Statistical Indistinguishability of Learning Algorithms

Alkis Kalavasis [1] **Amin Karbasi** [2 3] **Shay Moran** [4 3] **Grigoris Velegkas** [2]

Abstract

When two different parties use the same learning rule on their own data, how can we test whether the distributions of the two outcomes are similar? In this paper, we study the similarity of outcomes of learning rules through the lens of the Total Variation (TV) distance of distributions. We say that a learning rule is TV indistinguishable if the expected TV distance between the posterior distributions of its outputs, executed on two training data sets drawn independently from the same distribution, is small. We first investigate the learnability of hypothesis classes using TV indistinguishable learners. Our main results are information-theoretic equivalences between TV indistinguishability and existing algorithmic stability notions such as replicability and approximate differential privacy. Then, we provide statistical amplification and boosting algorithms for TV indistinguishable learners.

1. Introduction

Lack of replicability in experiments has been a major issue, usually referred to as the *reproducibility crisis*, in many scientific areas such as biology and chemistry. Indeed, the results of a survey that appeared in Nature (Baker, 2016) are very worrisome: more than 70% of the researchers that participated in it could not replicate other researchers' experimental findings while over half of them were not able to even replicate their own conclusions. In the past few years the number of scientific publications in the Machine Learning (ML) community has increased exponentially. Significant concerns and questions regarding replicability have also recently been raised in the area of ML. This can be witnessed by the establishment of various reproducibility challenges in major ML conferences such as the ICLR 2019 Reproducibility Challenge (Pineau et al., 2019) and the NeurIPS 2019 Reproducibility Program (Pineau et al., 2021).

Reproducibility of outcomes in scientific research is a necessary condition to ensure that the conclusions of the studies reflect inherent properties of the underlying population and are not an artifact of the methods that scientists used or the random sample of the population that the study was conducted on. In its simplest form, it requires that if two different groups of researchers carry out an experiment using the same methodologies but *different* samples of the *same* population, it better be the case that the two outcomes of their studies are *statistically indistinguishable*. In this paper, we investigate this notion in the context of ML (cf. Definition 1.1), and characterize for which learning problems statistically indistinguishable learning algorithms exist. Furthermore, we show how statistical indistinguishability, as a property of learning algorithms, is naturally related to various notions of algorithmic stability such as replicability of experiments, and differential privacy.

While we mainly focus on the fundamental ML task of binary classification to make the presentation easier to follow, many of our results extend to other statistical tasks (cf. Appendix A.2). More formally, the objects of interest are *randomized* learning rules $A : (\mathcal{X} \times \{0,1\})^n \to \{0,1\}^{\mathcal{X}}$. These learning rules take as input a sequence S of n pairs from $\mathcal{X} \times \{0,1\}$, i.e., points from a domain $\mathcal{X}$ along with their labels, and map them to a binary classifier in a randomized manner. We assume that this sequence S is generated i.i.d. from a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$. We denote by $\{0,1\}^{\mathcal{X}}$ the space of binary classifiers and by $A(S)$ the random variable that corresponds to the output of A on input S[1]. We also adopt a more algorithmic viewpoint for A where we denote it as a *deterministic* mapping $(\mathcal{X} \times \{0,1\})^n \times \mathcal{R} \to \{0,1\}^{\mathcal{X}}$, which takes as input a training set S of size n made of instance-label pairs and a random string $r \sim \mathcal{R}$ (we use $\mathcal{R}$ for both the probability space and the distribution) corresponding to the algorithm's *internal randomness*, and outputs a hypothesis $A(S,r) \in \{0,1\}^{\mathcal{X}}$. Thus, $A(S)$ corresponds to a random variable while $A(S,r)$ is a deterministic object. To make the distinction clear,

[1]Department of Computer Science, NTUA, Athens, Greece [2]Department of Computer Science, Yale University, New Haven, United States [3]Google Research [4]Department of Computer Science, Technion, Haifa, Israel. Correspondence to: Alkis Kalavasis <kalavasisalkis@mail.ntua.gr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]We identify with $A(S)$ the posterior distribution of A on input S when there is no confusion.

we refer to $A(S)$ as (the image of) a *learning rule* and to $A(S, r)$ as (the image of) a *learning algorithm*.

Indistinguishability. We measure how much two distributions over hypotheses differ using some notion of **statistical dissimilarity** d, which can belong to a quite general class; we could let it be either an Integral Probability Metric (IPM) (e.g., TV or Wasserstein distance, see Definition A.2) or an f-divergence (e.g., KL or Rényi divergence). For further details, see (Sriperumbudur et al., 2009). We are now ready to introduce the following general definition of *indistinguishability of learning rules*.

Definition 1.1 (Indistinguishability)**.** Let d be a statistical dissimilarity measure. A learning rule A is n-sample ρ-indistinguishable with respect to d if for any distribution $\mathcal{D}$ over inputs and two independent sets $S, S' \sim \mathcal{D}^n$ it holds that

$$\mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n} \left[d\left(A(S), A(S') \right) \right] \leq \rho \, .$$

In words, Definition 1.1 states that the expected dissimilarity of the outputs of the learning rule when executed on two training sets that are drawn independently from $\mathcal{D}$ is small. We view Definition 1.1 as a general information-theoretic way to study indistinguishability as a property of learning rules. In particular, it captures the property that the distribution of outcomes of a learning rule being *indistinguishable* under the resampling of its inputs. Definition 1.1 provides the flexibility to define the dissimilarity measure according to the needs of the application domain. For instance, it captures as a special case the global stability property (Bun et al., 2020) (see Appendix A.2).

Replicability. Since the issue of replicability is omnipresent in scientific disciplines it is important to design a formal framework through which we can argue about the replicability of experiments. Recently, various works proposed algorithmic definitions of replicability in the context of learning from samples (Impagliazzo et al., 2022; Bun et al., 2023), optimization (Ahn et al., 2022), bandits (Esfandiari et al., 2022) and clustering (Esfandiari et al., 2023), and designed algorithms that are provably replicable under these definitions. A notion that is closely related to Definition 1.1 was introduced by (Impagliazzo et al., 2022): reproducibility or replicability[2] of learning algorithms is defined as follows:

Definition 1.2 (Replicability (Impagliazzo et al., 2022))**.** Let $\mathcal{R}$ be a distribution over random strings. A learning algorithm A is n-sample ρ-replicable if for any distribution $\mathcal{D}$ over inputs and two independent sets $S, S' \sim \mathcal{D}^n$ it holds that

$$\Pr_{S,S' \sim \mathcal{D}^n, r \sim \mathcal{R}} [A(S, r) \neq A(S', r)] \leq \rho \, .$$

The existence of a shared random seed r in the definition of replicability is one of the main distinctions between Definition 1.1 and 1.2. This shared random string can be seen as a way to achieve a *coupling* (see Definition A.1) between two executions of the algorithm A. An interesting aspect of this definition is that replicability is verifiable; replicability under Definition 1.2 can be tested using polynomially many samples, random seeds r and queries to A. We remark that the work of (Ghazi et al., 2021b) introduced the closely related notion of pseudo-global stability (see Definition 1.7); the definitions of replicability and pseudo-global stability are equivalent up to polynomial factors in the parameters.

Differential Privacy. The notions of algorithmic indistinguishability and replicability that we have discussed so far have close connections with the classical definition of approximate differential privacy (Dwork & Roth, 2014). For $a, b, \varepsilon, \delta \in [0, 1]$, let $a \approx_{\varepsilon,\delta} b$ denote the statement $a \leq e^{\varepsilon} b + \delta$ and $b \leq e^{\varepsilon} a + \delta$. We say that two probability distributions P, Q are (ε, δ)-indistinguishable if $P(E) \approx_{\varepsilon,\delta} Q(E)$ for any measurable event E.

Definition 1.3 (Approximate Differential Privacy (Dwork et al., 2006))**.** A learning rule A is an n-sample (ε, δ)-differentially private if for any pair of samples $S, S' \in (\mathcal{X} \times \{0, 1\})^n$ that disagree on a single example, the induced posterior distributions $A(S)$ and $A(S')$ are (ε, δ)-indistinguishable.

We remind the reader that, in the context of PAC learning, any hypothesis class $\mathcal{H}$ can be PAC-learned by an approximate differentially-private algorithm if and only if it has a finite Littlestone dimension $\mathrm{Ldim}(\mathcal{H})$ (see Definition A.3), i.e., there is a qualitative equivalence between online learnability and private PAC learnability (Alon et al., 2019; Bun et al., 2020; Ghazi et al., 2021a; Alon et al., 2022).

Broader Perspective. Our work lies in the fundamental research direction of responsible ML. Basic concepts in this area, such as DP, replicability, and different forms of fairness, are formalized using various forms of stability. Therefore, it is natural and important to formally study the interrelations between different types of algorithmic stability. Our main purpose is to study statistical indistinguishability and replicability as properties of algorithms and, under the perspective of stability, investigate rigorous connections with DP.

1.1. TV Indistinguishable Learning Rules

As we discussed, our Definition 1.1 captures the property of a learning rule having *indistinguishable* outcomes under

[2] This property was originally defined as "reproducibility" in (Impagliazzo et al., 2022), but later it was pointed out that the correct term for this definition is "replicability" (see also (Bun et al., 2023)). We use the term replicability throughout our work.

the resampling of its inputs from the same distribution. In what follows, we instantiate Definition 1.1 with d being the total variation (TV) distance, probably the most well-studied notion of statistical distance in theoretical computer science. Total variation distance between two distributions P and Q over the probability space (Ω, Σ_Ω) can be expressed as

$$\begin{aligned} d_{\mathrm{TV}}(P,Q) &= \sup_{A \in \Sigma_\Omega} P(A) - Q(A) \\ &= \inf_{(X,Y)\sim\Pi(P,Q)} \mathbf{Pr}[X \neq Y]\,, \end{aligned} \tag{1}$$

where the infimum is over all couplings between P and Q so that the associated marginals are P and Q respectively. A *coupling* between the distributions P and Q is a set of variables (X, Y) on some common probability space with the given marginals, i.e., $X \sim P$ and $Y \sim Q$. We think of a coupling as a construction of random variables X, Y with prescribed laws.

Setting $d = d_{\mathrm{TV}}$ in Definition 1.1, we get the following natural definition. For simplicity, we use the term TV indistinguishability to capture indistinguishability with respect to the TV distance.

Definition 1.4 (Total Variation Indistinguishability)**.** A learning rule A is n-sample ρ-TV indistinguishable if for any distribution over inputs $\mathcal{D}$ and two independent sets $S, S' \sim \mathcal{D}^n$ it holds that

$$\mathop{\mathbf{E}}_{S,S'\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S), A(S'))] \leq \rho\,.$$

For some equivalent definitions, we refer to Appendix A.3. Moreover, for some extensive discussion about the motivation of this definition, see Appendix A.5. We emphasize that the notion of TV distance has very strong connections with statistical indistinguishability of distributions. If two distributions P and Q are close in TV distance, then, intuitively, no statistical test can distinguish whether an observation was drawn from P or Q. In particular, if $d_{\mathrm{TV}}(P, Q) = \rho$, then $\rho/2$ is the maximum advantage an analyst can achieve in determining whether a random sample X came from P or from Q (where P or Q is used with probability 1/2 each). In what follows, we focus on this notion of dissimilarity.

As a warmup, we start by proving a generalization result for TV indistinguishable learners. Recall that if we *fix* some binary classifier we can show, using standard concentration bounds, that its performance on a sample is close to its performance on the underlying population. However, when we train an ML algorithm using a dataset S to output a classifier h we cannot just use the fact that it has small loss on S to claim that its loss on the population is small because h depends on S. The following result shows that we can get such generalization bounds if A is a ρ-TV indistinguishable algorithm. We remark that a similar result regarding replicable algorithms appears in (Impagliazzo et al., 2022). The formal proof, stated in a slightly more general way, is in Appendix F.

Proposition 1.5 (TV Indistinguishability Implies Generalization)**.** *Let $\delta, \rho \in (0,1)^2$. Let $\mathcal{D}$ be a distribution over inputs and $S = \{(x_i, y_i)\}_{i\in[n]}$ be a sample of size n drawn i.i.d. from $\mathcal{D}$. Let $h : \mathcal{X} \to \{0,1\}$ be the output of an n-sample ρ-TV indistinguishable learning rule A with input S. Then, with probability at least $1 - \delta - 4\sqrt{\rho}$ over S, it holds that,*

$$\left| \mathop{\mathbf{E}}_{h\sim A(S)}[L(h)] - \mathop{\mathbf{E}}_{h\sim A(S)}\left[\widehat{L}(h)\right] \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}} + \sqrt{\rho}\,,$$

where $L(h) \triangleq \mathbf{Pr}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$ and $\widehat{L}(h) \triangleq \frac{1}{n}\sum_{(x,y)\in S} 1\{h(x) \neq y\}$.

1.2. Summary Of Contributions

In this work, we investigate the connections between TV indistinguishability, replicability and differential privacy.

- In Section 2, we show that TV indistinguishability and replicability are equivalent. This equivalence holds for countable domains[3] and extends to general statistical tasks (cf. Appendix C.2).
- In Section 3, we show that TV indistinguishability and (ε, δ)-DP are statistically equivalent. This equivalence holds for countable[4] domains in the context of PAC learning. As an intermediate result, we also show that replicability and (ε, δ)-DP are statistically equivalent in the context of PAC learning, and this holds for general domains.
- In Section 4, we provide statistical amplification and boosting algorithms for TV indistinguishable learners for countable domains. En route, we improve the sample complexity of some routines provided in (Impagliazzo et al., 2022).

1.3. Related Work

Our work falls in the research agenda of replicable algorithm design, which was initiated by (Impagliazzo et al., 2022). In particular, (Impagliazzo et al., 2022) introduced the notion of replicable learning algorithms, established that any statistical query algorithm can be made replicable, and designed replicable algorithms for various applications such as halfspace learning. Next, (Ahn et al., 2022) studied reproducibility in optimization and (Esfandiari et al., 2022) provided replicable bandit algorithms.

[3]We remark that the direction replicability implies TV indistinguishability holds for general domains.

[4]We remark that the direction (ε, δ)-DP implies TV indistinguishability holds for general domains.

The most closely related prior work to ours is the recent paper by (Bun et al., 2023). In particular, as we discuss below in greater detail, an alternative proof of the equivalence between TV indistinguishability, replicability, and differential privacy follows from (Bun et al., 2023). In contrast with our equivalence, the transformations by (Bun et al., 2023) are restricted to finite classes. On the other hand, (Bun et al., 2023) give a constructive proof whereas our proof is purely information-theoretic. In more detail, (Bun et al., 2023) establish a variety of equivalences between different notions of stability such as differential privacy, replicability, and one-way perfect generalization, and the latter contains TV indistinguishability as a special case:

Definition 1.6 ((One-Way) Perfect Generalization (Cummings et al., 2016; Bassily & Freund, 2016))**.** A learning rule $A : \mathcal{X}^n \to \mathcal{Y}$ is $(\beta, \varepsilon, \delta)$-perfectly generalizing if, for every distribution $\mathcal{D}$ over $\mathcal{X}$, there exists a distribution $\mathcal{P}_{\mathcal{D}}$ such that, with probability at least $1-\beta$ over S consisting of n i.i.d. samples from $\mathcal{D}$, and every set of outcomes $\mathcal{O} \subseteq \mathcal{Y}$

$$e^{-\varepsilon}\left(\Pr_{\mathcal{P}_{\mathcal{D}}}[\mathcal{O}] - \delta\right) \leq \mathbf{Pr}[A(S) \in \mathcal{O}] \leq e^{\varepsilon} \Pr_{\mathcal{P}_{\mathcal{D}}}[\mathcal{O}] + \delta\,.$$

Moreover, A is $(\beta, \varepsilon, \delta)$-one-way perfectly generalizing if $\mathbf{Pr}[A(S) \in \mathcal{O}] \leq e^{\varepsilon}\, \mathbf{Pr}_{\mathcal{P}_{\mathcal{D}}}[\mathcal{O}] + \delta$.

Note indeed that plugging $\varepsilon = 0$ to the definition of perfect generalization specializes the above definition to an equivalent variant of TV indistinguishability (see also Definition A.9). (Bun et al., 2023) derives an equivalence between replicability and one-way perfect generalization with $\varepsilon > 0$. However, in a personal communication they pointed out to us that their argument also applies to the case $\varepsilon = 0$, and hence to TV indistinguishability. In more detail, an intermediate step of their proof shows that any $(\beta, \varepsilon, \delta)$-perfectly generalizing algorithm A is also $(\beta, 0, 2\varepsilon + \delta)$-perfectly generalizing, which is qualitatively equivalent with our main definition (see Definition 1.4). As noted earlier our proof applies more generally to infinite countable domains but is non-constructive.

Differential Privacy. Differential privacy (Dwork, 2008; Dwork et al., 2010; Vadhan, 2017; Dwork & Roth, 2014) is quite closely related to replicability. The first connection between replicability and DP in the context of PAC learning was, implicitly, established by (Ghazi et al., 2021b) (for finite domains $\mathcal{X}$), via the technique of correlated sampling (see Appendix A.4) and the notion of pseudo-global stability (which is equivalent to replicability as noticed by (Impagliazzo et al., 2022)):

Definition 1.7 (Pseudo-Global Stability (Ghazi et al., 2021b))**.** Let $\mathcal{R}$ be a distribution over random strings. A learning algorithm A is said to be n-sample (η, ν)-pseudo-globally stable if for any distribution $\mathcal{D}$ there exists a hypothesis h_r for every $r \in \mathrm{supp}(\mathcal{R})$ (depending on $\mathcal{D}$) such that

$$\Pr_{r \sim \mathcal{R}}\left[\Pr_{S \sim \mathcal{D}^n}[A(S, r) = h_r] \geq \eta\right] \geq \nu\,.$$

The high-level connection between these notions appears to boil down to the notion of stability (Bousquet & Elisseeff, 2002; Poggio et al., 2004; Dwork et al., 2015; Abernethy et al., 2017; Bassily et al., 2016; Livni & Moran, 2020) (see (Alon et al., 2022) for further details between stability, online learnability and differential privacy). In particular, (Ghazi et al., 2021a) showed that a class of finite Littlestone dimension admits a list-globally stable learner (see Theorem 18 in (Ghazi et al., 2021b)). The work of (Ghazi et al., 2021b) (among other things) showed (i) how to perform a reduction from list-global stability to pseudo-global stability via correlated sampling in finite domains (see Theorem 20 in (Ghazi et al., 2021b)) and (ii) how to perform a reduction from pseudo-global stability to approximate DP via DP selection (see Theorem 25 in (Ghazi et al., 2021b)). We highlight that this equivalence between differential privacy and replicability for finite domains was made formal by (Bun et al., 2023) and was extended to general tasks.

TV Stability. The definition of TV indistinguishability has close connections with the definition of TV stability. This notion has appeared in the context of adaptive data analysis. The work of (Bassily et al., 2016) studied the following problem: suppose there is an unknown distribution P and a set S of n independent samples drawn i.i.d. from P. The goal is to design an algorithm that, with input S, will accurately answer a sequence of adaptively chosen queries about the unknown distribution P. The main question is how many samples must one draw from the distribution, as a function of the type of queries, the number of queries, and the desired level of accuracy to perform well? (Bassily et al., 2016) provide various results that rely on the connections between algorithmic stability, differential privacy and generalization. To this end, they think of differential privacy as max-KL stability and study the performance of other notions of stability such as TV stability. Crucially, in their definition, TV stability considers any pair of neighboring datasets S, S' and not two independent draws from P. More concretely, they propose the following definition.

Definition 1.8 (Total Variation Stability (Bassily et al., 2016))**.** A learning rule A is n-sample ρ-TV stable if for any pair of samples $S, S' \in (\mathcal{X} \times \{0,1\})^n$ that disagree on a single example, it holds that $d_{\mathrm{TV}}(A(S), A(S')) \leq \rho$.

We underline that for any constant ρ it is not challenging to obtain a ρ-TV stable algorithm in the learning setting we are interested in. It suffices to just sub-sample a small enough subset of the data. Hence, any class with finite VC dimension is TV stably learnable under this definition. As it is evident from our results (cf. Theorem 3.2), this is in stark contrast with the definition we propose. We remind

the readers that just sub-sampling the dataset is not enough to achieve differential privacy. This is because it is required that $\delta = o(1/n)$. We remark that the definition of total variation stability à la (Bassily et al., 2016) also appears in (Raginsky et al., 2016). The above definition of TV stability has close connections to machine unlearning. This problem refers to the ability of a user to delete their data that were used to train a ML algorithm. When this happens, the machine learning algorithm has to move to a state as if it had never used that data for training, hence the term *machine unlearning*. One can see that Definition 1.8 is suitable for this setting since it states that if one point of the dataset is deleted, the distribution of the algorithm should not be affected very much. For convex risk minimization problems, (Ullah et al., 2021) design TV stable algorithms based on noisy Stochastic Gradient Descent (SGD). Such approaches lead to the design of efficient unlearning algorithms, which are based on sub-sampling the dataset and constructing a maximal coupling of Markov chains for the noisy SGD procedure.

KL Stability and PAC-Bayes. In Appendix A.3 we provide some equivalent definitions to TV indistinguishability. In particular, Definition A.9 has connections with the line of work that studies distribution-dependent generalization bounds. To be more precise, if instead of the TV distance we use the KL divergence to measure the distance between the prior and the output of the algorithm we get the definition of the quantity that is used to derive PAC-Bayes generalization bounds. Interestingly, (Livni & Moran, 2020) show that the PAC-Bayes framework cannot be used to derive distribution-free PAC learning bounds for classes that have infinite Littlestone dimension; they show that for any algorithm that learns 1-dimensional linear classifiers (thresholds), there exists a realizable distribution for which PAC-Bayes bounds are trivial. Recently, a similar PAC-Bayes framework was proposed in (Amit et al., 2022), where the KL divergence is replaced with a general family of Integral Probability Metrics (cf. Definition A.2).

Probably Eventually Correct Learning. The work of (Malliaris & Moran, 2022) introduced the *Probably Eventually Correct* (PEC) model of learning. In this model, a learner outputs the same hypothesis[5], with probability one, after a uniformly bounded number of revisions. Intuitively, this corresponds to the property that the global stability parameter is close to 1. Interestingly, prior work on global stability (Bun et al., 2020; Ghazi et al., 2021a) had characterized *Littlestone classes* as being PAC learnable by an algorithm which outputs some fixed hypothesis with nonzero probability. However, the frequency of this hypothesis was typically very small and its loss was a priori non-zero. (Malliaris & Moran, 2022) give a new characterization to Littlestone classes by identifying them with the classes that can be PEC learned in a stable fashion. Informally, this means that the learning rule for $\mathcal{H}$ stabilizes on some hypothesis after changing its mind at most L times, where L is the Littlestone dimension of $\mathcal{H}$ (cf. Definition A.3). Interestingly, (Malliaris & Moran, 2022) manage to show that the well-known *Standard Optimal Algorithm* (SOA) (Littlestone, 1988) is a stable PEC learner, using tools from the theory of universal learning (Bousquet et al., 2021; 2022; Kalavasis et al., 2022; Hanneke et al., 2022). Moreover, they list various different notions of algorithmic stability and show that they all have something in common: a class $\mathcal{H}$ is learnable by such learners if and only if its Littlestone dimension is finite. Our main result shows that, indeed, classes that are learnable by TV indistinguishable learners fall into that category.

[5]Except maybe for a subset of $\mathcal{X}$ that has measure zero under the data-generating distribution.

2. TV Indistinguishability and Replicability

Our information-theoretic definition of TV indistinguishability seems to put weaker restrictions on learning rules than the notion of replicability in two ways: (i) it allows for *arbitrary* couplings between the two executions of the algorithm (recall the coupling definition of TV distance, see Eq.(1)), and, (ii) it allows for *different* couplings between every pair of datasets S, S' (the optimal coupling in the definition of TV distance will depend on S, S' of Definition 1.4). In short, our definition allows for *arbitrary data-dependent* couplings, instead of just sharing the randomness across two executions. TV indistinguishability can be viewed as a statistical generalization of replicability (cf. Definition 1.2) since it describes a property of *learning rules* rather than *learning algorithms*.

In this section, we will show that TV indistinguishability and replicability are (perhaps surprisingly) equivalent in a rather strong sense: under a mild measure-theoretic condition, every TV indistinguishable algorithm can be converted into an *equivalent* replicable one by *re-interpreting* its internal randomness. This will be made formal shortly.

First, we show that any replicable algorithm is TV indistinguishable.

Theorem 2.1 (Replicability $\Rightarrow$ TV Indistinguishability)**.** *If a learning rule A is n-sample ρ-replicable, then it is also n-sample ρ-*TV *indistinguishable.*

Proof. Fix some distribution $\mathcal{D}$ over inputs. Let A be n-sample ρ-replicable with respect to $\mathcal{D}$. For the random variables $A(S), A(S')$ where $S, S' \sim \mathcal{D}^n$ are two independent samples and using Eq.(1), we have that $\mathbf{E}_{S,S'\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S), A(S'))]$ equals $\mathbf{E}_{S,S'\sim\mathcal{D}^n}\left[\inf_{(h,h')\sim\Pi(A(S),A(S'))} \mathbf{Pr}[h \neq h']\right]$.
Let $\mathcal{R}$ be the source of randomness that A uses. The expected optimal coupling is at most

$\mathbf{E}_{S,S'\sim\mathcal{D}^n}[\mathbf{Pr}_{r\sim\mathcal{R}}[A(S,r)\neq A(S',r)]]$. This inequality follows from the fact that using shared randomness between the two executions of A is a particular way to couple the two random variables. To complete the proof, it suffices to notice that this upper bound is equal to $\mathbf{Pr}_{S,S'\sim\mathcal{D}^n,r\sim\mathcal{R}}[A(S,r)\neq A(S',r)]\leq\rho$. The last inequality follows since A is ρ-replicable. □

We now deal with the opposite direction, i.e., we show that TV indistinguishability implies replicability. We first provide some measure-theoretic definitions. Let us recall the definition of absolute continuity for two measures.

Definition 2.2 (Absolute Continuity)**.** Consider two measures P, Q on a σ-algebra $\mathcal{B}$ of subsets of Ω. We say that P is absolutely continuous with respect to Q if for any $E\in\mathcal{B}$ such that $Q(E)=0$, it holds that $P(E)=0$.

Since the learning rules induce posterior distributions over hypotheses, this definition extends naturally to such rules.

Definition 2.3. Given learning rule A, distribution over inputs $\mathcal{D}$ and reference probability measure $\mathcal{P}$, we say that A is absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$ if, for almost every sample S drawn from $\mathcal{D}$, the distribution $A(S)$ is absolutely continuous with respect to $\mathcal{P}$.

In the previous definition, we fixed the data-generating distribution $\mathcal{D}$. We next consider its distribution-free version.

Definition 2.4. Given learning rule A and reference probability measure $\mathcal{P}$, we say that A is absolutely continuous with respect to $\mathcal{P}$ if, for any distribution over inputs $\mathcal{D}$, A is absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$.

If $\mathcal{X}$ is finite, then one can take $\mathcal{P}$ to be the uniform probability measure over $\{0,1\}^{\mathcal{X}}$ and any learning rule is absolutely continuous with respect to $\mathcal{P}$. We now show how we can find such a prior $\mathcal{P}$ in the case where $\mathcal{X}$ is countable.

Claim 2.5 (Reference Probability Measure for Countable Domains)**.** Let $\mathcal{X}$ be a countable domain and A be a learning rule. Then, there is a reference probability measure $\mathcal{P}$ such that A is absolutely continuous with respect to $\mathcal{P}$.

Proof. Since $\mathcal{X}$ is countable, for a fixed n, we can consider an enumeration of all the n-tuples $\{S_i\}_{i\in\mathbb{N}}$. Then, we can take $\mathcal{P}$ to be a countable mixture of these probability measures, i.e., $\mathcal{P}=\sum_{i=1}^{\infty}\frac{1}{2^i}A(S_i)$. Notice that since, each $A(S_i)$ is a measure and $1/2^i>0$ for $i\in\mathbb{N}$, and, $\sum_{i=1}^{\infty}1/2^i=1$, we have that $\mathcal{P}$ is indeed a probability measure. We now argue that each $A(S_i)$ is absolutely continuous with respect to $\mathcal{P}$. Assume towards contradiction that this is not the case and let $E\in\mathcal{B}$ be a set such that $\mathcal{P}(E)=0$ but $A(S_j)(E)\neq 0$, for some $j\in\mathbb{N}$. Notice that $A(S_j)$ appears with coefficient $1/2^j>0$ in the mixture that we consider, hence if $A(S_j)(E)>0 \implies 1/2^j A(S_j)(E)>0$. Moreover $A(S_i)(E)\geq 0, \forall i\in\mathbb{N}$, which means that $\mathcal{P}(E)>0$, so we get a contradiction. □

We next define when two learning rules A, A' are equivalent.

Definition 2.6 (Equivalent Learning Rules)**.** Two learning rules A, A' are equivalent if for every sample S it holds that $A(S)=A'(S)$, i.e., for the same input they induce the same distribution over hypotheses.

In the next result, we show that for every TV indistinguishable algorithm A, that is absolutely continuous with respect to some reference probability measure $\mathcal{P}$, there exists an equivalent learning rule which is replicable.

Theorem 2.7 (TV Indistinguishability ⇒ Replicability)**.** *Let $\mathcal{P}$ be a reference probability measure over $\{0,1\}^{\mathcal{X}}$, and let A be a learning rule that is n-sample ρ-*TV *indistinguishable and absolutely continuous with respect to $\mathcal{P}$. Then, there exists an equivalent learning rule A' that is n-sample $\frac{2\rho}{1+\rho}$-replicable.*

In this section, we only provide a sketch of the proof and we refer the reader to Appendix C.1 for the complete one. Let us first state how we can use the previous result when $\mathcal{X}$ is countable.

Corollary 2.8. *Let $\mathcal{X}$ be a countable domain and let A be a learning rule that is n-sample ρ-*TV *indistinguishable. Then, there exists an equivalent learning rule A' that is n-sample $\frac{2\rho}{1+\rho}$-replicable.*

The proof of this result follows immediately from Claim 2.5 and Theorem 2.7.

Proof Sketch of Theorem 2.7. Let us consider a learning rule A satisfying the conditions of Theorem 2.7. Fix a distribution $\mathcal{D}$ over inputs. The crux of the proof is that given two random variables X, Y whose TV distance is bounded by ρ, we can couple them using only a carefully designed source of shared randomness $\mathcal{R}$ so that the probability that the realizations of these random variables differ is at most $2\rho/(1+\rho)$. We can instantiate this observation with $X=A(S)$ and $Y=A(S')$. Crucially, in the countable $\mathcal{X}$ setting, we can pick the shared randomness $\mathcal{R}$ in a way that only depends on the learning rule A, but not on S or S'. Let us now describe how this coupling works. Essentially, it can be thought of as a generalization of the von Neumann rejection-based sampling which does not necessarily require that the distribution has bounded density. Following (Angel & Spinka, 2019), we pick $\mathcal{R}$ to be a Poisson point process which generates points of the form (h,y,t) with intensity[6]

[6]Roughly speaking, a point process is a Poisson point process with intensity λ if (i) the number of points in a bounded Borel set E is a Poisson random variable with mean $\lambda(E)$ and (ii) the numbers of points in n disjoint Borel sets forms n independent random variables. For details, we refer to (Last & Penrose, 2017).

$\mathcal{P} \times \text{Leb} \times \text{Leb}$, where $\mathcal{P}$ is a reference probability measure with respect to which A is absolutely continuous and Leb is the Lebesgue measure over $\mathbb{R}_+$. Intuitively, $h \sim \mathcal{P}$ lies in the hypotheses' space, y is a non-negative real value and t corresponds to a time value. The coupling mechanism performs *rejection sampling* for each distribution we would like to couple (here $A(S)$ and $A(S')$): it checks (in the ordering indicated by the time parameter) for each point (h, y, t) whether $f(h) > y$ (i.e., if y falls below the density curve f at h) and accepts the first point that satisfies this condition. In the formal proof, there will be two density functions; f (resp. f') for the density function of $A(S)$ (resp. $A(S')$). We also refer to Figure 1. One can show (see Theorem A.13) that $\mathcal{R}$ gives rise to a coupling between $A(S)$ and $A(S')$ under the condition that both measures are absolutely continuous with respect to $\mathcal{P}$. This coupling technique appears in (Angel & Spinka, 2019). We can then apply it and get $\mathbf{Pr}_{r\sim\mathcal{R}}[A(S,r) \neq A(S',r)] \leq \frac{2d_{\text{TV}}(A(S),A(S'))}{1+d_{\text{TV}}(A(S),A(S'))}$. Taking the expectation with respect to the draws of S, S', we show (after some algebraic manipulations) that $\mathbf{Pr}_{S,S'\sim\mathcal{D}^n r\sim\mathcal{R}}[A(S,r) \neq A(S',r)] \leq 2\rho/(1+\rho)$. We end this section with the following remarks.

Remark 2.9 (General Equivalence). In Appendix C.2, we show that this equivalence actually holds for general statistical tasks. We first generalize the notions of indistinguishability, replicability and TV indistinguishability for general input spaces $\mathcal{I}$ and output spaces $\mathcal{O}$. We then discuss that replicability and TV indistinguishability remain equivalent (under the same measure theoretic conditions) in these more general abstract learning scenarios.

Remark 2.10 (Implementation of the Coupling). We note that, in order to implement algorithm A' of Theorem 2.7, we need sample access to a Poisson point process with intensity $\mathcal{P}\times\text{Leb}\times\text{Leb}$, where $\mathcal{P}$ is the reference probability measure from Claim 2.5 and Leb is the Lebesgue measure over $\mathbb{R}_+$. Importantly, $\mathcal{P}$ depends only on A. Moreover, we need full access to the values of the density f_i of the distribution $A(S_i)$ with respect to the reference probability measure $\mathcal{P}$, for any sample S_i. We underline that these quantities do not depend on the data-generating distribution $\mathcal{D}$ (since we iterate over any possible sample).

Remark 2.11 (TV Indistinguishability vs. Replicability). Notice that in the definition of replicability (cf. Definition 1.2) the source of randomness $\mathcal{R}$ needs to be specified and by changing it we can observe different behaviors for coupled executions of the algorithm. On the other hand, the definition of TV indistinguishability (cf. Definition 1.4) does not require the specification of $\mathcal{R}$ as it states a property of the posterior distribution of the learning rule.

3. TV Indistinguishability and DP

In this section we investigate the connections between TV indistinguishability and approximate DP in binary classification. Consider a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. We will say that $\mathcal{H}$ is learnable by a ρ-TV indistinguishable learning rule A if this rule satisfies the notion of learnability under the standard realizable PAC learning model and is ρ-TV indistinguishable (see Definition A.5).

The main result of this section is an equivalence between approximate DP and TV indistinguishability for countable domains $\mathcal{X}$, in the context of PAC learning. We remark that the equivalence of differential privacy with the notion of replicability is formally stated for finite outcome spaces (i.e., under the assumption that $\mathcal{X}$ is finite) due to the use of a specific correlated sampling strategy for the direction that "DP implies replicability" in the context of classification (Ghazi et al., 2021b). Moreover, (Bun et al., 2023) gave a constructive way to transform a DP algorithm to a replicable one for general statistical tasks and for finite domains. Thus, combining our results in Section 2 and the result of (Ghazi et al., 2021b; Impagliazzo et al., 2022; Bun et al., 2023), the equivalence of TV indistinguishability and DP for *finite* domains is immediate. We will elaborate more on the differences of our approach and (Ghazi et al., 2021b; Bun et al., 2023) later on. We also discuss our coupling and correlated sampling in Appendix A.4.

Recall that a learner is (α, β)-accurate if its misclassification probability is at most α with probability at least $1 - \beta$.

Theorem 3.1 ((ε, δ)-DP $\Rightarrow$ TV Indistinguishability)**.** *Let $\mathcal{X}$ be a (possibly infinite) domain and $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. Let $\gamma \in (0, 1/2), \alpha, \beta, \rho \in (0,1)^3$. Assume that $\mathcal{H}$ is learnable by an n-sample $(1/2-\gamma, 1/2-\gamma)$-accurate $(0.1, 1/(n^2\log(n)))$-differentially private learner. Then, it is also learnable by an (α, β)-accurate ρ-TV indistinguishable learning rule.*

Proof Sketch of Theorem 3.1. The proof goes through the notion of global stability (cf. Definition A.8). The existence of an (ε, δ)-DP learner implies that the hypothesis class $\mathcal{H}$ has finite Littlestone dimension (Alon et al., 2019) (cf. Theorem D.3). Thus, we know that there exists a ρ-globally stable learner for $\mathcal{H}$ (Bun et al., 2020) (cf. Theorem D.4). The next step is to use the replicable heavy-hitters algorithm (cf. Algorithm 1, (Impagliazzo et al., 2022)) with frequency parameter $O(\rho)$ and replicability parameter $O(\rho')$, where $\rho' \in (0,1)$ is the desired TV indistinguishability parameter of the learning rule. The global stability property implies that the list of heavy-hitters will be non-empty and it will contain at least one hypothesis with small error rate, with high probability. Finally, since the list of heavy-hitters is finite and has bounded size, we feed the output into the replicable agnostic learner (cf. Algorithm 2). Thus, we have designed a replicable learner for $\mathcal{H}$, and Theorem 2.1 shows

that this learner is also TV indistinguishable.

The formal proof of Theorem 3.1 is deferred to Appendix D.2. We also include a result which shows that *list-global* stability implies TV indistinguishability for general domains and general statistical tasks, which could be of independent interest (cf. Proposition D.12).

We proceed to the opposite direction where we provide an algorithm that takes as input a TV indistinguishable learning rule for $\mathcal{H}$ and outputs a learner for $\mathcal{H}$ which is (ε, δ)-DP. In this direction countability of $\mathcal{X}$ is crucial.

Theorem 3.2 (TV Indistinguishability $\Rightarrow$ (ε, δ)-DP)**.** *Let $\mathcal{X}$ be a countable domain. Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be learnable by an (α, β)-accurate ρ-TV indistinguishable learner A, for some $\rho \in (0,1), \alpha \in (0, 1/2), \beta \in \left(0, \frac{1-\rho}{1+\rho}\right)$. Then, for any $(\alpha', \beta', \varepsilon, \delta) \in (0,1)^4$, it is learnable by an $(\alpha + \alpha', \beta')$-accurate (ε, δ)-differentially private learner A'.*

We refer to Appendix D.4 for the proof. In the above statements, we omit the details about the sample complexity. We refer to Proposition D.12 and Proposition D.16 for these details. Let us now comment on the differences between (Ghazi et al., 2021b; Bun et al., 2023) which establish a transformation from a replicable learner to an approximately DP learner and our result. The high-level idea to obtain both of these results is similar. Essentially, the proof of (Ghazi et al., 2021b; Bun et al., 2023) can be viewed as a coupling between sufficiently many posteriors of the replicable learning rule using *shared randomness* in order to achieve this coupling. In our proof, instead of using shared randomness we use the reference measure we described in previous sections to achieve this coupling. We remark that we could have obtained the same qualitative result, i.e., that TV indistinguishability implies approximate DP, by using the transformation from replicability to approximate DP of (Ghazi et al., 2021a; Bun et al., 2023) in a black-box manner along with our result that TV indistinguishability implies replicability (cf. Theorem 2.7). However, this leads to worse guarantees in terms of the range of the parameters $\alpha, \beta, \delta, \varepsilon, \rho$ than the ones stated in Theorem 3.2. Thus, we have chosen to do a more careful analysis based on the coupling we proposed that leads to a stronger quantitative result. More concretely, the proof in (Ghazi et al., 2021b; Bun et al., 2023) starts by sampling many random strings independently of the dataset $\{S_i\}_{i\in[k]}$ and considers many executions of the algorithm using the same random strings but different data. In our algorithm we first sample the sets $\{S_i\}_{i\in[k]}$ and then we consider an optimal coupling along the $\{A(S_i)\}_{i\in[k]}$ which is also independent of the dataset, thus it satisfies the DP requirements. Moreover, our procedure covers a wider range of parameters α, β, ρ compared to (Ghazi et al., 2021b). The reason we need countability of $\mathcal{X}$ is because it allows us to design a *data-independent* reference probability measure $\mathcal{P}$, the same one as in Claim 2.5. Then, using this reference probability measure for the coupling helps us establish the DP properties. Nevertheless, we propose a simple change to our approach which we conjecture applies to general domains $\mathcal{X}$ and we leave it open as an interesting future direction. For a more detailed discussion, we refer the reader to Appendix D.5.

Remark 3.3 (Dependence on the Parameters). In the case of TV indistinguishability $\Rightarrow$ DP, the blowup in the sample complexity is stated explicitly in Proposition D.16. For the direction DP $\Rightarrow$ TV indistinguishability it is a bit trickier to state the exact sample complexity blow-up because we do not make explicit use of the DP learner. Instead, we use the fact that the existence of a non-trivial DP learner implies that the class has finite Littlestone dimension and then we use an appropriate algorithm that is known to work for such classes. In this case, it suffices to let the parameters of the DP learner to be $\varepsilon \in (0, 0.1), \delta \in \left(0, \frac{1}{n^2 \log(n)}\right), \alpha \in (0, 1/2), \beta \in (0, 1/2)$ and the parameters of the desired TV indistinguishable (α', β')-accurate learner are unconstrained, i.e., $\rho \in (0,1), \alpha' \in (0,1), \beta' \in (0,1)$. If we denote the Littlestone dimension of the class by L, then, as shown in Proposition D.12 the sample complexity of the TV indistinguishable learner is $\mathrm{poly}(L, 1/\rho, 1/\alpha', \log(1/\beta'))$[7].

4. Amplification and Boosting

In this section we study the following fundamental question: given a weak TV indistinguishable learning rule in terms of the indistinguishability parameter and the accuracy, can we amplify its indistinguishability and boost its accuracy?

In the context of approximate differential privacy, a series of works has lead to (constructive) algorithms that boost the accuracy and amplify the privacy guarantees (e.g., (Dwork et al., 2010; Bun et al., 2020; 2023)). This result builds upon the equivalence of online learnability and approximate differential privacy. Our result relating DP to TV indistinguishability implies the following existential result.

Corollary 4.1. *Let $\mathcal{X}$ be a countable domain. Suppose that for some sample size n_0, there exists an (α_0, β_0)-accurate ρ_0-TV indistinguishable learner A for a class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ with $\alpha_0 \in (0, 1/2), \rho_0 \in (0,1), \beta_0 \in \left(0, \frac{1-\rho_0}{1+\rho_0}\right)$. Then, for any $(\alpha, \beta, \rho) \in (0,1)^3$, $\mathcal{H}$ admits an (α, β)-accurate ρ-TV indistinguishable learner A'.*

This result relies on connections between learnability by TV indistinguishable learners and finiteness of the Littlestone dimension of the underlying hypothesis class that were discussed in Section 3. In particular, Corollary D.17

[7] This requires that uniform convergence holds for Littlestone classes. Otherwise, we get $\mathrm{poly}(2^{2^L}, 1/\rho, 1/\alpha', \log(1/\beta'))$ sample complexity (Corollary D.9).

shows that the existence of such a non-trivial TV indistinguishable learner implies that the $\mathcal{H}$ has finite Littlestone dimension, and Proposition D.12, states that the finiteness of the Littlestone dimension of $\mathcal{H}$ implies the existence of an (α, β)-accurate ρ-TV indistinguishable learner, for arbitrarily small choices of α, β, ρ. It is not hard to see that we need to constrain $\alpha \in (0, 1/2)$, because the algorithm needs to have an advantage compared to the random classifier. Moreover, it should be the case that $\beta \in (0, 1-\rho)$. If $\beta \geq 1-\rho$ then the algorithm which outputs a constant classifier with probability β and an α-good one with the remaining probability is ρ-TV indistinguishable and (α, β)-accurate. An interesting open problem is to investigate what happens when $\beta \in \left(\frac{1-\rho}{1+\rho}, 1-\rho\right)$.

We underline that Corollary 4.1 is existential and does not make actual use of the weak TV indistinguishable learner that is given as input. Hence, it is natural to try to come up with sample-efficient and constructive approaches that utilize the weak learner through black-box oracle calls to it during the derivation of the strong one. In what follows, we aim to design such algorithms. We remind the reader that if we constrain ourselves to work in the setting where $\mathcal{X}$ is countable, then the absolute continuity requirement in the next theorems comes immediately, due to Claim 2.5.

Indistinguishability Amplification. We first consider the amplification of the indistinguishability guarantees of an algorithm. An important ingredient of our approach is a replicable algorithm for finding heavy hitters of a distribution, i.e., elements whose frequency is above some given threshold. This algorithm has appeared in (Ghazi et al., 2021b; Impagliazzo et al., 2022). However, the dependence of the number of samples in the confidence parameter in these works is polynomial. We present a new variant of this algorithm that has polylogarithmic dependence on the confidence parameter. Moreover, using a stronger concentration inequality, we improve the dependence of the number of samples on the error parameter. We believe that this result could be of independent interest. We also design an agnostic learner for finite hypothesis classes. However, the dependence of the number of samples on $|\mathcal{H}|$ is polynomial. We believe that an interesting question is to design agnostic learners with polylogarithmic dependence on $|\mathcal{H}|$. We refer the reader to Appendix E.

Theorem 4.2 (Indistinguishability Amplification)**.** *Let $\mathcal{P}$ be a reference probability measure over $\{0,1\}^{\mathcal{X}}$ and $\mathcal{D}$ be a distribution over inputs. Consider the source of randomness $\mathcal{R}$ to be a Poisson point process with intensity $\mathcal{P} \times \mathrm{Leb} \times \mathrm{Leb}$, where Leb is the Lebesgue measure over $\mathbb{R}_+$. Consider a weak learning rule A that is (i) ρ-TV indistinguishable with respect to $\mathcal{D}$ for some $\rho \in (0,1)$, (ii) (α, β)-accurate for $\mathcal{D}$ for some $(\alpha, \beta) \in (0,1)^2$, such that $\beta < \frac{2\rho}{\rho+1} - 2\sqrt{\frac{2\rho}{\rho+1}} + 1$, and, (iii) absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$. Then, for any $\rho', \varepsilon, \beta' \in (0,1)^3$, there exists a learner $\textsc{Ampl}(A, \mathcal{R}, \beta', \varepsilon, \rho')$ that is ρ'-TV indistinguishable with respect to $\mathcal{D}$, and $(\alpha+\varepsilon, \beta')$-accurate for $\mathcal{D}$.*

We remark that the above result makes strong use of the equivalence between replicability and TV indistinguishability. Our algorithm is a variant of the amplification algorithm that appeared in (Impagliazzo et al., 2022), which (i) works for a wider range of parameters and (ii) its sample complexity is polylogarithmic in the parameter β'.

Accuracy Boosting. Next, we design an algorithm that boosts the accuracy of an n-sample ρ-TV indistinguishable algorithm and preserves its TV indistinguishability guarantee. Our algorithm is a variant of the boosting mechanism provided in (Impagliazzo et al., 2022). Similarly as in the case of amplification, our variant improves upon the dependence of the number of samples on the parameter β'.

Theorem 4.3 (Accuracy Boosting)**.** *Let $\mathcal{P}$ be a reference probability measure over $\{0,1\}^{\mathcal{X}}$ and $\mathcal{D}$ be a distribution over inputs. Consider the source of randomness $\mathcal{R}$ to be a Poisson point process with intensity $\mathcal{P} \times \mathrm{Leb} \times \mathrm{Leb}$, where Leb is the Lebesgue measure over $\mathbb{R}_+$. Consider a weak learning rule A that is (i) ρ-TV indistinguishable with respect to $\mathcal{D}$ for some $\rho \in (0,1)$, (ii) $(1/2-\gamma, \beta)$-accurate for $\mathcal{D}$ for some $(\gamma, \beta) \in (0,1)^2$, and, (iii) absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$. Then, for any $\beta', \varepsilon, \rho' \in (0,1)^3$, there exists a learner $\textsc{Boost}(A, \mathcal{R}, \varepsilon)$ that is ρ'-TV indistinguishable with respect to $\mathcal{D}$ and (ε, β')-accurate for $\mathcal{D}$.*

5. Acknowlegdements

Amin Karbasi acknowledges funding in direct support of this work from NSF (IIS-1845032), ONR (N00014- 19-1-2406), and the AI Institute for Learning-Enabled Optimization at Scale (TILOS). Shay Moran is a Robert J. Shillman Fellow; he acknowledges support by ISF grant 1225/20, by BSF grant 2018385, by an Azrieli Faculty Fellowship, by Israel PBC-VATAT, by the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by the the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Grigoris Velegkas is supported by NSF (IIS-1845032), an Onassis Foundation PhD Fellowship, and a Bodossaki Foundation PhD Fellowship. We thank Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell for illustrating discussions regarding the connection of this work with their recent paper (Bun et al., 2023).

References

Abernethy, J. D., Lee, C., McMillan, A., and Tewari, A. Online learning via differential privacy. 2017.

Ahn, K., Jain, P., Ji, Z., Kale, S., Netrapalli, P., and Shamir, G. I. Reproducibility in optimization: Theoretical framework and limits. *arXiv preprint arXiv:2202.04598*, 2022.

Alon, N., Livni, R., Malliaris, M., and Moran, S. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 852–860, 2019.

Alon, N., Bun, M., Livni, R., Malliaris, M., and Moran, S. Private and online learnability are equivalent. *ACM Journal of the ACM (JACM)*, 2022.

Amit, R., Epstein, B., Moran, S., and Meir, R. Integral probability metrics pac-bayes bounds. *arXiv preprint arXiv:2207.00614*, 2022.

Angel, O. and Spinka, Y. Pairwise optimal coupling of multiple random variables. *arXiv preprint arXiv:1903.00632*, 2019.

Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 2016.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bassily, R. and Freund, Y. Typicality-based stability and privacy. *arXiv preprint arXiv:1604.03336*, 2016.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059, 2016.

Bavarian, M., Ghazi, B., Haramaty, E., Kamath, P., Rivest, R. L., and Sudan, M. Optimality of correlated sampling strategies. *arXiv preprint arXiv:1612.01041*, 2016.

Ben-David, S. 2 notes on classes with vapnik-chervonenkis dimension 1. *arXiv preprint arXiv:1507.05307*, 2015.

Blum, A., Kalai, A., and Wasserman, H. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bousquet, O., Hanneke, S., Moran, S., Van Handel, R., and Yehudayoff, A. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 532–541, 2021.

Bousquet, O., Hanneke, S., Moran, S., Shafer, J., and Tolstikhin, I. Fine-grained distribution-dependent learning curves. *arXiv preprint arXiv:2208.14615*, 2022.

Broder, A. Z. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pp. 21–29. IEEE, 1997.

Bun, M., Nissim, K., and Stemmer, U. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 369–380, 2016.

Bun, M., Livni, R., and Moran, S. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 389–402. IEEE, 2020.

Bun, M., Gaboardi, M., Hopkins, M., Impagliazzo, R., Lei, R., Pitassi, T., Sivakumar, S., and Sorrell, J. Stability is stable: Connections between replicability, privacy, and adaptive generalization. *arXiv preprint arXiv:2303.12921*, 2023.

Charikar, M. S. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 380–388, 2002.

Chen, S., Koehler, F., Moitra, A., and Yau, M. Classification under misspecification: Halfspaces, generalized linear models, and evolvability. *Advances in Neural Information Processing Systems*, 33:8391–8403, 2020.

Cummings, R., Ligett, K., Nissim, K., Roth, A., and Wu, Z. S. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pp. 772–814. PMLR, 2016.

Dwork, C. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pp. 1–19. Springer, 2008.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 117–126, 2015.

Esfandiari, H., Kalavasis, A., Karbasi, A., Krause, A., Mirrokni, V., and Velegkas, G. Reproducible bandits. *arXiv preprint arXiv:2210.01898*, 2022.

Esfandiari, H., Karbasi, A., Mirrokni, V., Velegkas, G., and Zhou, F. Replicable clustering. *arXiv preprint arXiv:2302.10359*, 2023.

Fotakis, D., Kalavasis, A., Kontonis, V., and Tzamos, C. Efficient algorithms for learning from coarse labels. In *Conference on Learning Theory*, pp. 2060–2079. PMLR, 2021.

Ghazi, B., Golowich, N., Kumar, R., and Manurangsi, P. Sample-efficient proper pac learning with approximate differential privacy. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 183–196, 2021a.

Ghazi, B., Kumar, R., and Manurangsi, P. User-level private learning via correlated sampling. *arXiv preprint arXiv:2110.11208*, 2021b.

Goel, S., Gollakota, A., and Klivans, A. Statistical-query lower bounds via functional gradients. *Advances in Neural Information Processing Systems*, 33:2147–2158, 2020.

Gupta, A., Hardt, M., Roth, A., and Ullman, J. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 803–812, 2011.

Hanneke, S., Karbasi, A., Moran, S., and Velegkas, G. Universal rates for interactive learning. *Advances in Neural Information Processing Systems*, 35:28657–28669, 2022.

Holenstein, T. Parallel repetition: simplifications and the no-signaling case. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 411–419, 2007.

Impagliazzo, R., Lei, R., Pitassi, T., and Sorrell, J. Reproducibility in learning. *arXiv preprint arXiv:2201.08430*, 2022.

Kalavasis, A., Velegkas, G., and Karbasi, A. Multiclass learnability beyond the pac framework: Universal rates and partial concept classes. *arXiv preprint arXiv:2210.02297*, 2022.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Kearns, M. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Kleinberg, J. and Tardos, E. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM (JACM)*, 49(5):616–639, 2002.

Korolova, A., Kenthapadi, K., Mishra, N., and Ntoulas, A. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pp. 171–180, 2009.

Last, G. and Penrose, M. *Lectures on the Poisson process*, volume 7. Cambridge University Press, 2017.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

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

Livni, R. and Moran, S. A limitation of the pac-bayes framework. *Advances in Neural Information Processing Systems*, 33:20543–20553, 2020.

Malliaris, M. and Moran, S. The unstable formula theorem revisited. *arXiv preprint arXiv:2212.05050*, 2022.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Pineau, J., Sinha, K., Fried, G., Ke, R. N., and Larochelle, H. Iclr reproducibility challenge 2019. *ReScience C*, 5 (2):5, 2019.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *Journal of Machine Learning Research*, 22, 2021.

Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. General conditions for predictivity in learning theory. *Nature*, 428 (6981):419–422, 2004.

Raginsky, M., Rakhlin, A., Tsao, M., Wu, Y., and Xu, A. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pp. 26–30. IEEE, 2016.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Servedio, R. A. Smooth boosting and learning with malicious noise. *The Journal of Machine Learning Research*, 4:633–648, 2003.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Ullah, E., Mai, T., Rao, A., Rossi, R. A., and Arora, R. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142. PMLR, 2021.

Vaart, A. v. d. and Wellner, J. A. Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608, 1997.

Vadhan, S. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pp. 347–450. Springer, 2017.

Valiant, L. G. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Measures of complexity: festschrift for alexey chervonenkis*, pp. 11–30, 2015.

A. Preliminaries and Additional Definitions

A.1. Preliminaries

Probability Theory. We first review some standard definitions from probability theory.

Definition A.1 (Coupling)**.** A coupling of two probability distributions P and Q is a pair of random variables (X, Y), defined on the same probability space, such that the marginal distribution of X is P and the marginal distribution of Y is Q.

Definition A.2 (Integral Probability Metric)**.** The Integral Probability Metric (IPM) between two probability measures P and Q over $\mathcal{O}$ is defined as

$$d_{\mathcal{F},\mathcal{O}}(P,Q) = \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{O}} f dP - \int_{\mathcal{O}} f dQ \right| = \sup_{f\in\mathcal{F}} \left| \mathop{\mathbf{E}}_{x\sim P}[f(x)] - \mathop{\mathbf{E}}_{x\sim Q}[f(x)] \right| ,$$

where $\mathcal{F}$ is a set of real-valued bounded functions $\mathcal{O} \to \mathbb{R}$.

IPM distance measures are symmetric and non-negative. Note that the KL-divergence is not a special case of IPM, rather it belongs to the family of f-divergences, that intersect with IPM only at the TV distance. Such measures were recently used in order to derive PAC-Bayes style generalization bounds (Amit et al., 2022). The definition of an f-divergence will not be useful in this work and we refer the interested reader to e.g., (Sason & Verdú, 2016).

Learning Theory. We next review some standard definitions in statistical learning theory. We start with the definition of the Littlestone dimension (Littlestone, 1988).

Definition A.3 (Littlestone Dimension (Littlestone, 1988))**.** Consider a complete binary tree T of depth $d+1$ whose internal nodes are labeled by points in $\mathcal{X}$ and edges by $\{0,1\}$, when they connect the parent to the right, left child, respectively. We say that $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ Littlestone-shatters T if for every root-to-leaf path $x_1, y_1, \ldots, x_d, y_d, x_{d+1}$ there exists some $h \in \mathcal{H}$ such that $h(x_i) = y_i, 1 \le i \le d$. The Littlestone dimension is denoted by $\mathrm{Ldim}(\mathcal{H})$ is defined to be the largest d such that $\mathcal{H}$ Littlestone-shatters such a binary tree of depth $d+1$. If this happens for every $d \in \mathbb{N}$ we say that $\mathrm{Ldim}(\mathcal{H}) = \infty$.

We work under the well-known PAC learning model that was introduced in (Valiant, 1984). Let us denote the misclassification probability of a classifier h by $\mathrm{err}_{\mathcal{D}}(h) = \mathbf{Pr}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$. Also, we say that $\mathcal{D}$ is realizable with respect to $\mathcal{H}$ if there exists some $h^* \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(h^*) = 0$. Below, we slightly abuse notation and use the misclassification probability for distributions over classifiers.

Definition A.4 (PAC Learnability (Valiant, 1984; Shalev-Shwartz & Ben-David, 2014))**.** An algorithm A is n-sample (α,β)-accurate for a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ if, for any realizable distribution $\mathcal{D}$, it holds that $\mathbf{Pr}_{S\sim\mathcal{D}^n}[\mathrm{err}_{\mathcal{D}}(A(S)) > \alpha] \le \beta$. A hypothesis class $\mathcal{H}$ is PAC learnable if, for any $\alpha,\beta \in (0,1)^2$, there exist some $n_0(\alpha,\beta) \in \mathbb{N}$ and an algorithm A such that A is n-sample (α,β)-accurate for $\mathcal{H}$, for any $n \ge n_0(\alpha,\beta)$.

For the purposes of this work, an algorithm A should be thought of as a mapping from samples to a *distribution* over hypotheses. We want to design algorithms that satisfy two desiderata: they are PAC learners for some given hypothesis class $\mathcal{H}$ and they are total variation indistinguishable. In particular, we consider the following learning setting combining Definition 1.4 and A.4.

Definition A.5 (Realizable Learnability by TV Indistinguishable Learner)**.** An algorithm A is n-sample (α,β)-accurate ρ-TV indistinguishable for a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ if, for any realizable distribution $\mathcal{D}$, it holds that (i) A is n-sample ρ-TV indistinguishable and (ii) $\mathbf{Pr}_{S\sim\mathcal{D}^n}[\mathrm{err}_{\mathcal{D}}(A(S)) > \alpha] \le \beta$. A hypothesis class $\mathcal{H}$ is learnable by a TV indistinguishable algorithm if, for any $\alpha,\beta,\rho \in (0,1)$, there exist some $n_0(\alpha,\beta,\rho) \in \mathbb{N}$ and an algorithm A such that A is n-sample (α,β)-accurate ρ-TV indistinguishable for $\mathcal{H}$ for any $n \ge n_0(\alpha,\beta,\rho)$.

In the above definition, n depends on α,β,ρ (and $\mathcal{H}$), but *not* on the distribution.

Definition A.6 (Uniform Convergence Property)**.** We say that a domain $\mathcal{X}$ and a class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ satisfy the uniform convergence property if there exists a function $m^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that for any $\varepsilon,\delta \in (0,1)$, and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ it holds that if $S \sim \mathcal{D}^m$ and $m \ge m^{\mathrm{UC}}(\varepsilon,\delta)$, it holds that $\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \le \varepsilon$, with probability at least $1-\delta$, where L_S (resp. $L_{\mathcal{D}}$) is the empirical (resp. population) loss.

The fundamental theorem of learning theory (Vapnik & Chervonenkis, 2015; Blumer et al., 1989) states that the uniform convergence property is equivalent to the finiteness of the VC dimension of $\mathcal{H}$. However, one needs to make some (standard)

measurability assumptions on $\mathcal{X}, \mathcal{H}$ to rule out pathological cases. For instance, it is known that there classes with VC dimension 1 where uniform convergence does not hold (Ben-David, 2015)[8]. It is known that when $\mathcal{H}$ is countable and has finite VC dimension uniform convergence holds (Bartlett & Mendelson, 2002).

A.2. General Definition of Indistinguishability

While in the main body of the paper, we focused on binary classification, (most of) our proofs extend to general learning problems and so we first present a general abstract framework.

For general learning tasks, we can view learning rules (or algorithms) as randomized mappings $A : \mathcal{I} \to \Delta_{\mathcal{O}}$ which take as input instances from a domain $\mathcal{I}$ and map them to an element of the output space $\mathcal{O}$. We assume that there is a distribution μ on $\mathcal{I}$ that generates instances.

A second way to view the learning algorithm is via the mapping $A : \mathcal{I} \times \mathcal{R} \to \mathcal{O}$. Then A takes as input an instance $I \sim \mu$ and a random string $r \sim \mathcal{R}$ (we use $\mathcal{R}$ for both the probability space and the distribution) corresponding to the algorithm's *internal randomness* and outputs $A(I, r) \in \mathcal{O}$. Thus, $A(I)$ is a distribution over $\mathcal{O}$ whose randomness comes from the random variable r, while $A(I, r)$ is a deterministic object.

The space $\Delta_{\mathcal{O}}$ is endowed with some statistical dissimilarity measure.

Definition A.7 (Indistinguishability)**.** Let $\mathcal{I}$ be an input space, $\mathcal{O}$ be an output space and d be some statistical dissimilarity measure. A learning rule A satisfies ρ-indistinguishability with respect to d if for any distribution μ over $\mathcal{I}$ and two independent instances $I, I' \sim \mu$, it holds that

$$\mathop{\mathbf{E}}_{I,I' \sim \mu}[d\left(A(I), A(I')\right)] \leq \rho\,.$$

To illustrate the generality of our definition, we now show how we can instantiate $\mathcal{I}, \mathcal{O}, \mu, d$ to recover other definitions about stability of learning algorithms appearing in prior work.

Global Stability. Global stability (Bun et al., 2020) is a fundamental property of learning algorithms that was recently used to establish an equivalence between online learnability and approximate differential privacy in binary classification. We show how we can recover the definition of global stability. Let us first recall the definition.

Definition A.8 (Global Stability (Bun et al., 2020))**.** Let $\mathcal{R}$ be a distribution over random strings. A learning rule A is n-sample η-globally stable if for any distribution $\mathcal{D}$ there exists a hypothesis $h_{\mathcal{D}}$ such that

$$\mathop{\mathbf{Pr}}_{S \sim \mathcal{D}^n, r \sim \mathcal{R}}[A(S, r) = h_{\mathcal{D}}] \geq \eta\,.$$

In order to recover Definition A.8 using Definition A.7 we let $(S, r) \in \mathcal{I}, \mu = \mathcal{D}^n \times \mathcal{R}$ and $d(A(I, r), A(I', r')) = \mathbb{1}_{A(I,r) \neq A(I',r')}$. Thus, we have that

$$\mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n, r, r' \sim \mathcal{R}}[\mathbb{1}_{A(S,r) \neq A(S',r')}] \leq \rho \implies$$
$$\mathop{\mathbf{Pr}}_{S,S' \sim \mathcal{D}^n, r, r' \sim \mathcal{R}}[A(S, r) \neq A(S', r')] \leq \rho.$$

Notice that this gives us a two-sided version of the definition of global-stability. So far we have established that $\mathbf{Pr}_{S,S' \sim \mu, r, r' \sim \mathcal{R}}[A(S, r) = A(S', r')] \geq 1 - \rho > 0$. Since two independent draws of the random variable $A(S, r)$ are the same with non-zero probability it means that it must have point masses. Moreover, there are countably many such

[8] We note that the proof of the existence of such a class holds under the continuum hypothesis.

point masses. Let $\mathcal{H}_m = \{h \in \mathcal{H} : \mathbf{Pr}_{S\sim\mathcal{D}^n, r\sim\mathcal{R}}[A(S,r) = h]\}$. Then,

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{S,S'\sim\mu, r,r'\sim\mathcal{R}}[A(S,r) = A(S',r')] &= \sum_{h\in\mathcal{H}_m} \left(\mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n, r\sim\mathcal{R}}[A(S,r) = h]\right)^2 \\
&\leq \max_{h\in\mathcal{H}_m} \mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n, r\sim\mathcal{R}}[A(S,r) = h] \cdot \sum_{h\in\mathcal{H}_m} \mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n, r\sim\mathcal{R}}[A(S,r) = h] \\
&\leq \max_{h\in\mathcal{H}_m} \mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n, r\sim\mathcal{R}}[A(S,r) = h] \\
&= \max_{h\in\mathcal{H}} \mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n, r\sim\mathcal{R}}[A(S,r) = h]
\end{aligned}$$

Thus, by chaining the two inequalities we have established, we get that $\max_{h\in\mathcal{H}_m} \mathbf{Pr}_{S\sim\mathcal{D}^n, r\sim\mathcal{R}}[A(S,r) = h] \geq 1 - \rho$, so the algorithm A satisfies the notion of global stability.

A.3. Alternative Definitions of TV Indistinguishability

We now discuss alternative ways to define TV indistinguishability.

A.3.1. TV Indistinguishability with Fixed Prior

First, observe that the definition we propose is two-sided in the sense that we require drawing two sets of i.i.d. samples. A different way to view TV indistinguishability is by requiring that the output of the algorithm is close, in TV distance, to some *prior distribution*, which depends on the data-generating process $\mathcal{D}$ but is independent of the sample. Notice that we could introduce a similar one-sided general definition as a second viewpoint of Definition 1.1 (named Indistinguishability with Fixed Prior).

Definition A.9 (TV Indistinguishability with Fixed Prior)**.** A learning rule A is n-sample ρ-fixed prior TV indistinguishable if for any distribution over inputs $\mathcal{D}$, there exists some prior $\mathcal{P}_\mathcal{D}$ such that for $S \sim \mathcal{D}^n$ it holds that

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S), \mathcal{P}_\mathcal{D})] \leq \rho\,.$$

Notice that, using the triangle inequality, we can see that this definition is equivalent to Definition 1.4, up to a factor of 2. Formally, we have the following result.

Lemma A.10. *If A is ρ-*TV *indistinguishable then it is ρ-fixed prior* TV *indistinguishable. Conversely, if A is ρ-fixed prior* TV *indistinguishable then it is 2ρ-*TV *indistinguishable.*

We remark that if A is TV indistinguishable with respect to a distribution over inputs $\mathcal{D}$, one can show that it is also fixed prior TV indistinguishable with respect to $\mathcal{D}$ where the fixed prior is equal to $\mathcal{P}_\mathcal{D} = \int_S A(S)d(\mathcal{D}^n)$.

Proof. For the first direction, we let $\mathcal{P}_{S,S'}$ be a distribution with the property that $d_{\mathrm{TV}}(A(S), \mathcal{P}_{S,S'}) = d_{\mathrm{TV}}(A(S'), \mathcal{P}_{S,S'}) = d_{\mathrm{TV}}(A(S), A(S'))/2$, e.g., $\mathcal{P}_{S,S'} = 1/2 \cdot (A(S) + A(S'))$, for every $S, S' \sim \mathcal{D}^n$. We now define $\mathcal{P}_\mathcal{D}$ to be the average of $\mathcal{P}_{S,S'}$ with respect to the measure of the product distribution of S, S'. We have that

$$\begin{aligned}
\mathcal{P}_\mathcal{D} &= \int_{S,S'} \mathcal{D}^n(S)\mathcal{D}^n(S')\frac{A(S) + A(S')}{2} dS dS' \\
&= \int_T \left(\mathcal{D}^n(T)1\{S = T\}\frac{A(T)}{2}\left(\int_{S'} \mathcal{D}^n(S')\right) + \mathcal{D}^n(T)1\{S' = T\}\frac{A(T)}{2}\left(\int_S \mathcal{D}^n(S)\right)\right) dS dS' = \\
&= \int_T \mathcal{D}^n(T)A(T)dT\,.
\end{aligned}$$

This means that $\mathbf{E}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S), \mathcal{P}_\mathcal{D})] = \int_S \mathcal{D}^n(S) d_{\mathrm{TV}}\left(A(S), \int_T \mathcal{D}^n(T)A(T)dT\right) dS \leq \rho$.

For the converse, notice that

$$\begin{aligned}\mathop{\mathbf{E}}_{S,S'\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),A(S'))] &\le \mathop{\mathbf{E}}_{S,S'\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})+d_{\mathrm{TV}}(A(S'),\mathcal{P}_{\mathcal{D}})]\\ &= \mathop{\mathbf{E}}_{S,S'\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})]+\mathop{\mathbf{E}}_{S,S'\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S'),\mathcal{P}_{\mathcal{D}})]\\ &= 2\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})]\\ &\le 2\rho.\end{aligned}$$

□

A.3.2. With High Probability TV Indistinguishability

A different direction in which we can extend the definition of total variation indistinguishability has to do with replacing the expectation with a high-probability style of bound. We remark that (Impagliazzo et al., 2022) provide a similar alternative definition in the context of their work.

Definition A.11 (High-Probability TV Indistinguishability)**.** A learning rule A is n-sample high-probability (η,ν)-TV indistinguishable if for any distribution $\mathcal{D}$ there exists some prior $\mathcal{P}_{\mathcal{D}}$ such that

$$\mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})\le\eta]\ge 1-\nu\,.$$

Notice that in the above definition we have used the fixed prior version of TV indistinguishability to reduce the number of parameters, but it can also be stated in its the two-sided version. It is not hard to see that the "in expectation" and the "with high probability" versions of the definition are qualitatively equivalent. Moreover, we can establish a quantitative connection as follows.

Lemma A.12. *If a learning rule A is an n-sample ρ-fixed prior* TV *indistinguishable learner (cf. Definition A.9) then it is an n-sample high-probability $(\rho/\nu,\nu)$-*TV *indistinguishable learning rule (cf. Definition A.11), for any $\rho\le\nu<1$. Conversely, if a learnigng rule A is an n-sample high-probability (η,ν)-*TV *indistinguishable learner then it is an n-sample $(\eta+\nu-\eta\cdot\nu)$-fixed prior* TV *indistinguishable learning rule.*

Proof. The proof of the first part of claim is a direct consequence of Markov's inequality. Notice that $d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})$ is random variable whose expected value is bounded by ρ. Thus, we have that

$$\mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})\ge\rho/\nu]\le\nu\,.$$

Hence, we can see that A is a high-probability $(\rho/\nu,\nu)$-TV indistinguishable learning rule.

We now move to the second part of the claim. Let $\mathcal{E}$ be the event that $d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})\ge\eta$. Then, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})] &= \mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})|\mathcal{E}]\,\mathbf{Pr}[\mathcal{E}]+\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}[d_{\mathrm{TV}}(A(S),\mathcal{P}_{\mathcal{D}})|\mathcal{E}^c]\,\mathbf{Pr}[\mathcal{E}^c]\\ &\le 1\cdot\nu+\eta\cdot(1-\nu)\\ &= \eta+\nu-\eta\cdot\nu.\end{aligned}$$

□

A.4. Coupling and Correlated Sampling

Coupling is a fundamental notion in probability theory with many applications (Levin & Peres, 2017). The correlated sampling problem, which has applications in various domains, e.g., in sketching and approximation algorithms (Broder, 1997; Charikar, 2002), is described in (Bavarian et al., 2016) as follows: Alice and Bob are given probability distributions P and Q, respectively, over a finite set Ω. *Without any communication, using only shared randomness* as the means to coordinate, Alice is required to output an element x distributed according to P and Bob is required to output an element y distributed according to Q. Their goal is to minimize the disagreement probability $\mathbf{Pr}[x\neq y]$, which is comparable with $d_{\mathrm{TV}}(P,Q)$. Formally, a correlated sampling strategy for a finite set Ω with error $\varepsilon:[0,1]\to[0,1]$ is specified

by a probability space $\mathcal{R}$ and a pair of functions $f, g : \Delta_\Omega \times \mathcal{R} \to \Omega$, which are measurable in their second argument, such that for any pair $P, Q \in \Delta_\Omega$ with $d_{\mathrm{TV}}(P, Q) \leq \delta$, it holds that (i) the push-forward measure $\{f(P, r)\}_{r\sim\mathcal{R}}$ (resp. $\{g(Q, r)\}_{r\sim\mathcal{R}}$) is P (resp. Q) and (ii) $\mathbf{Pr}_{r\sim\mathcal{R}}[f(P, r) \neq g(Q, r)] \leq \varepsilon(\delta)$. We underline that a correlated sampling strategy is *not* the same as a coupling, in the sense that the latter requires a single function $h : \Delta_\Omega \times \Delta_\Omega \to \Delta_{\Omega\times\Omega}$ such that for any P, Q, the marginals of $h(P, Q)$ are P and Q respectively. It is known that for any coupling function h, it holds that $\mathbf{Pr}_{(x,y)\sim h(P,Q)}[x \neq y] \geq d_{\mathrm{TV}}(P, Q)$ and that this bound is attainable. Since $\{(f(P, r), g(Q, r))\}_{r\sim\mathcal{R}}$ induces a coupling, it holds that $\varepsilon(\delta) \geq \delta$ and, perhaps surprisingly, there exists a strategy with $\varepsilon(\delta) \leq \frac{2\delta}{1+\delta}$ (Broder, 1997; Kleinberg & Tardos, 2002; Holenstein, 2007) and this result is tight (Bavarian et al., 2016). A second difference between coupling and correlated sampling has to do with the size of Ω: while correlated sampling strategies can be extended to infinite spaces Ω, it remains open whether there exists a correlated sampling strategy for general measure spaces $(\Omega, \mathcal{F}, \mu)$ with any non-trivial error bound (Bavarian et al., 2016). On the other hand, coupling applies to spaces Ω of any size.

(Ghazi et al., 2021b) studied user-level privacy and introduced the notion of pseudo-global stability, which is essentially the same as replicability as observed by (Impagliazzo et al., 2022). (Ghazi et al., 2021b) showed that pseudo-global stability is qualitatively equivalent to approximate differential privacy. Their main technique was the use of correlated sampling that allowed users to output the same learned hypothesis (stability) employing shared randomness. We mention that (Ghazi et al., 2021b) provide their results for finite outcome space (i.e., $\mathcal{X}$ is finite and thus $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ is too). In particular, they need finiteness of the domain in order to apply correlated sampling which is used during their "DP implies pseudo-global stability" reduction. They mention that their results can be extended to the case where $\mathcal{X}$ is infinite and that this does require non-trivial generalization of tools such as correlated sampling and some measure-theoretic details to that setting[9]; we refer to a discussion in Section 5.3 of (Bavarian et al., 2016) about the assumptions needed in order to achieve correlated sampling in infinite spaces. Similarly, the last step of the constructive transformation of a DP algorithm to a replicable one provided in (Bun et al., 2023) uses correlated sampling and is hence also given for finite domains. For further comparisons between our coupling and the correlated sampling problem of (Bavarian et al., 2016), we refer to the discussion in (Angel & Spinka, 2019) after Corollary 4.

A very useful tool for our derivations is a coupling protocol that can be found in (Angel & Spinka, 2019).

Theorem A.13 (Pairwise Optimal Coupling (Angel & Spinka, 2019))**.** *Let $\mathcal{S}$ be any collection of random variables that are absolutely continuous with respect to a common probability measure[10] μ. Then, there exists a coupling of the variables in $\mathcal{S}$ such that, for any $X, Y \in \mathcal{S}$,*

$$\mathbf{Pr}[X \neq Y] \leq \frac{2d_{\mathrm{TV}}(X, Y)}{1 + d_{\mathrm{TV}}(X, Y)}\,.$$

Moreover, this coupling requires sample access to a Poisson point process with intensity $\mu \times \mathrm{Leb} \times \mathrm{Leb}$, where Leb is the Lebesgue measure over $\mathbb{R}_+$, and full access to the densities of all the random variables in $\mathcal{S}$ with respect to μ.

An intuitive illustration of how it works can be found in Figure 1.

A.5. Discussion on Definition 1.4

We discuss more extensively the TV Indistinguishability definition. One important motivation for the definition of TV indistinguishability is to show that replicability can be equivalently defined using the same high-level template like the well-studied PAC-Bayes framework, where one shows that the outputs of the algorithms are close, under the KL divergence, with some data-independent priors. In other words, our results show how to organize and view different well-studied notions of stability using the same template.

Moreover, an interpretation of the replicability definition is that two executions of the algorithm over independent datasets should be coupled using just shared internal randomness. However, this is one of potentially infinite ways to couple the two executions. Our definition, which we find quite natural, captures exactly this observation and allows for general couplings between two random runs. It is also worth noting that, to the best of our knowledge, all the notions of algorithmic stability that have been proposed in the past do not depend on the source of internal randomness of the algorithm. However, this is not the case with replicability.

[9] To be more specific, the proof of Theorem 20 in (Ghazi et al., 2021b) requires to define the correlated sampling strategy over the space $2^{\mathcal{X}}$ a priori (independently of the observed samples and input algorithm). Hence while the strategy is applied to distributions with finite support, an extension to infinite domain in that proof would require some modifications.

[10] This result extends to the setting where μ is a σ-finite measure, but it is not needed for the purposes of our work.

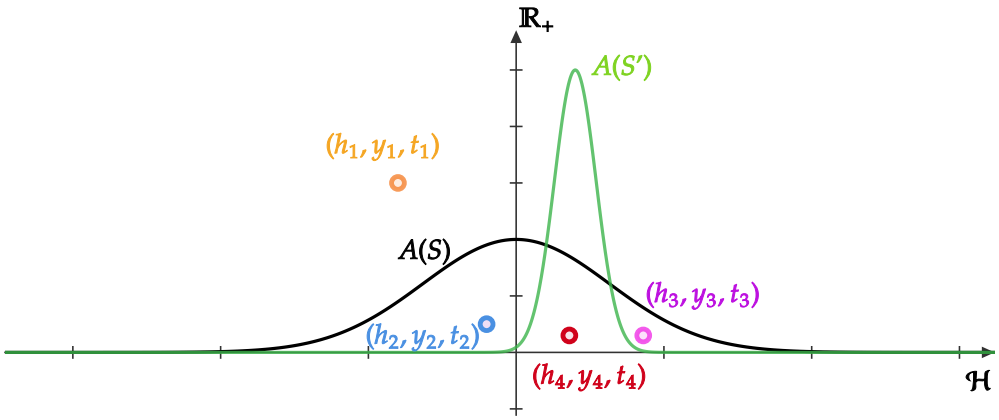

Figure 1. Our goal is to couple $A(S)$ with $A(S')$, where these two distributions are absolutely continuous with respect to the reference probability measure $\mathcal{P}$. A sequence of points of the form (h, y, t) is generated by the Poisson point process with intensity $\mathcal{P} \times \text{Leb} \times \text{Leb}$ where $h \sim \mathcal{P}$, $(y, t) \in \mathbb{R}_+^2$ and Leb is the Lebesgue measure over $\mathbb{R}_+$ (note that we do not have upper bounds for the densities). Intuitively, h lies in the hypotheses' space, y is a non-negative real value and t corresponds to a time value. Let f be the Radon-Nikodym derivate of $A(S)$ with respect to $\mathcal{P}$. We assign the first (the one with minimum t) value h to $A(S)$ that satisfies the property that $f(h) > y$, i.e., y falls below the density curve of $A(S)$. We assign a hypothesis to $A(S')$ in a similar manner. This procedure defines a data-independent way to couple the two random variables and naturally extends to multiple ones. In the figure's example, we set $A(S) = h_2$ and $A(S') = h_4$ given that $t_1 < t_2 < t_3 < t_4$.

Let us now present a concrete algorithm whose stability property is easier to prove under the new definition. (Ghazi et al., 2021b) presented a procedure that transforms a list-globally stable algorithm to a replicable one (Algorithm 1, page 9 in (Ghazi et al., 2021b)). Crucially, in the last step of this algorithm the authors use a correlated sampling procedure to prove the replicability property. This procedure induces a computational overhead to the overall algorithm, and it is not clear even if it is computable beyond finite domains. On the other hand, the TV indistinguishability property is immediate. Thus, the transformation from list-global stability to TV indistinguishability is computationally efficient and holds for general domains whereas the transformation from list-global stability to replicability is not.

To the best of our knowledge, most of the replicable algorithms that have been developed use their internal randomness over data-independent distributions. To make this point more clear let us consider the replicable SQ oracle of (Impagliazzo et al., 2022). In this work, the authors use randomness over distributions that are independent of the input sample S. Thus, no matter how the internal randomness is implemented, when one shares it across two executions the internal random choices of the algorithm are the same.

However, there are algorithms, like Algorithm 1 in (Ghazi et al., 2021b), that use internal randomness over a data-dependent distribution. If the algorithm makes random choices over data-dependent quantities like in (Ghazi et al., 2021b), when one shares the randomness across two executions the internal random choices are not necessarily the same even if the TV distance between the two distributions is small, unless one specifies carefully the source of internal randomness (i.e., using some coupling). This can lead to significant computational overhead when the domain is finite, computability issues when the domain is countable, and for general domains it is not clear yet that going from TV indistinguishability to replicability is possible. Hence, one advantage of TV indistinguishability is that it provides a relaxation over the stronger definition of replicability, which is the notion that our definition builds upon.

B. Useful Replicable Subroutines

In this section we present various replicable subroutines that will be useful in the derivation of our results.

B.1. Replicability Preliminaries

Recall the Statistical Query (SQ) model that was introduced by (Kearns, 1998) and is a restriction of the PAC learning model, appearing in various learning theory contexts (Blum et al., 2003; Gupta et al., 2011; Chen et al., 2020; Goel et al., 2020; Fotakis et al., 2021). In the SQ model, the learner interacts with an oracle in the following way: the learner submits a statistical query to the oracle and the oracle returns its expected value, after adding some noise to it. More formally, we have the following definition.

Definition B.1 ((Kearns, 1998))**.** Let $\tau, \delta \in (0,1)^2, \mathcal{D}$ be a distribution over the domain $\mathcal{X}$ and $\phi : \mathcal{X} \to [0,1]$ be a query. Let S be an i.i.d. sample of size $n = n(\tau, \delta)$. Then, the statistical query oracle outputs a value v such that $|v - \mathbf{E}_{x\sim\mathcal{D}}[\phi(x)]| \leq \tau$, with probability at least $1 - \delta$.

Essentially, using a large enough number of samples, the SQ oracle returns an approximation of the expected value of a statistical query whose range is bounded. (Impagliazzo et al., 2022) provide a replicable implementation of an SQ oracle with a mild blow-up in the sample complexity.

Theorem B.2 (Replicable SQ Learner (Impagliazzo et al., 2022))**.** *Let $\tau, \delta, \rho \in (0,1)^3, \delta \leq \rho/3, \mathcal{D}$ be a distribution over some domain $\mathcal{X}$, and $\phi : \mathcal{X} \to [0,1]$ be a query. Let S be an i.i.d. sample of size*

$$n = O\left(\frac{1}{\tau^2\rho^2}\log(1/\delta)\right).$$

Then there exists a ρ-replicable SQ oracle for ϕ.

The interpretation of the previous theorem is that we can estimate replicably statistical queries whose range is bounded.

The following result that was proved in (Impagliazzo et al., 2022) is useful for our derivations.

Claim B.3 (ρ-Replicability $\implies$ (η,ν)-Replicability (Impagliazzo et al., 2022))**.** Let A be a ρ-replicable algorithm and $\mathcal{R}$ be its source of randomness. Then for any $\nu \in [\rho, 1)$, it holds that

$$\Pr_{r\sim\mathcal{R}}\left[\left\{\exists h \in \mathcal{H} : \Pr_{S\sim\mathcal{D}^n}[A(S,r) = h] \geq 1 - \frac{\rho}{\nu}\right\}\right] \geq 1 - \nu\,.$$

Notice that in the definition of replicability (Definition 1.2), the learner shares all the internal random bits across its two executions. A natural extension is to consider learners that share only *part* of their random bits, i.e., they have access to private random bits that are not shared across its executions and public random bits that are shared. A result in (Impagliazzo et al., 2022) shows that these learners are, essentially, equivalent to the ones that use only private bits. To be more precise, we say that a learner A is ρ-replicable with respect to r_{pub} if

$$\Pr_{S,S'\sim\mathcal{D}^n, r_{priv}, r'_{priv}, r_{pub}\sim\mathcal{R}}[A(S, r_{priv}, r_{pub}) = A(S', r'_{priv}, r_{pub})] \geq 1 - \rho\,.$$

The following result states this property formally.

Lemma B.4 (Public, Private Replicability $\implies$ Replicability (Impagliazzo et al., 2022))**.** *Let A be an n-sample ρ-replicable learner with respect to r_{pub}. Then, A is a n-sample ρ-replicable learner with respect to (r_{pub}, r_{priv}).*

This result allows us to think of a replicable learner as having access to two different sources of randomness, one that is private to its execution and one that is shared across the executions. We will make use of it in transformations from DP learners to replicable learners and some boosting results.

B.2. Replicable Heavy-Hitters

In the analysis of the replicable heavy-hitter algorithm (cf. Algorithm 1) we will use the Bretagnolle-Huber-Carol inequality that bounds the estimation error of the parameters of a multinomial distribution from samples.

Lemma B.5 (Bretagnolle-Huber-Carol Inequality (Vaart & Wellner, 1997))**.** *Let $p = (p_1, \ldots, p_k)$ multinomial distribution supported on k elements. Then, given access to n i.i.d. samples from p we have that*

$$\Pr\left[\sum_{i=1}^{k} |\hat{p}_i - p_i| \geq \varepsilon\right] \leq 2^k e^{-n\varepsilon^2/2}\,,$$

for every $\varepsilon \in (0,1)$, where $\hat{p}_i$ is the empirical frequency of item i in the sample S.

The replicable heavy-hitters algorithm is depicted in Algorithm 1. As we alluded before, this approach is very similar to (Ghazi et al., 2021b; Impagliazzo et al., 2022). However, in our approach we treat the confidence parameter and the reproducibility parameters differently. Moreover, since we make use of Lemma B.5, we are able to reduce the sample complexity of the algorithm.

Algorithm 1 Replicable Heavy-Hitters

Input: Sample access to a distribution $\mathcal{D}$ over some domain $\mathcal{X}$
Parameters: Threshold v, error ε, confidence δ, replicability ρ
Output: List of elements L in $\mathcal{X}$
$n_1 \leftarrow \frac{\log(2/(\min\{\delta,\rho\}(v-\varepsilon)))}{v-\varepsilon}$
$S_1 \leftarrow n_1$ i.i.d. samples. from $\mathcal{D}$
$\mathcal{X}_h \leftarrow$ unique elements of S_1 {Notice that $|\mathcal{X}_h| \leq n_1$.}
$n_2 \leftarrow \frac{32(\ln(2/\min\{\delta,\rho\})+|\mathcal{X}|+1)}{\rho^2\varepsilon^2}$
$S_2 \leftarrow n_2$ i.i.d. samples from $\mathcal{D}$
$\hat{p}_x \leftarrow \mathrm{freq}_S(x), \forall x \in \mathcal{X}_h$ {$\hat{p}_x$ is the empirical frequency of every potential heavy hitter}
$v' \leftarrow U[v-\varepsilon/2, v+\varepsilon/2]$ {Set the threshold for acceptance of a heavy-hitter.}
$L \leftarrow \{x \in \mathcal{X}_h : \hat{p}_x \geq v'\}$ {Drop the elements of $\mathcal{X}_h$ that fall below the threshold.}
Output L

Lemma B.6. *Let $\mathcal{D}$ be distribution supported on some domain $\mathcal{X}$ and denote by $\mathcal{D}(x)$ the mass that it puts on $x \in \mathcal{X}$. For any $\varepsilon, \delta, \rho, v \in (0,1)^4$ such that $(v-\varepsilon, v+\varepsilon) \subseteq (0,1)$, Algorithm 1 is ρ-replicable and outputs a list L such that, with probability $1-\delta$, for all $x \in \mathcal{X}$:*

- *If $\mathcal{D}(x) < v-\varepsilon$ then $x \notin L$.*
- *If $\mathcal{D}(x) > v+\varepsilon$ then $x \in L$.*

Its sample complexity is at most $O\left(\frac{\log(1/(\min\{\delta,\rho\}(v-\varepsilon)))}{(v-\varepsilon)\rho^2\varepsilon^2}\right)$.

Proof. We first prove the correctness of the algorithm with the desired accuracy ε and confidence δ. For simplicity, let us assume that $\delta \leq \rho/4$. Otherwise, we can simply set $\delta = \rho/4$. After we pick n_1 points, the probability that a $(v-\varepsilon)$-heavy-hitter of the distribution is not included in S_1 is at most

$$(1-(v-\varepsilon))^{n_1} \leq e^{-(v-\varepsilon)\cdot n_1} \leq \frac{\delta\cdot(v-\varepsilon)}{2}.$$

Since there are at most $1/(v-\varepsilon)$ such heavy-hitters, we can see that with probability at least $\delta/2$ all of the are included in S_1. Let us call this event $\mathcal{E}_1$ and condition on it for the rest of the proof.

Let us consider a distribution $\widehat{\mathcal{D}}$ that puts the same mass on every element of $\mathcal{X}_h$ as $\mathcal{D}$ and the remaining mass on a new special element e. We can sample from $\widehat{\mathcal{D}}$ in the following way: we draw a sample from $\mathcal{D}$ and if it falls in $\mathcal{X}_h$ we return it, otherwise we return e. Thus, we can see that if we draw n samples from $\widehat{\mathcal{D}}$, they are distributed according to a multinomial distribution supported on $\mathcal{X}_h \cup \{e\}$. Thus, Lemma B.5 applies to this setting which means that if we draw n_2 i.i.d. samples from $\widehat{\mathcal{D}}$ we have that

$$\mathbf{Pr}\left[\sum_{i=1}^{k}|\hat{p}_i - p_i| \geq \frac{\varepsilon\rho}{4}\right] \leq 2^k e^{-n_2\varepsilon^2\rho^2/32} \leq e^k e^{-n_2\varepsilon^2\rho^2/32} = e^{k-n_2\varepsilon^2\rho^2/32},$$

where $k = |\mathcal{X}_h|+1$. Thus, $e^{k-n_2\varepsilon^2\rho^2/32} = e^{-\ln(2/\delta)} \leq \frac{\delta}{2}$. We call this event $\mathcal{E}_2$ and condition on it for the rest of the proof. Notice that under this event we have that $|\hat{p}_x - p_x| \leq \frac{\varepsilon\rho}{4} < \frac{\varepsilon}{2}, \forall x \in \mathcal{X}_h$. Since $v' \geq v-\varepsilon/2$ it means that if $\hat{p}_x \geq v' \geq v-\varepsilon/2 \implies p_x + \varepsilon/2 > v-\varepsilon/2 \implies p_x > v-\varepsilon$. Similarly, we get that if $\hat{p}_x < v' \implies p_x < v+\varepsilon$. Hence, we see that the algorithm is correct with probability at least $1-\delta/2-\delta/2 = 1-\delta$. This concludes the correctness proof.

We now focus on the replicability of the algorithm. Let $\mathcal{X}_h^1$ be the unique elements at Algorithm 1 of the algorithm in the first run and $\mathcal{X}_h^2$ in the second run. Notice that if $x \in (\mathcal{X}_h^1 \setminus \mathcal{X}_h^2) \cup (\mathcal{X}_h^2 \setminus \mathcal{X}_h^1)$ then, with probability at least $1-\delta/2-\delta/2 = 1-\delta$,

the element x is not a $(v-\varepsilon)$-heavy-hitter, so, with probability at least $1-\delta/2$, it will not be included in the output of the execution that it appears in. Let $E = \mathcal{X}_h^1 \cap \mathcal{X}_h^2$ and denote by L_1, L_2, the outputs of the first, second execution, respectively. We need to bound the probability of the event $\mathcal{E} = \{\exists x \in E : x \in L_1 \setminus L_2 \cup L_2 \setminus L_1\}$. Let $\hat{p}_x^1, \hat{p}_x^2$ the empirical frequencies of x in the first, second execution, respectively. Due to the concentration inequality we have used, we have that

$$\sum_{x \in \mathcal{X}_1 \cap \mathcal{X}_2} |\hat{p}_x^i - p_x| \leq \frac{\varepsilon\rho}{4}, i \in \{1,2\}\,,$$

with probability at least $1-\delta$. Under this event, using the triangle inequality, this means that

$$\sum_{x \in \mathcal{X}_1 \cap \mathcal{X}_2} |\hat{p}_x^1 - \hat{p}_x^2| \leq \frac{\varepsilon\rho}{2}, i \in \{1,2\}\,,$$

Notice that since pick a number uniformly at random from an interval with range ε, for some given $x \in \mathcal{X}_1 \cap \mathcal{X}_2$, we have that $\mathbf{Pr}[x \in L_1 \setminus L_2 \cup L_2 \setminus L_1] \leq |\hat{p}_x^1 - \hat{p}_x^2|/\varepsilon$. Thus, taking a union bound over $x \in \mathcal{X}_1 \cap \mathcal{X}_2$, we see that

$$\mathbf{Pr}[\mathcal{E}] \leq \frac{\sum_{x \in \mathcal{X}_1 \cap \mathcal{X}_2} |\hat{p}_x^1 - \hat{p}_x^2|}{2\varepsilon} \leq \frac{\varepsilon\rho}{2\varepsilon} = \frac{\rho}{2}\,.$$

Putting everything together, we see that the probability that the two outputs of the algorithm differ is at most $\delta + \delta/2 + \rho/2 < \rho$. □

B.3. Replicable Agnostic PAC Learner for Finite $\mathcal{H}$

In this section we present a replicable agnostic PAC learner for finite hypothesis classes, i.e., a learner whose output is a hypothesis that has error rate close to the best one in the class. Our construction relies on the replicable SQ oracle from (Impagliazzo et al., 2022) (see Theorem B.2). The idea is simple: since the error rate of every $h \in \mathcal{H}$ can be replicably estimated using Theorem B.2, we do that for every $h \in \mathcal{H}$ and then we return the one that has the smallest estimated value.

Algorithm 2 Replicable Agnostic Learner for Finite $\mathcal{H}$

```
Input: Hypothesis class H, sample access to a distribution D over X × {0,1}
Parameters: accuracy ε, confidence δ, replicability ρ
Output: Classifier h that is ε-close to the best one in H and its estimated
error on D
```

$\hat{\alpha}_h \leftarrow \text{ReprErrorEst}(\varepsilon/2, \delta/|\mathcal{H}|, \rho/|\mathcal{H}|), \forall h \in \mathcal{H}$ {Theorem B.2.}
$\hat{h}^* \leftarrow \arg\min_{h \in \mathcal{H}} \hat{a}_h$ {Break ties arbitrarily in a consistent manner.}
Output $(\hat{h}^*, \hat{\alpha}_{\hat{h}^*})$

It is not hard to see that Algorithm 2 is ρ-replicable and returns a hypothesis whose error is ε-close to the best one.

Claim B.7. Let $\mathcal{H}$ be a finite hypothesis class and $\varepsilon, \delta, \rho \in (0,1)^3$. Given $O\left(\frac{|\mathcal{H}|^3}{\varepsilon^2\rho^2}\log\left(\frac{|\mathcal{H}|}{\delta}\right)\right)$ i.i.d. samples from $\mathcal{D}$, Algorithm 2 is ρ-replicable and returns a classifier $\hat{h}^*$ with $\text{err}(\hat{h}^*) < \min_{h \in \mathcal{H}} \text{err}(h) + \varepsilon$, with probability at least $1-\delta$.

Proof. The replicability of the algorithm follows from the fact that we estimate each $\hat{a}_h$ replicably with parameter $\rho/|\mathcal{H}|$ and we make $|\mathcal{H}|$ such calls.

Notice that for each call to the replicable error estimator we need $n_h = O\left(\frac{|\mathcal{H}|^2}{\varepsilon^2\rho^2}\log\left(\frac{|\mathcal{H}|}{\delta}\right)\right)$ samples and we make $|\mathcal{H}|$ such calls.

Since the accuracy parameter of the statistical query oracle is $\varepsilon/2$, using the triangle inequality, we have that $|\hat{a}_{\hat{h}^*} - \min_{h \in \mathcal{H}} a_h| \leq \varepsilon$.

Finally, the correctness of the algorithm follows from a union bound over the correctness of every call to the oracle. □

C. TV Indistinguishability and Replicability

In this section, we will study the connection between TV indistinguishability and replicability.

C.1. The Proof of Theorem 2.7

We are now ready to establish the connection between TV indistinguishability and replicability. The upcoming result is particularly useful because it provides a *data-independent* way to couple the random variables.

Proof of Theorem 2.7. Let $\mathcal{R}$ be Poisson point process with intensity $\mathcal{P} \times \mathrm{Leb} \times \mathrm{Leb}$, where Leb is the Lebesgue measure over $\mathbb{R}_+$ (cf. Theorem A.13, Figure 1). The learning rule A' is defined in the following way. For every $S \in (\{\mathcal{X} \times \{0,1\})^n$, let $r = \{(h_i, y_i, t_i)\}_{i \in \mathbb{N}}$ be an infinite sequence of the Poisson point process $\mathcal{R}$ and let $j = \arg\min_{i \in \mathbb{N}} \{t_i : f_S(h_i) > y_i\}$. The output of A' is h_j and we denote it by $A'(S, r)$. We will shortly explain why this is well-defined, except for a measure zero event. The fact that A' is equivalent to A follows from the coupling guarantees of this process (cf. Theorem A.13). In particular, we can instantiate this result with the single random variable $\{A(S)\}$. We can now observe that, except for a measure zero event, (i) since A is absolutely continuous with respect to $\mathcal{P}$, there exists such a density f_S, (ii) the set over which we are taking the minimum is not empty, (iii) the minimum is attained at a unique point. This means that A' is well-defined, except for a measure zero event[11], and, by the correctness of the rejection sampling process (Angel & Spinka, 2019), $A'(S)$ has the desired probability distribution.

We now prove that A' is replicable. Since A is ρ-TV indistinguishable, it follows that

$$\mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n} [d_{\mathrm{TV}}(A(S), A(S'))] \leq \rho.$$

We have shown that A' is equivalent to A, so we can see that $\mathbf{E}_{S,S' \sim \mathcal{D}^n}[d_{\mathrm{TV}}(A'(S), A'(S'))] \leq \rho$. Thus, using the guarantees of Theorem A.13, we have that for any datasets S, S'

$$\mathop{\mathbf{Pr}}_{r \sim \mathcal{R}}[A'(S, r) \neq A'(S', r)] \leq \frac{2 d_{\mathrm{TV}}(A'(S), A'(S'))}{1 + d_{\mathrm{TV}}(A'(S), A'(S'))}\,.$$

By taking the expectation over S, S', we get that

$$\begin{aligned}
\mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n} \left[\mathop{\mathbf{Pr}}_{r \sim \mathcal{R}}[A'(S, r) \neq A'(S', r)] \right] &\leq \mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n} \left[\frac{2 d_{\mathrm{TV}}(A'(S), A'(S'))}{1 + d_{\mathrm{TV}}(A'(S), A'(S'))} \right] \\
&\leq \frac{2\, \mathbf{E}_{S,S' \sim \mathcal{D}^n} \left[d_{\mathrm{TV}}(A'(S), A'(S')) \right]}{1 + \mathbf{E}_{S,S' \sim \mathcal{D}^n} \left[d_{\mathrm{TV}}(A'(S), A'(S')) \right]} \\
&\leq \frac{2\rho}{1 + \rho},
\end{aligned}$$

where the first inequality follows from Theorem A.13 and taking the expectation over S, S', the second inequality follows from Jensen's inequality, and the third inequality follows from the fact that $f(x) = 2x/(1+x)$ is increasing. Now notice that since the source of randomness $\mathcal{R}$ is independent of S, S', we have that

$$\mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n} \left[\mathop{\mathbf{Pr}}_{r \sim \mathcal{R}}[A'(S, r) \neq A'(S', r)] \right] = \mathop{\mathbf{Pr}}_{S,S' \sim \mathcal{D}^n, r \sim \mathcal{R}}[A'(S, r) \neq A'(S', r)]\,.$$

Thus, we have shown that

$$\mathop{\mathbf{Pr}}_{S,S' \sim \mathcal{D}^n, r \sim \mathcal{R}}[A'(S, r) \neq A'(S', r)] \leq \frac{2\rho}{1 + \rho}\,,$$

so the algorithm A' is n-sample $\frac{2\rho}{1+\rho}$-replicable, which concludes the proof. □

C.2. A General Equivalence Result

In this section, we focus on the following two stability/replicability definitions.

Definition C.1 (Replicability (Impagliazzo et al., 2022))**.** Let $\mathcal{R}$ be a distribution over random strings. A learning rule A is ρ-replicable if for any distribution μ over $\mathcal{I}$ and two independent instances $I, I' \sim \mu$ it holds that

$$\mathop{\mathbf{Pr}}_{I, I' \sim \mu, r \sim \mathcal{R}}[A(I, r) \neq A(I', r)] \leq \rho\,.$$

[11] Under the measure zero event that at least one of these three conditions does not hold, we let $A'(S, r)$ be some arbitrary classifier.

Definition C.2 (Total Variation Indistinguishability)**.** A learning rule A is ρ-TV indistinguishable if for any distribution μ and two independent instances $I, I' \sim \mu$ it holds that

$$\mathop{\mathbf{E}}_{I,I'\sim\mu}[d_{\mathrm{TV}}(A(I), A(I'))] \leq \rho\,.$$

A learning rule A is ρ-fixed prior TV indistinguishable if for any distribution μ, there exists some prior $\mathcal{P}_\mu$ such that for $I \sim \mu$ it holds that

$$\mathop{\mathbf{E}}_{I\sim\mu}[d_{\mathrm{TV}}(A(I), \mathcal{P}_\mu)] \leq \rho\,.$$

Definition C.3 (Pseudo-Global Stability)**.** Let $\mathcal{R}$ be a distribution over random strings. A learning rule A is said to be (η, ν)-pseudo-globally stable if for any distribution μ there exists an element $o_r \in \mathcal{O}$ for every $r \in \mathrm{supp}(\mathcal{R})$ (depending on μ) such that

$$\mathop{\mathbf{Pr}}_{r\sim\mathcal{R}}\left[\mathop{\mathbf{Pr}}_{I\sim\mu}[A(I, r) = o_r] \geq \eta\right] \geq \nu\,.$$

Our general equivalence result follows.

Proposition C.4 (TV Indistinguishability $\equiv$ Replicability)**.** *Let $\mathcal{I}$ be an input space and $\mathcal{O}$ be an output space.*

- *If a learning rule A is ρ-replicable, then it is also ρ-TV indistinguishable.*
- *Consider a prior distribution $\mathcal{P}$ over $\mathcal{O}$. Consider a learning rule A that is ρ-TV indistinguishable and absolutely continuous with respect to $\mathcal{P}$. Then, there exists a learning rule A' that is equivalent to A and A' is $2\rho/(1+\rho)$-replicable.*

We remark that one can adapt the proofs of Theorem 2.1 and Theorem 2.7 by setting $\mathcal{I} = (\mathcal{X} \times \{0,1\})^n$, $\mu = \mathcal{D}^n$ and $\mathcal{O} = \{0,1\}^{\mathcal{X}}$. Moreover, when $\mathcal{I}$ is countable, the design of the reference probability measure works in a similar way. Hence, we get the following corollary.

Corollary C.5. *Let $\mathcal{I}$ be a countable domain and let A be a learning rule that is ρ-TV indistinguishable. Then, there exists a $\frac{2\rho}{1+\rho}$-replicable learning rule A' that is equivalent to A.*

D. TV Indistinguishability and Differential Privacy

D.1. DP Preliminaries

We introduce some standard tools from the DP literature. We start with the Stable Histograms algorithm (Korolova et al., 2009; Bun et al., 2016). Let $\mathcal{X}$ be some domain and let $S \in \mathcal{X}^n$ be a (multi)set of its elements. We denote by $\mathrm{freq}_S(x) = \frac{1}{n} \cdot |\{i \in [n] : x_i = x\}|$, i.e., the fraction of times that x appears in S. The following result holds. It essentially allows us to privately publish a short list of elements that appear with high frequency in a dataset.

Lemma D.1 (Stable Histograms (Korolova et al., 2009; Bun et al., 2016))**.** *Let $\mathcal{X}$ be some domain. For*

$$n \geq O\left(\frac{\log(1/(\eta\beta\delta))}{\eta\varepsilon}\right)$$

there exists an (ε, δ)-differentially private algorithm `StableHist` *which, with probability at least $1-\beta$, on input $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$, outputs a list $L \subseteq \mathcal{X}$ and a sequence of estimates $a \in [0,1]^{|L|}$ such that*

- *Every x with $\mathrm{freq}_S(x) \geq \eta$ appears in L.*
- *For every $x \in L$, the estimate a_x satisfies $|a_x - \mathrm{freq}_S(x)| \leq \eta$.*

We also recall the agnostic private learner for finite classes that was proposed in (Kasiviswanathan et al., 2011) and is based on the Exponential Mechanism of (McSherry & Talwar, 2007).

Lemma D.2 (Generic Private Learner (Kasiviswanathan et al., 2011)). *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. There is an $(\varepsilon, 0)$-differentially private algorithm* `GenPrivLearner` *which given*

$$n = O\left(\log(|\mathcal{H}|/\beta) \cdot \max\left\{\frac{1}{\varepsilon\alpha}, \frac{1}{\alpha^2}\right\}\right)$$

samples from $\mathcal{D}$, outputs a hypothesis h such that

$$\mathbf{Pr}\left[\mathrm{err}_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h') + \alpha\right] \geq 1 - \beta\,.$$

Finally, we state a result relating weak learners and privacy.

Theorem D.3 (Weakly Accurate Private Learning $\implies$ Finite Littlestone Dimension (Alon et al., 2019)). *Let $\mathcal{X}$ be some domain and $H \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class with Littlestone dimension $d \in \mathbb{N} \cup \{\infty\}$ and let A be a weakly accurate learning algorithm (i.e., (α, β)-accurate with $\alpha = 1/2 - \gamma, \beta = 1/2 - \gamma$) for H with sample complexity n that satisfies (ε, δ)-differential privacy with $(\varepsilon, \delta) = (0.1, 1/(n^2 \log(n)))$. Then, $n \geq \Omega(\log^\star(d))$.*

In particular any class that is privately weakly-learnable has a finite Littlestone dimension.

We remark that this theorem appears in (Alon et al., 2019) with accuracy constant 0.1. However, it is known from (Dwork et al., 2010) that a DP algorithm with error $1/2 - \gamma$ can be boosted to one with arbitrarily small error with negligible loss in the privacy guarantees.

The following result that appears in (Bun et al., 2020) shows that if $\mathcal{H}$ has finite Littlestone dimension, then there exists a ρ-globally stable learner for this class.

Theorem D.4 (Finite Littlestone Dimension $\implies$ Global Stability (Bun et al., 2020)). *Let $\mathcal{X}$ be some domain and $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class with Littlestone dimension $d < \infty$. Let $\alpha > 0$ be the accuracy parameter and define $n = 2^{2^{d+2}+1} 4^{d+1} \cdot \left\lceil \frac{2^{d+2}}{\alpha} \right\rceil$. Then, there exists a randomized algorithm $A : (\mathcal{X} \times \{0,1\})^n \times \mathcal{R} \to \{0,1\}^X$ such that for any realizable distribution $\mathcal{D}$ there exists a hypothesis $f_{\mathcal{D}}$ for which*

$$\mathop{\mathbf{Pr}}_{S \sim \mathcal{D}^n, r \sim \mathcal{R}}[A(S, r) = f_{\mathcal{D}}] \geq \frac{1}{(d+1)2^{2^{d+1}}}, \quad \mathop{\mathbf{Pr}}_{(x,y) \sim \mathcal{D}}[f_{\mathcal{D}}(x) \neq y] \leq \alpha\,,$$

where $\mathcal{R}$ is the source of internal randomness of A.

We also include a result from (Ghazi et al., 2021b; Bun et al., 2023) which states that replicability implies differential privacy under general input domains[12].

Theorem D.5 (Replicability $\implies$ Differential Privacy (Ghazi et al., 2021b; Bun et al., 2023)). *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, where $\mathcal{X}$ is some input domain. If $\mathcal{H}$ is learnable by an n-sample (α, β)-accurate ρ-replicable learner A, for $\alpha \in (0, 1/2), \rho \in (0,1), \beta \in \left(0, \frac{2\rho}{\rho+1} - 2\sqrt{\frac{2\rho}{\rho+1}} + 1\right)$, then, for any $(\alpha', \beta', \varepsilon, \delta) \in (0,1)^4$ it is learnable by an $(\alpha + \alpha', \beta')$-accurate (ε, δ)-differentially private learner. Moreover, its sample complexity is*

$$n \cdot \mathrm{poly}(1/\alpha', 1/\varepsilon, \log(1/\delta), \log(1/\beta'))\,.$$

D.2. The Proof of Theorem 3.1

In this section we show that Global Stability (cf. Definition A.8) implies TV indistinguishability in the context of PAC learning. In particular, we show that given black-box access to a ρ-globally stable learner A whose stable output is α-accurate, e.g., the one described in Theorem D.4, we can transform it to a ρ-TV indistinguishable learner which is $(\alpha + \alpha', \beta)$-accurate, with a multiplicative $\mathrm{poly}(1/\rho, 1/\alpha', \log(1/\beta)$ blow-up in its sample complexity. We remark that this transformation is not restricted to countable domains $\mathcal{X}$. As an intermediate result, we show that global stability implies replicability.

[12] In fact, this result holds for general statistical tasks. The parameters stated in (Bun et al., 2023) are slightly looser, but using our boosting results we can generalize them and use the ones that appear in the statement.

Lemma D.6 (Global Stability $\implies$ Replicability)**.** *Let A be an n-sample ρ-globally stable learner whose stable hypothesis is α-accurate. Then, for every $\rho', \alpha', \beta \in (0,1)^3$, there exists a learner A' (Algorithm 3) that is ρ'-replicable and $(\alpha + \alpha', \beta)$-accurate. Moreover, A' needs*

$$\widetilde{O}\left(\frac{\log(1/\beta)}{\rho'^2\rho^3}\right)$$

oracle calls to A and uses

$$\widetilde{O}\left(\frac{\log(1/\beta)}{\rho^2\rho'^3} \cdot \left(n + \frac{1}{\alpha'^2}\right)\right)$$

samples.

Proof. We first argue about the accuracy and the confidence of the algorithm. Let h_A be the hypothesis such that $\mathbf{Pr}_{S\sim\mathcal{D}^n}[A(S) = h] \geq 1 - \rho$. The replicable heavy hitters algorithm (Algorithm 1) guarantees that, with probability at least $1 - \beta/2$, h_A will be contained in the output list L (Lemma B.6). We call this event E_0 and we condition on it. In the next step, we call the replicable agnostic learner on L (Algorithm 2). Since there is a hypothesis whose error rate is at most α, we know that the output of the agnostic learner will have error rate at most $\alpha + \alpha'$, with probability at least $1 - \beta/2$ (Lemma D.2). Let us call this event E_1. Thus, we see that by taking a union bound over the probabilities of these two events, the error rate of the output of our algorithm will be at most $\alpha + \alpha'$, with probability at least $1 - \beta$.

We now shift our focus to the replicability of our algorithm. First, notice that because of the guarantees of the replicable heavy hitters (Lemma B.6) the list L will be the same across two executions when the randomness is shared, with probability at least $1 - \rho'/2$. Let us call this event E_2. Similarly, under the event E_2, the output of the agnostic learner will be the same across two executions with probability $1 - \rho'/2$. Let us call this event E_3. By taking a union bound over E_2, E_3, we see that the algorithm is ρ'-replicable.

The sample complexity of the algorithm follows by the sample complexity of the replicable heavy hitters and the replicable agnostic learner (Lemma B.6, Lemma D.2). In particular, we need

$$\widetilde{O}\left(n \cdot \frac{\log(1/\beta)}{\rho^2\rho'^3}\right),$$

samples for this step and since the list has size $O(1/\rho')$ we need

$$\widetilde{O}\left(\frac{\log(1/\beta)}{\alpha'^2\rho^2\rho'^3}\right),$$

for the replicable agnostic learner. □

Algorithm 3 From Global Stability to Replicability

1: `Input:` Black-box access to a n-sample ρ-globally stable learner A with α-accurate stable hypothesis, sample access to distribution $\mathcal{D}$
2: `Parameters:` $\rho', \alpha', \beta \in (0,1)^3$
3: `Output:` Classifier $h : \mathcal{X} \to \{0,1\}$
4: $\mathcal{D}' \leftarrow$ distribution induced by drawing $S \sim \mathcal{D}^n$ and running $A(S)$
5: $L \leftarrow$ output of `ReplicableHeavyHitters` (Algorithm 1) with threshold $\rho/2$, error $\rho/4$, confidence $\beta/2$, replicability $\rho'/4$
6: Output `AgnosticReplicableLearner` (Algorithm 2) on hypothesis class L, with accuracy α', confidence $\beta'/2$ and replicability $\rho'/2$

Corollary D.7. *Let A be an n-sample ρ-globally stable learner whose stable hypothesis is α-accurate. Then, for every $\rho', \alpha', \beta \in (0,1)^3$, there exists a learner A' (Algorithm 3) that is ρ'-TV indistinguishable and $(\alpha + \alpha', \beta)$-accurate. Moreover, A' needs*

$$\widetilde{O}\left(\frac{\log(1/\beta)}{\rho'^2\rho^3}\right)$$

oracle calls to A and uses

$$\widetilde{O}\left(\frac{\log(1/\beta)}{\rho^2\rho'^3}\cdot\left(n+\frac{1}{\alpha^2}\right)\right)$$

samples.

Proof. The proof follows immediately from Lemma D.6 and the fact that a ρ'-replicable algorithm is ρ'-TV indistinguishable (Theorem 2.1). □

We now explain how we can use the previous results in the previous section to design a replicable algorithm for a class $\mathcal{H}\subseteq\{0,1\}^{\mathcal{X}}$ when we know that $\mathcal{H}$ admits a DP learner, for general domains $\mathcal{X}$. Formally, we prove the following result.

Lemma D.8 (Differential Privacy $\implies$ Replicability in General Domains)**.** *Let $\mathcal{H}\subseteq\{0,1\}^{\mathcal{X}}$ be a hypothesis class, where $\mathcal{X}$ is some input domain. Let A be an n-sample $(0.1,1/(n^2\log(n)))$-differentially private $(1/2-\gamma,1/2-\gamma)$-accurate learner for $\mathcal{H}$, for some $\gamma\in(0,1/2]$. Then, for every $\rho,\alpha,\beta\in(0,1)^3$ there exists a learner A' that is ρ-replicable and (α,β)-accurate. Moreover, A' uses*

$$\widetilde{O}\left(\frac{(d+1)^3 2^{3\cdot(2^d+1)}\log(1/\beta)}{\rho^2}\cdot\left(2^{2^{d+2}+1}4^{d+1}\cdot\left\lceil\frac{2^{d+2}}{\alpha}\right\rceil+\frac{1}{\alpha^2}\right)\right)$$

samples, where d is the Littlestone dimension of $\mathcal{H}$.

Proof. The first step in the proof is to notice that the existence of such a DP learner for $\mathcal{H}$ implies that its Littlestone dimension d is finite ((Alon et al., 2019), Theorem D.3). Then, we instantiate Algorithm 3 with the globally stable algorithm from (Bun et al., 2020) (Theorem D.4) with accuracy $\alpha/2$. Notice that since the random bits for the globally stable need to be different across two executions of the algorithm, we use two different sources of randomness, one that is public, i.e., shared across two executions, and one that is private, i.e., not shared across two executions. Due to Lemma B.4, this is equivalent to the original definition of replicability (Definition 1.2). For the remaining two steps, i.e., the replicable heavy-hitters and the replicable agnostic learner, we use public random bits. The sample complexity of the algorithm follows from the sample complexity of Theorem D.4 and Lemma D.6. □

Corollary D.9 (Differential Privacy $\implies$ TV Indistinguishability in General Domains)**.** *Let $\mathcal{H}\subseteq\{0,1\}^{\mathcal{X}}$ be a hypothesis class, where $\mathcal{X}$ is some input domain. Let A be an n-sample $(0.1,1/(n^2\log(n)))$-differentially private $(1/2-\gamma,1/2-\gamma)$-accurate learner for $\mathcal{H}$, for some $\gamma\in(0,1/2]$. Then, for every $\rho,\alpha,\beta\in(0,1)^3$ there exists a learner A' that is ρ-TV indistinguishable and (α,β)-accurate. Moreover, A' uses*

$$\widetilde{O}\left(\frac{(d+1)^3 2^{3\cdot(2^d+1)}\log(1/\beta)}{\rho^2}\cdot\left(2^{2^{d+2}+1}4^{d+1}\cdot\left\lceil\frac{2^{d+2}}{\alpha}\right\rceil+\frac{1}{\alpha^2}\right)\right)$$

samples, where d is the Littlestone dimension of $\mathcal{H}$.

Proof. The proof of this result follows immediately by Lemma D.8 and the fact that replicable learners are also TV indistinguishable learners (Theorem 2.1). □

D.3. List-Global Stability $\implies$ TV Indistinguishability

In this section we provide a different TV indistinguishable learner for classes with finite Littlestone dimension that has polynomial sample complexity dependence on the Littlestone dimension of the class. This learner builds upon the results of (Ghazi et al., 2021a;b). In particular, (Ghazi et al., 2021a) show that a class with finite Littlestone dimension admits a list-globally stable learner (Definition D.10). This learner constructs a sequence of hypothesis classes whose Littlestone dimension is at most that of $\mathcal{H}$, and part of the proof requires that uniform convergence (Definition A.6) holds for all of them. In order to avoid making measurability assumptions on the domain $\mathcal{X}$ and the hypothesis class $\mathcal{H}$ that would imply such a claim, we only state the results for countable $\mathcal{H}$. Nevertheless, we emphasize that they hold for more general settings.

We underline that the result in (Ghazi et al., 2021b) which designs a pseudo-globally stable learner for classes with finite Littlestone dimension, holds in the setting where $\mathcal{X}$ is finite because it relies on correlated sampling. The reason behind

this fact is that they have to convert a DP learner to a pseudo-globally stable one. In our case, we have to show that if $\mathcal{H}$ is learnable by a DP algorithm, it also admits a TV indistinguishable one. The proof of Theorem 3.1 follows almost directly from a result appearing in (Ghazi et al., 2021b).

Definition D.10 (List-Global Stability (Ghazi et al., 2021b))**.** A learning algorithm A is said to be m-sample α-accurate (L, η)-list-globally stable if A outputs a set of at most L hypotheses and there exists a hypothesis h (that depends on $\mathcal{D}$) such that $\mathbf{Pr}_{(x_1,y_1),\ldots,(x_m,y_m)\sim\mathcal{D}^n}[h \in A((x_1, y_1), \ldots, (x_m, y_m))] \geq \eta$ and $\mathrm{err}_{\mathcal{D}}(h) \leq \alpha$.

(Ghazi et al., 2021b) showed the following result regarding list m-list-globally stable learners, which is a modification of a result of (Ghazi et al., 2021a).

Lemma D.11 (Finite Littlestone $\Rightarrow$ List-Global Stability (Ghazi et al., 2021b;a))**.** *Let $\alpha, \zeta > 0$. and $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a countable hypothesis class with $\mathrm{Ldim}(\mathcal{H}) = d < \infty$, where $\mathcal{X}$ is an arbitrary domain. Then, there is a $(d \log(1/\zeta)/\alpha)^{O(1)}$-sample α-accurate $\left(\exp\left((d/\alpha)^{O(1)}\right), \Omega(1/d)\right)$-list-globally stable learner for $\mathcal{H}$ such that, with probability at least $1 - \zeta$, every hypothesis h' in the output list satisfies $\mathrm{err}_{\mathcal{D}}(h') \leq 2\alpha$.*

Algorithm 4 List-Global Stability $\implies$ TV Indistinguishability (Essentially Algorithm 1 in (Ghazi et al., 2021b))

Input: Black-box access to list-globally stable learner A
Parameters: $\alpha, \beta, \rho, \eta, L$
Output: Classifier $h : \mathcal{X} \to \{0, 1\}$
$\tau \leftarrow 0.5\eta$
$\gamma \leftarrow \frac{10^6 \log(L/(\rho\tau))}{\tau}$
$k_1 \leftarrow \frac{10^6 \log(L/(\rho\tau))}{\tau^2}$
$k_2 \leftarrow \left\lceil \frac{10^6 \gamma^2 \log(L/(\rho\tau))}{\rho^2} \right\rceil$
$m \leftarrow (d \log(k_1/\beta)/\alpha)^{O(1)}$ {Number of samples to run list-globally stable learner with parameters $(\alpha, \beta/k_1)$ (Lemma D.11)}
for $i \leftarrow 1$ to k_1 **do**
    Draw $S_i \sim \mathcal{D}^m$, run A on S_i to get a set H_i
end for
Let H be the set of all $h \in \mathcal{H}$ that appear in at least $\tau \cdot k_1$ of the sets $H_1, \ldots, H_{k_1}$
for $j \leftarrow 1$ to k_2 **do**
    Draw $T_j \sim \mathcal{D}^m$, run A on T_j to get a set G_j
end for
for $h \in H$ **do**
    Let $\hat{Q}_{H,G_1,\ldots,G_{k_2}}(h) = \frac{|\{j \in [k_2] | h \in G_j\}|}{k_2}$
end for
Let $\hat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}}$ be the probability distribution on $\mathcal{H}$ defined by

$$\hat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}}(h) = \begin{cases} \frac{\exp(\gamma \hat{Q}_{H,G_1,\ldots,G_{k_2}}(h))}{\sum_{h' \in H} \exp(\gamma \hat{Q}_{H,G_1,\ldots,G_{k_2}}(h'))}, & h \in H, \\ 0, & \text{otherwise.} \end{cases}$$

Output $h \sim \hat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}}$

Proposition D.12 (Adaptation from (Ghazi et al., 2021b))**.** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a countable hypothesis class with $\mathrm{Ldim}(\mathcal{H}) = d < \infty$ and $\mathcal{X}$ be an arbitrary domain. Then, for all $\alpha, \beta, \rho \in (0,1)^3$, there exists an n-sample ρ-TV indistinguishable algorithm (Algorithm 4) that is (α, β)-accurate with respect to the data-generating distribution $\mathcal{D}$, where*

$$n = \mathrm{poly}(d, 1/\alpha, 1/\rho, \log(1/\beta)) .$$

Proof. First, Lemma D.11 guarantees the existence of a list-globally stable learner A for $\mathcal{H}$. We will borrow some notation from (Ghazi et al., 2021b). We remark that the proof is a simple adaptation of the proof of Theorem 20 in (Ghazi et al., 2021b) but we include it for completeness. We will use Algorithm 4 essentially appearing in (Ghazi et al., 2021b) (this algorithm is the same as Algorithm 1 in (Ghazi et al., 2021b); their algorithm has an additional last step which performs

correlated sampling). We will show that Algorithm 4 satisfies the conclusion of Proposition D.12 and is the desired TV indistinguishable learner.

Sample Complexity. The number of samples used by Algorithm 4 is $m \cdot (k_1 + k_2)$, where m is the number of samples used for the black-box list-globally stable learner A. In particular, we have that

$$n(\alpha, \beta, \rho) = \text{poly}(d, 1/\alpha, 1/\rho, \log(1/\beta)) \,.$$

Accuracy Analysis. By the guarantees of algorithm A, we get that the output of A consists only of hypotheses with distributional error at most α with probability $1 - \beta/k_1$, a union bound implies that this holds for all hypotheses in H with probability $1 - \beta$. This implies the accuracy guarantee for Algorithm 4.

TV Indistinguishability Analysis. Let us set $Q(h) = \mathbf{Pr}_{S \sim \mathcal{D}^n}[h \in A(S)]$, let $H_{\geq 0.9\tau} = \{h \in 2^{\mathcal{X}} : Q(h) \geq 0.9\tau\}$ and $H_{\geq 1.1\tau} = \{h \in 2^{\mathcal{X}} : Q(h) \geq 1.1\tau\}$. First, Algorithm 4 creates the set H that contains all $h \in \mathcal{H}$ that appear in at least $\tau \cdot k_1$ of the realizations $A(S_1), \ldots, A(S_{k_1})$.

The first lemma controls the probability that H contains hypotheses that are "heavy hitters" for A and does not contain hypotheses h whose $Q(h)$ is small.

Lemma D.13 (Adaptation of Lemma 22 in (Ghazi et al., 2021b))**.** *Let $\mathcal{E}$ denote the good event that $H_{1.1\tau} \subseteq H \subseteq H_{0.9\tau}$. Then $\mathbf{Pr}[\mathcal{E}] \geq 1 - \rho$, where the randomness is over the datasets $S_1, ..., S_{k_1}$ and A.*

Proof. We will first show that $\mathbf{Pr}[H_{1.1\tau} \subseteq H] \geq 1 - \rho/2$. Since A outputs a list of size at most L, $H_{1.1\tau} \leq \frac{L}{1.1\tau} \leq L/\tau$. For any $f \in H_{1.1\tau}$, we have that $1\{f \in H\}$ is an i.i.d. Bernoulli random variable with success probability $Q(f) \geq 1.1\tau$. Hoeffding's inequality implies that

$$\mathbf{Pr}[f \notin H] \leq \exp(-0.02\tau^2 k_1) \leq 0.01\rho\tau/L \,.$$

A union bound over all hypotheses in $H_{1.1\tau}$, implies the desired inequality. The other direction follows by a similar argument and we refer to (Ghazi et al., 2021b) for the complete argument. □

The next step is to define the distribution $\mathcal{P}$ with density $\mathcal{P}(h) \propto \exp(\gamma Q(h)) 1\{h \in H_{\geq 0.9\tau}\}$. We can also define $\mathcal{P}_H(h) \propto \exp(\gamma Q((h))) 1\{h \in H\}$, where γ is as in Algorithm 4. The next lemma relates the two distributions.

Lemma D.14 (Adaptation of Lemma 23 in (Ghazi et al., 2021b))**.** *Under the event $\mathcal{E}$, it holds that $d_{\text{TV}}(\mathcal{P}, \mathcal{P}_H) \leq \rho/2$.*

Proof. The proof is exactly the same as the one of Lemma 23 in (Ghazi et al., 2021b) with the single modification that we pick γ to be of different value, indicated by Algorithm 4. □

Given a list-globally stable learner A (which exists thanks to Lemma D.11), we can construct the distribution over hypotheses $\widehat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}}$ appearing in Algorithm 4. We can then relate the empirical distribution $\widehat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}}$ with its population analogue $\mathcal{P}_H$.

Lemma D.15 (Adaptation of Lemma 24 in (Ghazi et al., 2021b))**.** *It holds that $\mathbf{E}[d_{\text{TV}}(\mathcal{P}_H, \widehat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}})] \leq \rho/2$, where the expectation is over the sets $T_1, ..., T_{k_2}$ and the randomness of A.*

Proof. The proof is exactly the same as the one of Lemma 24 in (Ghazi et al., 2021b) with the single modification that we pick k_2 to be of different value, indicated by Algorithm 4. □

Combining the above lemmas (as in (Ghazi et al., 2021b)), we immediately get that $\mathbf{E}[d_{\text{TV}}(\mathcal{P}, \widehat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}})] \leq \rho$, where the expectation is over all the sets $S_1, ..., S_{k_1}$ and $T_1, ..., T_{k_2}$ given as input to the learner A and A's internal randomness. Note that $\mathcal{P}$ is independent of the data and depends only on A. Both $\mathcal{P}$ and $\widehat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}}$ are supported on a finite domain. Note that Algorithm 4 that, given a training set S, outputs the distribution over hypotheses $\widehat{\mathcal{P}}_{H,G_1,\ldots,G_{k_2}}$ (obtained by Algorithm 4) satisfies TV indistinguishiability with parameter 2ρ using triangle inequality. Hence, two independent runs of Algorithm 4 will be 2ρ-close in total variation in expectation and the algorithm is TV indistinguishable, as promised.

□

D.4. The Proof of Theorem 3.2

We are now ready to show that TV indistinguishability implies approximate DP. We first start by showing that a non-trivial TV indistinguishable learner for a class $\mathcal{H}$ gives rise to a non-trivial DP learner for $\mathcal{H}$. The algorithm is described in Algorithm 5. The result then follows from the fact that classes which admit non-trivial DP learners have finite Littlestone dimension (Theorem D.3).

Algorithm 5 From TV Indistinguishability to Differential Privacy

Input: Black-box access to (α, β)-accurate ρ-TV Indistinguishable Learner A, Sample S
Parameters: $\alpha', \beta', \varepsilon, \delta$
Output: Classifier $h : \mathcal{X} \to \{0, 1\}$
$k \leftarrow O_{\beta,\rho}\left(\frac{\log(\log(1/\beta')/(\beta'\delta))}{\varepsilon}\right), k' \leftarrow O_{\beta,\rho}(\log(1/\beta'))$
Break S into disjoint $\{S_i^j\}_{i\in[k],j\in[k']}$ with $|S_i^j| = n, \forall i \in [k], j \in [k']$
$\mathcal{P} \leftarrow$ data-independent reference probability measure from Claim 2.5
$(X_1^j, \ldots, X_k^j) \leftarrow \Pi_{\mathcal{R}}(A(S_1^j), \ldots, A(S_k^j)), \forall j \in [k']$ using the Poisson point process $\mathcal{R}$ with intensity $\mathcal{P} \times$ Leb $\times$ Leb $\{A(S_i^j)$ is a distribution over classifiers, the coupling $\Pi_{\mathcal{R}}$ is described in Theorem A.13.$\}$
Compute list $L^j \leftarrow$ StableHist$(X_1^j, \ldots, X_k^j)$, with $\eta = O_{\beta,\rho}(1/\log(1/\beta'))$, correctness $\beta'/3$, privacy $(\varepsilon/2, \delta), \forall j \in [k']$ {Lemma D.1}
$\widetilde{L}^j \leftarrow$ Remove elements from L^j that appear less than $\eta/2$ times, $\forall j \in [k']$
Output GenPrivLearner$(\widetilde{L}^1, \ldots, \widetilde{L}^{k'})$ with accuracy $(\alpha'/2, \beta'/3)$, privacy $(\varepsilon/2, 0)$ {Lemma D.2}

Before we state the result formally, let us first provide some intuition behind the approach. On a high level it resembles the approaches of (Bun et al., 2020; Ghazi et al., 2021b; Bun et al., 2023) to show that (pseudo-)global stability implies differential privacy. We consider k' different batches of k datasets of size n. For each such batch, our goal is to couple the k different executions of the algorithm on an input of size n, so that most of these outputs are, with high probability, the same. One first approach would be to use a random variable as a "pivot" element in each batch: we first draw $A(S_1^1)$ according to its distribution and the remaining $\{A(S_i^1)\}_{i\in[k]\setminus\{1\}}$ from their optimal coupling with $A(S_1^1)$, given its realized value. Even though this coupling has the property that, in expectation, most of the outputs will be the same, it is not robust at all. If the adversary changes a point of S_1^1, then the values of all the outputs will change! This is not privacy preserving. For this reason, we use the coupling that is described in Theorem A.13. We use the fact that $\mathcal{X}$ is countable to design a reference probability measure $\mathcal{P}$ that is independent of the data. This is the key step that leads to privacy-preservation. Then, we can argue that if we follow this approach for multiple batches, there will be a classifier whose frequency and performance are non-trivial. The next step is to feed all these hypotheses into the Stable Histograms algorithm (cf. Lemma D.1), which will output a list of frequent hypotheses that includes the non-trivial one we mentioned above. Finally, we feed these hypotheses into the Generic Private Learner (cf. Lemma D.2) and we get the desired result.

Proposition D.16. *Let $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ where $\mathcal{X}$ is countable. Assume that $\mathcal{H}$ is learnable by an (α, β)-accurate ρ-TV indistinguishable learner A using n_{TV} samples, where $\rho \in (0, 1), \alpha \in (0, 1/2), \beta \in (0, (1-\rho)/(1+\rho))$. Then, for any $(\alpha', \beta', \varepsilon, \delta) \in (0, 1)^4$, it is also learnable by an $(\alpha + \alpha', \beta')$-accurate (ε, δ)-differentially private learner and the sample complexity is*

$$n_{\mathrm{DP}} = O_{\beta,\rho}\left(\frac{\log(1/\beta') \cdot \log(\log(1/\beta')/(\beta'\delta))}{\varepsilon} + \log(1/\eta\beta') \cdot \max\left\{\frac{1}{\varepsilon\alpha'}, \frac{1}{\alpha'^2}\right\}\right) \cdot n_{\mathrm{TV}}\,.$$

Proof. Let A be the TV indistinguishable algorithm. We need to argue that the output of Algorithm 5 is $(\alpha + \alpha', \beta')$-accurate and (ε, δ)-DP. We start with the former property.

Performance Guarantee. Let us consider the following experiment. We draw k samples, each one of size $n = n_{\mathrm{TV}}$. Let $S_1^1, \ldots, S_k^1$ be these samples and $A(S_1^1), \ldots, A(S_k^1)$ be the distributions of the outputs of the algorithm on these samples. We denote by X_i^1 the random variable that follows the distribution $A(S_i^1)$. Let us consider a coupling of this collection of

variables. Then, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\text{coupling}}\left[\min_{j\in[k]}\sum_{i=1}^{k}\mathbb{1}_{X_i^1\neq X_j^1}\right] &\leq \mathop{\mathbf{E}}_{\text{coupling}}\left[\sum_{i=1}^{k}\mathbb{1}_{X_i^1\neq X_1^1}\right]\\ &= \sum_{i=1}^{k}\mathop{\mathbf{E}}_{\text{coupling}}\left[\mathbb{1}_{X_i^1\neq X_1^1}\right]\\ &= \sum_{i=1}^{k}\mathop{\mathbf{Pr}}_{\text{coupling}}[X_i^1\neq X_1^1]\,.\end{aligned}$$

Note that the above hold for any coupling between the random variables $(X_i^1)_{i\in[n]}$. Let us fix the DP parameters (ε,δ). We will use the coupling protocol of Theorem A.13 with $\Omega=\{0,1\}^{\mathcal{X}}$, $\mathcal{P}$ the probability measure described in Claim 2.5, and $\mathcal{R}$ the Poisson point process with intensity $\mathcal{P}\times\mathrm{Leb}\times\mathrm{Leb}$. We remark that this choice of $\mathcal{P}$ satisfies two properties: the collection $A(S_i^1)$ is absolutely continuous with respect to $\mathcal{P}$ and $\mathcal{P}$ is data-independent, so it will help us establish the differential privacy guarantees. The guarantees of the coupling of Theorem A.13 imply that

$$\mathop{\mathbf{Pr}}_{\mathcal{R}}[X_i^1\neq X_1^1]\leq\frac{2d_{\mathrm{TV}}(X_i^1,X_1^1)}{1+d_{\mathrm{TV}}(X_i^1,X_1^1)}\,,$$

for all $i\in[k]$. Thus, we have that

$$\mathop{\mathbf{E}}_{\mathcal{R}}\left[\min_{j\in[k]}\sum_{i=1}^{k}\mathbb{1}_{X_i^1\neq X_j^1}\right]\leq\sum_{i=1}^{k}\frac{2d_{\mathrm{TV}}(X_i^1,X_1^1)}{1+d_{\mathrm{TV}}(X_i^1,X_1^1)}\,.$$

By taking the expectation over the random draws of the samples $S_1,\ldots,S_k$, we see that

$$\begin{aligned}\mathop{\mathbf{E}}_{S_1^1,\ldots,S_k^1,\mathcal{R}}\left[\sum_{i=1}^{k}\mathbb{1}_{X_i^1\neq X_1^1}\right] &\leq \mathop{\mathbf{E}}_{S_1^1,\ldots,S_k^1}\left[\sum_{i=1}^{k}\frac{2d_{\mathrm{TV}}(X_i^1,X_1^1)}{1+d_{\mathrm{TV}}(X_i^1,X_1^1)}\right]\\ &= \sum_{i=1}^{k}\mathop{\mathbf{E}}_{S_1^1,\ldots,S_k^1}\left[\frac{2d_{\mathrm{TV}}(X_i^1,X_1^1)}{1+d_{\mathrm{TV}}(X_i^1,X_1^1)}\right]\\ &\leq \sum_{i=1}^{k}\frac{2\,\mathbf{E}_{S_1^1,\ldots,S_k^1}[d_{\mathrm{TV}}(X_i^1,X_1^1)]}{1+\mathbf{E}_{S_1^1,\ldots,S_k^1}[d_{\mathrm{TV}}(X_i^1,X_1^1)]}\\ &\leq \frac{2\rho}{1+\rho}\cdot k,\end{aligned}$$

where the second to last step follows by Jensen's inequality since the function $f(x)=2x/(1+x)$ is concave in $(0,1)$ and the last step because $\mathbf{E}_{S_1^1,\ldots,S_k^1}[d_{\mathrm{TV}}(X_i^1,X_1^1)]\leq\rho$ and f is increasing in $(0,1)$. To make the notation cleaner, we let $\rho'=\frac{2\rho}{1+\rho}$. Notice that if $\rho<1$ then $\rho'<1$. Now using Markov's inequality we get that

$$\mathbf{Pr}\left[\sum_{i=1}^{k}\mathbb{1}_{X_i^1\neq X_1^1}\geq\nu k\rho'\right]\leq\frac{1}{\nu}\implies\mathbf{Pr}\left[\sum_{i=1}^{k}\mathbb{1}_{X_i^1=X_1^1}\geq(1-\nu\rho')k\right]\geq1-\frac{1}{\nu}\,,$$

where the probability is with respect to the randomness of the samples and the coupling.

We denote by $\mathcal{E}_\nu^1=\left\{\sum_{i=1}^{k}\mathbb{1}_{X_i^1=X_1^1}\geq(1-\nu\rho')k\right\}$ the event that a $(1-\nu\rho')$-fraction of the outputs has the same value. Let us now focus on the number of classifiers in a single experiment that are correct, i.e., their error rate is at most $\alpha<1/2$. Let $Y_i^1=\mathbb{1}_{\mathrm{err}(X_i^1)\geq1/2}$. Notice that because of the coupling we have used, $\{Y_i^1\}_{i=1}^k$ are not independent, so we cannot simply apply a Chernoff bound to get concentration. Let $\mathcal{E}_\beta^1$ be the event that the classifier X_1^1 is correct. We know that $\mathbf{Pr}[\mathcal{E}_\beta^1]\geq1-\beta$, where the probability is taken with respect to the random draws of the input and the randomness of the

algorithm. Now notice that under the event $\mathcal{E}_\nu^1 \cap \mathcal{E}_\beta^1$ at least $(1-\nu\rho')k$ classifiers are correct and have the same output. By a union bound we see that

$$\mathbf{Pr}[\mathcal{E}_\nu^1 \cap \mathcal{E}_\beta^1] \geq 1 - \beta - \frac{1}{\nu}\,.$$

We now pick ν so that

$$1 - \beta - \frac{1}{\nu} = \frac{1-\beta-\rho'}{2} > 0 \implies \nu = \frac{2}{\rho' - \beta + 1}\,.$$

Thus, under $\mathcal{E}_\nu^1 \cap \mathcal{E}_\beta^1$ there are $\frac{1-\beta-\rho'}{1-\beta+\rho'}k$ classifiers that are equal to one another and are correct. We let $q = \frac{1-\beta-\rho'}{1-\beta+\rho'}$. As we discussed, the probability of this event is at least $\frac{1-\beta-\rho'}{2} = p$, so if we execute it k' times we have that with probability at least $1 - e^{-pk'}$ it will occur at least once, i.e., $\mathbf{Pr}\left[\cup_{j\in[k']}\{\mathcal{E}_\nu^j \cap \mathcal{E}_\beta^j\}\right] \geq 1 - e^{-pk'}$. We pick $k' = 1/p \cdot \log(3/\beta')$. Thus, with probability at least $1 - \beta'/3$ there is a correct classifier that appears at least qk times. We condition on this event for the rest of proof and we let $S_i^j, X_i^j \sim A(S_i^j)$ be the i-th sample, classifier of the j-th batch, respectively.

The next step is to feed these classifiers into the Stable Histograms algorithm (cf. Lemma D.1). We have shown that there exists a good classifier whose frequency is at least $\eta = \frac{qk}{k\cdot k'} = \frac{q}{k'}$. Thus, our goal is to detect hypotheses with frequency at least $\eta/2$. We pick the correctness parameter of the algorithm to be $\beta'/3$ and the DP parameters to be $(\varepsilon/2, \delta)$. In total, we need

$$n' = O\left(\frac{\log(1/(\eta\beta'\delta))}{\eta\varepsilon}\right) = O\left(\frac{\log(1/\beta') \cdot \log\left(\log(1/\beta')/(qp\beta'\delta)\right)}{qp\varepsilon}\right)\,,$$

hypotheses in our list. Since $n' = k \cdot k'$ it suffices to pick

$$k = O\left(\frac{\log\left(\log(1/\beta')/(qp\beta'\delta)\right)}{q\varepsilon}\right)\,.$$

Hence, with probability at least $1 - \beta'/3$, the output of the algorithm will be a list L that contains all the hypotheses with frequency at least $\eta/2$ along with estimates a_x such that $|a_x - \mathrm{freq}_S(x)| \leq \eta/2$. Let x^* be the correct and frequent hypothesis whose existence we have established. We know that $a_{x^*} \geq \eta/2$. Since this algorithm is DP, we can drop from its output all the elements $x \in L$ for which $a_x < \eta/2$ without affecting the privacy guarantees. Thus, we end up with a new list L' whose size is $O(1/\eta)$.

The last step of the algorithm is to feed this list into the Generic Private Learner (cf. Lemma D.2) with privacy parameters $(\varepsilon/2, 0)$ and accuracy parameters $(\alpha'/2, \beta'/3)$. The total number of samples we need for this step is

$$n'' = O\left(\log(1/\eta\beta') \cdot \max\left\{\frac{1}{\varepsilon\alpha'}, \frac{1}{\alpha'^2}\right\}\right)\,.$$

Since there is an element in the list whose error is at most α, the guarantees of the algorithm give us that with probability at least $1 - \beta'/3$ the output has error at most $\alpha + \alpha'$.

Thus, by taking a union bound over the correctness of the three steps we described, we see that with probability $1 - \beta'$ the algorithm outputs a hypothesis whose error is at most $\alpha + \alpha'$. We now argue that the algorithm is $(\varepsilon, \delta)-$DP.

Privacy Guarantee. First we need to show that the coupling step is differentially private. This is a direct consequence of the coupling protocol that we have provided (cf. Theorem A.13) and the fact that the reference probability measure is data-independent. If the adversary changes an element in $S_i^j, i \in [k], j \in [k']$, then the coupling is robust, in the sense that if we fix the internal randomness, then at most one of the elements that the coupling outputs will change. The result for the privacy preservation of this step follows by integrating over the internal randomness.

For the remaining two steps, i.e., the Stable Histograms and the Exponential Mechanism the privacy guarantee follows from their definition. Using the privacy composition, we get that overall our algorithm is $(\varepsilon/2, \delta) + (\varepsilon/2, 0) = (\varepsilon, \delta)$-differentially private.

□

Corollary D.17. *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, where $\mathcal{X}$ is a countable domain. If $\mathcal{H}$ is learnable by a (α, β)-accurate ρ-TV indistinguishable learner using n_{TV} samples, where $\rho \in (0,1), \alpha \in (0, 1/2), \beta \in (0, (1-\rho)/(1+\rho))$, then* $\mathrm{Ldim}(\mathcal{H}) < \infty$.

Proof. The proof follows directly by combining Proposition D.16 and Theorem D.3. □

D.5. Going Beyond Countable $\mathcal{X}$

We now propose an approach that we believe can lead to a generalization of the algorithm beyond countable domains. The only change that we make in the algorithm has to do with Algorithm 5, where for every batch j we pick $\mathcal{P}_j = \frac{1}{k}\sum_{i=1}^{k} A(S_i^j)$. Notice that for every $j \in [k']$ the $\{A(S_i^j)\}_{i\in[k]}$ are absolutely continuous with respect to $\mathcal{P}_j$. However, it is not immediate now that the choice of $\{\mathcal{P}_j\}_{j\in[k']}$ leads to a DP algorithm. We believe that it is indeed the case that the algorithm is approximately differentially private and we leave it as in interesting open problem.

E. Amplification and Boosting

E.1. The Proof of Theorem 4.2

Let us first restate the theorem along with the sample complexity of the algorithm.

Theorem (Indistinguishability Amplification)**.** *Let $\mathcal{P}$ be a reference probability measure over $\{0,1\}^{\mathcal{X}}$ and $\mathcal{D}$ be a distribution over inputs. Consider the source of randomness $\mathcal{R}$ to be a Poisson point process with intensity $\mathcal{P}\times \mathrm{Leb}\times \mathrm{Leb}$, where Leb is the Lebesgue measure over $\mathbb{R}_+$. Consider a weak learning rule A that is (i) ρ-TV indistinguishable with respect to $\mathcal{D}$ for some $\rho \in (0,1)$, (ii) (α,β)-accurate for $\mathcal{D}$ for $(\alpha,\beta)\in(0,1)^2, \beta < \frac{2\rho}{\rho+1} - 2\sqrt{\frac{2\rho}{\rho+1}} + 1$, and, (iii) absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$. Then, for any $\rho', \varepsilon, \beta' \in (0,1)^3$, there exists an algorithm $\texttt{IndistAmpl}(A,\mathcal{R},\beta',\varepsilon,\rho')$ (Algorithm 6) that is ρ'-TV indistinguishable with respect to $\mathcal{D}$ and $(\alpha+\varepsilon,\beta')$-accurate for $\mathcal{D}$.*

Let $n_A(\alpha,\beta,\rho)$ denote the sample complexity of the weak learning rule A with input β',ε,ρ'. Then, the learning rule $\texttt{IndistAmpl}(A,\mathcal{R},\beta',\varepsilon,\rho')$ uses

$$\widetilde{O}\left(\frac{\log^3\left(\frac{1}{\beta'}\right)}{\left(\frac{2\rho}{\rho+1} - 2\sqrt{\frac{2\rho}{\rho+1}} + 1 - \beta\right)^2\left(1-\sqrt{\frac{2\rho}{\rho+1}}\right)\varepsilon^2\rho'^2}\cdot n_A(\alpha,\beta,\rho)\right)$$

i.i.d. samples from $\mathcal{D}$.

Algorithm 6 Amplification of Indistinguishability Guarantees

Input: Black-box access to (α,β)-accurate ρ-TV Indistinguishable Learner A, Sample access to $\mathcal{D}$, Access to Poisson point process $\mathcal{R}$ with intensity $\mathcal{P}\times\mathrm{Leb}\times\mathrm{Leb}$ {$\mathcal{P}$ is the reference probability measure from Claim 2.5.}
Parameters: β',ε,ρ'
Output: Classifier $h:\mathcal{X}\to\{0,1\}$
$\eta,\nu \leftarrow \sqrt{\frac{2\rho}{1+\rho}}, \sqrt{\frac{2\rho}{1+\rho}}$
$\mathcal{P} \leftarrow$ data-independent reference probability measure from Claim 2.5
$k \leftarrow \frac{\log(3/\beta')}{1-\nu-\beta/(1-\eta)}$
$r_i \leftarrow$ an infinite sequence of the Poisson Point Process $\mathcal{R}, \forall i\in[k]$ {cf. Theorem A.13.}
$\mathcal{D}_{r_i} \leftarrow$ the distribution of hypotheses that is induced by $A(S,r_i)$ when $S\sim\mathcal{D}^n, \forall i\in[k]$
$L_i \leftarrow \mathrm{HeavyHitters}\left(\mathcal{D}_{r_i}, \frac{3}{4}(1-\eta), \frac{1}{4}(1-\eta), \rho'/(2k), \beta'/(3k)\right), \forall i\in[k]$ {Algorithm 1.}
$\left(\hat{h}_i, \widehat{\mathrm{err}}(\hat{h}_i)\right) \leftarrow \mathrm{AgnosticLearner}(L_i, \varepsilon/2, \rho'/(2k), \beta'/(3k)), \forall i\in[k]$ {Algorithm 2.}
for $i\leftarrow 1$ to k **do**
  if $\widehat{\mathrm{err}}(\hat{h}_i) \le \alpha + \varepsilon/2$ **then**
    Output $\hat{h}_i$
  end if
end for
Output the all 1 classifier

Proof. Since A is ρ-TV indistinguishable there is an equivalent learning rule A' that is $\frac{2\rho}{1+\rho}$-replicable (cf. Theorem 2.7) and uses randomness $\mathcal{R}$, where $\mathcal{R}$ is a Poisson point process with intensity $\mathcal{P}\times\mathrm{Leb}\times\mathrm{Leb}$, with Leb being the Lebesgue

measure over $\mathbb{R}_+$. Let

$$\mathcal{R}_\eta = \left\{ r \in \mathcal{R} : \exists h \in \mathcal{H} \text{ s.t. } \mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n}[A'(S,r) = h] \geq 1-\eta \right\},$$

We have that $\mathbf{Pr}_{r\sim\mathcal{R}}[r \in \mathcal{R}_\eta] \geq 1-\nu$, for $\eta = \frac{\frac{2\rho}{1+\rho}}{\nu}, \nu \in \left[\frac{2\rho}{1+\rho}, 1\right)$ (cf. Claim B.3). For each $r \in \mathcal{R}_\eta$ let $h_r \in \mathcal{H}$ be an element that witnesses its inclusion in $\mathcal{R}_\eta$[13]. Notice that since A' is (α,β)-accurate there is at most a $\frac{\beta}{1-\eta}$-fraction of $r \in \mathcal{R}$ such that $r \in \mathcal{R}_\eta, \mathrm{err}(h_r) > \alpha$. Let $\mathcal{R}^*_\eta = \{r \in \mathcal{R}_\eta : \mathrm{err}(h_r) \leq \alpha\}$. Now notice that $\mathbf{Pr}_{r\sim\mathcal{R}}[r \in \mathcal{R}^*_\eta] \geq 1-\nu-\frac{\beta}{1-\eta}$. Thus, by picking $k = \frac{\log(3/\beta')}{1-\nu-\beta/(1-\eta)}$ i.i.d. samples from $\mathcal{R}$ we have that with probability at least $1-\beta'/3$ there will be some $r_{i^*} \in \mathcal{R}^*_\eta$. We denote this event by $\mathcal{E}_1$ and we condition on it for the rest of the proof.

Let us now focus on the call to the replicable heavy hitters subroutine. We have that, with probability at least $1-\beta'/(3k)$, every call will return a list that contains all the $(1-\eta)$-heavy-hitters and no elements whose mass is less than $(1-\eta)/2$. By a union bound, this happens with probability at least $1-\beta'/3$ for all the calls. Let us call this event $\mathcal{E}_2$ and condition on it for the rest of the proof. Notice that under these two events, the list L_{i^*} that corresponds to r_{i^*} will be non-empty and will contain a classifier whose error is at most α.

We now consider the calls to the replicable agnostic learner. Notice that every list that this algorithm takes as input has size at most $\frac{2}{1-\eta}$. Moreover, with probability at least $1-\beta/3'$, the estimated error of every classifier will be at most $\varepsilon/2$ away from its true error. We call this event $\mathcal{E}_3$ and condition on it. Hence, for any $\hat{h}_j, j \in [k]$, that passes the test in the "if" statement, we have that $\mathrm{err}(\hat{h}_j) \leq \alpha+\varepsilon$. In particular, the call to L_{i^*} will return $\hat{h}_{i^*}$, with estimated error $\widehat{\mathrm{err}}(\hat{h}^*_i) \leq \alpha+\varepsilon/2$, which means that $\mathrm{err}(\hat{h}^*_i) \leq \alpha+\varepsilon$. Hence, the algorithm will such a classifier and, by a union bound, the total probability that this event happens is at least $1-\beta'$.

The replicability of the algorithm follows from a union bound over the replicability of the calls to the heavy hitters and the agnostic learner (cf. Lemma B.6, Claim B.7). In particular, since we call the replicable heavy hitters algorithm k times with replicability parameter $\rho'/(2k)$ and the replicable agnostic leaner k times with replicability parameter $\rho'/(2k)$, we know that with probability at least $1-\rho'$ all these calls will return the same output across two executions of the algorithm.

For the sample complexity notice that each call to the replicable heavy hitters algorithm requires $O\left(\frac{k^2\log(k/\beta'(1-\eta))}{(1-\eta)^3\rho'^2}\right)$ (cf. Lemma B.6.) Under the events we have conditioned on, we see that $|L_i| = O(1/(1-\eta)), \forall i \in [k]$, hence each call to the agnostic learner requires $O\left(\frac{k^2}{(1-\eta)^3\varepsilon^2\rho'^2}\log\left(\frac{k(1-\eta)}{\beta'}\right)\right)$ (cf. Claim B.7). Substituting the value of k gives us that the sample complexity is at most

$$O\left(\frac{\log^3\left(\frac{\log(1/\beta')}{\beta'((1-\eta)(1-\nu)-\beta)}\right)}{\left((1-\eta)(1-\nu)-\beta\right)^2(1-\eta)\varepsilon^2\rho'^2}\right).$$

Plugging in the values of η, ν we get the stated bound. □

E.2. The Proof of Theorem 4.3

Let us first recall the result we need to prove along with its sample complexity.

Theorem (Accuracy Boosting)**.** *Let $\mathcal{P}$ be a reference probability measure over $\{0,1\}^{\mathcal{X}}$ and $\mathcal{D}$ be a distribution over inputs. Consider the source of randomness $\mathcal{R}$ to be a Poisson point process with intensity $\mathcal{P}\times\mathrm{Leb}\times\mathrm{Leb}$, where Leb is the Lebesgue measure over $\mathbb{R}_+$. Consider a weak learning rule A that is (i) ρ-TV indistinguishable with respect to $\mathcal{D}$ for some $\rho \in (0,1)$, (ii) $(1/2-\gamma,\beta)$-accurate for $\mathcal{D}$ for some $\gamma \in (0,1/2), \beta \in \left(0, \frac{2\rho}{\rho+1} - 2\sqrt{\frac{2\rho}{\rho+1}} + 1\right)$, and, (iii) absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$. Then, for any $\rho', \varepsilon, \beta' \in (0,1)^3$, there exists an algorithm $\texttt{IndistBoost}(A,\mathcal{R},\varepsilon)$ (Algorithm 7) that is ρ'-TV indistinguishable with respect to $\mathcal{D}$ and (ε,β')-accurate for $\mathcal{D}$.*

If $n_A(\gamma,\beta,\rho)$ is the sample complexity of the weak learning rule A with input γ,β,ρ, then $\texttt{IndistBoost}(A,\mathcal{R},\varepsilon)$ uses

$$\widetilde{O}\left(\frac{n_A(\gamma,\beta'\varepsilon\gamma^2/6,\rho\varepsilon\gamma^2/(3(1+\rho)))\log(1/\beta')}{\varepsilon^2\gamma^2} + \frac{\log(1/\beta')}{(2\rho/(1+\rho))^2\varepsilon^3\gamma^2}\right)$$

i.i.d. samples from $\mathcal{D}$.

[13] If there are multiple such elements then we pick an arbitrary one using a consistent rule.

Algorithm 7 Boosting of Accuracy Guarantee

Input: Black-box access to weak $(\frac{1}{2} - \gamma, \beta)$-accurate ρ-TV Indistinguishable Learner A, Sample $S \sim \mathcal{D}^n$, Access to Poisson point process $\mathcal{R}$ with intensity $\mathcal{P} \times \text{Leb} \times \text{Leb}$ {$\mathcal{P}$ is the reference probability measure from Claim 2.5.}
Target : ε, β'
Output: Classifier $h : \mathcal{X} \to \{0, 1\}$
IndistBoost() {This algorithm appears in (Impagliazzo et al., 2022)}
$\rho' = 2\rho/(1+\rho)$
$T = 100/(\varepsilon\gamma^2)$
$\mu_1(x) = 1$
$n_w = n_A\left(\gamma, \frac{\beta'}{3T}, \frac{\rho'}{6T}\right)$
for $t = 1..T$ **do**
    $\mathcal{D}_{\mu_t}(x) = \frac{\mu_t(x)\mathcal{D}_X(x)}{d(\mu_t)}$
    $S_t \leftarrow n_w/\varepsilon \cdot \log(T/\beta')$
    $S'_t \leftarrow \text{RejectionSampling}\left(S_t, n_w, \mu_t, \mathcal{R}_t^{(1)}\right)$
    $h_t \sim A\left(S'_t, \mathcal{R}_t^{(2)}\right)$
    Update $\mu_{t+1}(x)$ using smooth boosting trick of (Servedio, 2003).
    Draw $S''_t = O(1/(\rho'^2\varepsilon^3\gamma^2))$ i.i.d. samples from $\mathcal{D}$
    If IndistingTestMeasure $\left(\mu_{t+1}, S''_t, \mathcal{R}_t^{(3)}, \rho'/(3T), \beta'/(3T)\right) \leq 2\varepsilon/3$ then **output** sgn $\left(\sum_i h_i\right)$
end for
RejectionSampling(S_{in}, size_out, $\mu, \mathcal{R}$)
$S_{\text{out}} = \emptyset$
for $(x, y) \in S_{\text{in}}$ **do**
    Pick $b \in [0, 1]$ using $\mathcal{R}$
    If $\mu(x) \geq b$ then $S_{\text{out}} \leftarrow \text{append}(S_{\text{out}}, (x, y))$
    If $|S_{\text{out}}| >$ size_out then **output** S_{out}
end for
IndistingTestMeasure($\mu, S, \mathcal{R}, \rho', \beta$)
Call Algorithm 1 in (Impagliazzo et al., 2022) (see Theorem B.2) with source of randomness $\mathcal{R}$ and dataset S, error $\varepsilon/3$, confidence β, replicability ρ and query function μ

Proof of Theorem 4.3. In the sample complexity bound of Theorem 4.3, we remark that the first term is the number of samples used by the RejectionSampling mechanism (appearing in (Impagliazzo et al., 2022)) in the T rounds and the second term controls the number of samples used for the IndistingTestMeasure procedure (appearing in (Impagliazzo et al., 2022)) for the T rounds (see Algorithm 7). Let $[T] = \{1, ..., T\}$. As in (Impagliazzo et al., 2022)[Theorem 6.1], we consider that the shared randomness between the two executions consists of a collection of $3T$ tapes with uniformly random bits. We denote the j-th tape in round t by $\mathcal{R}_t^{(j)}$ for $j \in [3]$ and $t \in [T]$. Since A is n-sample ρ-TV indistinguishable there is an equivalent learning rule A' that is n-sample $\frac{2\rho}{1+\rho}$-replicable (cf. Theorem 2.7) and uses randomness $\mathcal{R}$, where $\mathcal{R}$ is a Poisson point process with intensity $\mathcal{P} \times \text{Leb} \times \text{Leb}$, with Leb being the Lebesgue measure over $\mathbb{R}_+$. Let us set $\rho' = 2\rho/(1+\rho)$. The boosting algorithm that we provide below interprets the random strings as follows: for any $t \in [T]$, we set $\mathcal{R}_t^{(2)} = \mathcal{R}$ (these will be the tapes used by the equivalent learning algorithm A') and the remaining tapes $\mathcal{R}_t^{(j)}$ corresponds to random samples from the uniform distribution in $[0, 1]$ for $j \in \{1, 3\}$ (these will be the tapes used by our sub-routines RejectionSampling and IndistingTestMeasure.

The boosting algorithm works as follows:

1. As in (Servedio, 2003), it uses a measure μ_t to assign different scores to points of $\mathcal{X}$. First, $\mu_1(x) = 1$ for any point. We will not delve into the details on how this step works. For details we refer to (Servedio, 2003) (as in (Impagliazzo et al., 2022) since this step is not crucial for the proof).

2. At every round t, the algorithm performs rejection sampling on a fresh dataset S_t using the routine

`RejectionSampling`. This algorithm is TV indistinguishable since it uses the source of randomness $\mathcal{R}_t^{(1)}$ that provides uniform samples in $[0,1]$ (it is actually replicable).

3. The part of the dataset that was accepted from this rejection sampling process is given to replicable learner A', which is equivalent to the TV indistinguishable weak learner A. This algorithm uses the shared Poisson point process $\mathcal{R}_t^{(2)}$ with intensity $\mathcal{P} \times \text{Leb} \times \text{Leb}$, where $\mathcal{P}$ is the reference probability measure from Claim 2.5, and outputs the same hypothesis with probability $1 - \rho'/(6T)$.

4. Then we use the smooth update rule of (Servedio, 2003) to design the new measure μ_{t+1} for the upcoming iteration. This step is deterministic.

5. Last we check whether the boosting procedure is completed. To this end, we check whether μ_t is in expectation small. This step again uses a uniformly random threshold in $[0,1]$ and so makes use of the source $\mathcal{R}_t^{(3)}$.

The algorithm runs for $T = \frac{C}{\varepsilon\gamma^2}$ rounds for some numerical constant $C > 0$. Hence, we will assume access to $3T$ tapes of randomness, T with points from the Poisson point process and $2T$ with uniform draws from $[0,1]$. The correctness of the algorithm follows from (Servedio, 2003) and (Impagliazzo et al., 2022)[Theorem 6.1]. As for the TV indistinguishability, this is implied by the replicability of the whole procedure. We have that the weak learner A' is called T times with TV indistinguishability parameter $\rho'/(6T)$, the rejection sampler is called T times so that it outputs $\perp$ with probability $\rho'/(6T)$ and the indistinguishable measure tester is $\rho'/(3T)$-TV indistinguishable and called T times. A union bound gives the desired result. For further details, we refer to (Impagliazzo et al., 2022) since the analysis is essentially the same.

For the failure probability β', the algorithm can fail if the rejection sampling algorithm outputs $\perp$, if the weak learner fails, and if the replicable SQ oracle (Theorem B.2) fails. We have that the probability that the rejection sampling gives $\perp$ using $n_w/\varepsilon \cdot \log(T/\beta')$ is at most β'/T (which can be considered much smaller than $\rho'/(6T)$). Since each one of the three probabilities are upper bounded by $\beta'/(3T)$, the indistinguishable boosting algorithm succeeds with probability $1 - \beta'$. □

E.3. Tight Bound Between β, ρ

As we alluded before, Proposition D.16 shows that if we have a ρ-TV indistinguishable (α, β)-accurate learner with $\rho \in (0,1), \alpha \in (0,1/2), \beta \in \left(0, \frac{1-\rho}{1+\rho}\right)$, then the class $\mathcal{H}$ has finite Littlestone dimension. The reason we need $\beta \in \left(0, \frac{1-\rho}{1+\rho}\right)$ is because, in expectation over the random draws of the samples and the randomness of the coupling, this is the fraction of the executions of the algorithm that will give the same output. The results of (Angel & Spinka, 2019) show that under certain conditions, if we want to couple k random variables whose pairwise TV distance is at most ρ, then under the pairwise optimal coupling the probability that the realization of a pair of them differs is $\frac{2\rho}{1+\rho}$. However, it is unclear what the implication of this result is in the setting we are interested in.

E.4. Beyond Countable $\mathcal{X}$

The barrier to push our approach beyond countable $\mathcal{X}$ is very closely related to the one we explained in the DP section. To be more precise, it is not clear how one can design a data-independent reference probability measure $\mathcal{P}$ when $\mathcal{X}$ is uncountable. Hence, one idea would be to use some *data-dependent* probability measure $\mathcal{P}$. This would affect our algorithm in the following way: instead of first sampling the random Poisson point process sequence independently of the data, we first sample $S_1, \ldots, S_k$ and let the reference probability measure be $\mathcal{P} = \frac{1}{k}\sum_{i=1}^{k} A(S_i)$. The difficult step is to show that this algorithm is TV indistinguishable. When we consider a different execution of the algorithm we let $S'_1, ..., S'_k$ be the new samples and $\mathcal{P}' = \frac{1}{k}\sum_{i=1}^{k} A(S'_i)$ be the new reference probability measure. A natural approach to establish the TV indistinguishability property of the algorithm is to try to couple $\mathcal{P}, \mathcal{P}'$ and show that under this coupling, the expected TV distance of two executions of the new algorithm is small. We leave this question open for future work.

E.5. Amplification and Boosting

We can combine the amplification and boosting results for a wide range of parameters and get the next corollary.

Corollary E.1. *Let $\mathcal{X}$ be a countable domain and A be an n-sample ρ-TV indistinguishable (α, β)-accurate algorithm, for some $\rho \in (0,1), \alpha \in (0,1/2), \beta \in \left(0, \frac{2\rho}{\rho+1} - 2\sqrt{\frac{2\rho}{\rho+1}} + 1\right)$. Then, for any $\rho', \alpha', \beta' \in (0,1)^3$, there exists a ρ'-TV indistinguishable (α', β')-accurate learner A' that requires at most $O\left(\text{poly}\left(1/\rho, 1/\alpha', \log(1/\beta')\right) \cdot n\right)$ samples from $\mathcal{D}$.*

The proof of this result follows immediately from Theorem 4.2, Theorem 4.3, and from the fact that we can design the reference probability measure $\mathcal{P}$ for countable domains (cf. Claim 2.5). This result leads to two natural questions: what is the tightest range of β for which we can amplify the stability parameter ρ and under what assumptions can we design such boosting and amplification algorithms for general domains $\mathcal{X}$? For a more detailed discussion, we refer the reader to Appendix E.3, Appendix E.4.

Remark E.2 (Dependence on the Parameters). We underline that the polynomial dependence on ρ in the boosting result is not an artifact of the algorithmic procedure or the analysis we provide, but it is rather an inherent obstacle in TV indistinguishability. (Impagliazzo et al., 2022) show that in order to estimate the bias of a coin ρ-replicably with accuracy τ one needs at least $1/(\tau^2\rho^2)$ coin tosses. Since ρ-TV indistinguishability implies $(2\rho/(1+\rho))$-replicability as we have shown (without any blow-up in the sample complexity), we also inherit this lower bound. Our main goal behind the study of the boosting algorithms is to identify the widest range of parameters α, ρ, β such that coming up with a ρ-TV indistinguishable algorithm switches from being trivial to being difficult. For example, in PAC learning we know that if the accuracy parameter is strictly less than $1/2$, then there are sample-efficient boosting algorithms that can drive it down to any $\varepsilon > 0$. In the setting we are studying, it is crucial to understand the relationship between β, ρ, see Appendix E.3.

F. TV Indistinguishability and Generalization

Recall that in Proposition 1.5 we claimed that the generalization bound can shave the dependence on the VC dimension by paying an overhead in the confidence parameter. A similar result appears in (Impagliazzo et al., 2022) relating replicability to generalization. We now present its proof.

Proof of Proposition 1.5. Let S be a sample from $\mathcal{D}^n$. Since A is ρ-TV indistinguishable, it is also ρ-fixed prior TV indistinguishable and let $\mathcal{P}_{\mathcal{D}}$ be the sample-independent prior. Consider two samples $h_1 \sim A(S)$ and $h_2 \sim \mathcal{P}_{\mathcal{D}}$. We consider the following quantities:

- $\widehat{L}(h_1) = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{h_1(x) \neq y\}$ is the empirical loss of h_1 in S.
- $\widehat{L}(h_2) = \frac{1}{n}\sum_{(x,y)\in S} \mathbb{1}\{h_2(x) \neq y\}$ is the empirical loss of h_2 in S.
- $L(h_1) = \mathbf{Pr}_{(x,y)\sim\mathcal{D}}[h_1(x) \neq y]$ is the population loss of h_1 with respect to $\mathcal{D}$.

We will show that all these three quantities are close to each other. First, let us consider the space of measurable functions $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$. We have that

$$d_{\mathrm{TV}}(P, Q) = \sup_{f\in\mathcal{F}} \left| \mathop{\mathbf{E}}_{x\sim P}[f(x)] - \mathop{\mathbf{E}}_{x\sim Q}[f(x)] \right| .$$

This means that the total variation distance between two distributions is essentially the worst case bounded distinguisher f. Since $\widehat{L} : \{0,1\}^{\mathcal{X}} \to [0,1]$, we have that

$$\left| \mathop{\mathbf{E}}_{h_1\sim A(S)} \left[\widehat{L}(h_1)\right] - \mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}} \left[\widehat{L}(h_2)\right] \right| \leq d_{\mathrm{TV}}(A(S), \mathcal{P}_{\mathcal{D}}) .$$

Similarly, we get that

$$\left| \mathop{\mathbf{E}}_{h_1\sim A(S)} [L(h_1)] - \mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}} [L(h_2)] \right| \leq d_{\mathrm{TV}}(A(S), \mathcal{P}_{\mathcal{D}}) .$$

Now, since A is ρ-fixed prior TV indistinguishable, using Markov's inequality, we have that $\forall \varepsilon_1 > 0$,

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n} \left[\left| \mathop{\mathbf{E}}_{h_1\sim A(S)} \left[\widehat{L}(h_1)\right] - \mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}} \left[\widehat{L}(h_2)\right] \right| \right] \leq \rho \Rightarrow \mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n} \left[\left| \mathop{\mathbf{E}}_{h_1\sim A(S)} \left[\widehat{L}(h_1)\right] - \mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}} \left[\widehat{L}(h_2)\right] \right| > \varepsilon_1 \right] \leq \frac{\rho}{\varepsilon_1} .$$

In a similar manner, we get

$$\mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n} \left[\left| \mathop{\mathbf{E}}_{h_1\sim A(S)} [L(h_1)] - \mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}} [L(h_2)] \right| > \varepsilon_1 \right] \leq \frac{\rho}{\varepsilon_1}$$

We note that, since $\mathcal{P}_{\mathcal{D}}$ is sample-independent, we have that the statistic

$$\mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}}[\widehat{L}(h_2)] = \frac{1}{n}\sum_{(x,y)\in S}\mathop{\mathbf{Pr}}_{h_2\sim\mathcal{P}_{\mathcal{D}}}[h_2(x)\neq y]$$

is a sum of independent random variables with expectation $\mathbf{E}_{h_2\sim\mathcal{P}_{\mathcal{D}}}[L(h_2)]$. We can use standard concentration of independent random variables and get

$$\mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n}\left[\left|\mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}}\left[\widehat{L}_S(h_2)\right] - \mathop{\mathbf{E}}_{h_2\sim\mathcal{P}_{\mathcal{D}}}[L_{\mathcal{D}}(h_2)]\right| \geq \varepsilon_2\right] \leq 2e^{-2n\varepsilon_2^2},$$

for any $\varepsilon_2 > 0$. This means that

$$\mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^n}\left[\left|\mathop{\mathbf{E}}_{h_1\sim A(S)}\left[\widehat{L}_S(h_1)\right] - \mathop{\mathbf{E}}_{h_1\sim A(S)}[L_{\mathcal{D}}(h_1)]\right| \geq 2\varepsilon_1+\varepsilon_2\right] \leq 2\rho/\varepsilon_1 + 2e^{-2n\varepsilon_2^2},$$

so we have that, with probability at least $1-4\rho/\varepsilon-\delta$,

$$\left|\mathop{\mathbf{E}}_{h_1\sim A(S)}\left[\widehat{L}_S(h_1)\right] - \mathop{\mathbf{E}}_{h_1\sim A(S)}[L_{\mathcal{D}}(h_1)]\right| \leq \varepsilon + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

We note that we obtain the result of Proposition 1.5 by taking $\varepsilon = \sqrt{\rho}$. □

G. Open Questions

Our work leaves the following open problems:

1. Does the equivalence between TV indistinguishability and replicability hold for general spaces, i.e., when the input domain is not countable?
2. Does the equivalence between TV indistinguishability and (ε,δ)-DP hold for general spaces?
3. How can we boost the correctness and amplify the indistinguishability parameter of a weak TV indistinguishable learner to a strong one in general spaces?
4. What is the minimal condition that characterizes TV indistinguishable PAC learnability? This is closely related to understanding the limits of TV indistinguishable boosting algorithms.