling 41
10.4 Bounds for Bernoulli Sampling via Uniform Sampling 43
10.5 Bounds for Uniform Sampling via Bernoulli Sampling 44
10.5.1 Proof of Lemma 6.3 45
10.5.2 Proof of Lemma 7.3 49

11 Continuous ϵ-Approximation **50**

12 Online Learning **51**
12.1 Formal Defintions 51
12.2 Statement and Proof 52

13 Lower Bounds **52**
13.1 Proofs 53

A Probabilistic Material **59**
A.1 Filtration and Martingales 59
A.2 Sampling Without Replacement 60

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 which 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 which assume 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 [SBM+18], adaptive data analysis [DFH+15, WFRS18], security [NY15], and theoretical analysis of algorithms [CGP+18]. Consequently, there has recently been a surge of works exploring such scenarios, a partial list includes [MNS11, GHR+12, GHS+12, HW13, NY15, BJWY20, CN20, HRS20, HKM+20, WZ20]. In this work, we focus on the sequential sampling model recently proposed by Ben-Eliezer and Yogev [BEY20].

1.1 The Adversarial Sampling Model

We next formally describe the sampling setting and the main question we investigate. Ben-Eliezer and Yogev [BEY20] 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).

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

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:

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

That is, $\overline{s}$ is an ϵ-approximation of $\overline{x}$ if the *true-frequencies* $|\overline{x} \cap E|/|\overline{x}|$ are uniformly approximated by the *empirical frequencies* $|\overline{s} \cap E|/|\overline{s}|$. 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 $\overline{x}$, with high probability?*

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. 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 [VC71].

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: [BEY20] 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 Section 5.1 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.*

The proof follows from Theorems 2.3 and 13.1 (and from the well-known equivalence between online learnability and finite Littlestone dimension [Lit88, BPS09]). 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 Section 2.3) 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 $\Omega(d/\epsilon^2)$

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

for ϵ-approximation and $\Omega(d \log(1/\epsilon)/\epsilon)$ for ϵ-nets. We state these results and prove them in Section 13.

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 [BPS09] 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 [RST10, RST15a, RST15b] 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 Equation (2). They also asked as an open question whether the logarithmic factor in Equation (2) can be removed and pointed on difficulties to achieve this using some known techniques [RS14, RST15b].

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: (See Section 12 for more details.)

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 [BPS09]. We prove the upper bound in Section 12.

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 $\bar{s}$ of the stream $\bar{x}$ is an ϵ-net if whenever $E \in \mathcal{E}$ satisfies $|E \cap \bar{x}| \geq \epsilon n$, then $\bar{s} \cap E \neq \emptyset$. I.e. the subsample $\bar{s}$ 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 Section 7 we show that, like 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* [Vit85], $\mathrm{Res}(n,k)$ (see Section 5 for the precise definition), 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 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 (see Theorem 11.1).

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]

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 in an appropriate way, then the performance of the algorithm will be satisfactory, and the gain in efficiency can be significant. In fact, our proof of Lemma 9.5 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]The reduction from bounds on ϵ-approximations to bounds on the sequential Rademacher complexity appear in Section 6. 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.

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 [LP17]. Section 10 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 [VC71]. 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 Equation (3) has the same order of magnitude like the deviation w.r.t. the test-set $\overline{x}_J$ in Equation (4) above. Hence, it suffices to bound (4).

In order to bound Equation (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 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 Equation (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* [RS15], 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 $\mathrm{Rad}_n(\mathcal{E}) \leq O(\sqrt{d \cdot n})$. 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 $\mathrm{Rad}_n(\mathcal{E}) \leq O(\sqrt{n \log |\mathcal{E}|})$.

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 [Sau72], which asserts that the number of distinct intersection-patterns of sets in $\mathcal{E}$ with $\{x_1, \ldots, x_n\}$ is at most $\binom{n}{\leq \mathrm{VC}(\mathcal{E})} \leq O(n^{\mathrm{VC}(\mathcal{E})})$. The proof then follows by union bounding the discrepancy over $\{\overline{x} \cap E \colon E \in \mathcal{E}\}$, resulting in a bound of

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

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 [Tal94] using a technique called *chaining* [Dud87]. It relies on the notion of *approximate covers*:

Definition 3.1 (Approximate Covers)**.** *A family $\mathcal{C}$ is an ϵ-cover of $\mathcal{E}$ with respect to $x_1, \ldots, x_n$ if for every $E \in \mathcal{E}$ there exists $C \in \mathcal{C}$ 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 $\mathcal{C}_0, \mathcal{C}_1, \ldots$ where $\mathcal{C}_i$ is a 2^{-i}-cover for $\mathcal{E}$ w.r.t. $\overline{x}$, then writing the telescopic sum

$$\mathrm{Disc}_n(\mathcal{E}, \overline{x}, I) = \mathrm{Disc}_n(\mathcal{C}_0, \overline{x}, I) + \sum_{i=1}^{\infty} (\mathrm{Disc}_n(\mathcal{C}_i, \overline{x}, I) - \mathrm{Disc}_n(\mathcal{C}_{i-1}, \overline{x}, I))$$

and bounding each summand using a union bound.

Note that the SSP Lemma provides a bound of $|\mathcal{C}| \leq \binom{n}{\leq \mathrm{VC}(\mathcal{E})}$ in the case of $\epsilon = 0$, where d is the VC-dimension of $\mathcal{E}$. For $\epsilon > 0$, a classical result by Haussler [Hau92] 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 $\mathrm{Rad}_n(\mathcal{E}) \leq O(\sqrt{\mathrm{VC}(\mathcal{E}) n})$.

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 [BPS09, RST10] 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 [RST15a]. To overcome this difficulty we use an adaptive variant of the SSP Lemma due to [BPS09], 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 [BPS09] proved that any family $\mathcal{E}$ whose Littlestone dimension is d can be covered by $\binom{n}{\leq d}$ dynamic sets. That is, for every n there exists a family $\mathcal{C}$ of $\binom{n}{\leq d}$ 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 $\mathbb{B} \in \mathcal{C}$ 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 $\mathrm{Rad}_n(\mathcal{E}) \leq O(\sqrt{dn \log n})$ which is off by a logarithmic factor.

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 $(1/\epsilon)^{O(d)}$ dynamic sets?

Unfortunately, we cannot answer this question and leave it for future work. In fact, [RST15b] 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 $\mathcal{C}$ is an ϵ-cover for $\mathcal{E}$ then the uniform distribution over $\mathcal{C}$ is an (ϵ, γ)-fractional cover for $\mathcal{E}$ with $\gamma = |\mathcal{C}|$.

Small Approximate Fractional-Covers Exist. We prove Lemma 9.5 which asserts 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 $\mathcal{C}_s$ 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 $\mathcal{C}_s$.

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

[4] [BPS09] refers to dynamic-sets as experts, which is compatible with the terminology of online learning.

exists $\mathbb{B}_{\bar{s}} \in \mathcal{C}_s$ that agrees with E on $\bar{s}$; and (2) it can be shown that with high probability over the selection of the subset $\bar{s}$, $\mathbb{B}_{\bar{s}}$ agrees with E on all but at most ϵn of the stream $\bar{x}$. We call such values of $\bar{s}$ as *good*, and conclude from the two steps above that

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

We further comment on the proof that $\bar{s}$ is *good* with high probability: the proof relies on analyzing a *lazy* online learner that updates its internal state only once encountering elements from $\bar{s}$. We show that if $\bar{s}$ 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. $\mathbb{B}_{\bar{s}}$ agrees with E on $(1-\epsilon)n$ stream elements. We refer the reader to Section 9.2.2 for the proof.

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 Section 9.3 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 Section 13.

Universal Lower Bound by Adversarial Arguments. The main lower bound in [BEY20] 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. For more details, see Theorem 13.1.

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. For more details, see Theorems 13.2 and 13.3 respectively.

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 [VC71, VC74, Dud84, Vap98].) 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. [VC71, BEHW89, Bou04, SSSSS10, SSBD14].) From a technical perspective, our approach for deriving sample complexity upper bound is based on the chaining technique [Dud73, Dud78, Dud87], which was analogously used to establish optimal sample complexity bounds in the statistical setting [Tal94]. (The initial bounds by [VC71] 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., [MWW93]; see the book by Matoušek [Mat09] for a text-book introduction.)

4.2 Online Learning

The first works in online learning can be traced back to [Rob51, Bla56, Bla54, Han57]. In terms of learning binary functions, Littlestone's dimension was first proposed in [Lit88] to characterize online learning in the realizable (noiseless) setting. The agnostic (noisy) setting was first proposed by [Hau92] in the statistical model and later extended to the online setting by [LW94] who studied function-classes of bounded cardinality and then by [BPS09] and [RST10] who provided both upper and lower bounds with only a logarithmic gap.

We note that Rakhlin, Sridharan, and Tewari [RST10, RST15a, RST15b], 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:

- [RST10] 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.

- [RST10] developed methods 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 [BPS09] 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 [AMS99]. 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 [BEY20] 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 [HW13] 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 [NY15] showed that Bloom filters are susceptible to attacks by an adversarial stream of queries. On the positive side, several recent works [BJWY20, HKM$^+$20, WZ20] 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.

5 Preliminaries

5.1 Basic Definitions: Littlestone Dimension and Sampling Schemes

Littlestone Dimension Let X be a domain and let $\mathcal{E}$ be a family of subsets of X. The definition of the *Littlestone Dimension* [Lit88], 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 $i \leq d$. $\mathsf{Ldim}(\mathcal{E})$ is the depth of the largest complete tree shattered by $\mathcal{E}$, with the convention that $\mathsf{Ldim}(\emptyset) = -1$. See Figure 1 for an illustration.

Sampling Algorithms Our results are achieved by three of the simplest and most commonly used sampling procedures: Bernoulli sampling, uniform sampling, and reservoir sampling.

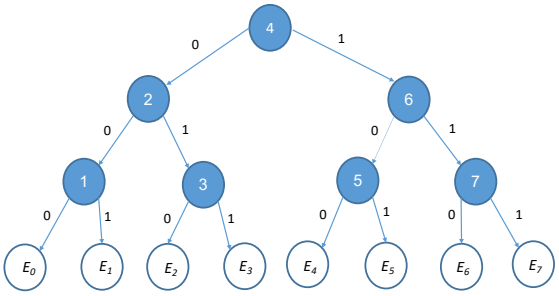

Figure 1: A tree shattered by the class $\mathcal{E}$ containing all thresholds $E_i \subseteq \{1, 2, \ldots, 7\}$, where $E_i = \{1, \ldots i\}$.

- **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 $i_1, \ldots, i_k$.[5]
- **Reservoir sampling**: $\mathrm{Res}(n, k)$ [Vit85] 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.

5.2 Notation

- **Random variables** are denoted in a bold font.
- **Universal constants:** let $C, c, C', \ldots$ denote universal numerical constants, that are independent of the problem parameters. Further, the values of these constants can change from the left-hand side to the right-hand side of some inequalities. We use $C_0, C_1, \ldots$ to denote universal constants whose values are fixed.
- **Sampling schemes:** We use $\boldsymbol{I}$ to denote the set of indices of elements sampled by the algorithm if they are sampled from a scheme with no deletions (e.g. Bernoulli or uniform sampling), and use $\overline{\boldsymbol{I}} = (\boldsymbol{I}_1, \ldots, \boldsymbol{I}_n)$ to denote a sampling scheme with deletions, where $\boldsymbol{I}_j$ is the set of indices of elements retained after round $j \in [n]$.
- **Adversaries:** We denote the set of adversaries that generate a stream of size n by Adv_n and commonly denote adversaries by $\mathcal{A}$. We assume the adversary to be deterministic: since the sampler has a fixed strategy, we can make this assumption without limiting the generality of the theorem.
- **The stream and subsets of it:** Let $\overline{x} = (x_1, \ldots, x_n)$ denote the stream where $x_i \in X$. We set $\overline{x}(\mathcal{A}, I)$ to denote the stream presented by the adversary when the sampler samples elements indexed by I. Notice that $\overline{x}(\mathcal{A}, I)_t$ depends only on $\mathcal{A}$ and $I \cap [t-1]$, since the t-th stream element is presented before the adversary knows if element t is added to the

[5]Note that the uniform sampler can be implement 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.

sample. Given a subset $J \subseteq [n]$ we let $\overline{x}_J = \{x_j : j \in J\}$. By abuse of notation, we may use $\overline{x}$ also to denote the multiset $\{x_1, \ldots, x_n\}$, allowing operations such as set intersection.

5.3 Additional Central Definitions

We define the central notions used in this proof, starting with the approximation rate in ϵ-approximations:

Definition 5.1. *Given an arbitrary family $\mathcal{E}$ over a domain X, an adversary $\mathcal{A} \in \mathrm{Adv}_n$ and a subset $I \subseteq [n]$, let $\overline{x} = \overline{x}(\mathcal{A}, I)$ and define*

$$\mathrm{App}_{\mathcal{A},I}(\mathcal{E}) = \mathrm{App}_{\mathcal{A},I} := \max_{E \in \mathcal{E}} \left| \frac{|E \cap \overline{x}|}{n} - \frac{|E \cap \overline{x}_I|}{k} \right|.$$

Secondly, define the notion of online discrepancy:

Definition 5.2. *Let $\mathcal{E}$ denote an arbitrary family over X, let $\mathcal{A} \in \mathrm{Adv}_n$, let $I \subseteq [n]$ and let $\overline{x} = \overline{x}(\mathcal{A}, I)$. The online discrepancy is defined by:*

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

The sequential Rademacher complexity is just the expected discrepancy:

Definition 5.3. *The* sequential Rademacher complexity *is defined as*

$$\mathrm{Rad}_T(\mathcal{E}) := \mathbb{E}_{\boldsymbol{I} \sim \mathrm{Ber}(n,1/2)}[\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E})].$$

The next definition is used in the proof for ϵ-nets. It defines an indicator to whether there exist $E \in \mathcal{E}$ that is well represented in the stream but not sufficiently represented in the sample.

Definition 5.4. *Fix $n, \overline{m}, \underline{m} \in \mathbb{N}$ such that $0 \leq \underline{m} \leq \overline{m} \leq n$, let $\mathcal{A} \in \mathrm{Adv}_n$ and $I \subseteq [n]$. Denote $\overline{x} = \overline{x}(\mathcal{A}, I)$. Define*

$$\mathrm{Net}^n_{\mathcal{A},I,\overline{m},\underline{m}}(\mathcal{E}) = \mathrm{Net}^n_{\mathcal{A},I,\overline{m},\underline{m}} = \begin{cases} 1 & \exists E \in \mathcal{E},\ |\overline{x} \cap E| \geq \overline{m} \text{ and } |\overline{x}_I \cap E| \leq \underline{m} \\ 0 & \text{otherwise} \end{cases}$$

Notice that $\mathrm{Net}^n_{\mathcal{A},I,\epsilon n,0}$ is an indicator to whether $\overline{x}_I$ fails to be an ϵ-net for $\overline{x}$.

5.4 Sampling Without Replacement

Here we present technical probabilistic lemmas for sampling without replacement that are used in the proof. The proofs of these lemmas appear in Appendix A.2.

For a sample $\boldsymbol{I} \sim \mathrm{Uni}(n,k)$ chosen without replacement, one would like to estimate the size of the intersection of $\boldsymbol{I}$ with any fixed set $U \subseteq [n]$. The next lemma bounds the variance of the intersection:

Lemma 5.5. *Let n, k such that $n \geq k$, let $U \subseteq [n]$, and let $\boldsymbol{I} \sim \mathrm{Uni}(n,k)$. Then, $\mathbb{E}[|U \cap \boldsymbol{I}|] = |U|k/n$ and $\mathrm{Var}(|U \cap \boldsymbol{I}|) \leq |U|k/n$.*

Further, exponential tail bounds can also be obtained:

Lemma 5.6 ([Cha05],[BM15]). *Let $n, k \in \mathbb{N}$ such that $n \geq k$. Let $U \subseteq [n]$. Then, the following holds:*

1. *For any $t \geq 0$,*
$$\Pr_{\boldsymbol{I} \sim \mathrm{Uni}(n,k)} \left[\left| \frac{|\boldsymbol{I} \cap U|}{k} - \frac{|U|}{n} \right| \geq t \right] \leq 2 \exp\left(-2t^2 k\right).$$
2. *For any $\alpha \in [0,1]$,*
$$\Pr_{\boldsymbol{I} \sim \mathrm{Uni}(n,k)} \left[\left| \frac{|\boldsymbol{I} \cap U|}{k} - \frac{|U|}{n} \right| \geq \alpha \frac{|U|}{n} \right] \leq 2 \exp\left(-\frac{\alpha^2 k |U|}{6n}\right).$$

6 Epsilon Approximations

Below, we prove Theorem 2.3. We start with a more formal statement of the theorem:

Theorem 6.1. *Let $\mathcal{E}$ denote a family of Littlestone dimension d, let $\delta, \epsilon \in (0, 1/2)$, $n \in \mathbb{N}$, $\mathcal{A} \in \mathrm{Adv}_n$ and $p \in [0,1]$. Define $k = \lfloor np \rfloor$. If $n \geq 3k$ then, for any $\delta > 0$,*
$$\Pr_{\boldsymbol{I}} \left[\mathrm{App}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E}) \geq C \sqrt{\frac{d + \log(1/\delta)}{k}} \right] \leq \delta,$$
where $\boldsymbol{I}$ is drawn either from $\mathrm{Uni}(n,k)$, $\mathrm{Ber}(n,p)$ or $\mathrm{Res}(n,k)$ and $C > 0$ is a universal constant.

Note that the requirement $n \geq 3k$ is merely technical; when n is smaller than that, one can just add all n elements to the sample and obtain a 0-approximation trivially.

Here we prove that the *uniform sample* $\mathrm{Uni}(n,k)$ is an ϵ-approximation and in Section 10 we show a reduction to Bernoulli and reservoir sampling. The bound for the uniform sampler consists of three steps (see Section 3). The first step utilizes the double sampling argument, to bound the approximation error $\mathrm{App}_{\mathcal{A},I}$ in terms of the discrepancy corresponding to a sampler $\mathrm{Uni}(2k,k)$:

Lemma 6.2. *Let $\mathcal{E}$ denote an arbitrary family of subsets from some universe. Fix $k, n \in \mathbb{N}$ such that $n \geq 2k$. Then, for any $t \geq 0$ and $\delta \in (0,1)$,*
$$\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I} \sim \mathrm{Uni}(n,k)} [\mathrm{App}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E}) > t] \leq 2 \max_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}' \sim \mathrm{Uni}(2k,k)} \left[\mathrm{Disc}_{\mathcal{A}',\boldsymbol{I}'}(\mathcal{E}) > tk - \sqrt{Ck} \right]. \tag{5}$$

The technique of double sampling is presented in Section 8 and the proof of Lemma 6.2 is in Section 8.1.

Applying Lemma 6.2, we are left with bounding $\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}}$ where $\mathcal{A} \in \mathrm{Adv}_{2k}$ and $\boldsymbol{I} \sim \mathrm{Uni}(2k,k)$. However, it is easier to analyze a random sample $\boldsymbol{I}' \sim \mathrm{Ber}(2k, 1/2)$ due to the independence of the coordinates. Hence, we prove the following lemma:

Lemma 6.3. *Let $\mathcal{E}$ denote an arbitrary family. Then, for any $t > 0$ and $\delta > 0$,*
$$\max_{\mathcal{A} \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I} \sim \mathrm{Uni}(2k,k)} [\mathrm{Disc}_{\mathcal{A},I}(\mathcal{E}) > t] \leq \max_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}' \sim \mathrm{Ber}(2k,1/2)} \left[\mathrm{Disc}_{\mathcal{A}',I'}(\mathcal{E}) > t - \sqrt{Ck \log(1/\delta)} \right] + \delta. \tag{6}$$

The technique of reducing between sampling schemes is in Section 10 and the proof of Lemma 6.3 is in Section 10.5.1.

The last step is to bound $\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}}$ for a sample $\boldsymbol{I}$ that is drawn Bernoulli $1/2$, for classes of bounded Littlestone dimension.

Lemma 6.4. *Let $\mathcal{E}$ be of Littleston dimension d and let $\mathcal{A} \in \mathrm{Adv}_{2k}$. Then, for all $\delta > 0$,*

$$\Pr_{\boldsymbol{I}\sim\mathrm{Ber}(2k,1/2)}\left[\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E}) > C\sqrt{k(d+\log(1/\delta))}\right] \leq \delta.$$

Lemma 6.4 is proved in Section 9.1.1. Combining Lemma 6.2, Lemma 6.3 and Lemma 6.4, We conclude that the uniform sample as an ϵ-approximation, as summarized below:

Theorem 6.5. *Let $\mathcal{E}$ denote a family of Littlestone dimension d, let $\delta, \epsilon \in (0, 1/2)$, $n \in \mathbb{N}$, $\mathcal{A} \in \mathrm{Adv}_n$ and $k \in \mathbb{N}$ such that $n \geq 2k$. Then, for any $\delta > 0$,*

$$\Pr_{\boldsymbol{I}\sim\mathrm{Uni}(n,k)}\left[\mathrm{App}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E}) \geq C\sqrt{\frac{d+\log(1/\delta)}{k}}\right] \leq \delta,$$

where $C > 0$ is a universal constant.

7 Epsilon Nets

We prove that the three sampling schemes discussed in this paper sample ϵ-nets with high probability, as stated below:

Theorem 7.1. *Let $\mathcal{E}$ denote a family of Littlestone dimension d, let $\epsilon \in (0, 1/2)$, $n \in \mathbb{N}$, $\mathcal{A} \in \mathrm{Adv}_n$ and $p \in [0, 1]$. Define $k = \lfloor np \rfloor$. If $n \geq 3k$ then*

$$\Pr_{\boldsymbol{I}}[\mathrm{Net}^n_{\mathcal{A},\boldsymbol{I},\epsilon n,0}(\mathcal{E}) = 1] \leq \left(\frac{Ck}{d}\right)^d \exp(-c\epsilon k)$$

where $\boldsymbol{I}$ is drawn either $\mathrm{Uni}(n,k)$, $\mathrm{Ber}(n,p)$ or $\mathrm{Res}(n,k)$ and $C > 0$ is a universal constant.

In this section, we prove Theorem 7.1 for the uniform sampler. Reductions to the other sampling schemes are given in Section 10.

Below the main lemmas are presented, starting with a reduction from a stream of size n to $2k$, via the technique of double sampling, presented in Section 8:

Lemma 7.2. *Let $\mathcal{E}$ be some family and $n, k \in \mathbb{N}$ be integers such that $n \geq 2k$ and $k \geq C/\epsilon$. Then,*

$$\max_{\mathcal{A}\in\mathrm{Adv}_n} \Pr_{\boldsymbol{I}\sim\mathrm{Uni}(n,k)}\left[\mathrm{Net}^n_{\mathcal{A},\boldsymbol{I},\epsilon\cdot n,0}(\mathcal{E}) = 1\right] \leq 2 \max_{\mathcal{A}'\in\mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'\sim\mathrm{Uni}(2k,k)}\left[\mathrm{Net}^{2k}_{\mathcal{A}',\boldsymbol{I}',\epsilon/4\cdot 2k,0}(\mathcal{E}) = 1\right].$$

The proof appears in Section 8.2. It would be desirable to replace the uniform sample $\mathrm{Uni}(2k,k)$ with a Bernoulli sample, since it selects each coordinate independently. In particular, we will show that the probability of $\mathrm{Uni}(2k,k)$ to fail to be an ϵ-net is bounded in terms of the probability of $\mathrm{Ber}(2k,1/8)$. Intuitively, this follows from the fact that a sample drawn $\mathrm{Uni}(2k,k)$ nearly contains a sample $\mathrm{Ber}(2k,1/8)$ in some sense. The formal statement is below:

Lemma 7.3. *Let $\mathcal{E}$ denote some family, $\epsilon \in (0,1)$, and $k \in \mathbb{N}$. Then,*

$$\max_{\mathcal{A}\in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}\sim \mathrm{Uni}(2k,k)}[\mathrm{Net}^{2k}_{\mathcal{A},\boldsymbol{I},\epsilon\cdot 2k,0}(\mathcal{E}) = 1] \\ \leq \max_{\mathcal{A}'\in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'\sim \mathrm{Ber}(2k,1/8)}[\mathrm{Net}^{2k}_{\mathcal{A}',\boldsymbol{I}',\epsilon\cdot 2k,\epsilon/16\cdot 2k}(\mathcal{E}) = 1] + 2\exp(-c\epsilon k).$$

The framework to reduce between sampling schemes is presented in Section 10 and the proof of Lemma 7.3 is presented in Section 10.5.2. Lastly, we bound the error probability corresponding to Bernoulli sampling, for classes of bounded Littelstone dimension, using the technique of covering numbers presented in Section 9.

Lemma 7.4. *Let $\mathcal{E}$ denote a family of Littlestone dimension d, let $m,k \in \mathbb{N}$ such that $m \leq 2k$, and let $p \in (0,1)$. Then,*

$$\Pr_{\boldsymbol{I}\in \mathrm{Ber}(2k,p)}[\mathrm{Net}^{2k}_{\mathcal{A},\boldsymbol{I},m,mp/2}(\mathcal{E}) = 1] \leq \left(\frac{Ck}{d}\right)^d \exp(-cmp).$$

The three lemmas stated above imply the following theorem, that is a special case of Theorem 7.1 for the uniform sampler:

Theorem 7.5. *Let $\mathcal{E}$ denote a family of Littlestone dimension d, let $\epsilon \in (0,1/2)$, $n \in \mathbb{N}$, $\mathcal{A} \in \mathrm{Adv}_n$ and $k \in \mathbb{N}$ that satisfies $n \geq 2k$. Then*

$$\Pr_{\boldsymbol{I}\sim \mathrm{Uni}(n,k)}[\mathrm{Net}^{n}_{\mathcal{A},\boldsymbol{I},\epsilon n,0}(\mathcal{E}) = 1] \leq \left(\frac{Ck}{d}\right)^d \exp(-c\epsilon k),$$

where $C > 0$ is a universal constant.

Proof. By Lemma 7.2,

$$\max_{\mathcal{A}\in \mathrm{Adv}_n} \Pr_{\boldsymbol{I}\sim \mathrm{Uni}(n,k)} \left[\mathrm{Net}^{n}_{\mathcal{A},\boldsymbol{I},\epsilon\cdot n,0}(\mathcal{E}) = 1\right] \leq 2 \max_{\mathcal{A}'\in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'\sim \mathrm{Uni}(2k,k)} \left[\mathrm{Net}^{2k}_{\mathcal{A}',\boldsymbol{I}',\epsilon/4\cdot 2k,0}(\mathcal{E}) = 1\right].$$

Applying Lemma 7.3 while substituting ϵ with $\epsilon/4$,

$$\max_{\mathcal{A}'\in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'\sim \mathrm{Uni}(2k,k)} \left[\mathrm{Net}^{2k}_{\mathcal{A}',\boldsymbol{I}',\epsilon/4\cdot 2k,0}(\mathcal{E}) = 1\right] \leq \max_{\mathcal{A}'\in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'\sim \mathrm{Ber}(2k,1/8)}[\mathrm{Net}^{2k}_{\mathcal{A}',\boldsymbol{I}',\epsilon/4\cdot 2k,\frac{2k\epsilon}{64}} = 1] \\ + 2\exp(-c\epsilon k).$$

Applying Lemma 7.4 with $p = 1/8$ and $m = \epsilon/4 \cdot 2k$,

$$\max_{\mathcal{A}'\in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'\sim \mathrm{Ber}(2k,1/8)}[\mathrm{Net}^{2k}_{\mathcal{A}',\boldsymbol{I}',\epsilon/4\cdot 2k,\epsilon/64\cdot 2k} = 1] \leq \left(\frac{Ck}{d}\right)^d \exp(-c\epsilon k).$$

□

8 Double Sampling

Let $n \in \mathbb{N}$ denote the stream length and assume that the sample is of size $k \leq n/2$. If $n \gg k$, it may be difficult to analyze the sample directly, since each element is selected with small probability and the universe is very large. This section presents a framework to replace the stream of size n with a stream of size $2k$. Then, this framework is used to prove Lemma 6.2 and Lemma 7.2 in Section 8.1 and Section 8.2, respectively.

Let $\overline{x} \in X^n$ denote the stream and $I \subseteq [n]$ be the index set of the sample, that has cardinality $|I| = k$. Let $f\colon X^k \times X^n \to \{0,1\}$ denote some function and we view $f(\overline{x}_I, \overline{x})$ as some indicator of whether $\overline{x}_I$ fails to approximate the complete sample $\overline{x}$. For example, $f(\overline{x}_I, \overline{x})$ could indicate whether $\overline{x}_I$ fails to be an ϵ-approximation for $\overline{x}$ with respect to some family $\mathcal{E}$. Denote by $\overline{\boldsymbol{x}} = \overline{x}(\mathcal{A}, I)$ the stream generated by the adversary $\mathcal{A} \in \mathrm{Adv}_n$ when the sample is indexed by I and we would like to bound $\Pr_{\boldsymbol{I} \sim \mathrm{Uni}(n,k)}[f(\overline{\boldsymbol{x}}_I, \overline{\boldsymbol{x}}) = 1]$. We would like to bound it by a different term that corresponds to only $2k$ elements. For this purpose, let $f'\colon X^k \times X^k \to \{0,1\}$ be another function, where $f(\overline{x}_I, \overline{x}_J)$ is an indicator of whether $\overline{x}_I$ fails to approximate $\overline{x}_J$. For example, $f'(\overline{x}_I, \overline{x}_J)$ can indicate whether $\overline{x}_I$ fails to be an ϵ-approximation for $\overline{x}_J$.

Let $\mathcal{A}' \in \mathrm{Adv}_{2k}$, $\boldsymbol{I}' \sim \mathrm{Uni}(2k, k)$ and $\overline{\boldsymbol{x}}' = \overline{x}(\mathcal{A}', \boldsymbol{I}')$ denote the stream of size $2k$ generated by $\mathcal{A}'$ with sample-index $\boldsymbol{I}'$. The following lemma gives a condition under which the probability that $f(\overline{\boldsymbol{x}}_{\boldsymbol{I}}, \overline{\boldsymbol{x}}) = 1$ can be bounded in terms of the probability that $f(\overline{\boldsymbol{x}}'_{\boldsymbol{I}'}, \overline{\boldsymbol{x}}'_{[n]\setminus \boldsymbol{I}'}) = 1$.

Lemma 8.1. *Let $\boldsymbol{I} \sim \mathrm{Uni}(n,k)$ and let $\boldsymbol{J}$ be distributed uniformly over all subsets of $[n] \setminus \boldsymbol{I}$ of size k, conditioned on $\boldsymbol{I}$. Let $f\colon X^k \times X^n \to \{0,1\}$. Let $\boldsymbol{I}' \sim \mathrm{Uni}(2k,k)$ and $f'\colon X^k \times X^k \to \{0,1\}$.*

Assume that for every $\overline{x}$ and I that satisfy $f(\overline{x}_I, \overline{x}) = 1$, it further holds that

$$\Pr_{\boldsymbol{J}}\left[f'(\overline{x}_I, \overline{x}_{\boldsymbol{J}}) = 1 \mid \boldsymbol{I} = I\right] \geq 1/2. \tag{7}$$

Then,

$$\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I}}[f(\overline{x}(\mathcal{A}, \boldsymbol{I})_{\boldsymbol{I}}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] \leq 2 \max_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'}[f'(\overline{x}(\mathcal{A}', \boldsymbol{I}')_{\boldsymbol{I}'}, \overline{x}(\mathcal{A}', \boldsymbol{I}')_{[2k]\setminus \boldsymbol{I}'}) = 1].$$

To give some intuition on the condition (7), assume again that $f(\overline{x}_I, \overline{x})$ denotes an indicator of whether $\overline{x}_I$ fails to be an ϵ-approximation to $\overline{x}$ and $f'(\overline{x}_I, \overline{x}_J)$ denotes whether $\overline{x}_I$ fails to be an ϵ'-approximation to $\overline{x}_J$, where ϵ' is slightly larger than ϵ. By concentration properties, if $f(\overline{x}_I, \overline{x}) = 1$ then with high probability over $\boldsymbol{J}$, $f'(\overline{x}_I, \overline{x}_{\boldsymbol{J}}) = 1$ as well.

The proof of Lemma 8.1 consists of two steps. In the first step, an index-set $\boldsymbol{J}$ is drawn uniformly at random from all the subsets of $[n] \setminus \boldsymbol{I}$ of size k, conditioned on $\boldsymbol{I}$. The set $\boldsymbol{J}$ is called a *ghost sample*, as it is used only for the analysis and in particular, the adversary is unaware of $\boldsymbol{J}$. The following lemma shows that under the condition (7), we can bound $f(\overline{\boldsymbol{x}}_{\boldsymbol{I}}, \overline{\boldsymbol{x}})$ in terms of $f'(\overline{\boldsymbol{x}}_{\boldsymbol{I}}, \overline{\boldsymbol{x}}_{\boldsymbol{J}})$.

Lemma 8.2. *Let $f\colon X^k \times X^n \to \{0,1\}$, $f'\colon X^k \times X^k \to \{0,1\}$ and $\mathcal{A} \in \mathrm{Adv}_n$. Assume that for every $\overline{x}$ and I that satisfy $f(\overline{x}_I, \overline{x}) = 1$, it further holds that*

$$\Pr_{\boldsymbol{J}}\left[f'(\overline{x}_I, \overline{x}_{\boldsymbol{J}}) = 1 \mid \boldsymbol{I} = I\right] \geq 1/2. \tag{8}$$

Then,

$$\Pr_{\boldsymbol{I}}[f(\overline{x}(\mathcal{A}, \boldsymbol{I})_{\boldsymbol{I}}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] \leq 2 \Pr_{\boldsymbol{I},\boldsymbol{J}}[f'(\overline{x}(\mathcal{A}, \boldsymbol{I})_{\boldsymbol{I}}, \overline{x}(\mathcal{A}, \boldsymbol{I})_{\boldsymbol{J}}) = 1]. \tag{9}$$

Proof. By (8),

$$\begin{aligned}\Pr_{\boldsymbol{I},\boldsymbol{J}}\left[f'(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{J}})=1\right] &= \mathbb{E}_{\boldsymbol{I}}\left[\Pr_{\boldsymbol{J}}\left[f'(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{J}})=1 \mid \boldsymbol{I}\right]\right] \\ &\geq \mathbb{E}_{\boldsymbol{I}}\left[f(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I}))\Pr_{\boldsymbol{J}}\left[f'(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{J}})=1 \mid \boldsymbol{I}\right]\right] \\ &\geq \mathbb{E}_{\boldsymbol{I}}\left[f(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I}))/2\right] = \Pr_{\boldsymbol{I}}\left[f(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{J}})=1\right]/2.\end{aligned}$$

□

Notice that the right hand side of (9) corresponds to drawing two subsets of size k from a stream of size n. It is desirable to bound this with a quantity that corresponds to partitioning a sample of size $2k$ to two subsets of size k. Essentially, this amounts to ignoring the elements out of $\boldsymbol{I}\cup\boldsymbol{J}$. Formally:

Lemma 8.3. *Let $f'\colon X^k\times X^k\to\{0,1\}$, let $\boldsymbol{I}$ and $\boldsymbol{J}$ be random subsets of $[n]$ as defined above and let $\boldsymbol{I}'\sim\mathrm{Uni}(2k,k)$. Then,*

$$\max_{\mathcal{A}\in\mathrm{Adv}_n}\Pr_{\boldsymbol{I},\boldsymbol{J}}[f'(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{J}})]\leq\max_{\mathcal{A}'\in\mathrm{Adv}_{2k}}\Pr_{\boldsymbol{I}'}[f'(\overline{x}(\mathcal{A}',\boldsymbol{I}')_{\boldsymbol{I}'},\overline{x}(\mathcal{A}',\boldsymbol{I}')_{[2k]\setminus\boldsymbol{I}'})=1]. \tag{10}$$

Proof. Let $\mathcal{A}$ denote the maximizer on the left hand side of (10). Let $U\subseteq[2n]$ be a set of size $2k$, and we will prove that

$$\Pr_{\boldsymbol{I},\boldsymbol{J}}[f'(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{J}})\mid\boldsymbol{I}\cup\boldsymbol{J}=U]\leq\max_{\mathcal{A}'}\Pr_{\boldsymbol{I}'}[f'(\overline{x}(\mathcal{A}',\boldsymbol{I}')_{\boldsymbol{I}'},\overline{x}(\mathcal{A}',\boldsymbol{I}')_{[2k]\setminus\boldsymbol{I}'})]. \tag{11}$$

The proof of Lemma 8.3 will then follow by taking an expectation over U.

The main idea to proving (11) is to match the subsets $I\subseteq U$ with the subsets $I'\subseteq[2k]$. One can define an adversary $\mathcal{A}_U\in\mathrm{Adv}_{2k}$ that simulates the behavior of $\mathcal{A}$ on U, hence matching the probability that $f'=1$. In particular, $\mathcal{A}_U$ simulates the selections of $\mathcal{A}$ on the set U, while skipping all the elements not in U. Formally, denote $U=(i_1,\ldots,i_{2k})$ where $i_1<i_2\cdots<i_{2k}$ and $U_{I'}=\{i_j\colon j\in I'\}$. Then $\mathcal{A}_U$ is defined to satisfy $\overline{x}(\mathcal{A}_U,I'):=\overline{x}(\mathcal{A},U_{I'})_U$. This implies that

$$\overline{x}(\mathcal{A}_U,I')_{I'}=\overline{x}(\mathcal{A},U_{I'})_{U_{I'}};\quad\overline{x}(\mathcal{A}_U,I')_{[2k]\setminus I'}=\overline{x}(\mathcal{A},U_{I'})_{U_{[2k]\setminus I'}}.$$

Hence,

$$f'(\overline{x}(\mathcal{A},U_{I'})_{U_{I'}},\overline{x}(\mathcal{A},U_{I'})_{U_{[2k]\setminus I'}})=f'(\overline{x}(\mathcal{A}_U,I')_{I'},\overline{x}(\mathcal{A}_U,I')_{[2k]\setminus I'}). \tag{12}$$

Notice that the joint distribution of $(U_{\boldsymbol{I}'},U_{[n]\setminus\boldsymbol{I}'})$ taken over $\boldsymbol{I}'\sim\mathrm{Uni}(2k,k)$, is the same as the joint distribution of $(\boldsymbol{I},\boldsymbol{J})$, conditioned on $\boldsymbol{I}\cup\boldsymbol{J}=U$. In combination with (12), this implies that

$$\begin{aligned}&\Pr[f'(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}},\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{J}})=1\mid\boldsymbol{I}\cup\boldsymbol{J}=U]=\Pr[f'(\overline{x}(\mathcal{A},U_{\boldsymbol{I}'})_{U_{\boldsymbol{I}'}},\overline{x}(\mathcal{A},U_{\boldsymbol{I}'})_{U_{[2k]\setminus\boldsymbol{I}'}})=1] \\ &=\Pr[f'(\overline{x}(\mathcal{A}_U,\boldsymbol{I}')_{\boldsymbol{I}'},\overline{x}(\mathcal{A}_U,\boldsymbol{I}')_{[2k]\setminus\boldsymbol{I}'})=1]\leq\max_{\mathcal{A}'}\Pr[f'(\overline{x}(\mathcal{A}',\boldsymbol{I}')_{\boldsymbol{I}'},\overline{x}(\mathcal{A}',\boldsymbol{I}')_{[2k]\setminus\boldsymbol{I}'})=1].\end{aligned}$$

This proves (11), and concludes the proof. □

The proof of Lemma 8.1 follow directly from Lemma 8.2 and Lemma 8.3.

8.1 Proof of Lemma 6.2

We start with the following auxiliary probabilistic lemma:

Lemma 8.4. *Let $I \subseteq [n]$ of size $|I| = k$, let $\overline{x}$ and let $t \geq 0$ be such that*

$$\max_{E \in \mathcal{E}} \left| \frac{|E \cap \overline{x}_I|}{k} - \frac{|E \cap \overline{x}|}{n} \right| \geq t. \tag{13}$$

Let $\boldsymbol{J}$ be distributed uniformly over all subsets of $[n] \setminus I$ of size k. Then, with probability at least $1/2$,

$$\max_{E \in \mathcal{E}} \left| \frac{|E \cap \overline{x}_I|}{k} - \frac{|E \cap \overline{x}_{\boldsymbol{J}}|}{k} \right| \geq t - C/\sqrt{k}.$$

Proof. Let E_0 be a maximizer of (13). We will assume that

$$\frac{|E_0 \cap \overline{x}|}{n} \geq \frac{|E_0 \cap \overline{x}_I|}{k}. \tag{14}$$

and the proof follows similarly in the other case. Applying Lemma 5.6 (item 1) with $U = \{i \in [n] \setminus I \colon x_i \in E_0\}$, $n = n - k$ and $\boldsymbol{I} = \boldsymbol{J}$, we have that with probability at least $1/2$ over $\boldsymbol{J}$,

$$\frac{|E_0 \cap \overline{x}_{\boldsymbol{J}}|}{k} = \frac{|U \cap \boldsymbol{J}|}{k} \geq \frac{|U|}{n-k} - C\sqrt{1/k} = \frac{|E_0 \cap \overline{x}_{[n]\setminus I}|}{n-k} - C\sqrt{1/k} \geq \frac{|E_0 \cap \overline{x}_{[n]}|}{n} - C\sqrt{1/k},$$

where the last inequality follows from (14). In particular,

$$\frac{|E_0 \cap \overline{x}_J|}{k} - \frac{|E_0 \cap \overline{x}_I|}{k} \geq t - C/\sqrt{k},$$

which concludes the proof. □

Proof of Lemma 6.2. First, apply Lemma 8.1 with the function $f(\overline{x}_I, \overline{x})$ that is the indicator of the event

$$\max_{E \in \mathcal{E}} \left| \frac{|E \cap \overline{x}_I|}{k} - \frac{|E \cap \overline{x}|}{n} \right| \geq t$$

and $f'(\overline{x}_I, \overline{x}_J)$ that is the indicator of

$$\max_{E \in \mathcal{E}} \left| \frac{|E \cap \overline{x}_I|}{k} - \frac{|E \cap \overline{x}_J|}{k} \right| \geq t - C_0/\sqrt{k},$$

where $C_0 > 0$ corresponds to the constant C in Lemma 8.4. The condition in Lemma 8.1 is satisfied from Lemma 8.4 and one derives that

$$\begin{aligned}
\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I}}[\mathrm{App}_{\mathcal{A},\boldsymbol{I}} \geq t] &= \max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I}}[f(\overline{x}(\mathcal{A},\boldsymbol{I})_{\boldsymbol{I}}, \overline{x}(\mathcal{A},\boldsymbol{I})) = 1] \\
&\leq 2 \max_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr[f'(\overline{x}'(\mathcal{A}',\boldsymbol{I}')_{\boldsymbol{I}'}, \overline{x}'(\mathcal{A}',\boldsymbol{I}')_{[2k]\setminus \boldsymbol{I}'})] \\
&= 2 \max_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr[\mathrm{Disc}_{\mathcal{A}',\boldsymbol{I}'}/k \geq t - C_0/\sqrt{k}].
\end{aligned}$$

□

8.2 Proof of Lemma 7.2

We start with the following probabilistic lemma:

Lemma 8.5. *Let $\epsilon \in (0,1)$, $k \geq C/\epsilon$, $\overline{x} \in X^n$ and $I \subseteq [n]$, $|I| = k$, be such that*

$$\exists E \in \mathcal{E},\ |\overline{x} \cap E| \geq \epsilon n \text{ and } \overline{x}_I \cap E = \emptyset.$$

Let $\boldsymbol{J}$ be drawn uniformly from the subsets of $[n] \setminus I$ of size k. Then, with probability at least $1/2$,

$$\exists E \in \mathcal{E},\ |\overline{x}_{I \cup J} \cap E| \geq \epsilon k/2 \text{ and } \overline{x}_I \cap E = \emptyset.$$

Proof. Let $E_0 \in \mathcal{E}$ be any set such that $|\overline{x} \cap E_0| \geq \epsilon n$ and $\overline{x}_I \cap E_0 = \emptyset$. It suffices to show that with probability at least $1/2$, $|\overline{x}_{\boldsymbol{J}} \cap E_0| \geq \epsilon k/2$. To prove this, apply Lemma 5.6 (item 2) with $\boldsymbol{I} = \boldsymbol{J}$, $U = \{i \in [n] \setminus I \colon x_i \in E_0\}$, $n = n - k$ and $\alpha = 1/2$. Notice that $|U| \geq \epsilon n \geq \epsilon(n-k)$ and we derive that

$$\begin{aligned}
&\Pr\left[|\overline{x}_{\boldsymbol{J}} \cap E_0| < \frac{\epsilon k}{2}\right] = \Pr\left[\frac{|\boldsymbol{J} \cap U|}{k} < \frac{\epsilon}{2}\right] \leq \Pr\left[\frac{|\boldsymbol{J} \cap U|}{k} \leq \frac{|U|}{2(n-k)}\right] \\
&\leq \Pr\left[\left|\frac{|\boldsymbol{J} \cap U|}{k} - \frac{|U|}{n-k}\right| \geq \frac{|U|}{2(n-k)}\right] \leq 2\exp\left(-\frac{k|U|/4}{6(n-k)}\right) \leq 2\exp\left(-\frac{k\epsilon/4}{6}\right) \leq 1/2,
\end{aligned}$$

where the last inequality follows from the assumption that $k \geq C\epsilon$ for a sufficiently large $C > 0$. □

Proof of Lemma 7.2. We will apply Lemma 8.1 with $f(\overline{x}_I, \overline{x})$ being the indicator of

$$\exists E \in \mathcal{E},\ |\overline{x} \cap E| \geq \epsilon n \text{ and } \overline{x}_I \cap E = \emptyset$$

and $f'(\overline{x}_I, \overline{x}_J)$ the indicator of

$$\exists E \in \mathcal{E},\ |\overline{x}_{I \cup J} \cap E| \geq \epsilon k/2 \text{ and } \overline{x}_I \cap E = \emptyset.$$

The condition of Lemma 8.1 holds from Lemma 8.5 and the proof follows. □

9 Covering Numbers

9.1 Overview

While studying online algorithms, it is natural to consider the following objects:

Definition 9.1 (Dynamic-Set.)**.** *A dynamic set is an online algorithm $\mathbb{B}$ which is defined on input sequences $\overline{x} \in X^n$. At each time-step $t \leq n$, the algorithm decides whether to retain x_t or to discard it. The decision whether to retain/discard x_t may depend on the elements $x_1, \ldots, x_t$ which were observed up to time t. The trace of $\mathbb{B}$ with respect to an input sequence $\overline{x}$ is defined by*

$$\mathbb{B}(\overline{x}) = \{x_t : x_t \text{ was retained by } \mathbb{B}, t \leq n\}.$$

and is viewed either as a set or as an ordered set. Further, we say that $|\mathbb{B}| \leq m$ if $\mathbb{B}$ cannot retain more than m elements. We stress that the decision whether to retain/discard an item is not reversible: retained (discarded) items can not be discarded (retained) in the future.

Given a family $\mathcal{E}$, we would like to cover it using dynamic sets, as defined below:

Definition 9.2. *Let $\mathcal{E}$ be some family and let $\mathcal{N}$ denote some finite collection of dynamic sets. We say that $\mathcal{N}$ is an ϵ-cover for $\mathcal{E}$ if for every input sequence $\overline{x}$ and every $E \in \mathcal{E}$ there exists $\mathbb{B} \in \mathcal{N}$ such that*

$$|(E \cap \{\overline{x}\}) \triangle \mathbb{B}(\overline{x})| \leq \epsilon^2 n$$

where Δ is the symmetric difference of sets. Further, define the covering number at scale ϵ, *$N(\mathcal{E}, \epsilon)$, as the smallest cardinality of an ϵ-cover for $\mathcal{E}$.*

We can obtain bounds on the covering numbers at scale 0 for Littlestone families, via a known argument:

Lemma 9.3 (Covering Littlestone Families with Few Dynamic Sets)**.** *Let $\mathcal{E}$ be a family of subsets of X with $\mathsf{Ldim}(\mathcal{E}) = d$ and let $n \in \mathbb{N}$. Then, $N(\mathcal{E}, 0) \leq \binom{n}{\leq d}$. Moreover, this is tight for $n = d$, where $N(\mathcal{E}, 0) = 2^d = \binom{n}{\leq d}$.*

While covering numbers at scale 0 can be used to derive bounds for ϵ-approximation and ϵ-nets for the Bernoulli sampler $\mathrm{Ber}(n, p)$ with p constant, these bounds are sub-optimal for epsilon-approximations. Improved bounds can be obtained by computing covering numbers at scale $\epsilon > 0$. While we do not know how derive better than $\binom{n}{\leq d}$ even for scales $\epsilon > 0$, it is possible to obtain improved bounds on *fractional covering numbers*, which is a notion that we define below and can replace the covering numbers:

Definition 9.4. *Let μ denote a probability measure over dynamic sets. We say that μ is an (ϵ, γ)-*fractional cover *for $\mathcal{E}$ if for every sequence $\overline{x}$ and every $E \in \mathcal{E}$,*

$$\mu\left(\left\{\mathbb{B} \colon |(E \cap \{\overline{x}\}) \triangle \mathbb{B}(\overline{x})| \leq \epsilon^2 n\right\}\right) \geq 1/\gamma.$$

Define the fractional covering number at scale ϵ, *$N'(\mathcal{E}, \epsilon)$, as the minimal value of γ such that there exists an (ϵ, γ) fractional cover for $\mathcal{E}$.*

Notice that $N'(\mathcal{E}, \epsilon) \leq N(\mathcal{E}, \epsilon)$: if $\mathcal{C}$ is an ϵ-cover, then $N'(\mathcal{E}, \epsilon) \leq |\mathcal{C}|$, by taking μ to be the uniform distribution over $\mathcal{C}$.

We can obtain the following bound on the fractional covering numbers for Littlestone classes:

Lemma 9.5. *It holds that $N'(\mathcal{E}, \epsilon) \leq (C/\epsilon)^{2d}$, for some universal $C > 0$.*

Next, we apply bounds on covering numbers for epsilon approximation and epsilon nets:

9.1.1 Epsilon Approximation and Sequential Rademacher

The following bound can be derived based on 0-nets:

Lemma 9.6. *Let $\mathcal{A} \in \mathrm{Adv}_{2k}$, $\boldsymbol{I} \sim \mathrm{Ber}(2k, 1/2)$, $\delta \in (0, 1/2)$ and let $\mathcal{E}$ be any family over some universe. Then, with probability $1 - \delta$,*

$$\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E}) \leq C\sqrt{k(\log N(\mathcal{E}, 0) + \log 1/\delta)}.$$

The proof is via a simple union bound. In combination with the bound on the 0-cover of Littlestone classes (Lemma 9.3), this derives that with probability $1-\delta$,

$$\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}} \leq \sqrt{k(d\log k + \log(1/\delta))},$$

which implies a sample complexity of

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

To derive sharper bounds, one can use covering numbers at scales $\epsilon > 0$. Since we only have fractional covering numbers for Littlestone classes, we present the following lemma that derives a bound based on them:

Lemma 9.7. *Let $\mathcal{A} \in \mathrm{Adv}_{2k}$, $\boldsymbol{I} \sim \mathrm{Ber}(2k, 1/2)$, $\delta \in (0, 1/2)$ and let $\mathcal{E}$ be some family. Then, with probability $1-\delta$,*

$$\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E}) \leq C\sqrt{k}\left(\int_0^1 \sqrt{\log N'(\mathcal{E},\epsilon)}d\epsilon + \sqrt{\log 1/\delta}\right).$$

This has the same form as the celebrated Dudley's integral but here we extend it to fractional covering numbers.

Using Lemma 9.5 and Lemma 9.7, Lemma 6.4 immediately follows. Indeed,

$$\begin{aligned}\frac{\mathrm{Disc}_{\mathcal{A},I}}{\sqrt{k}} &\leq C\left(\int_0^1 \sqrt{\log N'(\mathcal{E},\epsilon)}d\epsilon + \sqrt{\log 1/\delta}\right)\\ &\leq C\left(\int_0^1 \sqrt{d\log 1/\epsilon}d\epsilon + \sqrt{\log 1/\delta}\right) \leq C\left(\sqrt{d}+\sqrt{\log 1/\delta}\right).\end{aligned}$$

9.1.2 Epsilon Nets

We have the following statement:

Lemma 9.8. *Let $\mathcal{E}$ be any family, $\mathcal{A} \in \mathrm{Adv}_{2k}$, $\boldsymbol{I} \sim \mathrm{Ber}(2k, p)$ and let $m \in [2k]$. Then,*

$$\Pr_{\boldsymbol{I}}\left[\mathrm{Net}^{2k}_{\mathcal{A},\boldsymbol{I},m,mp/2} = 1\right] \leq N(\mathcal{E},0)\exp(-cmp).$$

Lemma 7.4 follows directly by applying the bound on the covering numbers at scale 0 for Littlestone classes, presented in Lemma 9.3.

9.1.3 Organization

Section 9.2 contains the proofs of Lemma 9.3 and Lemma 9.5 on the covering numbers for Littlestone classes; Section 9.3 contains the proofs of Lemma 9.6 and Lemma 9.7 on proving ϵ-approximations via covering numbers; and Section 9.4 contains the proof of Lemma 9.8 on proving ϵ-nets via covering numbers.

9.2 Covering for Littlestone Classes

Section 9.2.1 contains the proof of Lemma 9.3 on the covering numbers at scale 0, that is based on a known arguments; and Section 9.2.2 contains the proof of Lemma 9.5, that builds on the machinery presented in Section 9.2.1.

9.2.1 Proof of Lemma 9.3

Before we prove Lemma 9.3, let us make a couple of comparisons to related literature.

Remark 9.9. *It is fair to note that the proof of this proposition exploits standard and basic ideas from the online learning literature. In particular, the constructed dynamic-sets hinge on variants of the Standard Optimal Algorithm by [Lit88], and utilize its property of being a mistake-driven algorithm*[6]*. However, for the benefit of readers who are less familiar with this literature, we provide here a self-contained proof and modify some of the terminology/notation from the language of online learning to the language of ϵ-nets/approximations.*

Remark 9.10. *This comment concerns a connection with the celebrated Sauer-Shelah-Perles (SSP) Lemma [Sau72]. Note that the SSP Lemma is equivalent to a variant of Lemma 9.3 in which two quantifiers are flipped. Indeed, the SSP Lemma asserts that for every $x_1, \ldots, x_n$ there are at most $\binom{n}{\leq d}$ sets that realize all possible intersection patterns of the sets in $\mathcal{E}$ with $\{x_1, \ldots, x_n\}$. That is, if one allows the sets $\mathbb{B}_i$ to be chosen after seeing the entire input-sequence $x_1, \ldots, x_n$ then the conclusion in Lemma 9.3 extends to VC classes (which can have an unbounded Littlestone dimension, as witnessed by the class of thresholds).*

Proof of Lemma 9.3. We begin with the upper bound. The definition of the dynamic sets $\mathbb{B}_i$ exploits the following property of Littlestone families. Let $\mathcal{E}$ be a family with $\mathsf{Ldim}(\mathcal{E}) < \infty$, let $x \in X$, and consider the two "half-families"

$$\mathcal{E}_{\not\ni x} = \{E \in \mathcal{E} : x \notin E\}, \quad \mathcal{E}_{\ni x} = \{E \in \mathcal{E} : x \in E\}.$$

The crucial observation is that if $\mathcal{E} \neq \emptyset$ then for every $x \in X$:

$$\mathsf{Ldim}(\mathcal{E}_{\not\ni x}) < \mathsf{Ldim}(\mathcal{E}) \text{ or } \mathsf{Ldim}(\mathcal{E}_{\ni x}) < \mathsf{Ldim}(\mathcal{E}). \tag{15}$$

Indeed, otherwise we have $\mathsf{Ldim}(\mathcal{E}_{\not\ni x}) = \mathsf{Ldim}(\mathcal{E}_{\ni x}) = \mathsf{Ldim}(\mathcal{E}) =: d$ which implies that $\mathcal{E}$ shatters the following tree of depth $d+1$: the root is labelled with x, and the left and right subtrees of the root are trees which witness that the dimensions of $\mathcal{E}_{\not\ni x}$ and $\mathcal{E}_{\ni x}$ equal d. However, since $\mathsf{Ldim}(\mathcal{E}) = d$, this is not possible.

Littlestone Majority Vote. Equation (15) allows to define a notion of majority-vote of a (possibly infinite) family $\mathcal{E}$ with a finite Littlestone dimension. The intuition is that if x is such that $\mathsf{Ldim}(\mathcal{E}_{\ni x}) = d$ then by Equation (15) it must be that $\mathsf{Ldim}(\mathcal{E}_{\not\ni x}) < d$ and therefore $\mathsf{Ldim}(\mathcal{E}_{\ni x}) > \mathsf{Ldim}(\mathcal{E}_{\not\ni x})$ which we interpret as if x is contained in a "majority" of the sets in $\mathcal{E}$. Similarly, $\mathsf{Ldim}(\mathcal{E}_{\not\ni x}) = d$ is interpreted as if most sets in $\mathcal{E}$ do not contain x. This motivates the following definition

$$\mathsf{Lmaj}(\mathcal{E}) = \{x : \mathsf{Ldim}(\mathcal{E}_{\ni x}) = d\}, \tag{16}$$

[6] A mistake-driven algorithm updates its internal state only when it makes mistakes.

with the convention that $\mathsf{Lmaj}(\emptyset) = \emptyset$. Observe that $\mathsf{Lmaj}(\mathcal{E})$ shares the following property with the standard majority-vote over finite families: let $x \in X$ and assume $\mathcal{E} \neq \emptyset$. Then,

$$\Big((\forall E \in \mathcal{E}) : x \in E\Big) \implies x \in \mathsf{Lmaj}(\mathcal{E}) \quad \text{and} \quad \Big((\forall E \in \mathcal{E}) : x \notin E\Big) \implies x \notin \mathsf{Lmaj}(\mathcal{E}). \tag{17}$$

That is, if the sets in $\mathcal{E}$ agree on x unanimously, then $\mathsf{Lmaj}(\mathcal{E})$ agrees with them on x. We comment that this definition is the basis of the *Standard Online Algorithm* which witnesses the online-learnability of Littlestone classes in the mistake-bound model [Lit88].

We are now ready to define the required family of $\binom{n}{\leq d}$ dynamic sets. Each dynamic set $\mathbb{B}_I$ is indexed by a subset $I \subseteq [n]$ of size $|I| \leq d$. (Hence there are $\binom{n}{\leq d}$ dynamic sets.) Below is the pseudo-code of $\mathbb{B}_I$ for $I \subseteq [n]$.

The Dynamic Set $\mathbb{B}_I$

Let $\mathcal{E}$ be a family with $\mathsf{Ldim}(\mathcal{E}) = d$, and let $I \subseteq [n]$.
Let $x_1, \ldots, x_n$ denote the (adversarially-produced) input sequence.

1. Initialize $\mathcal{E}_0^I = \mathcal{E}$.
2. For $t = 1, \ldots, n$:
   (a) If $t \notin I$ then set $\mathcal{E}_t^I = \mathcal{E}_{t-1}^I$.
   (b) Else, set
   $$\mathcal{E}_t^I = \begin{cases} (\mathcal{E}_{t-1}^I)_{\not\ni x_t} & x_t \in \mathsf{Lmaj}(\mathcal{E}_{t-1}^I), \\ (\mathcal{E}_{t-1}^I)_{\ni x_t} & x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^I). \end{cases}$$
   (c) Retain x_t if and only if $x_t \in \mathsf{Lmaj}(\mathcal{E}_t^I)$.

Observe the following useful facts regarding $\mathbb{B}_I$:

1. The sequence of families $\{\mathcal{E}_t^I\}_{t=0}^n$ is a chain: $\mathcal{E}_0^I \supseteq \mathcal{E}_1^I \supseteq \ldots \supseteq \mathcal{E}_n^I$.
2. A strict containment $\mathcal{E}_{t-1}^I \supsetneq \mathcal{E}_t^I$ occurs only if $t \in I$.
3. Whenever a strict containment $\mathcal{E}_{t-1}^I \supsetneq \mathcal{E}_t^I$ occurs then also $\mathsf{Ldim}(\mathcal{E}_{t-1}^I) > \mathsf{Ldim}(\mathcal{E}_t^I)$. (By Equations (15) and (16).)

To complete the proof it remains to show that for every $\overline{x}$ and every $E \in \mathcal{E}$ there exists $I \subseteq [n]$, with $|I| \leq d$ such that

$$(\forall t \leq n) : x_t \in E \iff x_t \in \mathbb{B}_I(\overline{x}). \tag{18}$$

We construct the set $I = I(E)$ in a parallel fashion to the above process:

The Index Set $I = I(E)$

Let $E \in \mathcal{E}$ and let $x_1, \ldots, x_n$ denote the input sequence.

1. Initialize $\mathcal{E}_0^E = \mathcal{E}$ and $I = \emptyset$.
2. For $t = 1, \ldots, n$:
   (a) If E and $\mathsf{Lmaj}(\mathcal{E}_{t-1}^E)$ agree on x_t (i.e. $x_t \in E \iff x_t \in \mathsf{Lmaj}(\mathcal{E}_{t-1}^E)$) then set $\mathcal{E}_t^E = \mathcal{E}_{t-1}^E$.
   (b) Else, add t to I and set

$$\mathcal{E}_t^E = \begin{cases} (\mathcal{E}_{t-1}^E)_{\not\ni x_t} & x_t \in \mathsf{Lmaj}(\mathcal{E}_{t-1}^E) \wedge x_t \notin E, \\ (\mathcal{E}_{t-1}^E)_{\ni x_t} & x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^E) \wedge x_t \in E. \end{cases}$$

3. Output $I = I(E)$.

Note that by construction, $\mathcal{E}_t^E = \mathcal{E}_t^I$ for every $t \leq n$, and $E \in \mathcal{E}_t^E$ for all t. We need to show that the constructed set I satisfies Equation (18) and that $|I| \leq d$. For the first part, note that for every $t \leq n$:

$$\begin{aligned} x_t \in \mathbb{B}_I(\overline{x}) &\iff x_t \in \mathsf{Lmaj}(\mathcal{E}_t^I) && \text{(by definition of } \mathbb{B}_I) \\ &\iff x_t \in \mathsf{Lmaj}(\mathcal{E}_t^E) && \text{(since } \mathcal{E}_t^E = \mathcal{E}_t^I) \\ &\iff x_t \in E, && \text{(see below)} \end{aligned}$$

where the last step follows because all the sets $E' \in \mathcal{E}_t^E$ agree with E on x_t. Thus, by[7] Equation (17) also $\mathsf{Lmaj}(\mathcal{E}_t^E)$ agrees with E on x_t, which amounts to the last step.

To see that $|I| \leq d$, consider the chain

$$\mathcal{E}_0^E \supseteq \mathcal{E}_1^E \supseteq \ldots \supseteq \mathcal{E}_n^E.$$

Note that strict containments $\mathcal{E}_{t-1}^E \supsetneq \mathcal{E}_t^E$ occurs only if $t \in I$, and that whenever such a strict containment occurs, we have $\mathsf{Ldim}(\mathcal{E}_{t-1}^E) > \mathsf{Ldim}(\mathcal{E}_t^E)$. Therefore, since $\mathsf{Ldim}(\mathcal{E}_0^E) = d$ and $\mathsf{Ldim}(\mathcal{E}_n^E) \geq 0$, it follows that $|I| \leq d$ as required.

It remains to prove the lower bound. Let $D = \{\mathbb{B}_i : 1 \leq i < 2^d\}$ be a family of less than 2^d dynamic sets. Pick a tree $\mathcal{T}$ of depth d which is shattered by $\mathcal{E}$ and define an adversarial sequence $x_1, \ldots, x_d$ as follows:

[7] Note that $\mathcal{E}_t^E \neq \emptyset$ because $E \in \mathcal{E}_t^E$.

The Adversarial Sequence $x_1, \ldots, x_d$

1. Set $\mathcal{T}_1 = \mathcal{T}, D_1 = D, \mathcal{E}_1 = \mathcal{E}$, and $i = 1$.
2. For $i = 1, \ldots, d$
   (i) Set x_i to be the item labelling the root of $\mathcal{T}_i$.
   (ii) If less than half of the dynamic sets $\mathbb{B}_j \in D_i$ retain x_i then continue the next iteration with $\mathcal{T}_{i+1}$ being the right subtree of $\mathcal{T}_i$ (which corresponds to the sets containing x_i), and with $\mathcal{E}_{i+1} = \{E \in \mathcal{E}_i : x \in E\}$ and $D_{i+1} = \{\mathbb{B}_j \in D_i : x_i \in \mathbb{B}_j\}$.
   (iii) Else, continue to the next iteration with $\mathcal{T}_{i+1}$ being the left subtree of $\mathcal{T}_i$, and with $\mathcal{E}_{i+1} = \{E \in \mathcal{E}_i : x \notin E\}$ and $D_{i+1} = \{\mathbb{B}_j \in D_i : x_i \notin \mathbb{B}_j\}$.

Note that $\mathcal{E}_i$ contains all the sets in $\mathcal{E}$ that are consistent[8] with the path corresponding to $x_1, \ldots, x_{i-1}$, and similarly D_i contains all dynamic sets in D which are consistent with that path. Thus, since $|D_1| = |D| < 2^d$, it follows by construction that $|D_i| < 2^i$ for every $i < d$, and in particular that $D_d = \emptyset$ at the end of the process. Thus, the set $E \in \mathcal{E}$ which is consistent with the path corresponding to $x_1, \ldots, x_d$ satisfies $E \cap \{x_1, \ldots x_d\} \neq \mathbb{B}_i(x_1, \ldots x_d)$ for every $i < 2^d$, as required. □

9.2.2 Proof of Lemma 9.5

For convenience, let us bound $N'(\mathcal{E}, \sqrt{\epsilon}) \leq (C/\epsilon)^d$. We start by defining the fractional cover $\mathcal{B}$ and then prove its validity. Let $p = 3d/(\epsilon n)$ for a sufficiently large universal constant C_0, and $\mathbb{B} \sim \mathcal{B}$ is sampled as follows:

1. Select a random subset $\boldsymbol{I}' \subseteq [n]$, where each $i \in [n]$ is selected independently with probability p.
2. Select a subset $\boldsymbol{I} \subseteq \boldsymbol{I}'$ of size $|\boldsymbol{I}| \leq d$, uniformly at random from the set of all $\binom{|\boldsymbol{I}'|}{\leq d}$ subsets.
3. Output $\mathbb{B} = \mathbb{B}_{\boldsymbol{I}}$.

To prove that $\mathcal{B}$ is an $(\epsilon, (C/\epsilon)^d)$-fractional cover, fix some $E \in \mathcal{E}$ and sequence $\overline{x}$. From the proof of Lemma 9.3, for any I', there exists $I^* = I^*(I') \subseteq I'$ of size $|I^*| \leq d$ such that $(\mathbb{B}_{I^*}(\overline{x}))_{I'} = E \cap \overline{x}_{I'}$. Denote $\boldsymbol{I}^* = I^*(\boldsymbol{I}')$ where $\boldsymbol{I}'$ is distributed as above.

We will bound for below the probability that $\boldsymbol{I}^*$ satisfies

$$|\mathbb{B}_{\boldsymbol{I}^*}(\overline{x}) \triangle (E \cap \overline{x})| \leq \epsilon n.$$

Further, we will bound from below the probability that $\boldsymbol{I} = \boldsymbol{I}^*$. Combining this two bounds, this will give a lower bound on the probability that

$$|\mathbb{B}_{\boldsymbol{I}}(\overline{x}) \triangle (E \cap \overline{x})| \leq \epsilon n,$$

[8] $\mathbb{B}_j$ is consistent with the path corresponding to $x_1, \ldots, x_d$ means that $\mathbb{B}_j(x_1, \ldots, x_n)$ contains x_i if and only if x_{i+1} labels the right child of the node labelled x_i.

that suffices to complete the proof.

To begin with the first step, notice that

$$|\mathbb{B}_{\boldsymbol{I}^*}(\overline{x}) \triangle (E \cap \overline{x})| = |\{t \in [n] \colon x_t \in \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E\}|. \tag{19}$$

To analyze the right hand side of the above quantity, place each time $t \in [n]$ in one of four categories:

1. $x_t \in \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$ and $t \in \boldsymbol{I}'$.
2. $x_t \in \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$ and $t \notin \boldsymbol{I}'$.
3. $x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$ and $t \notin \boldsymbol{I}'$
4. $x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$ and $t \in \boldsymbol{I}'$.

Notice that the above properties apply:

- It holds that $\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \neq \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*})$ if and only if t is in category (1).
  - For t in category (1) it holds that $t \in \boldsymbol{I}'$ which implies that $x_t \notin \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$, since $\mathbb{B}_{\boldsymbol{I}^*}$ is defined to agree with E on all $t \in \boldsymbol{I}'$. While any t in category (1) satisfies $x_t \in \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$, this implies that $\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \neq \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*})$.
  - For t in categories (2) and (3) it holds that $t \notin \boldsymbol{I}'$. Since $\boldsymbol{I}^* \subseteq \boldsymbol{I}'$, then $t \notin \boldsymbol{I}^*$. By definition of $\mathbb{B}_{\boldsymbol{I}^*}$ it holds that $\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) = \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*})$ whenever $t \notin \boldsymbol{I}^*$.
  - For t in category (4), it holds that $x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$. Since $\mathbb{B}_{\boldsymbol{I}^*}$ agrees with E on $\boldsymbol{I}'$ and since $t \in \boldsymbol{I}'$, it holds that $\mathbb{B}_{\boldsymbol{I}^*}$ agrees with E on x_t, namely, $x_t \notin \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$. This implies that $x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})$. By definition of the dynamic set $\mathbb{B}_{\boldsymbol{I}^*}$ it holds that $\mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) = \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})$ if and only if $x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})$. In particular, $\mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) = \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})$ as required.
- It holds that $x_t \in \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$ if and only if t is in category (2):
  - For categories (1) and (4) it holds that $t \in \boldsymbol{I}'$. By definition of $\boldsymbol{I}^*$, $\mathbb{B}_{\boldsymbol{I}^*}$ and E agree for any $\overline{x}_t$ for $t \in \boldsymbol{I}'$. This implies that $x_t \notin \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$.
  - For categories (2) and (3), it holds that $t \notin \boldsymbol{I}'$ hence $t \notin \boldsymbol{I}^*$ which implies by definition of $\mathbb{B}_{\boldsymbol{I}^*}$ that $\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) = \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*})$. Since for category (2) we have $x_t \in \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$, we further have $x_t \in \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$. Similarly, in category (3) we have $x_t \notin \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}) \triangle E$ hence $x_t \notin \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$.

Let $\boldsymbol{H}$ (hit) denote the set of all indices t that correspond to case (1) and $\boldsymbol{M}$ (miss) denote the set of indices in case (2). We view $\boldsymbol{H}$ and $\boldsymbol{M}$ as random variables that are functions of the random variable $\boldsymbol{I}'$ (where $\overline{x}$ and E are fixed). Since only the elements in $t \in \boldsymbol{M}$ satisfy $x_t \in \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$, the goal is to upper bound $|\boldsymbol{M}|$. In fact, we will upper bound its expected value and then use Markov's inequality to derive tail bounds.

Before bounding $\mathbb{E}[|\boldsymbol{M}|]$, notice that $|\boldsymbol{H}| \leq d$. This holds due to the fact that, as described above, for each $t \in \boldsymbol{H}$, $\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \neq \mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*})$. And this can happen at most d times, since $\mathsf{Ldim}(\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})) < \mathsf{Ldim}(\mathsf{Lmaj}(\mathcal{E}_{t-1}^{\boldsymbol{I}^*}))$ for any such t, as described in the proof of Lemma 9.3.

We proceed with proving that $\mathbb{E}|\boldsymbol{M}| \leq d/p$. Denote $\boldsymbol{H} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{|\boldsymbol{H}|}\}$ let $\boldsymbol{t}_0 = 0$, and let $\boldsymbol{y}_j$ denote the number of elements of $\boldsymbol{M}$ between $\boldsymbol{t}_{j-1}$ and $\boldsymbol{t}_j$: $\boldsymbol{y}_j = |\boldsymbol{M} \cap \{\boldsymbol{t}_{j-1}+1, \boldsymbol{t}_{t-1}+2, \ldots, \boldsymbol{t}_j - 1\}|$. Note that $|\boldsymbol{M}| = \sum_{j=1}^{d} \boldsymbol{y}_j$.

We claim that $\mathbb{E}[\boldsymbol{y}_j] \leq 1/p$. Define

$$\boldsymbol{U}_j := \{t\colon t > \boldsymbol{t}_{j-1}, x_j \in \mathsf{Lmaj}(\mathcal{E}_{\boldsymbol{t}_{j-1}}^{\boldsymbol{I}^*}) \triangle E\}; \quad \boldsymbol{U}_j^- = \boldsymbol{U}_j \cap [t-1].$$

Notice that for any t that satisfies $\boldsymbol{t}_{j-1} < t < \boldsymbol{t}_j$ it holds that $\mathsf{Lmaj}(\mathcal{E}_{\boldsymbol{t}_{j-1}}^{\boldsymbol{I}^*}) = \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})$, since, as stated above, $\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})$ only changes at iterations $t \in \boldsymbol{H}$. This will imply the following:

$$\boldsymbol{M} \cap \{\boldsymbol{t}_{j-1}+1, \ldots, \boldsymbol{t}_j - 1\} = \boldsymbol{U}_j^-. \tag{20}$$

For the first direction of (20), any $t \in \boldsymbol{U}_j^-$ satisfies $t \notin \boldsymbol{H}$ by definition, and further it satisfies $x_t \in \mathsf{Lmaj}(\mathcal{E}_{\boldsymbol{t}_{j-1}}^{\boldsymbol{I}^*}) \triangle E = \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E$ which implies that it is in $\boldsymbol{M} \cup \boldsymbol{H}$. However, it cannot be in $\boldsymbol{H}$ since $\boldsymbol{H} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{|\boldsymbol{H}|}\}$. Further, it satisfies $\boldsymbol{t}_{j-1} < t < \boldsymbol{t}_j$ by definition. For the second direction, any t in the left hand side satisfies $x_t \in \mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*}) \triangle E = \mathsf{Lmaj}(\mathcal{E}_{\boldsymbol{t}_{j-1}}^{\boldsymbol{I}^*}) \triangle E$ which implies that it is in $\boldsymbol{U}_j$. We derive (20) which implies that $\boldsymbol{y}_j = |\boldsymbol{U}_j^-|$.

To estimate $|\boldsymbol{U}_j^-|$, notice that the first element of $\boldsymbol{U}_j$ that is also in $\boldsymbol{I}'$ is $\boldsymbol{t}_j$. This follows from the fact that $\mathsf{Lmaj}(\mathcal{E}_t^{\boldsymbol{I}^*})$ changes only once an element of $\boldsymbol{H}$ is observed. Further, conditioned on $\boldsymbol{t}_{j-1}$ and $\boldsymbol{I}' \cap [\boldsymbol{t}_{j-1}]$, the set $\boldsymbol{U}_j$ is fixed, and conditionally, since any element of $\boldsymbol{U}_j$ is in $\boldsymbol{I}'$ with probability p, the expected number of elements in $\boldsymbol{U}_j$ that are encountered before the first element of $\boldsymbol{I}'$ is bounded by $1/p$. This quantity is exactly $\mathbb{E}[\boldsymbol{y}_j] = |\boldsymbol{U}_j^-| \leq 1/p$, and we derive that $\mathbb{E}[|\boldsymbol{M}|] \leq d/p$. From Markov's inequality, $\Pr[|\boldsymbol{M}| \leq 3d/p] \geq 2/3$.

We have proved that with probability $2/3$, $|\boldsymbol{M}| \leq 3d/p$. This, from (19) and from the definitions of $\boldsymbol{M}$ and p, implies that with probability $2/3$, $|\mathbb{B}_{\boldsymbol{I}^*}(x) \triangle (E \cap \overline{x})| \leq 3d/p \leq \epsilon n$. Further, we want to lower bound the probability that $\boldsymbol{I} = \boldsymbol{I}^*$. Notice that $\Pr[\boldsymbol{I} = \boldsymbol{I}^* \mid \boldsymbol{I}'] = 1/\binom{|\boldsymbol{I}'|}{\leq d}$, hence it is desirable to show that $|\boldsymbol{I}'|$ is small with high probability. Indeed, since $\mathbb{E}|\boldsymbol{I}'| = np$, by Markov's inequality, $\Pr[|\boldsymbol{I}'| \leq 3np] \geq 2/3$. By a union bound,

$$\Pr[|\boldsymbol{I}'| \leq 3np, |\mathbb{B}_{\boldsymbol{I}^*}(x) \triangle (E \cap \overline{x})| \leq \epsilon n] \geq 1/3.$$

We conclude that

$$\begin{aligned}
&\Pr\left[|\mathbb{B}_{\boldsymbol{I}}(\overline{x}) \triangle (E \cap \overline{x})| \leq \epsilon n\right] \geq \Pr\left[|\boldsymbol{I}'| \leq 3np, |\mathbb{B}_{\boldsymbol{I}^*} \triangle (E \cap \overline{x})| \leq \epsilon n, \boldsymbol{I} = \boldsymbol{I}^*\right] \\
&= \Pr\left[|\boldsymbol{I}'| \leq 3np, |\mathbb{B}_{\boldsymbol{I}^*} \triangle (E \cap \overline{x})| \leq \epsilon n\right] \Pr\left[\boldsymbol{I} = \boldsymbol{I}^* \mid |\boldsymbol{I}'| \leq 3np, |\mathbb{B}_{\boldsymbol{I}^*} \triangle (E \cap \overline{x})|\right] \geq \frac{1}{3} \cdot \binom{3np}{\leq d}^{-1} \\
&= \frac{1}{3} \cdot \binom{9d/\epsilon}{\leq d}^{-1} \geq \left(\frac{C}{\epsilon}\right)^{-d},
\end{aligned}$$

using the fact that $\boldsymbol{I}^*$ is a function of $\boldsymbol{I}'$, and conditioned on any value of $\boldsymbol{I}'$, the probability that $\boldsymbol{I} = \boldsymbol{I}^*$ is $\binom{|\boldsymbol{I}'|}{\leq d}^{-1}$; and further, that $\binom{n}{k} \leq (Cn/k)^k$ for a universal $C > 0$.

9.3 Deriving Bounds on ϵ-Approximation via Fractional Covering Numbers

In this section we prove the concentration results based on covering numbers, starting with results based on deterministic 0-covers and moving to fractional ϵ-covers. The following definition will be useful: for any $E \in \mathcal{E}$ define $\boldsymbol{Y}_E = |E \cap \overline{x}_{\boldsymbol{I}}| - |E \cap \overline{x}_{[2k]\setminus \boldsymbol{I}}|$. Similarly, for any $\mathbb{B}$, define $\boldsymbol{Y}_{\mathbb{B}} = |\mathbb{B} \cap \overline{x}_{\boldsymbol{I}}| - |\mathbb{B} \cap \overline{x}_{[2k]\setminus \boldsymbol{I}}|$. Notice that

$$\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}} = \max_{E\in\mathcal{E}} |\boldsymbol{Y}_E|.$$

9.3.1 Basic Lemmas for Deterministic Covers and Proof of Lemma 9.6

We start with concentration of a single dynamic set:

Lemma 9.11. *Let $\mathbb{B}$ be a dynamic set with $|\mathbb{B}| \le m$. Let $\boldsymbol{I} \sim \mathrm{Ber}(n, 1/2)$. Then, for any $t \ge 0$,*

$$\Pr\left[|\boldsymbol{Y}_{\mathbb{B}}| \ge t\right] \le 2\exp(-t^2/(2m)).$$

Consequently, for any $\delta \in (0, 1/2)$, with probability $1-\delta$ it holds that

$$|\boldsymbol{Y}_{\mathbb{B}}| \le C\sqrt{m\log(1/\delta)}.$$

For the proof of Lemma 9.11, we need the following Martingale lemma (notice that an overview on Martingales is given in Section A.1)

Lemma 9.12 ([dlPn99], Theorem 6.1)**.** *Let $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_n$ be a Martingale adapted to the filtration $F_0, \ldots, F_n$, such that $\sum_{i=1}^n |\boldsymbol{y}_i - \boldsymbol{y}_{i-1}|^2 \le s$ holds almost surely for some $s > 0$. Assume that for all $i \in [n]$, conditioned on F_{i-1}, $\boldsymbol{y}_i - \boldsymbol{y}_{i-1}$ is a symmetric random variable (namely, it has the same conditional distribution as $\boldsymbol{y}_{i-1} - \boldsymbol{y}_i$). Then, for any $t > 0$,*

$$\Pr\left[|\boldsymbol{y}_n - \boldsymbol{y}_0| \ge t\right] \le 2\exp(-t^2/(2s)).$$

Proof of Lemma 9.11. This follows directly from Lemma 9.12. Indeed, we apply this lemma with $\boldsymbol{y}_i = |\mathbb{B} \cap I \cap [i]| - |\mathbb{B} \cap ([n] \setminus I) \cap [i]|$ and $s = m$. □

We are ready to prove Lemma 9.6.

Proof of Lemma 9.6. Notice that it suffices to prove that for any $t \ge 0$,

$$\Pr[\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}} > t] \le 2N(\mathcal{E}, 0)\exp(-t^2/(4k)).$$

Let $\mathcal{N}$ be a minimal 0-net for $\mathcal{E}$. Applying Lemma 9.11 with $m = 2k$, for any $\mathbb{B} \in \mathcal{N}$,

$$\Pr[|\boldsymbol{Y}_E| > t] \le 2\exp(-t^2/4k)$$

Applying a union bound over $\mathbb{B} \in \mathcal{N}$,

$$\Pr[\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}} > t] = \Pr[\max_{E\in\mathcal{E}} |\boldsymbol{Y}_E| > t] \le \Pr[\max_{\mathbb{B}\in\mathcal{N}} |\boldsymbol{Y}_{\mathbb{B}}| > t] \le 2N(\mathcal{E}, 0)\exp(-t^2/(4k)).$$

□

9.3.2 Basic Lemmas for Fractional Covers

To give intuition about fractional covers, we prove a variant of Lemma 9.6. First, we start with an auxiliary lemma the replaces the union bound:

Lemma 9.13. *Let $\{\boldsymbol{y}_j\}_{j\in J}$ denote random variables over $\{0,1\}$ where J is some index set, and assume that for any $j \in J$, $\Pr[\boldsymbol{y}_j = 1] \leq p$, for some $p > 0$. Let μ denote some probability measure over J and let $\alpha > 0$. Then,*

$$\Pr_{\{\boldsymbol{y}_j\}_{:\, j\in J}}\left[\mu(\{j\colon \boldsymbol{y}_j = 1\}) \geq \alpha\right] \leq p/\alpha.$$

Proof. Notice that by linearity of expectation,

$$\mathbb{E}[\mu(\{j\colon \boldsymbol{y}_j = 1\})] = \mathbb{E}\int \boldsymbol{y}_j d\mu = \int \mathbb{E}[\boldsymbol{y}_j] d\mu \leq \int p d\mu = p.$$

The proof follows by Markov's inequality. □

The following lemma applies Lemma 9.13 specifically for distributions over dynamic sets.

Lemma 9.14. *Let μ be a probability measure over dynamic sets with $|\mathbb{B}| \leq m$ for all $\mathbb{B}$ in the support of μ. Then, for any $\delta' > 0$, with probability at least $1-\delta'$ over $\boldsymbol{I} \sim \mathrm{Ber}(n, 1/2)$ it holds that*

$$\mu\left(\left\{\mathbb{B}\colon\ |\boldsymbol{Y}_{\mathbb{B}}| \leq C\sqrt{m\log(1/(\delta'\alpha))}\right\}\right) \geq 1-\alpha.$$

Proof. Apply Lemma 9.13 with $\boldsymbol{y}_{\mathbb{B}}$ being the indicator that $|\boldsymbol{Y}_{\mathbb{B}}| \leq C\sqrt{m\log(1/(\delta'\alpha))}$ and $p = \alpha\delta'$. If C is a sufficiently large constant, it follows from Lemma 9.11 that $\Pr[\boldsymbol{y}_{\mathbb{B}} = 1] \leq \alpha\delta'$ and Lemma 9.13 can be applied to derive the desired result. □

Using Lemma 9.14, one can derive bounds for epsilon approximation based on fractional covering numbers at scale 0:

Lemma 9.15. *Let $\mathcal{A} \in \mathrm{Adv}_n$, $\boldsymbol{I} \sim \mathrm{Bin}(n, 1/2)$ and $\delta > 0$. Then, with probability $1-\delta$,*

$$\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}} \leq C\sqrt{n\left(\log N'(\mathcal{E},0) + \log(1/\delta)\right)}.$$

Proof. Let μ be a $(0, N'(\mathcal{E},0))$-fractional cover for $\mathcal{E}$. We apply Lemma 9.14 with $m = n$, $\alpha = 1/(2N'(\mathcal{E},0))$, $\delta' = \delta$ and $\mu = \mu$ to get that with probability $1-\delta$,

$$\mu\left(\mathbb{B}\colon |\boldsymbol{Y}_{\mathbb{B}}| \leq C\sqrt{n(\log N'(\mathcal{E},0) + \log(1/\delta))}\right) \geq \frac{1}{2N'(\mathcal{E},0)}.$$

Whenever this holds, for every $E \in \mathcal{E}$ there exists $\mathbb{B} \in \mathcal{E}$ that $E \cap \overline{x} = \mathbb{B}(\overline{x})$ which implies that $\boldsymbol{Y}_{\mathbb{B}} = \boldsymbol{Y}_E$. Hence,

$$\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}} = \max_{E\in\mathcal{E}} |\boldsymbol{Y}_E| \leq \max_{\mathbb{B}\in\mathcal{N}} |\boldsymbol{Y}_{\mathbb{B}}| \leq C\sqrt{n(\log N'(\mathcal{E},0) + \log(1/\delta))}.$$

□

9.3.3 Chaining for Non-Fractional Covers

The proof of Lemma 9.7 follows the technique of chaining. Before presenting the proof for fractional covers, we start by presenting an outline of the proof for non-fractional covers, while obtaining a similar bound with $N(\mathcal{E},\epsilon)$ instead of $N'(\mathcal{E},\epsilon)$. The proof follows from standard techniques. Some technicalities are ignored for the sake of presentation.

Let $1=\epsilon_0>\epsilon_1>\cdots$ be a non-increasing sequence of values with $\lim_{i\to\infty}\epsilon_i=0$. We take nets $\mathcal{N}_0,\mathcal{N}_1,\ldots$, where each $\mathcal{N}_i$ is an ϵ_i net for $\mathcal{E}$. Each $E\in\mathcal{E}$ we approximate using elements from the different nets: for any $E\in\mathcal{E}$, $i\geq 0$ and $\overline{x}\in X^n$, let $\mathbb{B}_{E,i,\overline{x}}$ denote an arbitrarily chosen dynamic set $\mathbb{B}\in\mathcal{N}_i$ that satisfies $|\mathbb{B}_{E,i,\overline{x}}(\overline{x})\triangle(E\cap\overline{x})|\leq\epsilon_i^2 n$. Further, define the random variable over dynamic sets $\boldsymbol{B}_{E,i}=\mathbb{B}_{E,i,\overline{\boldsymbol{x}}}$ and notice that $|\boldsymbol{B}_{E,i}(\overline{\boldsymbol{x}})\triangle(E\cap\overline{\boldsymbol{x}})|\leq\epsilon_i^2 n$. Since $\epsilon_i\to 0$, we have $\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}\to\boldsymbol{Y}_E$, hence,

$$|\boldsymbol{Y}_E|=\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,0}}+\sum_{i=1}^{\infty}(\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}-\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}})\right|\leq\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,0}}\right|+\sum_{i=1}^{\infty}\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}-\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}}\right|. \tag{21}$$

The hope is for the sum in the right hand side of Equation (21) to converge. Indeed, notice that

$$|\boldsymbol{B}_{E,i}\triangle\boldsymbol{B}_{E,i-1}|\leq|\boldsymbol{B}_{E,i}\triangle E|+|E\triangle\boldsymbol{B}_{E,i-1}|\leq(\epsilon_i^2+\epsilon_{i-1}^2)n\leq 2\epsilon_{i-1}^2 n, \tag{22}$$

hence, as i increases, the differences $\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}-\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}}\right|$ tend to decrease. Taking a maximum over $E\in\mathcal{E}$ in (21), we have

$$\max_{E\in\mathcal{E}}|\boldsymbol{Y}_E|\leq\max_{E\in\mathcal{E}}\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,0}}\right|+\sum_{i=1}^{\infty}\max_{E\in\mathcal{E}}\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}-\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}}\right|. \tag{23}$$

We show how to bound the summand corresponding to any $i\geq 1$ while term $\max_{E\in\mathcal{E}}\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,0}}\right|$ can be similarly bounded. Since $\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}\in\mathcal{N}_i$ and $\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}}\in\mathcal{N}_{i-1}$, there can be at most $|\mathcal{N}_{i-1}||\mathcal{N}_i|=N(\mathcal{E},\epsilon_{i-1})N(\mathcal{E},\epsilon_i)$ distinct differences $\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}-\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}}$. Intuitively, as i increases, the maximum is taken over more elements, however, the individual differences are smaller, and the hope is that the sum in Equation (23) would converge.

Using Equation (22), we can apply Lemma 9.11 with $m=2\epsilon_{i-1}^2 n$ to obtain that each distance $\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}-\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}}$ is bounded by $C\epsilon_i\sqrt{n\log(1/\delta')}$ with probability $1-\delta'$. Taking δ' smaller than $1/N(\mathcal{E},\epsilon_{i-1})N(\mathcal{E},\epsilon_i)$ and applying a union bound over at most $N(\mathcal{E},\epsilon_{i-1})N(\mathcal{E},\epsilon_i)$ elements, we derive that with high probability,

$$\max_{E\in\mathcal{E}}\left|\boldsymbol{Y}_{\boldsymbol{B}_{E,i}}-\boldsymbol{Y}_{\boldsymbol{B}_{E,i-1}}\right|\leq O\left(\epsilon_{i-1}\sqrt{n\log N(\mathcal{E},\epsilon_i)}\right). \tag{24}$$

We further take a union bound over $i\geq 0$ and derive by (23) and (24) that w.h.p,

$$\max_{E\in\mathcal{E}}|\boldsymbol{Y}_E|\leq O\left(\sqrt{n}\sum_{i=1}^{\infty}\epsilon_{i-1}\sqrt{\log N(\mathcal{E},\epsilon_i)}\right).$$

A standard choice for ϵ_i is $\epsilon_i=2^{-i}$, and this yields

$$\max_{E\in\mathcal{E}}|\boldsymbol{Y}_E|\leq O\left(\sqrt{n}\sum_{i=1}^{\infty}2^{-i}\sqrt{\log N(\mathcal{E},2^{-i})}\right)\leq O\left(\sqrt{n}\int_{\epsilon=0}^{1}\sqrt{\log N(\mathcal{E},\epsilon)}\right), \tag{25}$$

where the last inequality is by approximating the sum with an integral. This, in fact, is the celebrated *Dudley's integral*.

Remark 9.16. *The choice of $\epsilon_i = 2^{-i}$ can generally only yield bounds on $\mathbb{E}\max_{E\in\mathcal{E}}|\boldsymbol{Y}_E|$, rather than high probability bounds. It is common in literature to obtain high probability bounds by first bounding the expectation $\mathbb{E}\max_{E\in\mathcal{E}}|\boldsymbol{Y}_E|$ and then showing that this maximum concentrates around its expectation, via McDiarmid-like inequalities. However, such concentration inequalities cannot be applied in the adversarial setting. Hence we use instead a different well-studied choice of ϵ_i that directly gives high probability bounds.*

9.3.4 Proof of Lemma 9.7

The proof is by the standard technique of chaining, with adaptations to handle fractional covering numbers. Our goal is to bound $\max_{E\in\mathcal{E}}\boldsymbol{Y}_E$. The main ideas is to create finer and finer approximations for $\mathcal{E}$ using fractional epsilon nets. Formally, let ϵ_j denote the minimal value of ϵ such that $N'(\mathcal{E},\epsilon_j)\le 2^{2^j}$. We will derive the following bound: with probability $1-\delta$,

$$\max_{E\in\mathcal{E}}|\boldsymbol{Y}_E| \le C\sqrt{n}\left(\sqrt{\log(1/\delta)}+\sum_{j=0}^{\infty}\epsilon_j\sqrt{\log N'(\mathcal{E},\epsilon_j)}\right) \le C\sqrt{n}\left(\sqrt{\log(1/\delta)}+\sum_{j=0}^{\infty}\epsilon_j 2^{j/2}\right). \quad (26)$$

This series in the right hand side is known to be equal, up to constant factors, to the following integral, which is known as Dudley's integral. We include the proof for completeness.

Lemma 9.17.

$$\sum_{j=0}^{\infty}\epsilon_j 2^{j/2} \le \frac{\sqrt{2}}{\sqrt{2}-1}\int_0^1\sqrt{\log N'(\mathcal{E},\epsilon)}d\epsilon.$$

Proof. Notice that

$$\begin{aligned}\sum_{j=0}^{\infty}\epsilon_j 2^{j/2} &\le \sum_{j=0}^{\infty}\sum_{i=j}^{\infty}(\epsilon_i-\epsilon_{i+1})2^{j/2} = \sum_{i=0}^{\infty}\sum_{j\le i}(\epsilon_i-\epsilon_{i+1})2^{j/2}\\ &= \sum_{i=0}^{\infty}(\epsilon_i-\epsilon_{i+1})\frac{\sqrt{2}^{i+1}-1}{\sqrt{2}-1} \le \frac{\sqrt{2}}{\sqrt{2}-1}\sum_{i=0}^{\infty}(\epsilon_i-\epsilon_{i+1})2^{i/2}.\end{aligned}$$

Further, by definition of ϵ_i we have that for any $\epsilon<\epsilon_i$, $N'(\mathcal{E},\epsilon)>2^{2^i}$, hence

$$\sqrt{\log N'(\mathcal{E},\epsilon)} > 2^{i/2}.$$

This implies that

$$\sum_{i=0}^{\infty}(\epsilon_i-\epsilon_{i+1})2^{i/2} \le \sum_{i=0}^{\infty}\int_{\epsilon_{i+1}}^{\epsilon_i}\sqrt{\log N'(\mathcal{E},\epsilon)} \le \int_0^1\sqrt{\log N'(\mathcal{E},\epsilon)}.$$

□

We start with some definitions; for any dynamic sets $\mathbb{B}$ and $\mathbb{B}'$ and $E\in\mathcal{E}$:

- Let $\mathbb{B}\setminus\mathbb{B}'$ be defined by $(\mathbb{B}\setminus\mathbb{B}')(x_1,\ldots,x_n) = \mathbb{B}(x_1,\ldots,x_n)\setminus\mathbb{B}'(x_1,\ldots,x_n)$. Notice that $\mathbb{B}\setminus\mathbb{B}'$, as defined, is a dynamic set.

- Define $\mathbb{B}_{\le m}$ as the dynamic sets that simulates $\mathbb{B}$ up to the point where $\mathbb{B}$ has m elements, and then it stops adding elements.
- Let $\overline{\boldsymbol{x}} = \overline{x}(\mathcal{A}, \boldsymbol{I})$ denote the stream that is output by the adversary $\mathcal{A}$ in interaction with the sampler that samples the coordinates from $\boldsymbol{I}$.

Let μ_j be a probability measure over dynamic sets that is a fractional $(\epsilon_j, N'(\mathcal{E}, \epsilon_j))$-cover for $\mathcal{E}$. We will approximate each $E \in \mathcal{E}$ using dynamic sets $\pi_{E,0}, \pi_{E,1}, \pi_{E,2}, \ldots$, where $\pi_{E,j} \in \text{support}(\mu_j)$ is an ϵ-approximation for E, namely,

$$|(E \cap \overline{\boldsymbol{x}}) \triangle \pi_{E,j}(\overline{\boldsymbol{x}})| \le \epsilon^2 n,$$

Notice that by definition of (fractional) covers, $\pi_{E,j}$ may depend on the stream $\overline{\boldsymbol{x}}$, which is a random variable, hence $\pi_{E,j}$ is also a random variable. The following lemma shows that for a sufficiently large j, $\pi_{E,j}(\overline{\boldsymbol{x}})$ equals $E \cap \overline{\boldsymbol{x}}$.

Lemma 9.18. *Assume that $N'(\mathcal{E}, \epsilon) < \infty$ for all $\epsilon > 0$. Then there exists $j_1 > 0$ such that for all $E \in \mathcal{E}$ and all $j \ge j_1$, $E \cap \overline{\boldsymbol{x}} = \pi_{E,j}(\overline{\boldsymbol{x}})$ holds with probability 1 over $\overline{\boldsymbol{x}}$. Consequently, $\boldsymbol{Y}_E = \boldsymbol{Y}_{\pi_{E,j}}$ for all $j \ge j_1$.*

Proof. The assumption of the lemma implies that for some j, $\epsilon_j < n^{-1/2}$ and let j_1 be the minimal such value. By definition of $\pi_{E,j}$ we have that for all $j \ge j_1$, $|(E \cap \overline{\boldsymbol{x}}) \triangle \pi_{E,j}(\overline{\boldsymbol{x}})| < 1$, hence $E \cap \overline{\boldsymbol{x}} = \pi_{E,j}(\overline{\boldsymbol{x}})$ as required. □

We can assume that $N'(\mathcal{E}, \epsilon) < \infty$ for all $\epsilon > 0$ otherwise Dudley's integral (appearing in Lemma 9.7) would diverge. Hence, by Lemma 9.18, for any $j_0 \ge 0$,

$$|\boldsymbol{Y}_E| := \left| \boldsymbol{Y}_{\pi_{E,j_0}} + \sum_{j=j_0}^{\infty} \left(\boldsymbol{Y}_{\pi_{E,j+1}} - \boldsymbol{Y}_{\pi_{E,j}} \right) \right| \le \left| \boldsymbol{Y}_{\pi_{E,j_0}} \right| + \sum_{j=j_0}^{\infty} \left| \boldsymbol{Y}_{\pi_{E,j+1}} - \boldsymbol{Y}_{\pi_{E,j}} \right| \tag{27}$$

$$= \left| \boldsymbol{Y}_{\pi_{E,j_0}} \right| + \sum_{j=j_0}^{\infty} \left| \boldsymbol{Y}_{\pi_{E,j+1} \setminus \pi_{E,j}} - \boldsymbol{Y}_{\pi_{E,j} \setminus \pi_{E,j+1}} \right| \le \left| \boldsymbol{Y}_{\pi_{E,j_0}} \right| + \sum_{j=j_0}^{\infty} \left| \boldsymbol{Y}_{\pi_{E,j+1} \setminus \pi_{E,j}} \right| + \sum_{j=j_0}^{\infty} \left| \boldsymbol{Y}_{\pi_{E,j} \setminus \pi_{E,j+1}} \right|. \tag{28}$$

We bound the supremum over $E \in \mathcal{E}$ by taking the supremum over each term separately:

$$\sup_{E \in \mathcal{E}} |\boldsymbol{Y}_E| \le \sup_{E \in \mathcal{E}} \left| \boldsymbol{Y}_{\pi_{E,j_0}} \right| + \sum_{j=j_0}^{\infty} \sup_{E \in \mathcal{E}} \left| \boldsymbol{Y}_{\pi_{E,j+1} \setminus \pi_{E,j}} \right| + \sum_{j=j_0}^{\infty} \sup_{E \in \mathcal{E}} \left| \boldsymbol{Y}_{\pi_{E,j} \setminus \pi_{E,j+1}} \right|. \tag{29}$$

Each supremum will be bounded using the generalized union bound Lemma 9.13.

Next, we define measures over differences of dynamic sets, that will be used to bound the right hand side of (29). For any $j \ge 1$ we let $\mu_{j,0}$ denote a probability measure over dynamic sets such that $\mathbb{B} \sim \mu_{j,0}$ is drawn by first drawing $\mathbb{B}_j \sim \mu_j$ and $\mathbb{B}_{j+1} \sim \mu_{j+1}$ and then outputting $\mathbb{B} = (\mathbb{B}_j \setminus \mathbb{B}_{j+1})_{\le 2\epsilon_j^2}$. Similarly, let $\mu_{j,1}$ denote the measure that outputs $(\mathbb{B}_{j+1} \setminus \mathbb{B}_j)_{\le 2\epsilon_j^2 n}$. By the generalized union bound, we have the following:

Lemma 9.19. *Let $j_0 = \lceil \log_2 \log_2(1/\delta) \rceil$. With probability at least $1 - \delta$, the following holds:*

- *It holds that*

$$\mu_{j_0} \left(\left\{ \mathbb{B} \in \text{support}(\mu_{j_0}) \colon |\boldsymbol{Y}_{\mathbb{B}}| > C_0 \sqrt{\log(1/\delta)} \sqrt{n} \right\} \right) \le \frac{1}{3 \cdot 2^{2^{j_0}}}.$$

- *For all $j \geq j_0$ and all $b \in \{0,1\}$,*

$$\mu_{j,b}\left(\left\{\mathbb{B} \in \text{support}(\mu_{j,b})\colon |\boldsymbol{Y}_{\mathbb{B}}| > C_0\epsilon_j 2^{j/2}\sqrt{n}\right\}\right) \leq \frac{1}{18 \cdot 2^{2^j} \cdot 2^{2^{j+1}}}.$$

Proof. The proof follows directly from Lemma 9.14. First, we show that the first item holds with probability $1 - \delta/2$: it follows by substituting in Lemma 9.14 the values $m = n$, $\delta' = \delta/2$ and $\alpha = \frac{1}{3 \cdot 2^{2^{j_0}}}$, and notice that

$$\log_2(1/\alpha) = \log_2 3 + 2^{j_0} \leq C \cdot \log(1/\delta),$$

by definition of j_0.

For the second item, we show that the term corresponding to a specific $j \geq j_0$ and $b \in \{0,1\}$ holds with probability $1 - 2^{-j-3}\delta$. Indeed, we can substitute $m = 2\epsilon_j^2 n$, $\delta' = 2^{-j-3}\delta$ and $\alpha = \frac{1}{18 \cdot 2^{2^j} \cdot 2^{2^{j+1}}}$, and notice that

$$\log_2 \frac{1}{\delta'} = \log_2 \frac{1}{\delta} + j + 3 \leq C \cdot 2^j,$$

since $j \geq j_0$ and by definition of j_0. Further,

$$\log_2 \frac{1}{\alpha} = \log_2 18 + 2^{j+2} \leq C \cdot 2^j.$$

By a union bound, the failure probability is bounded by

$$\delta/2 + 2 \cdot \sum_{j=j_0}^{\infty} 2^{-j-3}\delta \leq \delta/2 + 2^{-j_0-1}\delta \leq \delta/2 + \delta/2 = \delta.$$

□

For the remainder of the proof, we fix some stream $\overline{x}$ such that the condition in Lemma 9.19 holds when $\overline{\boldsymbol{x}} = \overline{x}$. This fixes values $\{Y_E\}_{E \in \mathcal{E}}$ and $\{Y_{\mathbb{B}}\}_{\mathbb{B} \text{ dynamic set}}$ such that $\boldsymbol{Y}_E = Y_E$ and $\boldsymbol{Y}_{\mathbb{B}} = Y_{\mathbb{B}}$ for all E and $\mathbb{B}$. We will show that for any $E \in \mathcal{E}$

$$|\boldsymbol{Y}_E| \leq C\sqrt{n \log(1/\delta)} + C \sum_{j=0}^{\infty} 2^{j/2}\epsilon_j \sqrt{n},$$

which suffices to complete the proof from Lemma 9.17. Fix $E \in \mathcal{E}$; we show how to define $\pi_{E,j}$. First, for any $j \geq j_0$, let A_j denote the set of elements $\mathbb{B} \in \text{support}(\mu_j)$ such that $|(E \cap \overline{x}) \triangle \mathbb{B}(\overline{x})| \leq \epsilon_j^2 n$. By the property of the fractional cover μ_j, we know that

$$\mu_j(A_j) \geq 1/2^{2^j}. \tag{30}$$

The set A_j contains the candidates for $\pi_{E,j}$. Notice that in order to bound (29), we would like to bound $Y_{\pi_{E,j} \setminus \pi_{E,j-1}}$ and $Y_{\pi_{E,j-1} \setminus \pi_{E,j}}$. For this purpose, we now define for any $j \geq j_0$ the function $R\colon \text{support}(\mu_j) \times \text{support}(\mu_{j+1}) \to \{0,1\}$, that indicates which pairs of elements $\mathbb{B}_j, \mathbb{B}_{j+1}$ are not suitable to be defined as $\pi_{E,j}$ and $\pi_{E,j+1}$:

$$R(\mathbb{B}_j, \mathbb{B}_{j+1}) = \begin{cases} 1 & \max\left(|Y_{\mathbb{B}_j \setminus \mathbb{B}_{j+1}}|, |Y_{\mathbb{B}_{j+1} \setminus \mathbb{B}_j}|\right) > C_0\epsilon_j 2^{j/2}\sqrt{n},\ \mathbb{B}_j \in A_j,\ \mathbb{B}_{j+1} \in A_{j+1} \\ 0 & \textit{otherwise} \end{cases},$$

where C_0 is the constant from Lemma 9.19. Next, we further restrict the set of candidates by creating a set $A'_j \subseteq A_j$, that contains only dynamic sets $\mathbb{B}_j \in A_j$ such that for many elements $\mathbb{B}_{j+1} \in A_{j+1}$, the pair $(\mathbb{B}_j, \mathbb{B}_{j+1})$ is suitable, which is formally defined as:

$$A'_j = \left\{ \mathbb{B}_j \in A_j \colon \mu_{j+1}\left(\left\{\mathbb{B}_{j+1} \colon R(\mathbb{B}_j, \mathbb{B}_{j+1}) = 1\right\}\right) \leq \frac{1}{3 \cdot 2^{2^{j+1}}} \right\}.$$

Next, we lower bound the measure $\mu_j(A'_j)$:

Lemma 9.20. *Assume that the high probability event from Lemma 9.19 holds. Then, for any $j \geq j_0$,*

$$\mu_j(A'_j) \geq \frac{2}{3 \cdot 2^{2^j}}.$$

Proof. Fix some stream $\overline{x}$ such that the high probability event of Lemma 9.19 holds; this fixes the values of $Y_{\mathbb{B}}$ for all $\mathbb{B}$. If $\mathbb{B}_j \sim \mu_j$ and $\mathbb{B}_{j+1} \sim \mu_{j+1}$ are drawn independently, then

$$\begin{aligned}
&\mathbb{E}_{\mathbb{B}_j \sim \mu_j} \mathbb{E}_{\mathbb{B}_{j+1} \sim \mu_{j+1}}[R(\mathbb{B}_j, \mathbb{B}_{j+1})] = \Pr_{\mathbb{B}_j \sim \mu_j, \mathbb{B}_{j+1} \sim \mu_{j+1}}[R(\mathbb{B}_j, \mathbb{B}_{j+1}) = 1] \\
&= \Pr_{\mathbb{B}_j, \mathbb{B}_{j+1}}\left[\max\left(|Y_{\mathbb{B}_j \setminus \mathbb{B}_{j+1}}|, |Y_{\mathbb{B}_{j+1} \setminus \mathbb{B}_j}|\right) > C_0 \epsilon_j 2^{j/2}\sqrt{n},\ \mathbb{B}_j \in A_j,\ \mathbb{B}_{j+1} \in A_{j+1}\right] \\
&\leq \Pr_{\mathbb{B}_j, \mathbb{B}_{j+1}}\left[|Y_{\mathbb{B}_j \setminus \mathbb{B}_{j+1}}| > C_0 \epsilon_j 2^{j/2}\sqrt{n},\ \mathbb{B}_j \in A_j,\ \mathbb{B}_{j+1} \in A_{j+1}\right] \qquad (31) \\
&\quad + \Pr_{\mathbb{B}_j, \mathbb{B}_{j+1}}\left[|Y_{\mathbb{B}_{j+1} \setminus \mathbb{B}_j}| > C_0 \epsilon_j 2^{j/2}\sqrt{n},\ \mathbb{B}_j \in A_j,\ \mathbb{B}_{j+1} \in A_{j+1}\right], \qquad (32)
\end{aligned}$$

We focus on bounding (31); (32) is bounded in a similar fashion. For any $\mathbb{B}_j \in A_j$ and $\mathbb{B}_{j+1} \in A_{j+1}$ we have that by definition of A_j and A_{j+1},

$$\begin{aligned}
|(\mathbb{B}_j \setminus \mathbb{B}_{j+1})(\overline{x})| &= |\mathbb{B}_j(\overline{x}) \setminus \mathbb{B}_{j+1}(\overline{x})| \leq |\mathbb{B}_j(\overline{x}) \triangle \mathbb{B}_{j+1}(\overline{x})| \leq |\mathbb{B}_j(\overline{x}) \triangle E| + |E \triangle \mathbb{B}_{j+1}(\overline{x})| \\
&\leq \epsilon_j^2 n + \epsilon_{j+1}^2 n \leq 2\epsilon_j^2 n,
\end{aligned}$$

This implies that

$$|(\mathbb{B}_j \setminus \mathbb{B}_{j+1})(\overline{x})| = |(\mathbb{B}_j \setminus \mathbb{B}_{j+1})_{\leq 2\epsilon_j^2 n}(\overline{x})|,$$

hence by definition of $\mu_{j,0}$ and by Lemma 9.19,

$$\begin{aligned}
&\Pr_{\mathbb{B}_j \mathbb{B}_{j+1}}\left[|Y_{\mathbb{B}_j \setminus \mathbb{B}_{j+1}}| > C_0 \epsilon_j 2^{j/2}\sqrt{n},\ \mathbb{B}_j \in A_j,\ \mathbb{B}_{j+1} \in A_{j+1}\right] \\
&\leq \Pr_{\mathbb{B}_j, \mathbb{B}_{j+1}}\left[\left|Y_{(\mathbb{B}_j \setminus \mathbb{B}_{j+1})_{\leq 2\epsilon_j^2}}\right| > C_0 \epsilon_j 2^{j/2}\sqrt{n}\right] = \Pr_{\mathbb{B} \sim \mu_{j,0}}\left[|Y_{\mathbb{B}}| > C_0 \epsilon_j 2^{j/2}\sqrt{n}\right] \\
&\leq \frac{1}{18 \cdot 2^{2^j} \cdot 2^{2^{j+1}}}.
\end{aligned}$$

Similarly, (32) is bounded by the same quantity, hence by (31) and (32) above,

$$\mathbb{E}_{\mathbb{B}_j \sim \mu_j} \mathbb{E}_{\mathbb{B}_{j+1} \sim \mu_{j+1}}[R(\mathbb{B}_j, \mathbb{B}_{j+1})] \leq \frac{1}{9 \cdot 2^{2^j} \cdot 2^{2^{j+1}}}.$$

By Markov's inequality,

$$\begin{aligned}\Pr_{\mathbb{B}_j\sim\mu_j}\left[\mathbb{B}_j \in A_j \setminus A'_j\right] &= \Pr_{\mathbb{B}_j\sim\mu_j}\left[\mu_{j+1}\left(\left\{\mathbb{B}_{j+1}\colon R(\mathbb{B}_j,\mathbb{B}_{j+1})=1\right\}\right) > \frac{1}{3\cdot 2^{2^{j+1}}}\right]\\ &= \Pr_{\mathbb{B}_j\sim\mu_j}\left[\mathbb{E}_{\mathbb{B}_{j+1}\sim\mu_{j+1}}[R(\mathbb{B}_j,\mathbb{B}_{j+1})] > \frac{1}{3\cdot 2^{2^{j+1}}}\right]\\ &\le \frac{\mathbb{E}_{\mathbb{B}_j\sim\mu_j}\mathbb{E}_{\mathbb{B}_{j+1}\sim\mu_{j+1}}[R(\mathbb{B}_j,\mathbb{B}_{j+1})]}{1/(3\cdot 2^{2^{j+1}})} \le \frac{1}{3\cdot 2^{2^j}}.\end{aligned}$$

By (30),

$$\mu_j(A'_j) \ge \mu_j(A_j) - \mu_j(A_j\setminus A'_j) \ge \frac{1}{2^{2^j}} - \frac{1}{3\cdot 2^{2^j}} = \frac{2}{3\cdot 2^{2^j}}.$$

□

To complete the proof, we show how to define $\pi_{E,j}$ inductively on j such that $R(\pi_{E,j},\pi_{E,j+1}) = 0$ for all $j \ge j_0$. First, we select π_{E,j_0} to be any element of A'_j such that $|Y_{\pi_{E,j_0}}| \le C_0\sqrt{\log(1/\delta)}\sqrt{n}$; such an element exists from Lemma 9.19 and Lemma 9.20. Next, assume for $j \ge j_0$ that $\pi_{E,j}$ was already selected and select $\pi_{E,j+1}$ to be any element $\mathbb{B}_{j+1} \in A'_{j+1}$ such that $R(\pi_{E,j},\mathbb{B}_{j+1}) = 0$. Such an element exists since $\pi_{E,j} \in A'_j$, hence by definition of A'_j,

$$\mu_{j+1}\left(\left\{\mathbb{B}_{j+1}\colon R(\pi_{E,j},\mathbb{B}_{j+1})=1\right\}\right) \le \frac{1}{3\cdot 2^{2^{j+1}}},$$

while $\mu_{j+1}(A'_{j+1}) \ge 2/(3\cdot 2^{2^{j+1}})$ by Lemma 9.20. By (28) and since we defined $\pi_{E,j}$ such that $|Y_{\pi_{E,j_0}}| \le C_0\sqrt{\log(1/\delta)}\sqrt{n}$, $R(\pi_{E,j},\pi_{E,j+1}) = 0$ and $\pi_{E,j} \in A_j$ for all j, we have that

$$|Y_E| \le \left|Y_{\pi_{E,j_0}}\right| + \sum_{j=j_0}^{\infty}\left|Y_{\pi_{E,j+1}\setminus\pi_{E,j}}\right| + \sum_{j=j_0}^{\infty}\left|Y_{\pi_{E,j}\setminus\pi_{E,j+1}}\right| \le C_0\sqrt{n}\left(\sqrt{\log(1/\delta)} + 2\sum_{j=j_0}^{\infty}\epsilon_j 2^{j/2}\right).$$

This proves (26) as required.

9.4 Bounds on ϵ-Nets via Fractional Covering Numbers

The goal of this section is to prove Lemma 9.8. First, we use the following Martingale bound (see Section A.1 for an introduction to Martingales):

Lemma 9.21 (Freedman's inequality [Fre75])**.** *Let $\boldsymbol{y}_0,\ldots,\boldsymbol{y}_n$ be a Martingale adapted to the filtration $F_0,\ldots,F_n$, such that $\sum_{i=1}^n \mathbb{E}[(\boldsymbol{y}_i-\boldsymbol{y}_{i-1})^2 \mid F_{i-1}] \le s$ holds almost surely for some $s>0$. Further, assume that $|\boldsymbol{y}_i-\boldsymbol{y}_{i-1}| \le M$ for all i. Then, for any $t>0$,*

$$\Pr\left[\boldsymbol{y}_n - \boldsymbol{y}_0 \ge t\right] \le \exp\left(-\frac{t^2}{2(s+Mt)}\right).$$

We derive the following concentration bound for a dynamic set:

Lemma 9.22. *Let $\mathbb{B}$ be a dynamic set with $|\mathbb{B}| \le m$. Let $\boldsymbol{I} \sim \mathrm{Ber}(n,p)$. Let $\overline{\boldsymbol{x}} = \overline{\boldsymbol{x}}(\mathcal{A},\boldsymbol{I})$. Then, for any $t \ge 0$,*

$$\Pr\left[|\mathbb{B}(\overline{\boldsymbol{x}}) \cap \overline{\boldsymbol{x}}_{\boldsymbol{I}}| \le p|\mathbb{B}(\overline{\boldsymbol{x}})| - t\right] \le \exp(-ct^2/(mp+t)).$$

Proof. This follows from Lemma 9.21. We apply this lemma with $\boldsymbol{y}_i = |\mathbb{B}(\overline{\boldsymbol{x}}) \cap \overline{\boldsymbol{x}}_{\boldsymbol{I}\cap[i]}| - p|\mathbb{B}(\overline{\boldsymbol{x}}) \cap \overline{\boldsymbol{x}}_{[i]}|$ and $F_i = \sigma(\boldsymbol{I} \cap [i])$. We can substitute $s = c'mp$, due to the following reason: conditioned on F_{i-1}, we know $\overline{\boldsymbol{x}}_{[i]}$, which implies that we know whether $\boldsymbol{x}_i \in \mathbb{B}(\overline{\boldsymbol{x}})$. If this holds true, then,

$$\boldsymbol{y}_i - \boldsymbol{y}_{i-1} = \begin{cases} 1-p & \text{if } i \in \boldsymbol{I} \text{ (holds with probability } p) \\ -p & \text{otherwise (holds with probability } 1-p) \end{cases}.$$

By simple calculations we have $\mathbb{E}[(\boldsymbol{y}_i - \boldsymbol{y}_{i-1})^2 \mid F_{i-1}] \leq c'p$ in this case. If $\boldsymbol{x}_i \notin \mathbb{B}(\overline{\boldsymbol{x}})$, then $\boldsymbol{y}_i = \boldsymbol{y}_{i-1}$ and we have $\mathbb{E}[(\boldsymbol{y}_i - \boldsymbol{y}_{i-1})^2 \mid F_{i-1}] = 0$. Since $\overline{x}_i \in \mathbb{B}(\overline{\boldsymbol{x}})$ can hold true for at most m values of i, then $\sum_{i=1}^{n} \mathbb{E}[(\boldsymbol{y}_i - \boldsymbol{y}_{i-1})^2 \mid F_{i-1}] \leq c'mp$. Further, we can substitute $M = 1$ in Lemma 9.21, and the result follows. □

We are ready to prove Lemma 9.8:

Proof of Lemma 9.8. Let $\mathcal{N}$ be a 0-net for $\mathcal{E}$ with minimal cardinality. For each $\mathbb{B} \in \mathcal{N}$, let $\mathbb{B}_{\leq m}$ be the dynamic set that simulates $\mathbb{B}$ up to the point that it has retained m elements, and then it discards all the remaining elements. We apply Lemma 9.22 on $\mathbb{B}_{\leq m}$ to obtain that

$$\begin{aligned}\Pr[|\mathbb{B}(\overline{\boldsymbol{x}})| \geq m,\ |\mathbb{B}(\overline{\boldsymbol{x}}) \cap \overline{\boldsymbol{x}}_{\boldsymbol{I}}| \leq mp/2] &\leq \Pr[|\mathbb{B}_{\leq m}(\overline{\boldsymbol{x}})| = m,\ |\mathbb{B}_{\leq m}(\overline{\boldsymbol{x}}) \cap \overline{\boldsymbol{x}}_{\boldsymbol{I}}| \leq mp/2] \\ &\leq \Pr[|\mathbb{B}_{\leq m}(\overline{\boldsymbol{x}}) \cap \overline{\boldsymbol{x}}_{\boldsymbol{I}}| \leq p|\mathbb{B}_{\leq m}(\overline{\boldsymbol{x}})| - mp/2] \leq \exp(-cmp).\end{aligned}$$

The proof follows by a union bound over $\mathbb{B} \in \mathcal{N}$. □

10 Reductions Between Different Sampling Schemes

This section establishes a framework that enables one to obtain bounds with respect to one sampler in terms of bounds with respect to a different sampler. In particular, this shows how, given bounds on ϵ-nets and ϵ-approximations for the uniform sampler, one can obtain bounds for the Bernoulli and the reservoir sampler. Section 10.1 presents an overview of an abstract method to reduce between two sampling schemes, that is formally presented in Section 10.2. Section 10.3 shows how to obtain bounds with respect to the reservoir sampler given bounds for the uniform sampler. Section 10.4 presents bounds on the Bernoulli sampler based on bounds for the uniform sampler. Finally, Section 10.5 provides bounds with respect to the uniform sampler based on the Bernoulli sampler, which proves the auxiliary Lemma 6.2 and Lemma 7.2. Both Section 10.3 and Section 10.5 use the abstract reduction method of Section 10.2 while Section 10.4 utilizes the fact that the Bernoulli sampler can be presented as a mixture of uniform samplers $\mathrm{Uni}(n,k)$ for different values of k.

10.1 Intuition for the Reduction Method

For convenience, we consider samplers with no deletions, that are characterized by some distribution over subsets of $[n]$, such as $\mathrm{Uni}(n,k)$, that is the uniform distribution over subsets of $[n]$ of size $[k]$. We use $\boldsymbol{I}$ and $\boldsymbol{I}'$ to denote such random variables over subsets of $[n]$ (e.g., they can be distributed $\mathrm{Ber}(n,p)$ or $\mathrm{Uni}(n,k)$).

In these reductions, our goal is to show that one sampling scheme $\boldsymbol{I}$ is at least as good as a different scheme $\boldsymbol{I}'$. For example, that $\boldsymbol{I}$ attains ϵ-approximations for values of ϵ smaller than

those attained by $\boldsymbol{I}'$. In other words, we would like to say that $\boldsymbol{I}$ is resilient to the adversary at least as well as $\boldsymbol{I}'$. The above is equivalent to saying that the *worst* adversary for $\boldsymbol{I}'$ *is at least as bad* as the worst adversary for $\boldsymbol{I}$. The above can be shown by reduction: given an adversary $\mathcal{A}$ that is *bad* for $\boldsymbol{I}$, we will construct an adversary $\mathcal{A}'$ that is bad for $\boldsymbol{I}'$.

Here we define $\mathcal{A}'$, that plays against a sampler that samples $\boldsymbol{I}'$. The general idea for $\mathcal{A}'$ is to simulate $\mathcal{A}$. However, $\mathcal{A}$ is known to be bad against $\boldsymbol{I}$ while $\mathcal{A}'$ plays against $\boldsymbol{I}'$. To tackle this issue, $\mathcal{A}'$ will simulate a sample $\boldsymbol{J}$ that has the same distribution as $\boldsymbol{I}$ and then simulate the actions of $\mathcal{A}$ against $\boldsymbol{J}$. Then, $\mathcal{A}'$ will output the same stream output by the simulated $\mathcal{A}$.

We would like to show that $\mathcal{A}'$ is bad against $\boldsymbol{I}'$. In order to show that, we will have to assume that the simulated sample $\boldsymbol{J}$ is very close to the true sample $\boldsymbol{I}'$ with high probability (say, in symmetric difference of sets). Since $\mathcal{A}$ is bad against $\boldsymbol{I}$ and $\boldsymbol{J} \sim \boldsymbol{I}$, the simulated actions of $\mathcal{A}$ are bad against the simulated sample $\boldsymbol{J}$. Since further $\boldsymbol{J}$ is very close to $\boldsymbol{I}'$, then the simulated $\mathcal{A}$ is bad also against $\boldsymbol{I}'$. Since $\mathcal{A}'$ outputs the same stream as the simulated $\mathcal{A}$, this implies that $\mathcal{A}'$ is bad against $\boldsymbol{I}'$, as required.

Notice that since $\mathcal{A}'$ would like the simulated sample $\boldsymbol{J}$ to be similar to the true sample $\boldsymbol{I}'$, it has to construct $\boldsymbol{J}$ based on $\boldsymbol{I}'$ and this defines a joint probability distribution between $\boldsymbol{I}'$ and $\boldsymbol{J}$. Such a joint distribution is called *coupling*. Further, the simulation has to be performed in an online fashion: once $\mathcal{A}'$ receives the actions taken by the sampler $\boldsymbol{I}'$ that it plays against, it has to immediately simulate the actions of the simulated sample $\boldsymbol{J}$. In particular, once $\mathcal{A}'$ knows whether $t \in \boldsymbol{I}'$, it has to decide whether $t \in \boldsymbol{J}$. Since the coupling between $\boldsymbol{J}$ and $\boldsymbol{I}'$ is constructed in an online fashion, we denote it an *online coupling*.

Using the notation above, the goal of $\mathcal{A}'$ is to construct an online coupling of $\boldsymbol{J}$ and $\boldsymbol{I}'$ such that $\boldsymbol{J} \sim \boldsymbol{I}$ and such that with high probability, the symmetric set difference between $\boldsymbol{J}$ and $\boldsymbol{I}'$ is small. This can be done, for example, if $\boldsymbol{J} \sim \mathrm{Uni}(2k, k)$ and $\boldsymbol{I}' \sim \mathrm{Ber}(2k, 1/2)$: there, $\mathcal{A}'$ will have to omit or add a approximately $O(\sqrt{k})$ elements to $\boldsymbol{I}'$ to create $\boldsymbol{J}$.

10.2 Abstract Reduction Method

We refer to sampling schemes that are oblivious to the adversary, namely that the choice to retain or discard an element is independent of the stream. Formally, we denote by $\boldsymbol{I}_t$ the set of indices of elements retained by the algorithm after seeing $\overline{x}_{[t]}$, for $t \in [n]$. An *oblivious sampling scheme* is one where $\overline{\boldsymbol{I}} = (\boldsymbol{I}_1, \ldots, \boldsymbol{I}_n)$ are random variables jointly distributed, that are not a function of the adversary $\mathcal{A}$. This section compares one oblivious sampling scheme $\overline{\boldsymbol{I}} := (\boldsymbol{I}_1, \ldots, \boldsymbol{I}_n)$ with another, $\overline{\boldsymbol{I}}' = (\boldsymbol{I}'_1, \ldots, \boldsymbol{I}'_n)$. It will be shown that if the adversary, given an input stream $\overline{\boldsymbol{I}}'$ can simulate a stream that is distributed according to $\overline{\boldsymbol{I}}$ such that with high probability, the input stream is close to the output stream in some sense, then the sampling scheme $\overline{\boldsymbol{I}}$ is at least as resilient to the adversary as $\overline{\boldsymbol{I}}'$. We begin with the following definition of online simulation:

Definition 10.1. *An* online simulator $\mathcal{S}$ *is an algorithm that receives a stream $\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_n$ of subsets of $[n]$ and an unlimited pool of independent random bits and outputs a stream $\boldsymbol{I}'_1, \ldots, \boldsymbol{I}'_n$, such that $\boldsymbol{I}'_t$ has to be computed before seeing $\boldsymbol{I}_{t+1}$, for $t \in [n]$. In other words, $\boldsymbol{I}'_t$ depends only on $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_t, \boldsymbol{I}'_1, \ldots, \boldsymbol{I}'_{t-1}$ and on the randomness of the simulator. The joint distribution of $\overline{\boldsymbol{I}}$ and $\overline{\boldsymbol{I}}'$ is called an* online coupling *of $\overline{\boldsymbol{I}}$ to $\overline{\boldsymbol{I}}'$. Equivalent, we can say that $\overline{\boldsymbol{I}}$ is online coupled to $\overline{\boldsymbol{I}}'$.*

Notice that an online coupling of $\overline{\boldsymbol{I}}$ and $\overline{\boldsymbol{I}}'$ is also a *coupling* of these two random variables, which is any joint distribution between them. Further, notice that an online coupling is not a

symmetric notion: an online coupling of $\overline{\boldsymbol{I}}$ to $\overline{\boldsymbol{I}}'$ is not necessarily an online coupling of $\overline{\boldsymbol{I}}'$ to $\overline{\boldsymbol{I}}$.

In cases that the sampler cannot delete elements from its sample, notice that $\boldsymbol{I}_t = \boldsymbol{I}_n \cap [t]$. To simplify the notation, we can write $\boldsymbol{I} = \boldsymbol{I}_n$ and $\boldsymbol{I} \cap [t] = \boldsymbol{I}_t$. Hence, we have the following definition of online coupling for no-deletion samplers:

Definition 10.2. *Let $\boldsymbol{I}$ and $\boldsymbol{I}'$ be jointly distributed random variables over $[n]$. We say that $\boldsymbol{I}$ is* online coupled *to $\boldsymbol{I}'$ if $(\boldsymbol{I} \cap [1], \boldsymbol{I} \cap [2], \ldots, \boldsymbol{I} \cap [n])$ is online coupled to $(\boldsymbol{I}' \cap [1], \boldsymbol{I}' \cap [2], \ldots, \boldsymbol{I}' \cap [n])$.*

In order to show that $\overline{\boldsymbol{I}}$ is more resilient to the adversary than $\overline{\boldsymbol{I}}'$, it suffices to find an online coupling of $\boldsymbol{I}$ to $\boldsymbol{I}'$ such that $\boldsymbol{I}_n$ is similar to $\boldsymbol{I}'_n$ in some sense. To be more formal, let $f(I_n, \overline{x})$ be some $\{0,1\}$-valued function that we view as an indicator denoting whether the sub-sample $\overline{x}_{I_n}$ fails to represent the full stream $\overline{x}$. For example, f can be an indicator of whether $\overline{x}_{I_n}$ is *not* an ϵ-approximation of $\overline{x}$. Our goal is to bound the failure probability with the worst adversary. Namely, to bound $\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\overline{\boldsymbol{I}}}[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}, \overline{\boldsymbol{I}})) = 1]$. Say that we already know how to bound a similar quantity for a different sampling scheme $\overline{\boldsymbol{I}}'$. If we can online couple $\overline{\boldsymbol{I}}$ to $\overline{\boldsymbol{I}}'$, then we can reduce between these two bounds:

Lemma 10.3. *Let $\overline{\boldsymbol{I}}$ be online coupled to $\overline{\boldsymbol{I}}'$. Let $f, g \colon \{0,1\}^n \times X^n \to \{0,1\}$. Then,*

$$\begin{aligned}
&\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\overline{\boldsymbol{I}}}[f(\boldsymbol{I}_n, \overline{\boldsymbol{x}}(\mathcal{A}, \overline{\boldsymbol{I}})) = 1] \\
&\quad \le \max_{\mathcal{A}' \in \mathrm{Adv}_n} \Pr_{\overline{\boldsymbol{I}}'}[g(\boldsymbol{I}'_n, \overline{\boldsymbol{x}}(\mathcal{A}', \overline{\boldsymbol{I}}')) = 1] + \Pr_{\overline{\boldsymbol{I}}, \overline{\boldsymbol{I}}'}[\exists \overline{x} \in X^n \; s.t \; f(\boldsymbol{I}_n, \overline{x}) = 1 \; and \; g(\boldsymbol{I}'_n, \overline{x}) = 0]. \qquad (33)
\end{aligned}$$

Notice the second term in the right hand side of (33): it equals zero if $\boldsymbol{I}_n = \boldsymbol{I}'_n$, and, it is expected to be small if $\boldsymbol{I}_n \approx \boldsymbol{I}'_n$ with high probability.

Proof. Let $\mathcal{A}_{\max}$ be the adversary that achieves the maximum on the left hand side of (33). Let $\mathcal{S}$ denote the simulation adversary that given $\overline{\boldsymbol{I}}'$ and some additional random string $\boldsymbol{r}$, outputs $\overline{\boldsymbol{I}} = \mathcal{S}(\overline{\boldsymbol{I}}', \boldsymbol{r})$. We will create the following adversary $\mathcal{A}'_{\boldsymbol{r}}$ that operates on the stream $\overline{\boldsymbol{I}}'$ and has additional randomness $\boldsymbol{r}$: it creates the sample $\overline{\boldsymbol{I}} = \mathcal{S}(\overline{\boldsymbol{I}}', \boldsymbol{r})$, simulates $\mathcal{A}_{\max}$ on this sample and outputs the same stream as the simulated $\mathcal{A}_{\max}$. In particular, we have

$$\overline{x}(\mathcal{A}_{\boldsymbol{r}}, \overline{\boldsymbol{I}}') = \overline{x}(\mathcal{A}_{\max}, \overline{\boldsymbol{I}}). \qquad (34)$$

We view $\mathcal{A}'_{\boldsymbol{r}}$ as a distribution over deterministic adversaries $\{\mathcal{A}'_r\}_{r \in \mathrm{support}(\boldsymbol{r})}$. Further, notice that each $\mathcal{A}'_r$ defines an appropriate adversary. By (34),

$$\begin{aligned}
&\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\overline{\boldsymbol{I}}} \left[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}, \overline{\boldsymbol{I}})) = 1 \right] = \Pr_{\overline{\boldsymbol{I}}} \left[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}_{\max}, \overline{\boldsymbol{I}})) = 1 \right] \\
&\quad = \Pr_{\overline{\boldsymbol{I}}, \overline{\boldsymbol{I}}', \boldsymbol{r}} \left[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}_{\max}, \overline{\boldsymbol{I}})) = 1 \text{ and } g(\boldsymbol{I}'_n, \overline{x}(\mathcal{A}'_{\boldsymbol{r}}, \overline{\boldsymbol{I}}')) = 1 \right] \\
&\qquad + \Pr_{\overline{\boldsymbol{I}}, \overline{\boldsymbol{I}}', \boldsymbol{r}} \left[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}_{\max}, \overline{\boldsymbol{I}})) = 1 \text{ and } g(\boldsymbol{I}'_n, \overline{x}(\mathcal{A}'_{\boldsymbol{r}}, \overline{\boldsymbol{I}}')) = 0 \right] \\
&\quad = \Pr_{\overline{\boldsymbol{I}}, \overline{\boldsymbol{I}}', \boldsymbol{r}} \left[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}_{\max}, \overline{\boldsymbol{I}})) = 1 \text{ and } g(\boldsymbol{I}'_n, \overline{x}(\mathcal{A}'_{\boldsymbol{r}}, \overline{\boldsymbol{I}}')) = 1 \right] \\
&\qquad + \Pr_{\overline{\boldsymbol{I}}, \overline{\boldsymbol{I}}', \boldsymbol{r}} \left[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}_{\max}, \overline{\boldsymbol{I}})) = 1 \text{ and } g(\boldsymbol{I}'_n, \overline{x}(\mathcal{A}_{\max}, \overline{\boldsymbol{I}})) = 0 \right] \\
&\quad \le \Pr_{\overline{\boldsymbol{I}}', \boldsymbol{r}} \left[g(\boldsymbol{I}'_n, \overline{x}(\mathcal{A}'_{\boldsymbol{r}}, \overline{\boldsymbol{I}}')) = 1 \right] + \Pr_{\overline{\boldsymbol{I}}, \overline{\boldsymbol{I}}'} \left[\exists \overline{x} \in X^n \text{ s.t } f(\overline{\boldsymbol{I}}, \overline{x}) = 1 \text{ and } g(\overline{\boldsymbol{I}}', \overline{x}) = 0 \right],
\end{aligned}$$

as required. □

10.3 Bounds for Reservoir Sampling via Uniform Sampling

Next, we show how to obtain bounds for reservoir sampling based on uniform sampling. The intuition is that the reservoir sampler gives less information than the uniform sampler: indeed, when the reservoir sampler selects an element, the adversary does not know whether this element will remain for the final sample, while this is not the case for the uniform sampler. The following holds:

Lemma 10.4. *The reservoir sampler $\overline{\boldsymbol{I}} = (\boldsymbol{I}_1, \ldots, \boldsymbol{I}_n) \sim \mathrm{Res}(n,k)$ can be online coupled to the Uniform sampler $\boldsymbol{I}' \sim \mathrm{Uni}(n,k)$ such that $\boldsymbol{I}_n = \boldsymbol{I}'$ with probability* 1.

Notice that we describe the uniform sampler using one index-set as it is an insertion-only scheme, while the reservoir sample has deletions hence we describe it using n index-sets. Lemma 10.4, in combination with Lemma 10.3, immediately implies that any high probability bound obtained for the uniform sampling, also holds true for the reservoir sampler:

Proof of Theorem 6.1, reservoir sampling. Let $f(I, \overline{x}) = g(I, \overline{x})$ denote an indicator of whether $\overline{x}_I$ fails to be an ϵ-approximation for $\overline{x}$. Let $\overline{\boldsymbol{I}}$ denote the reservoir sampler and let $\boldsymbol{I}'$ denote the uniform sampler. Then, by Lemma 10.4, we can online couple $\overline{\boldsymbol{I}}$ to $\boldsymbol{I}'$ such that $\boldsymbol{I}_n = \boldsymbol{I}'$. By Lemma 10.3, we have

$$\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\overline{\boldsymbol{I}}} \left[f(\boldsymbol{I}_n, \overline{x}(\mathcal{A}, \overline{\boldsymbol{I}})) = 1 \right] \leq \max_{\mathcal{A}' \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I}'} \left[g(\boldsymbol{I}', \overline{x}(\mathcal{A}', \boldsymbol{I}')) = 1 \right].$$

By Theorem 6.5, the right hand side is bounded by δ, for a suitable value of δ. Hence, the left hand side is bounded by the same quantity.

Lastly, notice that the assumption $n \geq 2k$ in Theorem 6.5 translates to $n \geq 3k$ in the reduction Lemma 10.4. □

Proof of Theorem 7.1, reservoir sampling. The proof follows the same steps as the proof for Theorem 6.1, while replacing ϵ-approximations with ϵ-nets and using Theorem 7.5 for the bound on the uniform sampler. □

Finally, we prove Lemma 10.4. First, an auxiliary lemma:

Lemma 10.5. *Let $\overline{\boldsymbol{I}} \sim \mathrm{Res}(n,k)$ and fix $t \in [n]$. Then, conditioned on $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{t-1}, \boldsymbol{I}_n \cap [t]$, it holds that $\boldsymbol{I}_t$ is independent of $\boldsymbol{I}_n$.*

Proof. The proof follows from the following steps:

- First, the conditional distribution of $\boldsymbol{I}_n$ conditioned on $\boldsymbol{I}_1 = I_1, \ldots, \boldsymbol{I}_t = I_t$ is only a function of I_t. That is due to the fact that $\boldsymbol{I}_1 \to \boldsymbol{I}_2 \to \cdots \to \boldsymbol{I}_n$ is a Markov chain.
- This implies that the conditional distribution of $\boldsymbol{I}_n$ conditioned on $\boldsymbol{I}_1 = I_1, \ldots, \boldsymbol{I}_t = I_t, \boldsymbol{I}_n \cap [t] = S$ is only a function of S and I_t.
- Conditioned on $\boldsymbol{I}_1 = I_1, \ldots, \boldsymbol{I}_t = I_t, \boldsymbol{I}_n \cap [t] = S$, we can write $\boldsymbol{I}_t = S \cup (I_t \setminus S)$. Notice that due to the symmetry of deletion, namely, that the deleted element is chosen uniformly at random, one derives that the conditional probability of $\boldsymbol{I}_n$ conditioned on $\boldsymbol{I}_1 = I_1, \ldots, \boldsymbol{I}_t = I_t, \boldsymbol{I}_n \cap [t] = S$ is not dependent on $I_t \setminus S$, hence it is only a function of S.

- The above implies that conditioned on $\boldsymbol{I}_n \cap [t] = S$, the random vector $(\boldsymbol{I}_1, \ldots, \boldsymbol{I}_t)$ is independent of $\boldsymbol{I}_n$.
- This further implies that conditioned on $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{t-1}, \boldsymbol{I}_n \cap [t]$, it holds that $\boldsymbol{I}_t$ is independent of $\boldsymbol{I}_n$.

□

The following is a well-known fact that can be proved by induction:

Lemma 10.6. *Let* $\overline{\boldsymbol{I}} = (\boldsymbol{I}_1, \ldots, \boldsymbol{I}_n) \sim \mathrm{Res}(n,k)$. *Then,* $\boldsymbol{I}_n \sim \mathrm{Uni}(n,k)$.

Using only Lemma 10.5 and Lemma 10.6, we can prove Lemma 10.4:

Proof of Lemma 10.4. Let $\boldsymbol{I}' \sim \mathrm{Uni}(n,k)$ and $\overline{\boldsymbol{I}} \sim \mathrm{Res}(n,k)$. We will define a random varible $\overline{\boldsymbol{J}}$ that is online coupled to $\boldsymbol{I}'$ and show that both $\boldsymbol{J}_n = \boldsymbol{I}'$ with probability 1 and that $\overline{\boldsymbol{J}}$ has the same distribution as $\overline{\boldsymbol{I}}$. The sample $\overline{\boldsymbol{J}}$ is created using the following inductive argument: for $t = 1, \ldots, n$, assume that we have already set $\boldsymbol{J}_1 = J_1, \ldots, \boldsymbol{J}_{t-1} = J_{t-1}$, and that $\boldsymbol{I}' = I'$, and recall that by definition of online simulation, we can set $\boldsymbol{J}_t$ to be any randomized function of $J_1, \ldots, J_{t-1}, I' \cap [t]$. Specifically, $\boldsymbol{J}_t$ is drawn from the following conditional distribution: for any J_t,

$$\begin{aligned}&\Pr\left[\boldsymbol{J}_t = J_t \mid \boldsymbol{J}_1 = J_1, \ldots, \boldsymbol{J}_{t-1} = J_{t-1}, \boldsymbol{I}' \cap [t] = I' \cap [t]\right]\\ &\quad = \Pr\left[\boldsymbol{I}_t = J_t \mid \boldsymbol{I}_1 = J_1, \ldots, \boldsymbol{I}_{t-1} = J_{t-1}, \boldsymbol{I}_n \cap [t] = I' \cap [t]\right]\end{aligned}$$

Using the fact that the random coins used by the algorithm are independent of the sample $\boldsymbol{I}'$, we derive that conditioned on $\boldsymbol{J}_1, \ldots, \boldsymbol{J}_{t-1}, \boldsymbol{I}' \cap [t]$, it holds that $\boldsymbol{J}_t$ is independent of $\boldsymbol{I}'$. In combination with Lemma 10.5, it follows that

$$\begin{aligned}&\Pr\left[\boldsymbol{J}_t = J_t \mid \boldsymbol{J}_1 = J_1, \ldots, \boldsymbol{J}_{t-1} = J_{t-1}, \boldsymbol{I}' = I'\right]\\ &\quad = \Pr\left[\boldsymbol{J}_t = J_t \mid \boldsymbol{J}_1 = J_1, \ldots, \boldsymbol{J}_{t-1} = J_{t-1}, \boldsymbol{I}' \cap [t] = I' \cap [t]\right]\\ &\quad = \Pr\left[\boldsymbol{I}_t = J_t \mid \boldsymbol{I}_1 = J_1, \ldots, \boldsymbol{I}_{t-1} = J_{t-1}, \boldsymbol{I}_n \cap [t] = I' \cap [t]\right]\\ &\quad = \Pr\left[\boldsymbol{I}_t = J_t \mid \boldsymbol{I}_1 = J_1, \ldots, \boldsymbol{I}_{t-1} = J_{t-1}, \boldsymbol{I}_n = I'\right]. \qquad (35)\end{aligned}$$

The following inductive argument shows that for $t \in \{0, \ldots, n\}$, the conditional distribution $\boldsymbol{J}_1, \ldots, \boldsymbol{J}_t$ conditioned on $\boldsymbol{I}' = I'$ equals the conditional distribution of $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_t$ conditioned on $\boldsymbol{I}_n = I'$. For the base of the induction, $t = 0$, there is nothing to prove. For the induction step, assume that the above holds for $t-1$ and we will prove for t. Indeed, by the chain rule, the induction hypothesis, and (35),

$$\begin{aligned}&\Pr[\boldsymbol{J}_1 = J_1, \ldots, \boldsymbol{J}_t = J_t \mid \boldsymbol{I}' = I'] && (36)\\ &\quad = \Pr[\boldsymbol{J}_1 = J_1, \ldots, \boldsymbol{J}_{t-1} = J_{t-1} \mid \boldsymbol{I}' = I'] \Pr[\boldsymbol{J}_t = J_t \mid \boldsymbol{J}_1 = J_1, \ldots, \boldsymbol{J}_{t-1} = J_{t-1}, \boldsymbol{I}' = I'] && (37)\\ &\quad = \Pr[\boldsymbol{I}_1 = J_1, \ldots, \boldsymbol{I}_{t-1} = J_{t-1} \mid \boldsymbol{I}_n = I'] \Pr[\boldsymbol{I}_t = J_t \mid \boldsymbol{I}_1 = J_1, \ldots, \boldsymbol{I}_{t-1} = J_{t-1}, \boldsymbol{I}_n = I'] && (38)\\ &\quad = \Pr[\boldsymbol{I}_1 = J_1, \ldots, \boldsymbol{I}_t = J_t \mid \boldsymbol{I}_n = I']. && (39)\end{aligned}$$

This concludes the induction. It follows that $\boldsymbol{J}_n = \boldsymbol{I}'$ with probability 1, since the conditional distribution of $\boldsymbol{J}_n$ conditioned on $\boldsymbol{I}' = I'$ equals the conditional distribution of $\boldsymbol{I}_n$ conditioned on $\boldsymbol{I}_n = I'$, which constantly equals I'. This proves one of the guarantees on $\overline{\boldsymbol{J}}$. Further, it implies

that $\boldsymbol{J}_n$ has the same distribution as $\boldsymbol{I}'$, which, by Lemma 10.6 implies that $\boldsymbol{J}_n$ is distributed as $\boldsymbol{I}_n$. In combination with (39), we derive that

$$\begin{aligned}\Pr[\boldsymbol{J}_1 = J_1, \dots, \boldsymbol{J}_n = J_n] &= \Pr[\boldsymbol{J}_n = J_n] \Pr[\boldsymbol{J}_1 = J_1, \dots, \boldsymbol{J}_{n-1} = J_{n-1} \mid \boldsymbol{J}_n = J_n] \\ &= \Pr[\boldsymbol{J}_n = J_n] \Pr[\boldsymbol{J}_1 = J_1, \dots, \boldsymbol{J}_{n-1} = J_{n-1} \mid \boldsymbol{I}' = J_n] \\ &= \Pr[\boldsymbol{I}_n = I_n] \Pr[\boldsymbol{I}_1 = J_1, \dots, \boldsymbol{I}_{n-1} = J_{n-1} \mid \boldsymbol{I}_n = J_n] \\ &= \Pr[\boldsymbol{I}_1 = J_1, \dots, \boldsymbol{I}_n = J_n],\end{aligned}$$

as required. □

10.4 Bounds for Bernoulli Sampling via Uniform Sampling

Here, our goal is to show that concentration guarantees on uniform sampling imply guarantees on Bernoulli sampling. Notice that the latter can be viewed as a mixture of uniform sampling schemes for different values of k. To be more precise, a Bernoulli sample $\mathrm{Ber}(n, p)$ can be obtained by first drawing $\boldsymbol{k} \sim \mathrm{Bin}(n, p)$ and then drawing a uniform sample $\mathrm{Uni}(n, \boldsymbol{k})$, where Bin denotes the binomial distribution. For this reason, it is *harder* to be adversarial against a Bernoulli sampler, because the adversary there does not know in advance what value of $\boldsymbol{k}$ is drawn.

To formalize this notion, we use $f(I, \overline{x})$ as some indicator of failure of the sample $\overline{x}_I$ to represent $\overline{x}$, for example, an indicator of whether $\overline{x}_I$ is not an ϵ-approximation of $\overline{x}$. One would like to minimize the probability that $f = 1$, against any adversary. The following lemma compares the failure probability of the Bernoulli sampling with that of the uniform sampling:

Lemma 10.7. *Let $n \in \mathbb{N}$ and $p \in [0, 1]$. Let $f \colon \{0,1\}^n \times X^n \to \{0,1\}$. Then,*

$$\begin{aligned}&\max_{\mathcal{A} \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I} \sim \mathrm{Ber}(n,p)} [f(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] \\ &\quad \le \max_{\substack{k \in \mathbb{N}: \\ np/2 \le k \le 3np/2}} \max_{\mathcal{A}^k \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I}^k \sim \mathrm{Uni}(n,k)} \left[f(\boldsymbol{I}^k, \overline{x}(\mathcal{A}^k, \boldsymbol{I}^k)) = 1 \right] + 2\exp(-cnp).\end{aligned}$$

The proof relies on a variant of Bernstein's inequality, on the tail of the binomial random variable:

Lemma 10.8 (Bernstein's inequality)**.** *Let $\boldsymbol{k} \sim \mathrm{Bin}(n, p)$. Then, for any $\epsilon \in [0, 1]$,*

$$\Pr[|\boldsymbol{k} - np| \ge \epsilon np] \le 2\exp(-\epsilon^2 np/3).$$

Proof of Lemma 10.7. Let $\mathcal{A}$ denote the maximizer with respect to the Bernoulli sample. First, we decompose the Bernoulli sample into a mixture of uniform samples. Let $\boldsymbol{k} \sim \mathrm{Bin}(n, p)$ drawn from a binomial distribution; we have

$$\Pr_{\boldsymbol{I} \sim \mathrm{Ber}(n,p)} [f(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] = \sum_{k=0}^{n} \Pr[\boldsymbol{k} = k] \Pr_{\boldsymbol{I}^k \sim \mathrm{Uni}(n,k)} \left[f(\boldsymbol{I}^k, \overline{x}(\mathcal{A}, \boldsymbol{I}^k)) = 1 \right].$$

First, summing the terms corresponding to $np/2 \le k \le 3np/2$, we have

$$\begin{aligned}&\sum_{k=\lceil np/2 \rceil}^{\lfloor 3np/2 \rfloor} \Pr[\boldsymbol{k} = k] \Pr_{\boldsymbol{I}^k \sim \mathrm{Uni}(n,k)} \left[f(\boldsymbol{I}^k, \overline{x}(\mathcal{A}, \boldsymbol{I}^k)) = 1 \right] \le \max_{k:\ np/2 \le k \le 3np/2} \Pr_{\boldsymbol{I}^k \sim \mathrm{Uni}(n,k)} \left[f(\boldsymbol{I}^k, \overline{x}(\mathcal{A}, \boldsymbol{I}^k)) = 1 \right] \\ &\quad \le \max_{k:\ np/2 \le k \le 3np/2} \max_{\mathcal{A}^k \in \mathrm{Adv}_n} \Pr_{\boldsymbol{I}^k \sim \mathrm{Uni}(n,k)} \left[f(\boldsymbol{I}^k, \overline{x}(\mathcal{A}^k, \boldsymbol{I}^k)) = 1 \right].\end{aligned}$$

Next, the sum in the remaining terms is bounded by the probability that $\boldsymbol{k} \notin [np/2, 3np/2]$, or equivalently, the probability that $|\boldsymbol{k} - np| > np/2$. From Lemma 10.8, this is bounded by $2\exp(-cnp)$. □

As a direct application, we derive Theorem 6.1 and Theorem 7.1 for the Bernoulli sampling, given the bounds corresponding to the uniform sampling:

Proof of Theorem 6.1, Bernoulli sampling. Let $\delta \in (0, 1/2)$. Define by $f(I, \overline{x})$ the indicator of whether $\overline{x}_I$ fails to be an ϵ-approximation for $\overline{x}$, where $\epsilon = C_0\sqrt{(d + \log(1/\delta))/(np)}$ and $C_0 > 0$ is a sufficiently large constant. From Theorem 6.5, for any k such that $np/2 \leq k \leq 3np/2$ and any $\mathcal{A} \in \mathrm{Adv}_n$, it holds that

$$\Pr_{\boldsymbol{I} \sim \mathrm{Uni}(n,k)}[f(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] \leq \delta/2.$$

By Lemma 10.7, it follows that for any $\mathcal{A} \in \mathrm{Adv}_n$,

$$\Pr_{\boldsymbol{I} \sim \mathrm{Uni}(n,k)}[f(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] \leq \delta/2 + \exp(-c_0 np), \tag{40}$$

for some universal constant $c_0 > 0$. Notice that if $\epsilon > 1$ then the result trivially follows and otherwise, we have that

$$np \geq C_0^2 \log(1/\delta).$$

Assuming that C_0 is sufficiently large, we have that

$$c_0 np \geq \log(2/\delta),$$

which implies that

$$\exp(-c_0 np) \leq \delta/2.$$

In combination with (40), this concludes the proof. Notice that the condition that $2k \leq n$ in Theorem 6.5 translates here to $3np \leq n$. □

Proof of Theorem 7.1, Bernoulli sampling. The proof follows similar steps as the proof for Theorem 6.1, while replacing ϵ-approximations with ϵ-nets and using Theorem 7.5 for the bound on the uniform sampler. □

10.5 Bounds for Uniform Sampling via Bernoulli Sampling

This section reduces bounds for the uniform sampler $\boldsymbol{I} \sim \mathrm{Uni}(2k, k)$ to bounds for the Bernoulli sampler $\boldsymbol{I}' \sim \mathrm{Ber}(2k, p)$. This is done via the method of online coupling, presented in Section 10.2. First, we present some simple well known properties of binary random variables, and then, we describe an online coupling of $\boldsymbol{I}$ to $\boldsymbol{I}'$:

Lemma 10.9. *If $\boldsymbol{y}$ and $\boldsymbol{z}$ are two random variables over $\{0, 1\}$, then there exists a coupling (i.e. joint distribution) of them such that:*

1. $\Pr[\boldsymbol{y} \neq \boldsymbol{z}] = |\Pr[\boldsymbol{y} = 1] - \Pr[\boldsymbol{z} = 1]|$.
2. *If* $\Pr[\boldsymbol{y} = 1] \geq \Pr[\boldsymbol{z} = 1]$ *then* $\boldsymbol{y} \geq \boldsymbol{z}$ *with probability* 1.

Proof. One can couple the following way: first, draw a random variable $\boldsymbol{\xi}$ uniformly in $[0,1]$ and set $\boldsymbol{y} = 0$ if $\boldsymbol{\xi} \le \Pr[\boldsymbol{y} = 0]$ and $\boldsymbol{z} = 0$ if $\boldsymbol{\xi} \le \Pr[\boldsymbol{z} = 0]$. This satisfies the requirements of the lemma. □

To online couple $\boldsymbol{I}$ to $\boldsymbol{I}'$ we use the coupling guaranteed from the next lemma:

Lemma 10.10. *Fix $k \in \mathbb{N}$ and $p \in (0,1)$, and let $\boldsymbol{I} \sim \mathrm{Uni}(2k,k)$ and $\boldsymbol{I}' \sim \mathrm{Ber}(2k,p)$. Then, $\boldsymbol{I}$ can be online-coupled to $\boldsymbol{I}'$ such that for any $t \in [n]$ and any $I_{t-1}, I'_{t-1} \subseteq [t-1]$, the following holds:*

1. $$\begin{aligned}&\Pr\left[t \in \boldsymbol{I} \triangle \boldsymbol{I}' \mid \boldsymbol{I} \cap [t-1] = I_{t-1}, \boldsymbol{I}' \cap [t-1] = I'_{t-1}\right] \\ &\quad = \left|\Pr\left[t \in \boldsymbol{I} \mid \boldsymbol{I} \cap [t-1] = I_{t-1}\right] - \Pr\left[t \in \boldsymbol{I}' \mid \boldsymbol{I}' \cap [t-1] = I'_{t-1}\right]\right| \\ &\quad = \left|\frac{k - |I_{t-1}|}{2k - (t-1)} - p\right|, \end{aligned} \tag{41}$$

   where $\triangle$ denotes the symmetric set difference $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

2. *For any $t \in [n]$ such that $p \le (k - |\boldsymbol{I} \cap [t-1]|)/(2k - t + 1)$, it holds that $\boldsymbol{I}' \cap \{t\} \subseteq \boldsymbol{I} \cap \{t\}$.*

Proof of Lemma 10.10. By Lemma 10.9, conditioned on $\boldsymbol{I} \cap [t-1] = I_{t-1}, \boldsymbol{I}' \cap [t-1] = I'_{t-1}$, there is a joint probability distribution between the indicators $\mathbb{1}_{t \in \boldsymbol{I}}$ and $\mathbb{1}_{t \in \boldsymbol{I}'}$ such that

$$\begin{aligned}&\Pr\left[t \in \boldsymbol{I} \triangle \boldsymbol{I}' \mid \boldsymbol{I} \cap [t-1] = I_{t-1}, \boldsymbol{I}' \cap [t-1] = I'_{t-1}\right] \\ &\quad = \Pr\left[\mathbb{1}_{t \in \boldsymbol{I}} \ne \mathbb{1}_{t \in \boldsymbol{I}'} \mid \boldsymbol{I} \cap [t-1] = I_{t-1}, \boldsymbol{I}' \cap [t-1] = I'_{t-1}\right] \\ &\quad = \left|\Pr\left[\mathbb{1}_{t \in \boldsymbol{I}} = 1 \mid \boldsymbol{I} \cap [t-1] = I_{t-1}\right] - \Pr\left[\mathbb{1}_{t \in \boldsymbol{I}'} = 1 \mid \boldsymbol{I}' \cap [t-1] = I'_{t-1}\right]\right| \\ &\quad = \left|\Pr\left[t \in \boldsymbol{I} \mid \boldsymbol{I} \cap [t-1] = I_{t-1}\right] - \Pr\left[t \in \boldsymbol{I}' \mid \boldsymbol{I}' \cap [t-1] = I'_{t-1}\right]\right|.\end{aligned}$$

Define the online sampler to obey this property, namely, while receiving the value of $\mathbb{1}_{t \in \boldsymbol{I}'}$, it can sample $\mathbb{1}_{t \in \boldsymbol{I}}$ from its conditional distribution, conditioned on the obtained value of $\mathbb{1}_{t \in \boldsymbol{I}'}$ in the above coupling. This proves property 1. Property 2 follows from property 2 in Lemma 10.9. □

Notice that there is a unique way to define a coupling that satisfies the properties above and we term it the *online monotone coupling*. We proceed with applying the monotone online coupling to prove Lemma 6.3 and Lemma 7.3.

10.5.1 Proof of Lemma 6.3

We will in fact prove a more general lemma. We let $\varphi(I, \overline{x})$ be some real valued function, that we view as some loss corresponding to how the sample $\overline{x}_I$ represents the complete stream $\overline{x}$. We have the following:

Lemma 10.11. *Let $\varphi\colon \{0,1\}^{2k} \times X^{2k} \to \mathbb{R}$ be a function that is L-Lipschitz in each coordinate of I, namely, for all $\overline{x} \in X^n$,*

$$|\varphi(I, \overline{x}) - \varphi(I', \overline{x})| \le L|I \triangle I'|.$$

Then, for any $t \ge 0$ and $\delta \in (0, 1/2)$,

$$\begin{aligned}&\sup_{\mathcal{A} \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I} \sim \mathrm{Uni}(2k,k)}[\varphi(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) \ge t] \\ &\quad \le \sup_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}' \sim \mathrm{Ber}(2k,1/2)}\left[\varphi(\boldsymbol{I}', \overline{x}(\mathcal{A}', \boldsymbol{I}')) \ge t - CL\sqrt{k \log(1/\delta)}\right] + \delta.\end{aligned}$$

This directly implies Lemma 6.3:

Proof of Lemma 6.3. Apply Lemma 10.11 with

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

This function is $L = 1/k$-Lipschitz with respect to each coordinate of I, as the maximum of L-Lipschitz functions is L-Lipschitz itself. This suffices to conclude the proof. □

To prove Lemma 10.11, we start with the following auxiliary property:

Lemma 10.12. *Let $\boldsymbol{I} \sim \mathrm{Uni}(2k, k)$ and $\boldsymbol{I}' \sim \mathrm{Ber}(2k, 1/2)$ be coupled according to the monotone online coupling. Then, for every $\delta \in (0, 1/2)$,*

$$\Pr\left[|\boldsymbol{I}' \triangle \boldsymbol{I}| \geq C\sqrt{k \log(1/\delta)} \right] \leq \delta.$$

First, one can bound the expected symmetric difference between $\boldsymbol{I}$ and $\boldsymbol{I}'$:

Lemma 10.13. *Let $\boldsymbol{I} \sim \mathrm{Uni}(2k, k)$ and $\boldsymbol{I}' \sim \mathrm{Ber}(2k, 1/2)$ be coupled according to the monotone online coupling. Then,*

$$\mathbb{E}\left[|\boldsymbol{I}' \triangle \boldsymbol{I}| \right] \leq C\sqrt{k}.$$

Proof. Summing (41) over all t and taking expectation:

$$\mathbb{E}[|\boldsymbol{I} \triangle \boldsymbol{I}'|] = \sum_{t=1}^{2k} \Pr[t \in \boldsymbol{I} \triangle \boldsymbol{I}'] = \sum_{t=1}^{2k} \mathbb{E}\left| \frac{1}{2} - \frac{k - |\boldsymbol{I} \cap [t-1]|}{2k - (t-1)} \right|.$$

By using Jenssen's inequality and applying Lemma 5.5 with $U = \{t, \ldots, 2k\}$, $k = k$ and $n = 2k$, we have

$$\begin{aligned}
\mathbb{E}\left| \frac{1}{2} - \frac{k - |\boldsymbol{I} \cap [t-1]|}{2k - (t-1)} \right| &\leq \sqrt{\mathbb{E}\left[\left(\frac{1}{2} - \frac{k - |\boldsymbol{I} \cap [t-1]|}{2k - (t-1)} \right)^2 \right]} \\
&= \sqrt{\mathrm{Var}\left[\frac{|\boldsymbol{I} \cap [t-1]|}{2k - (t-1)} \right]} = \sqrt{\mathrm{Var}\left[\frac{|\boldsymbol{I} \cap \{t, \ldots, 2k\}|}{2k - (t-1)} \right]} \\
&= \frac{1}{2k - (t-1)} \sqrt{\mathrm{Var}\left[|\boldsymbol{I} \cap \{t, \ldots, 2k\}| \right]} \leq \frac{1}{2k - (t-1)} \sqrt{\frac{(2k - (t-1))k}{2k}} \\
&= \frac{1}{2\sqrt{2k - (t-1)}}.
\end{aligned}$$

Summing over $t = 1, \ldots, 2k$, we have

$$\mathbb{E}[|\boldsymbol{I} \triangle \boldsymbol{I}'|] \leq \sum_{t=1}^{2k} \frac{1}{2\sqrt{2k - (t-1)}} = \sum_{t=1}^{2k} \frac{1}{2\sqrt{t}} \leq \sqrt{2k}.$$

□

The next step is to show that with high probability, $|\boldsymbol{I} \cap \boldsymbol{I}'|$ is close to its expectation. For that, the notion of Martingales is used. We use a standard notation, that is presented in Section A.1. In particular, we use the following commonly used corollary of Azuma's inequality (Lemma A.1):

Lemma 10.14. *Let $F_0 \subseteq F_1 \subseteq \cdots F_n$ be a filtration such that F_0 is the trivial σ-algebra. Let $\boldsymbol{y}$ be a random variable that is F_n measurable, and assume that $a_1, \ldots, a_n$ are numbers such that $|\mathbb{E}[\boldsymbol{y} \mid F_i] - \mathbb{E}[\boldsymbol{y} \mid F_{i-1}]| \leq a_i$ holds for all i with probability 1. Then, for all $t > 0$,*

$$\Pr[\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}] > t] \leq \exp\left(\frac{-t^2}{2\sum_i a_i^2}\right).$$

Notice that, as described in Section A.1, $\mathbb{E}[\boldsymbol{y} \mid F_i]$ is a random variable, and a bound of $|\mathbb{E}[\boldsymbol{y} \mid F_i] - \mathbb{E}[\boldsymbol{y} \mid F_{i-1}]| \leq a_i$ states that the information that is present in F_i and not in F_{i-1} does not significantly affect the conditional expectation of $\boldsymbol{y}$.

Proof of Lemma 10.14. We define the following Martingale, which is known as *Doob's Martingale*: $\boldsymbol{y}_i = \mathbb{E}[\boldsymbol{y} \mid F_i]$, for $i = 0, \ldots, n$. Then, $\boldsymbol{y}_0 = \mathbb{E}\boldsymbol{y}$ and $\boldsymbol{y}_n = \boldsymbol{y}$, and the proof follows directly from Lemma A.1. □

We are ready to bound the deviation of $|\boldsymbol{I} \triangle \boldsymbol{I}'|$:

Lemma 10.15. *Let $\boldsymbol{I} \sim \mathrm{Uni}(2k, k)$ and $\boldsymbol{I}' \sim \mathrm{Ber}(2k, 1/2)$ be coupled according to the monotone online coupling. Then, for every $\delta \in (0, 1/2)$,*

$$\Pr\left[|\boldsymbol{I}' \triangle \boldsymbol{I}| - \mathbb{E}[|\boldsymbol{I}' \triangle \boldsymbol{I}|] \geq C\sqrt{k \log(1/\delta)}\right] \leq \delta.$$

Proof. For any $t = 0, \ldots, n$, let F_t denote the σ-field that contains all the information up to (and including) round t, $F_t = \sigma(\boldsymbol{I} \cap [t], \boldsymbol{I}' \cap [t])$. In order to apply Lemma 10.14, it is desirable to bound the differences $|\mathbb{E}[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \mathcal{F}_t] - \mathbb{E}[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \mathcal{F}_{t-1}]|$ for all $t \in [2k]$. Notice that

$$\begin{aligned}
&\left|\mathbb{E}\left[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \mathcal{F}_t\right] - \mathbb{E}\left[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \mathcal{F}_{t-1}\right]\right| \\
&\quad = \left|\mathbb{E}\left[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \boldsymbol{I} \cap [t], \boldsymbol{I}' \cap [t]\right] - \mathbb{E}\left[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \boldsymbol{I} \cap [t-1], \boldsymbol{I}' \cap [t-1]\right]\right|. \qquad (42)
\end{aligned}$$

We would like to bound (42) for any realization of $\boldsymbol{I}$ and $\boldsymbol{I}'$, namely, bounding for any $S, S' \subseteq [t]$ the quantity

$$\begin{aligned}
&\left|\mathbb{E}\left[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \boldsymbol{I} \cap [t] = S, \boldsymbol{I}' \cap [t] = S'\right] \qquad (43) \right. \\
&\quad \left. -\mathbb{E}\left[|\boldsymbol{I} \triangle \boldsymbol{I}'| \mid \boldsymbol{I} \cap [t-1] = S \cap [t-1], \boldsymbol{I}' \cap [t-1] = S' \cap [t-1]\right]\right|. \qquad (44)
\end{aligned}$$

Define four random variables, $\boldsymbol{J}_t, \boldsymbol{J}'_t, \boldsymbol{J}_{t-1}, \boldsymbol{J}'_{t-1}$ in a joint probability space such that $(\boldsymbol{J}_t, \boldsymbol{J}'_t)$ is distributed according to the joint distribution of $(\boldsymbol{I}, \boldsymbol{I}')$ conditioned on $\boldsymbol{I} \cap [t] = S, \boldsymbol{I}' \cap [t] = S'$ and $(\boldsymbol{J}_{t-1}, \boldsymbol{J}'_{t-1})$ is distributed according to the joint distribution of $(\boldsymbol{I}, \boldsymbol{I}')$ conditioned on $\boldsymbol{I} \cap [t-1] = S \cap [t-1], \boldsymbol{I}' \cap [t-1] = S' \cap [t-1]$. Then, we derive that (43) equals

$$|\mathbb{E}[|\boldsymbol{J}_t \triangle \boldsymbol{J}'_t|] - \mathbb{E}[|\boldsymbol{J}_{t-1} \triangle \boldsymbol{J}'_{t-1}|]| \leq \mathbb{E}[||\boldsymbol{J}_t \triangle \boldsymbol{J}'_t| - |\boldsymbol{J}_{t-1} \triangle \boldsymbol{J}'_{t-1}||] \qquad (45)$$

and our goal is to bound the right hand side of (45). The joint distribution is defined by an inductive argument, defining for $j \geq t$ the intersection of the above four random variables with

$[j]$ given the intersection with $j-1$. Begin with $j=t$: here, $\boldsymbol{J}_t$ and $\boldsymbol{J}-\mathbf{1}_t$ are fixed to S and S', respectively. Further, the intersections of $\boldsymbol{J}_{t-1}$ and $\boldsymbol{J}'_{t-1}$ with $[t-1]$ are fixed and equal $S\cap[t-1]$, and their intersections with $[t]$ are random drawn according to the monotone online coupling. Further, for $j>t$, we start by drawing a random variable $\boldsymbol{\xi}$ uniformly in $[0,1]$, and for any $\boldsymbol{U}\in\{\boldsymbol{J}_t,\boldsymbol{J}'_t,\boldsymbol{J}_{t-1},\boldsymbol{J}'_{t-1}\}$ we set $j\in\boldsymbol{U}$ if and only if $\boldsymbol{\xi}\leq\Pr[j\in\boldsymbol{U}\mid\boldsymbol{U}\cap[j-1]]$. The above defined joint distribution satisfies the following properties:

- From the proofs of Lemma 10.9 and Lemma 10.10, it follows that $(\boldsymbol{J}_t,\boldsymbol{J}'_t)$ is distributed according to the joint distribution of $(\boldsymbol{I},\boldsymbol{I}')$ conditioned on $\{\boldsymbol{I}\cap[t]=S,\boldsymbol{I}'\cap[t]=S'\}$ and $(\boldsymbol{J}_{t-1},\boldsymbol{J}'_{t-1})$ is distributed according to the joint distribution of $(\boldsymbol{I},\boldsymbol{I}')$ conditioned on $\{\boldsymbol{I}\cap[t-1]=S\cap[t-1],\boldsymbol{I}'\cap[t-1]=S'\cap[t-1]\}$.
- For any $\boldsymbol{U},\boldsymbol{U}'\in\{\boldsymbol{J}_t,\boldsymbol{J}'_t,\boldsymbol{J}_{t-1},\boldsymbol{J}'_{t-1}\}$ and any $j>t$, if $\Pr[j\in\boldsymbol{U}\mid\boldsymbol{U}\cap[j-1]]\leq\Pr[j\in\boldsymbol{U}'\mid\boldsymbol{U}'\cap[j-1]]$ then $j\in\boldsymbol{U}$ implies $j\in\boldsymbol{U}'$.
- Notice that for any for any $j>t$ and any $\boldsymbol{U}\in\{\boldsymbol{J}'_t,\boldsymbol{J}'_{t-1}\}$ it holds that $\Pr[j\in\boldsymbol{U}\mid\boldsymbol{U}\cap[j-1]]=1/2$, hence $j\in\boldsymbol{J}'_t$ if and only if $j\in\boldsymbol{J}'_{t-1}$.
- For any $j>t$ and any $\boldsymbol{U}\in\{\boldsymbol{J}_t,\boldsymbol{J}_{t-1}\}$, it holds that

  $$\Pr[j\in\boldsymbol{U}\mid\boldsymbol{U}\cap[j-1]]=\frac{k-|\boldsymbol{U}\cap[j-1]|}{2k-(j-1)},$$

  which is a monotone decreasing function of $|\boldsymbol{U}\cap[j-1]|$. This implies the following properties:

  - For any $j>t$ such that $|\boldsymbol{J}_t\cap[j-1]|=|\boldsymbol{J}_{t-1}\cap[j-1]|$, it holds that $j\in\boldsymbol{J}_t$ if and only of $j\in\boldsymbol{J}_{t-1}$.
  - For any $j>t$ such that $|\boldsymbol{J}_t\cap[j-1]|\geq|\boldsymbol{J}_{t-1}\cap[j-1]|$, it holds that $\Pr[j\in\boldsymbol{J}_t\mid\boldsymbol{J}_t\cap[j-1]]\leq\Pr[j\in\boldsymbol{J}_{t-1}\mid\boldsymbol{J}_{t-1}\cap[j-1]]$, hence, $j\in\boldsymbol{J}_t$ implies $j\in\boldsymbol{J}_{t-1}$.
  - For any $j>t$ such that $|\boldsymbol{J}_t\cap[j-1]|\leq|\boldsymbol{J}_{t-1}\cap[j-1]|$, due to a similar argument, $j\in\boldsymbol{J}_{t-1}$ implies $j\in\boldsymbol{J}_t$.
- From the above arguments, for any j, if $||\boldsymbol{J}_t\cap[j-1]|-|\boldsymbol{J}_{t-1}\cap[j-1]||=1$ then either $j\notin\boldsymbol{J}_t\triangle\boldsymbol{J}_{t-1}$ or $|\boldsymbol{J}_t\cap[j]|=|\boldsymbol{J}_{t-1}\cap[j]|$.
- It holds that $||\boldsymbol{J}_t\cap[t]|-|\boldsymbol{J}_{t-1}\cap[t]||\leq 1$. From the above arguments, at the first $j>t$ such that $j\in\boldsymbol{J}_t\triangle\boldsymbol{J}_{t-1}$, it holds that $|\boldsymbol{J}_t\cap[j]|=|\boldsymbol{J}_{t-1}\cap[j]|$ and from that point onward, $j\notin\boldsymbol{J}_t\triangle\boldsymbol{J}_{t-1}$. In particular, there is at most one $j>t$ such that $j\in\boldsymbol{J}_t\triangle\boldsymbol{J}_{t-1}$.
- It follows that there are at most two values of j such that $j\in\boldsymbol{J}_t\triangle\boldsymbol{J}_{t-1}$: one (possibly) for $j=t$ and one (possibly) for $j>t$. Since the only possible j where $j\in\boldsymbol{J}'_t\triangle\boldsymbol{J}'_{t-1}$ is $j=t$, it follows that

  $$||\boldsymbol{J}_t\triangle\boldsymbol{J}'_t|-|\boldsymbol{J}_{t-1}\triangle\boldsymbol{J}'_{t-1}||\leq 2. \tag{46}$$

From (45) and (46) we derive that

$$\left|\mathbb{E}\left[|\boldsymbol{I}\triangle\boldsymbol{I}'|\mid\mathcal{F}_t\right]-\mathbb{E}\left[|\boldsymbol{I}\triangle\boldsymbol{I}'|\mid\mathcal{F}_{t-1}\right]\right|\leq 2.$$

Applying Lemma 10.14, the proof follows. □

Lemma 10.13 and Lemma 10.15 together imply Lemma 10.12.

Proof of Lemma 10.11. We apply Lemma 10.3 with

$$f(\overline{x}, I) = \mathbb{1}\left(\varphi(I, \overline{x}) \geq t\right)$$

and

$$g(\overline{x}, I') = \mathbb{1}\left(\varphi(I', \overline{x}) \geq t - C_0 L\sqrt{k\log(1/\delta)}\right).$$

Here $C_0 > 0$ is the universal constant guaranteed from Lemma 10.12 such that with probability $1-\delta$,

$$|\boldsymbol{I} \triangle \boldsymbol{I}'| \leq C_0\sqrt{k\log(1/\delta)}.$$

Notice that for any I and I' such that $|I \triangle I'| \leq C_0\sqrt{k\log(1/\delta)}$ and any $\overline{x} \in X^n$, it holds that

$$|\varphi(I, \overline{x}) - \varphi(I', \overline{x})| \leq L|I \triangle I'| \leq C_0 L\sqrt{k\log(1/\delta)}.$$

Hence, for any such I, I' and any $\overline{x}$ such that $f(I, \overline{x}) = 1$, it holds that $g(I', \overline{x}) = 1$. Applying Lemma 10.3, we derive that for $\boldsymbol{I} \sim \mathrm{Uni}(2k, k)$ and $\boldsymbol{I}' \sim \mathrm{Ber}(2k, 1/2)$,

$$\begin{aligned}
&\max_{\mathcal{A}\in\mathrm{Adv}_{2k}} \Pr[\varphi(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) \geq t] = \max_{\mathcal{A}\in\mathrm{Adv}_{2k}} \Pr[f(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] \\
&\quad\leq \max_{\mathcal{A}'\in\mathrm{Adv}_{2k}} \Pr[g(\boldsymbol{I}', \overline{x}(\mathcal{A}', \boldsymbol{I}')) = 1] + \delta \\
&\quad= \max_{\mathcal{A}'\in\mathrm{Adv}_{2k}} \Pr\left[\varphi(\boldsymbol{I}', \overline{x}(\mathcal{A}', \boldsymbol{I}')) \geq t - C_0 L\sqrt{k\log(1/\delta)}\right] + \delta.
\end{aligned}$$

□

10.5.2 Proof of Lemma 7.3

We prove the following property of the online monotone coupling:

Lemma 10.16. *Let $\boldsymbol{I} \sim \mathrm{Uni}(2k, k)$ be coupled to $\boldsymbol{I}' \sim \mathrm{Ber}(2k, 1/8)$ according to the monotone online coupling and fix $m \leq n$. Then,*

$$\Pr[\boldsymbol{I}' \subseteq \boldsymbol{I} \cup \{n-m+1, \ldots, n\}] \geq 1 - 2\exp(-cm).$$

Proof. From Lemma 10.10, it suffices to show that with probability $1 - e^{-cm}$, for all $t \leq 2k - m$,

$$1/8 \leq \frac{k - |\boldsymbol{I} \cap [t-1]|}{2k - (t-1)}. \tag{47}$$

Let $m_0, m_1, \ldots, m_r$, such that $m_i = m \cdot 2^i$ and $k < m_r \leq 2k$. For any $i = 0, \ldots, r$, apply Lemma 5.6 with $n = 2k$, $k = k$, $U = \{n - m_i + 1, \ldots, 2k\}$, $\alpha = 1/2$ and $\boldsymbol{I} = \boldsymbol{I}$, deriving

$$\Pr\left[|\boldsymbol{I}' \cap (\{n - m_i + 1, \ldots, 2k\})| \leq \frac{m_i}{4}\right] \leq \Pr\left[\left|\frac{|\boldsymbol{I} \cap U|}{k} - \frac{|U|}{2k}\right| \geq \frac{m_i}{4k}\right] \leq 2\exp\left(-cm_i\right). \tag{48}$$

Summing the failure probabilities over $i = 0, \ldots, r$, the sum is dominated by the first summand, and we derive that with probability at least $1 - 2\exp(-cm)$, (48) fails to hold for all $i = 0, \ldots, r$. Fix a realization I of $\boldsymbol{I}$ such that (48) fails for all i, and we will prove (47), to complete the proof. Fix $t \leq 2k - m$, and let i_0 be the maximal i such that $t \leq 2k - m_i + 1$. Then, $2k - (t-1) < m_{i_0+1} = 2m_{i_0}$. Further,

$$k - |I \cap [t-1]| = |I \cap \{t, \ldots, n\}| \geq |I \cap \{n - m_i + 1, \ldots, n\}| \geq m_i/4.$$

We derive that

$$\frac{k - |I \cap [t-1]|}{2k - (t-1)} \geq \frac{1}{8}$$

as required. □

Proof of Lemma 7.3. Let $\boldsymbol{I}$ be coupled to $\boldsymbol{I}'$ according to the monotone online coupling. We wish to apply Lemma 10.3 with

$$f(I, \overline{x}) = \begin{cases} 1 & \exists E \in \mathcal{E},\ |\overline{x}_{I \cup J} \cap E| \geq \epsilon \cdot 2k,\ \overline{x}_I \cap E = \emptyset \\ 0 & \text{otherwise} \end{cases},$$

and

$$g(I', \overline{x}) = \begin{cases} 1 & \exists E \in \mathcal{E},\ |\overline{x} \cap E| \geq \epsilon \cdot 2k \text{ and } |\overline{x}_I \cap E| \leq \epsilon/16 \cdot 2k \\ 0 & \textit{otherwise} \end{cases}$$

Applying Lemma 10.16 with $m = \lfloor \epsilon/16 \cdot 2k \rfloor$, it holds with probability $1 - 2\exp(-c\epsilon k)$ that

$$\boldsymbol{I}' \subseteq \boldsymbol{I} \cup \{2k - \lfloor \epsilon/16 \cdot 2k \rfloor + 1, n\}. \tag{49}$$

For any values I, I' such that (49) holds and any $\overline{x} \in X^n$ such that $f(I, \overline{x}) = 1$, it also holds that $g(I', \overline{x}) = 1$. Indeed, let $E \in \mathcal{E}$ be a set such that $|E \cap \overline{x}| \geq \epsilon \cdot 2k$ and $\overline{x}_I \cap E = \emptyset$. From (49),

$$|\overline{x}_{I'} \cap E| \leq |\overline{x}_I \cap E| + |\{2k - \lfloor \epsilon/16 \cdot 2k \rfloor + 1, n\} \cap E| \leq \epsilon/16 \cdot 2k,$$

which implies that $g(I', \overline{x}) = 1$. From Lemma 10.3,

$$\begin{aligned} \max_{\mathcal{A} \in \mathrm{Adv}_{2k}} \Pr\left[\mathrm{Net}^{2k}_{\mathcal{A}, \boldsymbol{I}, \epsilon \cdot 2k, 0}(\mathcal{E}) = 1\right] &= \max_{\mathcal{A} \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}}[f(\boldsymbol{I}, \overline{x}(\mathcal{A}, \boldsymbol{I})) = 1] \\ &\leq \max_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr_{\boldsymbol{I}'}[g(\boldsymbol{I}', \overline{x}(\mathcal{A}', \boldsymbol{I}')) = 1] + 2\exp(-c\epsilon k) \\ &= \max_{\mathcal{A}' \in \mathrm{Adv}_{2k}} \Pr\left[\mathrm{Net}^{2k}_{\mathcal{A}', \boldsymbol{I}', \epsilon \cdot 2k, \epsilon/16 \cdot 2k}(\mathcal{E}) = 1\right] + 2\exp(-c\epsilon k). \end{aligned}$$

□

11 Continuous ϵ-Approximation

In the adversarial model we discuss in this paper, the general goal is that in the end of the process (after all elements have been sent by the adversary), the obtained sample would be an ϵ-approximation of the entire adversarial sequence. However, in many practical scenarios of

interest, one might want the sample obtained so-far to be an ϵ-approximation of the current adversarial sequence *at any point along the sequence* (and not just at the end of the sequence). We call this condition a *continuous ϵ-approximation*. Note that such a requirement only makes sense for sampling procedures that allow deletions, like reservoir sampling. (For insertion-only samplers, like Bernoulli and uniform sampling, one cannot hope for the sample to approximate the stream until there is sufficient "critical mass" collected in the sample; this is not an issue with reservoir sampling, which overcomes this by sampling the first elements in the sequence with higher probability, but may also delete them later.)

Obtaining upper bounds for continuous ϵ-approximation can be done easily by plugging-in our upper bounds for reservoir sampling to a block-box argument by Ben-Eliezer and Yogev [BEY20, Section 6]. There, it is shown that if one ensures that the current sample approximates the current sequence at $O(\log n)$ carefully located "checkpoints" along the stream (while setting the error parameter to be $\delta' = \Theta(\delta / \log n)$), then with probability $1 - \delta$, the sample is a continuous ϵ-approximation for the sequence. That is, we have the following.

Theorem 11.1 (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 a continuous ϵ-approximation w.r.t $\mathcal{E}$ satisfies:*

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

This bound is attained by the reservoir sampler $\mathrm{Res}(n, k)$.

Compared with the standard setting (as summarized in Theorem 2.3), the bound here has an additional $\log\log n$ term in the numerator.

12 Online Learning

In this section, we prove an optimal bound on the regret of online classification. We first provide the formal definitions and then proceed with the formal statement and the proof.

12.1 Formal Defintions

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.

Notice that both the learner and the adversary are allowed to use private randomness.

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

$$R_T(\mathcal{E}, \mathcal{L}, \mathcal{A}) := \mathbb{E}\left[\sum_t \mathbb{1}(y_t \neq \hat{y}_t) - \min_{E \in \mathcal{E}} \sum_t \mathbb{1}\Big(y_t \neq \mathbb{1}(x_t \in E)\Big)\right].$$

The *optimal regret* is defined as the value of the the regret achieved by the best sampler against its worst adversary:

$$R_T(\mathcal{E}) = \min_{\mathcal{L}} \max_{\mathcal{A}} R_T(\mathcal{E}, \mathcal{L}, \mathcal{A}).$$

12.2 Statement and Proof

We prove the following theorem:

Theorem 12.1. *Let $\mathcal{E}$ denote a class of Littlestone dimension d. Then, the expected regret $R_T(\mathcal{E})$ for a T-round online learner is bounded by*

$$R_T(\mathcal{E}) \leq C\sqrt{dT} \ ,$$

where $C > 0$ is a universal constant.

We use a bound by [RST15a] on the regret based on the sequential Rademacher complexity:

Theorem 12.2 ([RST15a], Theorem 7)**.** *The expected regret satisfies*

$$R_T(\mathcal{E}) \leq 2\mathrm{Rad}_T(\mathcal{E}) \ .$$

We combine this with the bound on the sequential Rademacher complexity from Lemma 6.4, to complete the proof:

Proof of Theorem 12.1. By Theorem 12.2, by definition of the sequential Rademacher complexity and by Lemma 6.4,

$$R_T(\mathcal{E}) \leq 2\mathrm{Rad}_T(\mathcal{E}) = 2\mathbb{E}_{\boldsymbol{I}\sim\mathrm{Ber}(n,1/2)}[\mathrm{Disc}_{\mathcal{A},\boldsymbol{I}}(\mathcal{E})] \leq C\sqrt{dT} \ .$$

This concludes the proof. □

13 Lower Bounds

In this section we state and prove our lower bounds. Our first lower bound applies to *any* family $\mathcal{E}$, showing that the linear dependence of our upper bounds in the Littlestone dimension is universally tight.

Theorem 13.1 (A universal lower bound)**.** *Let $\mathcal{E}$ be a family with Littlestone dimension d. Then, there exists a (deterministic) adversary such that the following holds. For any algorithm that retains at most $k \leq d$ items (without deletions), the adversary presents d items $x_1, \ldots, x_d$ such that*

$$(\exists E \in \mathcal{E}) : \bar{s} \cap E = \emptyset \quad \textit{and} \quad \frac{|\bar{x} \cap E|}{|\bar{x}|} = 1 - \frac{k}{d},$$

with probability 1 over the algorithm's randomness, where $\bar{x}$ denotes the adversarial stream and $\bar{s}$ is the sample. In particular, any subset of the sample of k items retained by the algorithm does not form an ϵ-approximation with respect to $x_1, \ldots, x_n$ unless $\epsilon \geq 1 - \frac{k}{d}$.

Our second result in this section shows the existence of families $\mathcal{E}$ of Littlestone dimension d in which all ϵ-approximations are of size $\Omega(d/\epsilon^2)$, so long as $d = \Omega(\log(1/\epsilon))$. Interestingly, the requirement that d is large enough is necessary: classical results in discrepancy theory [MWW93, Mat95] imply that when $d = o(\log 1/\epsilon)$, smaller ϵ-approximations exist.

Theorem 13.2 (ϵ-approximation: quadratic lower bound)**.** *Let $d \in \mathbb{N}$ and $\epsilon > 0$ where $d \geq C \log(1/\epsilon)$ for a large absolute constant $C > 0$. Then, there exists a family $\mathcal{E}$ with Littlestone dimension at most d and a subset $\{x_1, \ldots, x_n\} \subset X$ for which no subset of size less than $c \cdot \frac{\mathsf{Ldim}(\mathcal{E})}{\epsilon^2}$ is an ϵ-approximation, where $c > 0$ is a small absolute constant.*

We also prove similar results for ϵ-nets (without the requirement that d is large enough).

Theorem 13.3 (ϵ-net: a super linear lower bound)**.** *Let $d \in \mathbb{N}$ and $\epsilon > 0$. Then, there exists a family $\mathcal{E}$ with Littlestone dimension $\leq d$ and a subset $\{x_1, \ldots, x_n\} \subset X$ for which no subset of length less than $c \cdot \frac{\mathsf{Ldim}(\mathcal{E}) \log(1/\epsilon)}{\epsilon}$ is an ϵ-net, where $c > 0$ is a small absolute constant.*

13.1 Proofs

Proof of Theorem 13.1. The proof generalizes the construction from [BEY20], which provided a lower bound for the family of one-dimensional thresholds.[9] Let $\mathcal{T}$ be a tree of depth d which is shattered by $\mathcal{E}$. The tree $\mathcal{T}$ can be thought of as a strategy for the adversary as follows:

1. Set $\mathcal{T}_1 = \mathcal{T}$ and $i = 1$.
2. For $i = 1, \ldots, d$
   (i) Pick x_i to be the item labelling the root of $\mathcal{T}_i$ and present it to the algorithm.
   (ii) If x_i was retained by the algorithm then continue to the next iteration with $\mathcal{T}_{i+1}$ being the left subtree of $\mathcal{T}_i$ (corresponding to the sets in $\mathcal{E}_{\not\ni x_i}$).
   (iii) Else, continue to the next iteration with $\mathcal{T}_{i+1}$ being the left subtree of $\mathcal{T}_i$ (corresponding to the sets in $\mathcal{E}_{\ni x_i}$).

Thus, the adversary picks the elements $x_1, \ldots, x_d$ according to a path on the tree such that whenever x_i is retained by the algorithm then a left turn is taken and whenever x_i is not retained by the algorithm then a right turn is taken. Thus, since the tree is shattered, there exists a set $E \in \mathcal{E}$ such that

$$E \cap \{x_1, \ldots, x_n\} = \{x_i : x_i \text{ was not sampled by the algorithm}\}.$$

In particular, $\bar{s} \cap E = \emptyset$, and if the algorithm samples $m \leq d$ points then $\frac{|\bar{x} \cap E|}{|\bar{x}|} = 1 - \frac{m}{d}$, as required. □

Proof of Theorem 13.2. The proof follows from standard probabilistic arguments, and shows that most families in a certain setting have bounded Littlestone dimension yet do not admit a small ϵ-approximation. Suppose that $d \geq \log(1/\epsilon)$ and let $n = d/6\epsilon^2$. Let F be a family of $2^d \cdot d/\epsilon^2$ subsets of $[n]$ of size $n/2$, picked uniformly at random among all such families, and note that (by definition and since $d \geq \log(1/\epsilon)$) the Littlestone dimension of F is at most $\log |F| = O(d)$.

[9] We note that the proof from [BEY20] would give a lower bound of $\Omega(\log d)$ for any family of Littlestone dimension d (as compared to the $\Omega(d)$ lower bound we prove here); this follows since, roughly speaking, any such family "contains" a class of thresholds of dimension logarithmic in d.

We now claim that with high probability, there is no ϵ-approximation of size less than $n/2$ for F. Indeed, fix any subset S of size $m \leq n/2$. By a simple counting argument, the number of sets A of size $n/2$ for which $|d_A(S) - d_A([n])| \geq \epsilon$ is at least

$$\binom{m}{\left(\frac{1}{2}-\epsilon\right)m}\binom{n-m}{\frac{n-m}{2}-\epsilon m} \geq \binom{m}{\frac{m}{2}}\binom{n-m}{\frac{n-m}{2}} \cdot (1-2\epsilon)^{2\epsilon m} \geq \frac{2^n}{2n} \cdot e^{-3\epsilon^2 n} = \frac{2^n}{2n} \cdot e^{-d/2},$$

where the second inequality holds for $\epsilon < 1/10$.

Plugging in the right hand side above, and noting the negative correlation between the events at hand, the probability that F does not contain any such A with $|d_A(S) - d_A([n])| \geq \epsilon$ is bounded by

$$\left(1 - \frac{2^n \cdot e^{-d/2}}{2n\binom{n}{n/2}}\right)^{|F|} \leq e^{-2^d \cdot e^{-d/2}} \leq e^{-1.2^d \cdot d/\epsilon^2}.$$

Taking a union bound over all (less than $2^n = 2^{d/6\epsilon^2}$) possible subsets $S \subseteq [n]$ of size at most $[n]/2$, it follows that with high probability (as a function of d), no ϵ-approximation exists. □

Proof of Theorem 13.3. The proof extends a simple probabilistic construction in the projective plane, suggested by Alon, Kalai, Matoušek, and Meshulam [AKMM02].

Consider the projective plane of order p, where we pick $p = C/\epsilon$ for a suitable constant C. Recall that this projective plane has $p^2 + p + 1$ points and lines, where each line consists of exactly $p + 1$ points, and every two points are contained in exactly one line. For each line L, pick uniformly at random (and independently from choices for other lines) a subset H_L containing exactly half the elements of L; we call such a subset a *half line*. Consider the family consisting of all such half lines H_L. As was shown in [AKMM02], with high probability every ϵ-net for this family has size $\Omega(p \log p)$, whilst the VC dimension is at most 2. We claim that the same bound also holds for the Littlestone dimension.

Claim 13.4. *The Littlestone dimension of the family consisting of all half lines as above is at most* 2.

Proof. Suppose to the contrary a depth-3 tree exists as in the definition of the Littlestone dimension, and consider the elements x, y, z appearing in the internal nodes of the all-1 branch in this tree. By definition, all three elements must belong to some half line H_L from the family. However, since any two lines $L_1 \neq L_2$ in the projective plane intersect in exactly one point, we have $|H_{L_1} \cap H_{L_2}| \leq 1$. It follows that there does not exist $L' \neq L$ where $x, y \in L'$, and thus, no half line corresponds to the $(1, 1, 0)$-branch of the tree, a contradiction. □

The proof that no small ϵ-net exists is a straightforward probabilistic proof similar in spirit to that of Theorem 13.2, and is given in detail in [AKMM02]. The proof bounds from above the probability of any fixed set of size (say) $0.1p \log p$ to intersect all half lines, and then takes a union bound over all such sets.

Next, we show how to generalize the above to get a lower bound with linear dependence in d. Let p be as above, consider d copies of the projective plane of order p and let $\mathcal{C}_1, \ldots, \mathcal{C}_d$ be collections of half lines generated as above, one in each plane. Now let $\mathcal{C}$ be the collection of all unions of exactly $d/2$ half lines coming from different planes, namely, $\mathcal{C}$ contains all sets of the form

$$H = H^{i_1}_{L_{j_1}} \cup H^{i_2}_{L_{j_2}} \cup \ldots \cup H^{i_{d/2}}_{L_{j_{d/2}}},$$

where $i_1 < i_2 < \ldots < i_{d/2} \in [d]$, $H^{i_t}_{L_{j_t}}$ is a half line from the i_t copy corresponding to the line L_{j_t} in that copy of the plane.

Consider the family $\mathcal{C}$ with the underlying universe with $d(p^2+p+1)$ points, containing all points from all d planes.

Claim 13.5. *The Littlestone dimension of $\mathcal{C}$ is at most d.*

Proof. The proof is a straightforward extension of the proof of Claim 13.4. Suppose to the contrary that the Littlestone dimension is $t > d$. Let T be a labeled tree of depth t as in the definition of Littlestone dimension and consider its all-1 branch. This branch corresponds to some set $H = H^{i_1}_{L_{j_1}} \cup H^{i_2}_{L_{j_2}} \cup \ldots \cup H^{i_{d/2}}_{L_{j_{d/2}}}$. In particular, all elements labeling nodes along the branch are contained in H.

By the pigeonhole principle, there exist three elements x, y, z along the branch (in this order) contained in the same half line H_L from one of the plane copies. We claim that there is no set in $\mathcal{C}$ that corresponds to any branch which is all-1 up until (and not including) z, and takes the value 0 at z. Indeed, such a set H, if exists, will contain x, y bot not z. However, this is a contradiction as in Claim 13.4: any set H that contains x, y must also contain all elements in the half line H_L containing them both, and thus $z \in H$. □

It remains to prove that there is no ϵ-net of size $o(d\epsilon^{-1}\log\epsilon^{-1})$. But this follows easily from the $\Omega(\epsilon^{-1}\log\epsilon^{-1})$ lower bound for each of the planes separately: there exists some absolute constant $C > 0$ so that for each of the planes at hand, no ϵ-net of size $C\epsilon^{-1}\log\epsilon^{-1}$ exists. Consider now any set S of less than $Cd\epsilon^{-1}\log\epsilon^{-1}/2$ points in our universe, the union of all planes; since each point belongs to exactly one plane, there exist $d/2$ planes with less than $C\epsilon^{-1}\log\{\epsilon^{-1}\}$ points. Let $i_1 < i_2 < \ldots < i_{d/2}$ denote their indices. It follows that there exists some set $H = H^{i_1}_{L_{j_1}} \cup H^{i_2}_{L_{j_2}} \cup \ldots \cup H^{i_{d/2}}_{L_{j_{d/2}}} \in \mathcal{C}$ not intersecting S. This completes the proof. □

References

[AKMM02] Noga Alon, Gil Kalai, Jiří Matoušek, and Roy Meshulam. Transversal numbers for hypergraphs arising in geometry. *Advances in Applied Mathematics*, 29(1):79 – 101, 2002.

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[BEY20] Omri Ben-Eliezer and Eylon Yogev. The adversarial robustness of sampling. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, page 49–62, 2020.

[BJWY20] Omri Ben-Eliezer, Rajesh Jayaram, David P. Woodruff, and Eylon Yogev. A framework for adversarially robust streaming algorithms. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pages 63–80, 2020.

[Bla54] David Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 336–338, 1954.

[Bla56] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[BM15] Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

[Bou04] Olivier Bousquet. Introduction to Statistical Learning Theory. In *Advanced lectures on machine learning*, volume 3176, pages 169–207. Springer, 2004.

[BPS09] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.

[CGP+18] Timothy Chu, Yu Gao, Richard Peng, Sushant Sachdeva, Saurabh Sawlani, and Junxing Wang. 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)*, pages 361–372, 2018.

[Cha05] Sourav Chatterjee. Concentration inequalities with exchangeable pairs (Ph.D. thesis). *arXiv preprint math/0507526*, 2005.

[CN20] Yeshwanth Cherapanamjeri and Jelani Nelson. On adaptive distance estimation. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[dlPn99] Victor H. de la Peña. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564, 1999.

[Dud73] Richard M. Dudley. Sample functions of the Gaussian process. *The Annals of Probability*, 1(1):66–103, 1973.

[Dud78] Richard M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978.

[Dud84] Richard M. Dudley. A course on empirical processes. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XII - 1982*, pages 1–142. Springer Berlin Heidelberg, 1984.

[Dud87] Richard M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4):1306–1326, 1987.

[Fre75] David A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[GHR+12] Anna C. Gilbert, Brett Hemenway, Atri Rudra, Martin J. Strauss, and Mary Wootters. Recovering simple signals. In *Information Theory and Applications Workshop (ITA)*, pages 382–391, 2012.

[GHS+12] Anna C. Gilbert, Brett Hemenway, Martin J. Strauss, David P. Woodruff, and Mary Wootters. Reusable low-error compressive sampling schemes through privacy. In *IEEE Statistical Signal Processing Workshop (SSP)*, pages 536–539, 2012.

[Han57] James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

[Hau92] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78 – 150, 1992.

[HKM+20] Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Yossi Matias, and Uri Stemmer. Adversarially robust streaming algorithms via differential privacy. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[HRS20] Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[HW13] Moritz Hardt and David P. Woodruff. How robust are linear sketches to adaptive inputs? In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, pages 121–130, 2013.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

[LP17] David A. Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, 2017.

[LW94] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[Mat95] Jiří Matoušek. Tight upper bounds for the discrepancy of half-spaces. *Discrete & Computational Geometry*, 13:593–601, 1995.

[Mat09] Jiří Matoušek. *Geometric Discrepancy: An Illustrated Guide*. Springer-Verlag Berlin Heidelberg, 2009.

[MNS11] Ilya Mironov, Moni Naor, and Gil Segev. Sketching in adversarial environments. *SIAM Journal on Computing*, 40(6):1845–1870, 2011.

[MWW93] Jiří Matoušek, Emo Welzl, and Lorenz Wernisch. Discrepancy and approximations for bounded VC-dimension. *Combinatorica*, 13(4):455–466, 1993.

[NY15] Moni Naor and Eylon Yogev. Bloom filters in adversarial environments. In *Advances in Cryptology - CRYPTO 2015 - 35th Annual Cryptology Conference*, pages 565–584, 2015.

[Rob51] Herbert Robbins. 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, 1951.

[RS14] Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction. *Book Draft*, 2014.

[RS15] Alexander Rakhlin and Karthik Sridharan. On Martingale Extensions of Vapnik–Chervonenkis Theory with Applications to Online Learning. In *Measures of Complexity*, pages 197–215. Springer, 2015.

[RST10] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *Advances in Neural Information Processing Systems*, pages 1984–1992, 2010.

[RST15a] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015.

[RST15b] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.

[Sau72] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[SBM+18] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. DARTS: deceiving autonomous cars with toxic signs. *CoRR*, abs/1802.06430, 2018.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning*. Cambridge university press, 2014.

[SSSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[Tal94] Michel Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.

[Vap98] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

[VC71] Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[VC74] Vladimir N. Vapnik and Alexey Y. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

[Vit85] Jeffrey S. Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985.

[WFRS18] Blake E. Woodworth, Vitaly Feldman, Saharon Rosset, and Nati Srebro. The everlasting database: Statistical validity at a fair price. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 6532–6541, 2018.

[WZ20] David P. Woodruff and Samson Zhou. Tight bounds for adversarially robust streams and sliding windows via difference estimators. *arXiv preprint arXiv:2011.07471*, 2020.

A Probabilistic Material

A.1 Filtration and Martingales

In this section we give a brief probability background to Martingales, considering only finite probability spaces. Recall that a probability space consists of a sample space Ω, a σ-field $F \subseteq \{0,1\}^\Omega$ that contains all measurable events and a probability measure μ over Ω. With finite probability spaces, it is possible for F to contain all subsets of Ω, however, smaller sets can be considered as well. For instance, if $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are random variables over the finite space Y, then $\sigma(\boldsymbol{y}_1)$, the *σ-field generated by $\boldsymbol{y}_1$*, contains all the events that depend only on $\boldsymbol{y}_1$. Formally, we have $\Omega = Y^n$ and $\sigma(\boldsymbol{y}_1) = \{\{\boldsymbol{y}_1 \in U\} \colon U \subseteq Y\}$, where $\{Y \in U\} = \{(y_1, \ldots, y_n) \colon y_1 \in U\}$ is the event that $\boldsymbol{y}_1 \in U$. Similarly, we can have sigma fields generated by multiple random variables, for instance, $\sigma(y_1, y_3, y_4)$, that contains all the events that depend only on these three random variables. It is in fact also possible to consider the σ-algebra generated by zero random variables $\sigma(\{\}) = \{0, \Omega\}$ which is called the *trivial σ-algebra*.

Conditioning on more random variables results in a larger σ-algebra, namely, if $i \leq j$ then $\sigma(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_i) \subseteq \sigma(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_j)$. Intuitively, larger σ-algebras contain more information. We say that a σ-field F is *$\boldsymbol{y}$-measurable* if $\sigma(\boldsymbol{y}) \subseteq F$, which intuitively holds whenever F it contains all the information on $\boldsymbol{y}$. Further, a *filtration* is a collection of nested σ-algebras $F_0 \subseteq F_1 \cdots \subseteq F_n$.

One can define *conditional expectation* with respect to a σ-algebra. In our application, each σ-algebra will be generated by a collection of random variables, and it holds that

$$\mathbb{E}[\cdot \mid \sigma(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k)] = \mathbb{E}[\cdot \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k]. \tag{50}$$

Notice that the quantity in (50) is a function of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$, hence it is also a random variable. Additionally, if F is the trivial σ-algebra then

$$\mathbb{E}[\cdot \mid F] = \mathbb{E}[\cdot \mid \sigma(\{\})] = \mathbb{E}[\cdot].$$

And if $\boldsymbol{y}$ is F-measurable, then $\mathbb{E}[\boldsymbol{y} \mid F] = \boldsymbol{y}$.

A collection of random variables $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_n$ defines a *Martingale adapted to the filtration* $F_0 \subseteq \cdots \subseteq F_n$ if $\boldsymbol{z}_i$ is F_i-measurable and if for any $i < j$, $\mathbb{E}[\boldsymbol{z}_j \mid F_i] = \boldsymbol{z}_i$. The simplest case is when $F_i = \sigma(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_i)$, and there, the martingale condition translates to $\mathbb{E}[\boldsymbol{z}_j \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_i] = \boldsymbol{z}_i$. However, in the general case F_i can have additional information on other random variables.

Remarkably, Martingales obey high probability bounds. Perhaps the most well known bound is Azuma's inequality, which is an adaptation of Chernoff's bound for Martingales:

Lemma A.1. *Let $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_n$ be a Martingale adapted to the filtration $F_0, \ldots, F_n$. Let $a_1, \ldots, a_n \geq 0$ be numbers such that almost surely, $|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}| \leq a_i$. Then, for any $t \geq 0$,*

$$\Pr[\boldsymbol{y}_n - \boldsymbol{y}_0 > t] \leq \exp\left(\frac{-t^2}{2\sum_i a_i^2}\right).$$

A.2 Sampling Without Replacement

Further, we have the following version of Chernoff without replacement:

Lemma A.2 ([BM15])**.** *Let $a_1, \ldots, a_N \in \mathbb{R}$ and let I denote a uniformly random subset of $[N]$ of size $n \in \mathbb{N}$. Let $R = \max_i a_i - \min_i a_i$. Then, for any $t > 0$,*

$$\Pr\left[\frac{1}{n}\sum_{i\in I} a_i - \frac{1}{N}\sum_{i=1}^{N} a_i > t\right] \leq \exp\left(\frac{-2nt^2}{R^2}\right).$$

Another without-replacement lemma:

Lemma A.3 ([Cha05], Proposition 3.10)**.** *Let $\{a_{ij}\}_{i,j\in[n]}$ be a collection of numbers from $[0,1]$. Let $Y = \sum_{i=1}^{n} a_{i\pi(i)}$ where π is drawn from the uniform distribution over the set of permutations of $\{1, \ldots, n\}$. Then for any $t \geq 0$,*

$$\Pr\left[|Y - \mathbb{E}Y| \geq t\right] \leq 2\exp(-t^2/(4\mathbb{E}Y + 2t)).$$

Proof of Lemma 5.6. First item follows directly from Lemma A.2. The second item follows from Lemma A.3 as described below. Define $m = |U|$. Let $\pi\colon [n] \to [n]$ be a uniformly random permutation and let $I = \{i\colon \pi(i) \leq k\}$. Define $\{a_{i,j}\}_{i,j\in[n]}$ by $a_{i,j} = 1$ if $i \in U$ and $j \leq k$. Notice that for all $i \in U$, $a_{i\pi(i)} = 1$ if $i \in I$ and for all $i \notin U$, $a_{i\pi(i)} = 0$. Hence,

$$Y := \sum_{i=1}^{n} a_{i\pi(i)}$$

equals $|I \cap U|$ and $\mathbb{E}Y = km/n$. From Lemma A.3 we derive that for any $t \geq 0$,

$$\Pr\left[|Y - \mathbb{E}Y| \geq t\right] \leq \exp\left(-\frac{t^2}{4\mathbb{E}Y + 2t}\right).$$

Substitute $t = \alpha\mathbb{E}Y$ and we get that

$$\begin{aligned}\Pr\left[\left|\frac{Y}{\mathbb{E}Y} - 1\right| \geq \alpha\right] = \Pr\left[|Y - \mathbb{E}Y| \geq t\right] &\leq \exp\left(-\frac{t^2}{4\mathbb{E}Y + 2t}\right) = \exp\left(-\frac{\alpha^2\mathbb{E}[Y]^2}{4\mathbb{E}Y + 2\alpha\mathbb{E}Y}\right) \\ &\leq \exp\left(-\frac{\alpha^2\mathbb{E}[Y]^2}{6\mathbb{E}Y}\right) = \exp\left(-\frac{\alpha^2 km}{6n}\right).\end{aligned}$$

□

Proof of Lemma 5.5. Denote $|U| = m$. For any $i \in U$, let $\boldsymbol{z}_i$ denote the indicator of whether $i \in \boldsymbol{I}$, and notice that $|U \cap \boldsymbol{I}| = \sum_{i\in U} \boldsymbol{z}_i$. Therefore, we have

$$\mathbb{E}[|U \cap \boldsymbol{I}|] = \mathbb{E}\left[\sum_{i\in U} \boldsymbol{z}_i\right] = \sum_{i\in U} \mathbb{E}[\boldsymbol{z}_i] = \sum_{i\in U} k/n = mk/n.$$

Next,

$$\mathbb{E}[|U \cap \boldsymbol{I}|^2] = \sum_{i,j \in U} \mathbb{E} z_i z_j = \sum_{i \in U} \mathbb{E} z_i^2 + 2 \sum_{i,j \in U \colon i<j} \mathbb{E} z_i z_j.$$

For the first term, since z_i is an indicator, we have

$$\sum_{i \in U} \mathbb{E} z_i^2 = \sum_{i \in U} \mathbb{E} z_i = mk/n.$$

For the second term, fix $i < j$, and we have

$$\begin{aligned}\mathbb{E} z_i z_j = \Pr[z_i = 1, z_j = 1] = \Pr[\{i,j\} \subseteq \boldsymbol{I}] &= \frac{1}{\binom{n}{k}} \left|\{I \subseteq [n] \colon |I| = k, i, j \in I\}\right| \\ &= \frac{\binom{n-2}{k-2}}{\binom{n}{k}} = \frac{(n-2)!k!(n-k)!}{n!(k-2)!(n-k)!} = \frac{k(k-1)}{n(n-1)} \leq \frac{k^2}{n^2}.\end{aligned}$$

We derive that

$$\mathbb{E}[|U \cap \boldsymbol{I}|^2] \leq \frac{mk}{n} + \frac{m^2k^2}{n^2}.$$

Hence,

$$\mathrm{Var}(|U \cap \boldsymbol{I}|) = \mathbb{E}[|U \cap \boldsymbol{I}|^2] - \mathbb{E}[|U \cap \boldsymbol{I}|] \leq \frac{mk}{n} + \frac{m^2k^2}{n^2} - \frac{m^2k^2}{n^2} = \frac{mk}{n},$$

as required. □