Realizable Learning is All You Need

Max Hopkins   NMHOPKIN@ENG.UCSD.EDU
Department of Computer Science and Engineering, UCSD, California, CA 92092.

Daniel Kane   DAKANE@ENG.UCSD.EDU
Department of Computer Science and Engineering / Department of Mathematics, UCSD, California, CA 92092.

Shachar Lovett   SLOVETT@CS.UCSD.EDU
Department of Computer Science and Engineering, UCSD, California, CA 92092.

Gaurav Mahajan   GMAHAJAN@ENG.UCSD.EDU
Department of Computer Science and Engineering, UCSD, California, CA 92092.

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

The equivalence of realizable and agnostic learnability is a fundamental phenomenon in learning theory. With variants ranging from classical settings like PAC learning and regression to recent trends such as adversarially robust and private learning, it's surprising we still lack a unified theory; traditional proofs of the equivalence tend to be disparate, and rely on strong model-specific assumptions like uniform convergence and sample compression.

In this work, we give the first model-independent framework explaining the equivalence of realizable and agnostic learnability: a three-line blackbox reduction that simplifies, unifies, and extends our understanding across a wide variety of settings. This includes models with no known characterization of learnability such as learning with arbitrary distributional assumptions or general loss, as well as a host of other popular settings such as robust learning, partial learning, fair learning, and the statistical query model.

More generally, we argue that the equivalence of realizable and agnostic learning is actually a special case of a broader phenomenon we call property generalization: any desirable property of a learning algorithm (e.g. noise tolerance, privacy, stability) that can be satisfied over finite hypothesis classes extends (possibly in some variation) to any learnable hypothesis class.

Keywords: PAC Learning, Realizable Learning, Agnostic Learning, Non-Uniform Covering

1. Introduction

The equivalence of realizable and agnostic learnability in Valiant (1984)'s Probably Approximately Correct (PAC) model is one of the best known results in learning theory, and numbers among its most surprising. Given a set X and a family of binary classifiers H, the result states that the ability to learn a classifier $h \in H$ from examples of the form $(x, h(x))$ is in fact sufficient for something much stronger: given samples from *any* distribution D over $X \times \{0, 1\}$, it is possible to learn the *best approximation to D in H*. This surprising equivalence stems from a classical result of Vapnik and Chervonenkis (1974) (VC), and independently Blumer, Ehrenfeucht, Haussler, and Warmuth (BEHW) (Blumer et al., 1989) and Haussler (1992), who equate both the former model (known as *realizable learning*) and the latter model (known as *agnostic learning*) to a strong property of pairs (X, H) called *uniform convergence*.[1]

---

1. Uniform convergence promises that a large enough sample gives a good approximation for loss of *every $h \in H$* simultaneously.

VC, BEHW, and Haussler's result was certainly a breakthrough in its own right, but its proof technique is too indirect to reveal any deeper connections between realizable and agnostic learning beyond the PAC setting. Further, recent years have seen both theory and practice shift away not only from this original formalization, but more generally from the "uniform convergence equals learnability" paradigm, often in favor of distributional or data-dependent assumptions like margin that are more applicable to the real world. The inability of VC, BEHW, and Haussler's proof technique to generalize to such scenarios raises a fundamental question: *is the equivalence of realizable and agnostic learning a fundamental property of learnability, or simply a happy coincidence derived from the original PAC framework?*

In the 30 years since these works, a mountain of evidence has amassed in favor of the former: almost every reasonable variant of learning shares some sort of similar equivalence. This includes a long list of popular settings such as regression (Bartlett et al., 1996), distribution-dependent learning (Benedek and Itai, 1991), multi-class learning (David et al., 2016), robust learning (Montasser et al., 2019), online learning (Ben-David et al., 2009), private learning (Beimel et al., 2014; Alon et al., 2020), and partial learning (Long, 2001; Alon et al., 2021b). What's more, the uniform convergence paradigm fails miserably in most of these models. In the distribution-dependent model, for instance, it is easy to build classes which are trivially learnable (even with one sample!) but completely fail to satisfy uniform convergence (Benedek and Itai, 1991). On the other hand, models such as private learning give well-known examples where uniform convergence fails to imply learnability (Alon et al., 2019b). In spite of this, we are really no closer today to a general understanding of this phenomenon than we were in the early 90s. Much like Vapnik and Chervonenkis (1974), Blumer, Ehrenfeucht, Haussler, and Warmuth (Blumer et al., 1989), and Haussler (1992)'s proofs, the above works often use indirect methods and tend to rely on powerful model-dependent assumptions.

In this work, we aim to offer a generic, unifying theory by way of the first direct reduction from agnostic to realizable learning. Unlike any previous work, our reduction is blackbox, relies on no additional assumptions, and, perhaps most importantly, is incredibly simple. In fact, the basic algorithm can be stated in three lines.

Algorithm 1: Agnostic to Realizable Reduction

Input: Realizable PAC-Learner $\mathcal{A}$, Unlabeled Sample Oracle $\mathcal{O}_U$, Labeled Sample Oracle $\mathcal{O}_L$
Algorithm:

1. Draw an unlabeled sample $S_U \sim \mathcal{O}_U$, and labeled sample $S_L \sim \mathcal{O}_L$.

2. Run $\mathcal{A}$ over all possible labelings of S_U to get:

$$C(S_U) \coloneqq \{\mathcal{A}(S_U, h(S_U)) \mid h \in H|_{S_U}\}.$$

3. **Return** the hypothesis in $C(S_U)$ with lowest empirical error over S_L.

This basic reduction simplifies and unifies classic results such as BEHW and Haussler's distribution-free equivalence and Benedek and Itai (1991)'s analogous result in the distribution-dependent set-

ting,[2] with no loss in sample-complexity.[3] Moreover, because Algorithm 1 doesn't rely on model-dependent properties like uniform convergence, it extends to learning regimes without known characterizations. One such example is the notoriously difficult distribution-family model, in which the adversary is given a restricted family of distributions $\mathscr{D}$ along with the pair (X, H). While no characterization of learnability is known in this model, Algorithm 1 can still be used to show that the realizable and agnostic settings are equivalent.

Unfortunately, while Algorithm 1 does avoid any significant blowup in sample complexity, it is inherently computationally inefficient. In fact, this is necessary unless $P = NP$. There are many basic classes (e.g. halfspaces) which are easy to learn in the realizable model, but NP-hard in the agnostic setting (see e.g. Feldman et al. (2009)).[4] As a result, we focus in this work only on information theoretic considerations, though building computationally efficient reductions under restricted settings remains a very interesting avenue of research.

Before moving on to analysis of Algorithm 1, it is worth mentioning that the reduction actually touches on a more general phenomenon than just the equivalence of realizable and agnostic learnability. Think of realizable learning as a "base learner," and of agnostic learning as an additional *property* we'd like to satisfy (a sort of noise tolerance). Algorithm 1 works in part because the agnostic property is easy to satisfy over *finite classes*. Call a property P (e.g. privacy, noise-tolerance) *finitely-satisfiable* if for every finite class, there exists a learner satisfying property P. The structure of our reduction suggests that any such property should in fact generalize to all "learnable" classes:

Guiding Principle *[Property Generalization] If $\mathcal{A}$ is a (sample-efficient) base learner and P is a finitely-satisfiable property, $\mathcal{A}$ can be used to build a (sample-efficient) learner with property P.*

We stress that the above is a guide, not a theorem, and indeed sometimes requires modification for a given application (e.g. while private and realizable learning are not equivalent, applying the principle can still lead to a number of useful weaker variants). There are many examples of property generalization in the literature (e.g. malicious noise (Kearns and Li, 1993), nasty noise (Bshouty et al., 2002), privacy (Beimel et al., 2013b), and robustness (Montasser et al., 2019)). While these works rely on strong model-specific assumptions, Algorithm 1 provides a more unified framework and helps extend such results beyond the PAC-model. This includes basic extensions such as general loss functions[5] and the distribution-family model, but also more involved modifications such as partial learning, robust learning, or even the statistical query model. To give a more concrete example, let's consider our principle in the setting where the "base learner" is an adversarially robust realizable learner over square loss, and the "finitely-satisfiable property" we'd like to generalize is privacy. The statement then claims that this base learner can be used as a subroutine to build an adversarially robust learner over square loss satisfying a variant of privacy (e.g. semi-privacy).

Along with providing a general framework, Algorithm 1's lack of reliance on strong assumptions like uniform convergence can actually be useful even in classical settings where they hold. To exhibit this, we focus on *semi-private learning*, a model in which the learner has access to a public

---

2. Formally Benedek and Itai only consider random classification noise, but it is clear that their analysis extends to the agnostic model.

3. In the distribution-free setting, the resulting sample complexity is optimal up to a log factor, though it was later discovered this could be removed through a more complicated chaining technique (Li et al., 2001).

4. We note this is only known to hold in the proper setting. This does not lose significant generality since our reduction is inherently proper, and even an improper reduction would break standard cryptographic hardness assumptions such as learning parity with noise.

5. Over infinite label spaces, we will require some weak assumptions on the loss.

unlabeled database and a private labeled database. While the goal is generally to learn in as few *public* samples as possible (since these are harder to collect in practice), uniform convergence based methods (Beimel et al., 2013b; Alon et al., 2019a) require extra unlabeled samples to achieve the condition. Algorithm 1 circumvents this entirely by appealing directly to a realizable PAC-learner, and achieves the information-theoretically optimal unlabeled sample complexity as a result.

2. The Basic Reduction

Since all of our results are derived from variants of Algorithm 1, it is instructive to start by considering its basic analysis in our simplest non-trivial setting: distribution-family classification. This framework captures learnability with arbitrary distributional assumptions, a well-studied relaxation of PAC learning in practice where worst-case distributional assumptions are often too strong, and encompasses both the distribution-free and distribution-dependent PAC settings. Unlike these models, however, the distribution-family setting has no known characterization of learnability: uniform convergence is not necessary (Benedek and Itai, 1991) as in the former, and finite coverability is not sufficient (Dudley et al., 1994) as in the latter. Indeed, it is plausible no combinatorial characterization of this model exists at all, as it shares characteristics with EMX learnability which is independent of the ZFC set theory axioms (Ben-David et al., 2017). As such, we cannot hope to prove the equivalence of agnostic and realizable learning in this model by finding a common characterization.

With this in mind, let's define distribution-family learning more formally. Let X be a set (called the *instance space*), $Y = \{0, 1\}$ the set of binary labels, $\mathscr{D}$ a family of distributions over X, and $H = \{h : X \to Y\}$ a family of binary classifiers. A tuple $(\mathscr{D}, X, H)$ is *realizably learnable* if there exists an algorithm[6] $\mathcal{A}$ and a function $n(\varepsilon, \delta)$ such that for every $\varepsilon, \delta > 0$, distribution $D \in \mathscr{D}$, and hypothesis $h \in H$, $\mathcal{A}$ outputs a good classifier with high probability on samples of size $n(\varepsilon, \delta)$:

$$\Pr_{S \sim D^{n(\varepsilon, \delta)}}[\mathrm{err}_{D \times h}(\mathcal{A}(S, h(S))) \leq \varepsilon] \geq 1 - \delta,$$

where $\mathrm{err}_{D \times h}(\mathcal{A}(S, h(S)))$ is commonly called the *error* or *risk* of h:

$$err_{D \times h}(h') = \Pr_{x \sim D}[h'(x) \neq h(x)].$$

Likewise, a tuple $(\mathscr{D}, X, H)$ is *agnostically learnable* if there exists an algorithm $\mathcal{A}$ which for every distribution D over $X \times Y$ whose marginal $D_X \in \mathscr{D}$ outputs h' close to the best hypothesis in H with probability $1 - \delta$:

$$\Pr_{S \sim D^{n(\varepsilon, \delta)}}[\mathrm{err}_D(\mathcal{A}(S)) \leq OPT + \varepsilon] \geq 1 - \delta,$$

where $OPT = \min_{h \in H}\{\mathrm{err}_D(h)\}$ is the error of the best hypothesis in the class and the risk $\mathrm{err}_D(\mathcal{A}(S))$ is similarly defined:

$$\mathrm{err}_D(h') = \Pr_{(x,y) \sim D}[h'(x) \neq y].$$

With this in mind, we can now state the most basic application of Algorithm 1: the equivalence of agnostic and realizable learning for distribution-family classification.

---

6. Note that $\mathcal{A}$ could be deterministic or randomized. This distinction has no effect on any of the arguments in this work.

Theorem 1 (Agnostic → Realizable (Distribution-Family Classification)) *Let $\mathcal{A}$ be a realizable learner for $(\mathscr{D}, X, H)$ using $n(\varepsilon, \delta)$ samples. Then Algorithm 1 is an agnostic learner for $(\mathscr{D}, X, H)$ using:*

$$m_U(\varepsilon, \delta) \leq n(\varepsilon/2, \delta/2)$$

unlabeled samples, and

$$m_L(\varepsilon, \delta) \leq O\left(\frac{n(\varepsilon/2, \delta/2) + \log(1/\delta)}{\varepsilon^2}\right)$$

labeled samples. Moreover, if (X, H) has finite VC dimension d, the latter can be improved to

$$m_L(\varepsilon, \delta) \leq O\left(\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2}\right).$$

Along with its novelty in the distribution-family setting (where no such equivalence was known), it is worth noting that in the distribution-free setting, Theorem 1 actually recovers the same sample complexity bound as standard analysis of uniform convergence.[7] We also note that while unlabeled sample complexity is not usually considered separately from labeled complexity in the PAC setting, this will become a useful distinction in semi-supervised extensions considered later in the work. As such, it is instructive to keep the complexities separate for the time being.

With this out of the way, let's prove Theorem 1. The analysis breaks naturally into two parts, corresponding respectively to Step 2 and Step 3 of Algorithm 1. In the first part, we'll show that $C(S_U)$, the set of outputs corresponding to running the realizable learner $\mathcal{A}$ across all possible labelings of the unlabeled sample S_U, is in some sense a "good approximation" of the class H. More formally, the crucial observation is that for any choice of the adversary's distribution, $C(S_U)$ will (almost) always contain a hypothesis close to the optimal solution.

Claim 2 *For any distribution D over $X \times Y$ whose marginal $D_X \in \mathscr{D}$, with probability $1 - \delta/2$, there exists $h' \in C(S_U)$ which is within $\varepsilon/2$ of the optimal risk:*

$$err_D(h') \leq OPT + \varepsilon/2.$$

Once we have this claim, the second step is to show that Step 3, an empirical risk minimization process on $C(S_U)$, gives the desired agnostic learner. This actually follows from standard arguments. In particular, given a hypothesis $h \in C(S_U)$, let

$$err_{S_L}(h) = \Pr_{(x,y) \sim S_L}[h(x) \neq y]$$

denote its *empirical risk* with respect to S_L. Since $C(S_U)$ is finite, a standard Chernoff+Union bound gives that with probability at least $1 - \delta/2$, the empirical risk of every hypothesis in $C(S_U)$ with respect to S_L is close to its true risk. Then as long as S_L is sufficiently large, empirical risk minimization returns a solution with at most $OPT + \varepsilon$ error with high probability.

It remains to prove Claim 2. The key observation lies in an equivalence between realizable PAC-learning and a weak type of randomized covering: for any fixed $h \in H$, $C(S_U)$ contains a hypothesis close to h with high probability.

---

7. Though it should be noted that the additional $\log(1/\varepsilon)$ factor can be removed by a more complicated chaining argument (Li et al., 2001).

Lemma 3 *For any distribution D over $X \times Y$ with marginal $D_X \in \mathscr{D}$ and any $h \in H$, with probability $1 - \delta/2$, there exists $h' \in C(S_U)$ which is within $\varepsilon/2$ of h in classification distance:*

$$\Pr_{x \sim D_X}[h'(x) \neq h(x)] \leq \varepsilon/2.$$

Proof The proof is essentially immediate from the definition of realizable PAC-learning. $\mathcal{A}$ promises that for any $h \in H$ and $D \in \mathscr{D}$, a $1 - \delta/2$ fraction of *labeled* samples $(S, h(S)) \sim D^{n(\varepsilon/2, \delta/2)}$ satisfy

$$\text{err}_{D \times h}[\mathcal{A}(S, h(S))] = \Pr_D[\mathcal{A}(S, h(S))(x) \neq h(x)] \leq \varepsilon/2.$$

Since $C(S_U)$ contains $\mathcal{A}(S_U, h(S_U))$ for every $h \in H$ by construction, the result follows. ∎

We call $C(S_U)$ a *non-uniform cover* (see Definition 16 for formal details). Note that Lemma 3 does not imply that $C(S_U)$ contains hypotheses close to every $h \in H$ *simultaneously*. This stronger object is called a *uniform* cover and takes provably more samples to construct (see Appendix J). In our case, a non-uniform cover is sufficient. Since the guarantee holds for every *fixed* $h \in H$, it must hold in particular for the optimal hypothesis h_{OPT}, so $C(S_U)$ contains some h' within $\varepsilon/2$ of optimal. Let's now formalize these ideas and put everything together to prove Theorem 1.

Proof [Proof of Theorem 1] Let D be the adversary's distribution over $X \times Y$, and let $h_{OPT} \in H$ be a hypothesis achieving the optimal error. By Lemma 3, with probability $1 - \delta/2$, $C(S_U)$ contains a hypothesis h' such that:

$$\Pr_{x \sim D_X}[h'(x) \neq h_{OPT}(x)] \leq \varepsilon/2.$$

This implies Claim 2 (that $C(S_U)$ contains a hypothesis with error at most $OPT + \varepsilon/2$) since

$$err_D(h') \leq \Pr_{(x,y) \sim D}[h_{OPT}(x) \neq y] + \Pr_{(x,y) \sim D}[h_{OPT}(x) \neq h'(x)] \leq OPT + \varepsilon/2.$$

We can now use standard empirical risk minimization bounds on $C(S_U)$ to find a hypothesis with error at most $OPT + \varepsilon$. Chernoff and union bounds imply that with probability at least $1 - \delta/2$, the empirical risk of every hypothesis in $C(S_U)$ on a sample of size $O\left(\frac{\log(|C(S_U)|/\delta)}{\varepsilon^2}\right)$ is at most $\varepsilon/4$ away from its true error. Since h' has error at most $OPT + \varepsilon/2$, its empirical risk is at most $OPT + 3\varepsilon/4$, and by the above guarantee any hypothesis in $C(S_U)$ with empirical risk at most $OPT + 3\varepsilon/4$ has true error at most $OPT + \varepsilon$.

Putting everything together, we have that with probability $1 - \delta$ over the entire process, the empirical risk minimizer of $C(S_U)$ has error at most $OPT + \varepsilon$ as desired. The sample complexity bounds follow from noting that $|C(S_U)|$ is at most $2^{n(\varepsilon/2, \delta/2)}$, and at most $\left(\frac{e \cdot n(\varepsilon/2, \delta/2)}{d}\right)^d$ if the class has VC dimension d. The sample complexity bound for the latter case then follows by plugging in the standard bound for distribution-free classification: $n(\varepsilon/2, \delta/2) \leq O\left(\frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right)$. ∎

3. Property Generalization Beyond Binary Classification

While simplifying and generalizing classic results on distribution-free (Blumer et al., 1989; Haussler, 1992) and distribution-dependent (Benedek and Itai, 1991) learning is nice, the real benefit of Algorithm 1 lies in its ability to adapt across a huge range of models. In this section we give a high-level overview of some extensions to popular settings beyond the PAC model and look at how Algorithm 1 extends to properties beyond agnostic learning (e.g. privacy, stability, malicious noise).

3.1. Generalizing Labels and Loss

We start with one of the most basic and important modifications in practice: loss functions beyond binary classification, encompassing classic settings such as multi-class classification, regression, and more. Let X be the instance space, and $\mathscr{D}$ a family of distributions over X. Instead of working over $\{0, 1\}$, we now consider classifiers over a generic *label space* Y, and *loss function* $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$. For simplicity, we assume that $\ell(y, y) = 0$, and say that ℓ satisfies the *identity of indiscernibles* if $\ell(y_1, y_2) = 0$ iff $y_1 = y_2$. Given a family of classifiers $H = \{h : X \to Y\}$ and a distribution D over $X \times Y$, the risk $err_{D,\ell}(h)$ of a hypothesis $h \in H$ is its expected loss $\underset{(x,y)\sim D}{\mathbb{E}}[\ell(h(x), y)]$, and learnability is defined analogously where $err_{D,\ell}(h)$ replaces the standard classification error.

Unfortunately, it is fairly clear that a naive application of Algorithm 1 will fail in this general model. If Y is infinite, the unlabeled sample S_U may have infinitely many labelings, causing $C(S_U)$ to be infinite in turn. Since we are not in the domain where uniform convergence is equivalent to learnability, empirical risk minimization is no longer guaranteed to work, and Algorithm 1 may therefore fail. In fact, this is not just an issue with our algorithm: it is an inherent barrier. Realizable and agnostic learning simply aren't equivalent for most reasonable losses over infinite label classes.

Proposition 4 (Proposition 20) *Let ℓ be any loss function over $\mathbb{R}$ satisfying the identity of indiscernibles that is continuous in the first variable. Then there exists a class $(\mathscr{D}, X, H, \ell)$ which is realizably learnable but not agnostically learnable.*

It is worth mentioning that $\mathscr{D}$ can be taken to be the set of all distributions over X in this construction, so the lower bound holds in the distribution-free setting as well.

The construction in Proposition 4 exploits the continuous label space by exactly encoding each hypothesis so that any labeled example uniquely determines its corresponding hypothesis. This is clearly learnable in the realizable case, but a small amount of noise erases the encoding so it can't be agnostically learned. On the other hand, a simple modification of Algorithm 1 shows this is essentially *the only barrier*. More formally, call a class *discretely learnable* if for all $\varepsilon > 0$, there exists an ε-discretization[8] of $(\mathscr{D}, X, H, \ell)$ that is learnable to $O(\varepsilon)$ error. Discrete learnability can be thought of as a weak type of noise tolerance, essentially acting only to rule out the above construction. We prove that discrete learnability is equivalent to agnostic learnability over two broad classes of loss. The first is a basic generalization of loss functions over finite label spaces.

Theorem 5 (Informal Theorem 44: Agnostic $\to$ Realizable (Doubly Bounded Loss)) *Let $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ be a loss function such that $\forall y_1 \neq y_2 \in Y$:*

$$\ell(y_1, y_2) \in [\alpha, \beta]$$

for some $\beta > \alpha > 0$. Then for any class $(\mathscr{D}, X, H, \ell)$, the following are equivalent:

1. *$(X, H, \mathscr{D}, \ell)$ is (properly) discretely-learnable.*

2. *$(X, H, \mathscr{D}, \ell)$ is (properly) agnostically-learnable.*

Theorem 5 implies that realizable and agnostic learning are equivalent over any loss function satisfying the identity of indiscernibles over finite Y, since realizable and discrete learnability are

---

8. A discretization is a class H' over a finite (or probably finite) label space such that every $h \in H$ is close to some $h' \in H'$. See Appendix C.1 for details.

equivalent in this case and every such loss is doubly bounded. We complement this with a lower bound showing a separation between realizable and agnostic learning for general loss on finite Y by exploiting a simple ternary loss function that fails the identity of indiscernibles (see Proposition 19).

While many reasonable loss functions on infinite label classes aren't bounded away from 0, they often come with other structure we can utilize. We'll prove a similar result under the weak assumption that our loss satisfies an approximate triangle inequality. Such loss functions, which we call approximate pseudometrics, can informally be thought of as generalizing distance-based loss.

Theorem 6 (Informal Theorem 24: Agnostic $\to$ Realizable (Approximate Pseudometrics)) *If ℓ : $Y \times Y \to \mathbb{R}_{\geq 0}$ is an (upper) bounded[9] loss function satisfying a c-approximate triangle inequality:*

$$\forall y_1, y_2, y_3 \in Y : \ell(y_1, y_3) \leq c(\ell(y_1, y_2) + \ell(y_2, y_3)),$$

then for any class $(\mathscr{D}, X, H, \ell)$ the following are equivalent:

1. *$(\mathscr{D}, X, H, \ell)$ is (properly) discretely-learnable.*

2. *$(\mathscr{D}, X, H, \ell)$ is (properly) c-agnostically learnable.*

where "c-agnostic" learnability only requires error $c \cdot OPT + \varepsilon$ rather than $OPT + \varepsilon$.

While c-agnostic learnability is a weaker guarantee than before, it is actually necessary for approximate pseudometrics. We prove there exist simple discretely-learnable classes over c-approximate pseudometrics which are not c'-agnostically learnable for any $c' < c$ (see Proposition 25). It is also worth noting that the sample complexity blowup in both Theorem 5 and Theorem 6 remains polynomial in ε^{-1} (indeed nearly-quadratic) in most scenarios.[10] Finally we remark that along with being novel in the distribution-family setting, to our knowledge these results are actually new to the distribution-free setting as well, where such an equivalence was only known for bounded Lipschitz (Bartlett et al., 1996; Wolf, 2018) or binary-valued (Bendavid et al., 1995; David et al., 2016) loss.

3.2. Beyond the PAC Setting

While distributional assumptions and general loss are a good step towards practice, today's trends have branched even further from the PAC setting. Here, we discuss a prototypical example of applying Algorithm 1 to a newer model: *adversarial robustness*. In the appendix, we cover extensions to Partial Learning (Appendix F), SQ-learning (Appendix H), and Fair Learning (Appendix I).

Robust learning is an extension of the PAC model handling adversarial perturbations at test time. Practically, this is meant to ensure that an integrated prediction system (e.g. self-driving car) cannot be tricked by small adversarial changes to the outside world. The model is formalized by modifying how we compute error. Given a function $\mathcal{U} : X \to P(X)$ mapping each $x \in X$ to a set of possible corruptions, the robust risk of a labeling $c : X \to Y$ with respect to a distribution D over $X \times Y$ is:

$$\text{R-err}_{\mathcal{U},D}(c) = \mathbb{E}_{(x,y) \sim D} \left[\max_{x' \in \mathcal{U}(x)} (\ell(c(x'), y)) \right].$$

---

9. By this we mean the range of ℓ is $[0, B]$ for some $B \in \mathbb{R}_{\geq 0}$. We note that this condition can be replaced with the weaker assumption that $(\mathscr{D}, X, C, \ell)$ is agnostically learnable for every *finite* subset $C \subseteq H$.

10. More formally, the blowup is at worst $\frac{\log(|Y_\varepsilon|)}{\varepsilon^2}$, where $|Y_\varepsilon|$ is the size of the ε-discretized label space.

Realizable and agnostic robust learning are then defined analogously to the PAC-model.[11] A basic modification to Algorithm 1 implies the equivalence of realizable and agnostic robust learning.

Theorem 7 (Informal Theorem 49: Agnostic → Realizable (Robust Classification)) *If $(\mathscr{D}, X, H)$ is robustly PAC-learnable in the realizable setting with sample complexity $n(\varepsilon, \delta)$, then (a modification of) Algorithm 1 robustly learns $(\mathscr{D}, X, H)$ in the agnostic setting in:*

$$m_U(\varepsilon, \delta) \leq O\left(\max_{\mu \in [0, 1-\varepsilon]} \left\{\frac{n(\varepsilon/(2(1-\mu)), \delta/3)}{1 - \mu}\right\}\right)$$

unlabeled samples and

$$m_L(\varepsilon, \delta) \leq O\left(\frac{m_U(\varepsilon, \delta) + \log(1/\delta)}{\varepsilon^2}\right)$$

labeled samples.

We note that this result can also be combined with our analysis for more general loss functions, albeit with a slightly worse c-agnostic parameter. In the classification setting, Theorem 7 generalizes recent work giving such an equivalence in the distribution-free model (Montasser et al., 2019, 2021), though the sample complexity of our algorithm suffers an extra factor of ε^{-1} in this special case.

3.3. Beyond Agnostic Learning

So far we have only considered using Algorithm 1 to reduce from agnostic to realizable learning, albeit in a number of extended settings. On the other hand, we claimed in the introduction that Algorithm 1 can be used to satisfy *any* "finitely-satisfiable" property. In this section, we'll discuss two such examples: privacy, and malicious noise. In Appendix G we cover a similar application to uniform stability. Note that since we are only modifying the learning *property* in this section, the base (realizable) learner remains the same and does not require any additional constraints.

We'll start with Kearns and Li (1993)'s *malicious noise*. In this model, the learner is given a faulty sample oracle $\mathcal{O}_M(\cdot)$ which returns a labeled sample from the adversary's true distribution with probability $1 - \eta$, and otherwise returns an adversarially chosen pair (x, y). A learner is tolerant to malicious noise if it achieves the standard PAC guarantees while drawing from the malicious oracle instead of the standard sample oracle. Like agnostic learning, tolerance to malicious noise is easy on finite hypothesis classes. As a result, a basic modification of Algorithm 1 gives a blackbox reduction from agnostic learning with malicious noise to realizable learning.

Theorem 8 (Informal Theorem 27: Malicious → Realizable) *If $(\mathscr{D}, X, H)$ is realizably PAC-learnable with sample complexity $n(\varepsilon, \delta)$, then for any $\eta < \frac{\varepsilon}{1+\varepsilon}$, (a modification of) Algorithm 1 gives an agnostic learner for $(\mathscr{D}, X, H)$ tolerant to η malicious noise. Furthermore letting $\Delta = \frac{\varepsilon}{1+\varepsilon} - \eta$, the sample complexity is at most $poly(\Delta^{-1}, \log(1/\delta), n(\Delta, \delta))$.*

This extends the result of Kearns and Li (1993) to the distribution-family setting, and is tight in the sense that $\frac{\varepsilon}{1+\varepsilon}$ is the best possible error tolerance in the malicious model (Kearns and Li, 1993). As before, the result can be combined with our prior techniques to give a similar equivalence for approximate pseudometric loss new to both the distribution-family and distribution-free settings.

---

11. We note the distribution-family model does require a slight twist, see Appendix E for details.

While malicious noise is certainly distinct from the agnostic model, both are examples of noise-tolerance properties. We'll end this section with a property of a different flavor: privacy. Informally, an algorithm is said to be α-*differentially private* if its output is not susceptible to small changes in the underlying sample (see Appendix C.3). Privacy is a very strong condition, even relaxed notions such as (α, δ)-differential privacy require finite Littlestone dimension in the distribution-free setting (Alon et al., 2019b), which rules out any direct reduction from realizable learning.

On the other hand, our reduction actually is able to recover a different relaxation known as *semi-private* learning (Beimel et al., 2013b). In this generalization of the well-studied 'label privacy' (Chaudhuri et al., 2006) model, the learning algorithm has access both to a private database of labeled samples and to a smaller *public* database of unlabeled data. This models a common practical scenario: a small portion of "opt-in" users are willing to release their participation, but still wish to hide sensitive label data. By replacing empirical risk minimization (ERM) in Step 3 of Algorithm 1 with a private algorithm known as the exponential mechanism (McSherry and Talwar, 2007), we give a direct reduction from semi-private to realizable learning in the distribution-family setting.

Theorem 9 (Informal Theorem 34: Semi-Private $\rightarrow$ Realizable) *If $(\mathscr{D}, X, H)$ is realizably learnable with sample complexity $n(\varepsilon, \delta)$, then (a modification of) Algorithm 1 gives an α-semi-private, agnostic learner using*

$$m_{pub}(\alpha, \varepsilon, \delta) \leq n(\varepsilon/2, \delta/2)$$

unlabeled (public) samples, and

$$m_{pri}(\alpha, \varepsilon, \delta) \leq O\left(\frac{n(\varepsilon/2, \delta/2) + \log(1/\delta)}{\varepsilon \cdot \min\{\varepsilon, \alpha\}} \right)$$

labeled (private) samples.

This generalizes Beimel, Nissim, and Stemmer's (Beimel et al., 2013b) original equivalence in the distribution-free setting, and unlike previous arguments extends to the general loss functions covered in Section 3.1. In fact, in this case our algorithm even gives a quantitative improvement over previous work (Beimel et al., 2013b; Alon et al., 2019a) in the distribution-free setting.

Corollary 10 (Informal Corollary 35: Semi-Private $\rightarrow$ Realizable (Distribution-Free Classification)) *If (X, H) has VC-dimension d, then (X, H) is α-semi-private, agnostically learnable in:*

$$m_{pub}(\varepsilon, \delta, \alpha) \leq O\left(\frac{d + \log(1/\delta)}{\varepsilon} \right)$$

unlabeled samples,[12] and

$$m_{pri}(\varepsilon, \delta, \alpha) \leq O\left(\frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon \cdot \min\{\varepsilon, \alpha\}} \right)$$

labeled samples.

---

12. We note that this algorithm is *improper*. A proper algorithm may need an additional $\log(1/\varepsilon)$ unlabeled samples for some classes.

For fixed d and δ, Corollary 10 resolves the unlabeled sample complexity of semi-private learning, as Alon, Bassily, and Moran (ABM) (Alon et al., 2019a) showed that any class which is not privately learnable requires at least $\Omega(\frac{1}{\varepsilon})$ public samples to (semi)-privately learn. The private sample complexity, on the other hand, remains off by a logarithmic factor from known lower bounds (Chaudhuri and Hsu, 2011). Resolving the latter requires improving the standard application of the exponential mechanism, and remains an interesting open problem.

It is also worth noting that Theorem 9 and Corollary 10 are robust to some amount of shift in distribution between the public and private databases. This problem, often called *covariate shift* in other contexts, is a commonly observed issue in machine learning practice, and is especially of concern in privacy where a distribution over "opt-in" public users could easily differ from the overall distribution of private data. We discuss covariate shift in the semi-private setting in Appendix C.4.

4. Proof Overview: Modification Archetypes

We now overview four general types of modification we use to extend Algorithm 1 to scenarios beyond the basic reduction for binary classification covered in Section 2.

Discretization. We start with our main technique for infinite label spaces. Here, since we cannot afford to run our learner over all labelings of S_U, the key is to run it only on those coming from some *discretization* of the class. As long as we have access to a learner for the discretization, the same arguments as in Section 2 then recover property generalization for infinite label spaces. We formalize these notions in Appendix C.1, where we use the technique to prove Theorem 6 (Theorem 5 is proved in Appendix D). Discretization can also be used to handle learning models such as the statistical query setting which output real-valued query responses (see Appendix H).

Subsampling. Another core limitation to Algorithm 1 is access to clean unlabeled data. Algorithm 1 works by running a realizable learner over a representative set of unlabeled data, but in practice such data may be corrupted, and data-dependent assumptions like margin might mean the optimal hypothesis isn't even well-defined on this set. We handle such cases by a simple subsampling procedure: run the learner over all labelings of *all subsets of S_U*. As long as S_U contains enough uncorrupted data, subsampling finds it and we can maintain the guarantees discussed in Section 2. We use this technique to prove property generalization for robust learning (Theorem 7 and Appendix E), partial learning (Appendix F), and malicious noise (Theorem 8 and Appendix C.2).

Replacing the Finite Learner. In the introduction, we proposed a general paradigm called property generalization: that a variant of any learning property which holds for finite classes should hold for *any* "learnable" class in the base model. The main idea relies on replacing the ERM process in Algorithm 1 with a generic learner for finite classes with the desired property. To prove Theorem 9, for example, we replace the ERM process with McSherry and Talwar (2007)'s exponential mechanism. We use a similar strategy in Appendix G to prove analogous results for uniform stability.

Replacing the Base Learner. Finally we note a very basic modification of Algorithm 1 that allows us to extend property generalization beyond the PAC setting: simply replace the input realizable PAC learner with a learner in the desired model. This is usually combined with one of the techniques above depending on the specific application, e.g. to prove property generalization for robust learning and the statistical query model. The same idea can also be used to analyze semi-private learning with covariate shift (Appendix C.4) and property generalization for fair learning (Appendix I).

5. Related Work

Agnostic learning is a very widely studied model across learning theory, and works across many different sub-areas have noted model specific equivalences with realizable learning. Here we'll survey a few representative examples, and discuss how they relate and differ from our approach.

5.1. Beyond Binary Classification

Uniform Convergence and Multiclass Classification. It is well known that the uniform convergence equals learnability paradigm continues to hold for $0/1$-valued loss functions over constant-size label spaces Blumer et al. (1989); Bendavid et al. (1995); Vapnik and Chervonenkis (1971); Natarajan (1989), and that agnostic and realizable learning are equivalent as a result. On the other hand, Daniely, Sabato, Ben-Devid, and Shalev-Shwartz Daniely et al. (2015) showed this is no longer the case as the number of labels grows large. In this regime, even basic multi-class learning is no longer equivalent to uniform convergence, so the connection between realizable and agnostic learning becomes non-trivial. A few years later, David, Moran, and Yehuyadoff (DMY) David et al. (2016) proved the equivalence nevertheless holds in the infinite multi-class setting through the weaker *sample compression equals learnability* paradigm. While more general than the uniform convergence paradigm, their proof remains model-specific and fails in many of the settings we consider, e.g. partial learning Alon et al. (2021b).

Discretization and General Loss Functions. Basic forms of discretization were also considered back in the mid 90s in work on characterizing the learnability of real-valued functions. In a seminal work, Bartlett, Long, and Williamson (BLW) Bartlett et al. (1996) proved that a scale-sensitive measure introduced by Kearns and Schapire Kearns and Schapire (1990) called fat-shattering dimension characterizes learnability under bounded Lipschitz loss functions.[13] BLW use a basic form of discrete learning (called quantization) to prove that fat-shattering dimension is a necessary condition, and use uniform convergence to prove sufficiency. We give a similar argument as BLW in the necessary direction, but show that uniform convergence is not necessary for the equivalence to hold, and instead use Algorithm 1 to appeal directly to discrete learnability. This allows us to extend BLW's result across a much more general set of loss functions and scenarios without strong model-specific assumptions.

5.2. Semi-supervised, Active, and Semi-Private Learning

Our reduction hinges on combining a realizable learner with unlabeled data to cut down the number of potential hypotheses in our class. The use of unlabeled samples to this effect is one of the core ideas in the field of *semi-supervised* learning Balcan and Blum (2010); Zhu (2017). Here, it is usually additionally assumed that the function to be learnt has some relation (or 'compatibility') to the underlying data distribution, for example it might have large margin on unlabelled data as in Transductive SVM Joachims et al. (1999) or redundant sufficient information as in Co-training Blum and Mitchell (1998); Dasgupta et al. (2002). In their seminal work on the topic, Balcan and Blum Balcan and Blum (2010) employed a similar strategy to Algorithm 1 in which they draw an unlabeled sample S_U and select hypotheses consistent with each possible labeling based upon compatibility. They argue via uniform convergence that this results in a uniform cover, and then use

---

13. Though the original work only considers $\ell 1$ loss, their techniques generalize to Lipschitz loss, see for instance Wolf (2018).

empirical risk minimization to select a good hypothesis in the cover. It is worth noting that around the same time a similar strategy independently found use in the online learning literature in work of Ben-David, Pal, and Shalev-Schwarz Ben-David et al. (2009) who simulated the so-called 'standard optimal algorithm' (SOA) over a sequence of examples and applied weighted majority Littlestone and Warmuth (1994) over the resulting set of hypotheses to obtain an agnostic online learner.

Similar strategies have also found use in the related *active learning* literature. Hanneke and Yang Hanneke and Yang (2015) use the same technique to build a cover from unlabeled samples (adding one hypothesis consistent with each possible labeling), and then apply active (adaptive) query algorithms to learn the best hypothesis in the cover in as few labeled samples as possible. This generalized earlier work of Dasgupta Dasgupta (2005), who assumed a priori that the cover was known to the learner ahead of time. Most recently, the approach has seen use in the study of semi-private learning. In their original work on the model, Beimel, Nissim, and Stemmer Beimel et al. (2013b) again apply the same trick for building a uniform cover, but then find the best hypothesis privately via the exponential mechanism (similar to our proof of Theorem 9). The analysis of this strategy was later improved by Alon, Bassily, and Moran (ABM) Alon et al. (2019a).

The above works differ from ours in two crucial senses. First, each work focuses solely on developing an algorithm for their specific framework (rather than working to understand a more general equivalence or reduction between settings). In this sense, one can view each of these prior results as a specific instance of our general framework where the "base learner" in our reduction is restricted to be an empirical risk minimizer (or SOA in the online setting), and a problem-specific learner for the relevant property (online, agnostic, active, or private) is then applied over the resulting cover. Second, and perhaps most importantly, these previous works all rely fundamentally on uniform convergence.[14] This means that their algorithms break down as soon as one moves away from the original PAC model (even to say the basic distribution-dependent setting), and can also lead to sub-optimal sample complexity bounds. In the analysis of semi-private learning, for instance, we show that avoiding uniform convergence leads to asymptotically better bounds, actually resolving the public sample complexity of the model altogether. Indeed, one can show that building a uniform cover requires asymptotically more unlabeled samples than a non-uniform one, and therefore cannot result in optimal semi-supervised algorithms.[15]

5.3. Non-Uniform Covering and Probabilistic Representations

Covering techniques have long been used in learning theory, and while almost all prior works focus on *uniform* notions (where all hypotheses are covered simultaneously), there is one notable exception. In 2013, Beimel, Nissim, and Stemmer (BNS) Beimel et al. (2013a) introduced *probabilistic representations*, a strong randomized form of covering used to characterize pure differentially private learning. In the language of our work, given a class (X, H), a probabilistic representation is a distribution over subsets of H which is a non-uniform cover *simultaneously over all distributions of* X. BNS prove that private learning is equivalent to the existence of a probabilistic representations for the class. Equivalently, this can be thought of as the ability to build a non-uniform cover without access to the underlying distribution at all. On the other hand, we are interested in the much weaker

---

14. The online setting is the only exception, but this model diverges much more substantially from the PAC-type learning we consider in this work than previous uniform convergence based methods.

15. Formally we only prove this for the distribution-family model, though we conjecture it holds in the distribution-free setting too.

setting where a non-uniform cover can be built from a bounded number of samples from the distribution (and crucially argue that this is equivalent to realizable learning). Thus in a sense, our core connection between realizable learning and non-uniform covering can be thought of as an analog of BNS' characterization of private learning by probabilistic representations.

Acknowledgments

References

Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. *arXiv preprint arXiv:1910.11519*, 2019a.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019b.

Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *Conference on Learning Theory*, pages 119–152. PMLR, 2020.

Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 447–455, 2021a.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes, 2021b.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. *journal of computer and system sciences*, 52(3):434–452, 1996.

Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan Ullman, and Steven Wu. Private query release assisted by public data. In *International Conference on Machine Learning*, pages 695–703. PMLR, 2020a.

Raef Bassily, Shay Moran, and Anupama Nandi. Learning from mixtures of private and public populations. *arXiv preprint arXiv:2008.00331*, 2020b.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 97–110, 2013a.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013b.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Learning privately with labeled and unlabeled examples. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 461–477. SIAM, 2014.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Shai Ben-David, Pavel Hrubes, Shay Moran, Amir Shpilka, and Amir Yehudayoff. A learning problem that is independent of the set theory zfc axioms. *arXiv preprint arXiv:1711.05195*, 2017.

S. Bendavid, N. Cesabianchi, D. Haussler, and P.M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

Gyora M Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

Avrim Blum, Steve Hanneke, Jian Qian, and Han Shao. Robust learning under clean-label attack. *arXiv preprint arXiv:2103.00671*, 2021.

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

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

Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Learning concept classes with privacy. *Manuscript*, 2006.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186. JMLR Workshop and Conference Proceedings, 2011.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

Yuval Dagan and Vitaly Feldman. PAC learning with stable and private predictions. In *Conference on Learning Theory*, pages 1389–1410. PMLR, 2020.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16:2377–2404, 2015.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. *Advances in neural information processing systems*, 18, 2005.

Sanjoy Dasgupta, Michael Littman, and David McAllester. PAC generalization bounds for co-training. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002. URL https://proceedings.neurips.cc/paper/2001/file/4c144c47ecba6f8318128703ca9e2601-Paper.pdf.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. *Advances in Neural Information Processing Systems*, 29:2784–2792, 2016.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

Richard M Dudley, Sanjeev R Kulkarni, T Richardson, and Ofer Zeitouni. A metric entropy bound is not sufficient for learnability. *IEEE Transactions on Information Theory*, 40(3):883–885, 1994.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Conference On Learning Theory*, pages 1693–1702. PMLR, 2018.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

S. Hanneke. Proper PAC learning VC dimension bounds. Theoretical Computer Science Stack Exchange, 2019. URL https://cstheory.stackexchange.com/q/44252. URL:https://cstheory.stackexchange.com/q/44252 (version: 2019-07-11).

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(12):3487–3602, 2015.

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

Thorsten Joachims et al. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, volume 99, pages 200–209, 1999.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

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

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *31st Annual Symposium on the Foundations of Computer Science,*, 1990.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Sanjeev R Kulkarni and Mathukumalli Vidyasagar. Learning decision rules for pattern classification under a family of probability measures. *IEEE Transactions on Information Theory*, 43(1):154–166, 1997.

Yi Li, Philip M Long, and Aravind Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.

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

Philip M Long. On agnostic learning with {0,*, 1}-valued and real-valued hypotheses. In *International Conference on Computational Learning Theory*, pages 289–302. Springer, 2001.

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

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Reducing adversarially robust learning to non-robust PAC learning. *arXiv preprint arXiv:2010.12039*, 2020.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. *arXiv preprint arXiv:2102.02145*, 2021.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Guy N Rothblum and Gal Yona. Probably approximately metric-fair learning. *arXiv preprint arXiv:1803.03242*, 5(2), 2018.

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

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Leslie G Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 1971.

M Vidyasagar, S Balaji, and Barbara Hammer. Closure properties of uniform convergence of empirical means and PAC learnability under a family of probability measures. *Systems & control letters*, 42(2):151–157, 2001.

Michael M Wolf. Mathematical foundations of supervised learning, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021.

Xiaojin Zhu. *Semi-supervised Learning*, pages 1142–1147. Springer US, 2017.

Appendix Organization

Our appendix is split into two main portions. Appendix A, Appendix B, and Appendix C make up the "main body" of the paper and are meant to be read in order as written. On the other hand, the remaining sections cover various applications of our reduction to a variety of learning models and properties. These sections are all self-contained, and are meant more as a reference text in the sense that the reader interested in some particular model or property should simply skip to the section covering that application.

The main body portion is organized as follows: we cover preliminary definitions in Appendix A, our base reduction from agnostic to realizable learning for finite label classes in Appendix B, and discuss the four modification archetypes along with a representative application in Sections C.1, C.2, C.3, and C.4. In more detail, these sections respectively cover: extensions to infinite label classes via discretization, malicious noise via sub-sampling, agnostic semi-private learning via replacing ERM, and covariate shift via replacing the base learner.

In the remaining sections we cover applications to doubly-bounded loss (Appendix D), robust learning (Appendix E), partial learning (Appendix F), uniformly-stable learning (Appendix G), the statistical query model (Appendix H), and fair learning (Appendix I), and discuss further connections of non-uniform covers to previous notions of covering (Appendix J).

Appendix A. Preliminaries

Before moving to a more formal discussion of our results, we'll cover the most basic learning models discussed in this work: standard (distribution-free) PAC-learning and distribution-family PAC-learning. Extended models we consider beyond these (e.g. malicious noise, robust learning, partial learning, etc.) will instead be introduced in their respective sections.

A.1. PAC-Learning

We'll start by reviewing the seminal PAC-learning model of Valiant (1984) and Vapnik and Chervonenkis (1974). We start with a few core definitions for the setting of general loss. Let X be an arbitrary set called the *instance space* (e.g. $\mathbb{R}^d$), Y a set called the *label space* (e.g. $\{0, 1\}$), and

H a family of labelings of X by Y (that is a family of functions of the form $h : X \to Y$). Given a class (X, H), it will often be useful to consider its *growth function* $\Pi_H(n)$ which measures the maximum size of H when restricted to a sample of size n:

$$\Pi_H(n) = \max_{h \in H, S \in X^n} (|H|_S|).$$

We note that the growth function is trivially bounded by $|Y|^n$, but one can often give stronger bounds when (X, H) satisfies some finite combinatorial dimension (e.g. VC-dimension for the binary case).

While PAC-learning is sometimes used to refer only to classification, we will study the model under general loss functions. With that in mind, we call a function $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ a *loss function* if $\ell(y, y) = 0$ for all $y \in Y$. We say a loss ℓ satisfies the *identity of indiscernibles* if $\ell(y_1, y_2) = 0$ iff $y_1 = y_2$. Given any distribution D over $X \times Y$ and loss ℓ, the *risk* of a labeling $h : X \to Y$ with respect to D and ℓ is its expected loss:

$$\mathrm{err}_{D,\ell}(h) = \mathbb{E}_{(x,y) \sim D}[\ell(h(x), y)].$$

The goal of learning is generally to find a classifier $h \in H$ that minimizes risk. More formally, there are two commonly studied variants of this problem. The original formulation, now called *realizable* learning, assumes the existence of a hypothesis in H with no loss.

Definition 11 ((Realizable) PAC-learning) *We say (X, H, ℓ) is realizable PAC-learnable if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ such that $\min_{h \in H} err_{D,\ell}(h) = 0$:*

$$\Pr_{S \sim D^{n(\varepsilon, \delta)}}[err_{D,\ell}(\mathcal{A}(S)) > \varepsilon] \leq \delta.$$

$\mathcal{A}$ *is called* **proper** *if it outputs only labels in H.*

Perhaps a more realistic variant of PAC-learning is to drop this restriction on the adversary, and let them choose an arbitrary distribution over $X \times Y$. This model, introduced by Haussler (1992) and Kearns, Schapire, and Sellie (Kearns et al., 1994), is known as *agnostic* learning.

Definition 12 ((Agnostic) PAC-learning) *We say (X, H, ℓ) is agnostic PAC-learnable if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$:*

$$\Pr_{S \sim D^{n(\varepsilon, \delta)}}[err_{D,\ell}(\mathcal{A}(S)) > OPT + \varepsilon] \leq \delta,$$

where $OPT = \min_{h \in H}\{err_{D,\ell}(h)\}$.

For some settings covered in this work, it will turn out that reaching $OPT + \varepsilon$ error is too stringent of a condition. However, we will show in these cases that it is sometimes possible to maintain a weaker guarantee and learn up to $c \cdot OPT + \varepsilon$ error for some constant $c > 1$. We call such classes *c-agnostic learnable*.

Finally, we note that for simplicity when ℓ is the standard "classification error:"

$$\ell(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 = y_2 \\ 1 & \text{else,} \end{cases}$$

we'll simply write (X, H) to mean (X, H, ℓ). Realizable and Agnostic Learning are well studied under many basic loss functions including binary classification, where both models are known to be characterized by a combinatorial parameter called VC-dimension.

A.2. Learning Under Distribution Families

The standard PAC-models described above are often called *distribution-free* due to the fact that no assumptions are made on the marginal distribution over X. In practice, however, this is usually too worst-case an assumption. We often expect distributions in nature to be "nice" in some way, or at least somewhat restricted. This is reflected in the fact that popular machine learning algorithms usually significantly outperform the PAC-model's worst-case generalization bounds. Indeed such niceness assumptions have long been popular in learning theory as well, where conditions such as tail bounds or anti-concentration are frequently used to build efficient algorithms.

These ideas are captured more generally by a simple (but notoriously difficult) extension to the PAC framework originally proposed by Benedek and Itai (1991), where the adversary is restricted to picking from a fixed, known set of distributions.

Definition 13 ((Realizable) Distribution-Family PAC-learning) *Let X be an instance space and $\mathscr{D}$ a family of distributions over X. We say $(\mathscr{D}, X, H, \ell)$ is realizable PAC-learnable if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ satisfying:*

1. The marginal $D_X \in \mathscr{D}$,

2. $\min_{h \in H} err_{D,\ell}(h) = 0$,

we have

$$\Pr_{S \sim D^{n(\varepsilon,\delta)}}[err_{D,\ell}(\mathcal{A}(S)) > \varepsilon] \leq \delta.$$

Agnostic learning is defined similarly. The adversary must still choose a marginal distribution in $\mathscr{D}$, but the conditional labeling can be arbitrary.

Definition 14 (Agnostic Distribution-Family PAC-learning) *We say $(\mathscr{D}, X, H, \ell)$ is agnostic PAC-learnable if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ satisfying:*

1. The marginal $D_X \in \mathscr{D}$,

$\mathcal{A}$ outputs a good hypothesis with high probability:

$$\Pr_{S \sim D^{n(\varepsilon,\delta)}}[err_{D,\ell}(\mathcal{A}(S)) > OPT + \varepsilon] \leq \delta$$

where $OPT = \min_{h \in H}\{err_{D,\ell}(h)\}$.

The weaker c-agnostic learning is defined analogously with OPT replaced by $c \cdot OPT$. Unlike the standard model, very little is known about distribution-family learnability. While a number of works have made some progress on this front (Benedek and Itai, 1991; Dudley et al., 1994; Kulkarni and Vidyasagar, 1997; Vidyasagar et al., 2001), a characterization of learnability remains elusive despite some 30 years of effort.

Appendix B. The Core Reduction: Agnostic to Realizable Learning

In this section, we give a more detailed exposition of our main reduction as covered in Section 2, including the more general setting of arbitrary loss on constant size label spaces (in the distribution-family model), matching lower bounds, and additional discussion of non-uniform covers. As mentioned previously, since there is no known combinatorial characterization of learnability in the distribution-family model, standard techniques (Blumer et al., 1989; Bendavid et al., 1995; Vapnik and Chervonenkis, 1971; Natarajan, 1989) cannot be used, and it is plausible that no combinatorial characterization of learnability exists for this model at all (Ben-David et al., 2017).

Before jumping into our reduction proper, it is worth discussing why we can't simply take the approach of prior works and rely on *uniform convergence*, a strong condition which promises that on a large enough sample, the empirical error of every hypothesis will be close to its true error. While uniform convergence was a very popular technique in the early years of learning, practitioners have since moved away from the paradigm which fails to capture learning rates seen in practice (Zhang et al., 2021; Nagarajan and Kolter, 2019). Indeed it soon became clear that the technique failed to capture even basic theoretical models such as the distribution-dependent setting.[16]

Proposition 15 (Benedek and Itai Benedek and Itai (1991)) *There exists a PAC-learnable class (D, X, H) over binary labels and classification loss without the uniform convergence property.*

Proof Let $X = [0, 1]$, D be the uniform distribution over X, $Y = \{0, 1\}$, and H consist of all indicator functions for finite sets $S \subset X$, as well as for X itself. It is not hard to see that (D, X, H) is realizably PAC-learnable by the following scheme in only a single sample: if the learner draws a sample labeled 1, output the all 1's function. Otherwise, output all 0s. When the adversary has chosen a finite set, with probability 1 the learner draws a sample labeled 0, and outputs a hypothesis with 0 error (since the finite set has measure 0). If the adversary chooses the all 1's function, the learner will always output the all 1's function.

On the other hand, it is clear that when the adversary chooses the all 1's function, no matter how many samples the learner draws, there will exist a hypothesis in the class that is poorly approximated by the sample. Namely the hypothesis whose support is given by the support of the sample itself has empirical measure 1, but true measure 0. As a result, this class fails to have the uniform convergence property despite its learnability. ∎

In later sections, we will even see distribution-free models where uniform convergence fails, such as the Partial PAC model (Long, 2001; Alon et al., 2021b) which captures realistic scenarios such as learning with margin. Since even the most basic modifications of PAC-learning fail to satisfy uniform convergence, it is clear we need to move beyond the condition to gain a more general understanding of the common phenomenon of equivalence between learning models.

Instead of relying on uniform convergence, our core observation is an equivalence between learning and sample access to a combinatorial object we call a *non-uniform cover*.

Definition 16 (Non-uniform Cover) *Let (X, H) be a class over label space Y and $L_{X,Y}$ denote the family of all labelings from X to Y. If C is a random variable over the power set $P(L_{X,Y})$ and $d : L_{X,Y} \times L_{X,Y} \to \mathbb{R}_{\geq 0}$ is a "distance" function between labelings, we call C a non-uniform (ε, δ)-cover of H with respect to d if for all $h \in H$:*

$$\Pr_{T \sim C} \left[\exists h' \in T : d(h', h) \leq \varepsilon \right] \geq 1 - \delta.$$

---

16. In this setting only a single marginal distribution D is allowed.

*We call C **bounded** if its support lies entirely on subsets of size at most some $k \in \mathbb{N}$, and we call the smallest such k its **size**.*

Non-uniform covers share a close connection to several notions of covering used throughout the learning literature such as uniform covers (Alon et al., 2019a) and fractional covers (Alon et al., 2021a). We discuss these connections in more detail in Appendix J. For the moment, we note only that previous works using the strictly stronger notion of uniform covering necessarily lose factors in the sample complexity as a result. We discuss this further in Appendix C.3 as well.

In Section 2, we argued (at least implicitly) that once we have sampling access to a bounded non-uniform cover, agnostic learnability follows from standard arguments. Namely since a sample T has bounded size and is guaranteed to contain a concept "close" to optimal, it suffices to run empirical risk minimization over about $\log(|T|/\delta)/\varepsilon^2$ samples. The key to our reduction therefore boils down to turning blackbox access to a realizable PAC-learner into sampling access to some relevant non-uniform cover. This is given by Step 2 of Algorithm 1, which we rewrite here as a subroutine called LEARNINGTOCOVER.

Algorithm 2: LEARNINGTOCOVER(H, $\mathcal{A}$, S_U)

Input: Hypothesis Class H, Realizable PAC-Learner $\mathcal{A}$, Unlabeled Sample size S_U.
Algorithm:

1. Run $\mathcal{A}$ over all possible labelings of S_U by H.

2. **Return** the set of responses $C(S_U)$:

$$C(S_U) \coloneqq \{\mathcal{A}(S_U, h(S_U)) \mid h \in H|_{S_U}\}.$$

In fact, we already argued in Section 2 that LEARNINGTOCOVER gives sampling access to a non-uniform cover, but we will re-write the result here in this formulation for convenience.

Lemma 17 (Core Lemma: Realizable Learning Implies Non-Uniform Covering) *Let $\mathcal{A}$ be an algorithm that (ε, δ)-PAC learns a class $(\mathscr{D}, X, H, \ell)$ in $n = n(\varepsilon, \delta)$ samples. Then for any $D \in \mathscr{D}$, running LEARNINGTOCOVER on $S_U \sim D^n$ returns a sample from a size $\Pi_H(n)$, non-uniform (ε, δ)-cover with respect to the standard distance between hypotheses:*

$$d(h, h') = \mathbb{E}_D[\ell(h'(x), h(x))].$$

Proof The proof is essentially immediate from the definition of realizable PAC-learning. $\mathcal{A}$ promises that for any $h \in H$ and $D \in \mathscr{D}$, a $1 - \delta$ fraction of *labeled* samples $(S_U, h(S_U)) \sim D^{n(\varepsilon, \delta)}$ satisfy

$$\mathrm{err}_{D \times h, \ell}(\mathcal{A}(S_U, h(S_U))) = \mathbb{E}_D[\ell(h'(x), h(x))] \leq \varepsilon,$$

where $h' = \mathcal{A}(S_U, h(S_U))$. Since $C(S_U)$ contains $\mathcal{A}(S_U, h(S_U))$ for every h by definition, the result follows. $\blacksquare$

This means that as long as we have blackbox access to a realizable PAC-learner and unlabeled samples from the adversary's distribution, we can simulate access to a non-uniform cover. Let's now

formalize our previous intuition that this is sufficient to turn a realizable learner into an agnostic one for any finite label class. We will generalize this result to doubly-bounded loss in Appendix D, but it is instructive to consider the setting of finite Y first.

Theorem 18 (Agnostic $\rightarrow$ Realizable (Finite Label Classes)) *Let $(\mathscr{D}, X, H, \ell)$ be any class on a finite label space Y with loss function $\ell : Y \rightarrow Y$ satisfying the identity of indiscernibles. Then Algorithm 1 is an agnostic learner with sample complexity:*

$$m(\varepsilon, \delta) \leq n(\eta_\ell \varepsilon, \delta/2) + O\left(\frac{\log\left(\frac{\Pi_H(n(\eta_\ell \varepsilon, \delta/2))}{\delta} \right)}{\varepsilon^2} \right)$$

where $\eta_\ell \geq \Omega\left(\frac{\min_{a \neq b}(\ell(a,b))}{\max_{a \neq b}(\ell(a,b))} \right)$ is a constant depending only on ℓ.

Proof Let $\mathcal{A}$ be the promised realizable learner for $(X, H, \mathscr{D}, \ell)$ with sample complexity $n(\varepsilon, \delta)$. Run LEARNINGTOCOVER with parameters $\varepsilon' = \eta_\ell \varepsilon$ and $\delta' = \delta/2$. We argue that the output contains some h' such that $\text{err}_{D,\ell}(h') \leq OPT + \varepsilon/2$. Since $C(S_U)$ is finite and ℓ is upper bounded, a standard Chernoff bound gives that choosing an empirical risk minimizer from $C(S_U)$ based on $O\left(\frac{\log\left(\frac{|C(S_U)|}{\delta} \right)}{\varepsilon^2} \right)$ additional samples gives the desired learner. The sample complexity then follows immediately from the definition of $C(S_U)$.

To see why $C(S_U)$ has this property, recall that for any $h \in H$, Lemma 17 states that $C(S_U)$ contains some h' such that:

$$\mathbb{E}_{x \sim D_X}[\ell(h'(x), h(x))] \leq \eta_\ell \varepsilon.$$

Because we assume that $\ell(a, b) = 0$ iff $a = b$, this actually implies a stronger relation—h and h' must be close in *classification* error:

$$\Pr_{x \sim D_X}[h(x) \neq h'(x)] \leq \frac{\eta_\ell \varepsilon}{\min_{a \neq b}(\ell(a,b))}.$$

Let $h_{OPT} \in H$ be an optimal hypothesis, and let h'_{OPT} denote the corresponding output of LEARNINGTOCOVER. Then by the above, we have that:

$$
\begin{aligned}
\text{err}_{D,\ell}(h'_{OPT}) &= \mathbb{E}_{(x,y) \sim D}[\ell(h'_{OPT}(x), y)] \\
&= \mathbb{E}_{(x,y) \sim D}[\ell(h'_{OPT}(x), y) - \ell(h_{OPT}(x), y) + \ell(h_{OPT}(x), y)] \\
&= OPT + \mathbb{E}_{(x,y) \sim D}[\ell(h'_{OPT}(x), y) - \ell(h_{OPT}(x), y)] \\
&\leq OPT + Pr_D[h_{OPT}(x) \neq h'_{OPT}(x)] \max_{a \neq b}(\ell(a,b)) \\
&\leq OPT + \varepsilon/2
\end{aligned}
$$

where we have used the assumption that we set $\eta_\ell = c \frac{\min_{a \neq b}(\ell(a,b))}{\max_{a \neq b}(\ell(a,b))}$ for some universal $c > 0$. ∎

It's worth spending a moment discussing our only assumption on the loss function ℓ, that it satisfies the identity of indiscernibles. This is not only a natural assumption for most cases in practice (that mislabeling has non-zero error), it is theoretically justified as well: realizable and agnostic learning aren't necessarily equivalent for ℓ without this property, even in the distribution-free setting.

Proposition 19 (Identity of Indiscernibles Lower Bound) *There exists a realizably learnable class* (X, H, ℓ) *over a finite label space* Y *which is not agnostically learnable.*

Proof Let the instance space $X = \mathbb{N}$ be the set of natural numbers, the label space $Y = \{0, 1\}^2$. We consider the hypothesis class H with all functions which output the first bit as 0, that is:

$$H = \{h : h(x) = (0, \cdot) \, \forall x \in X\}.$$

Furthermore, we define the loss function $\ell : Y \times Y \to \{0, 1, c\}$ as

$$\ell((b_1, r_1), (b_2, r_2)) = \begin{cases} 0 & b_1 = b_2 \\ 1 & b_1 \neq b_2 \text{ and } r_1 = r_2 \\ c & \text{otherwise.} \end{cases}$$

Note that (X, H, ℓ) is trivially learnable in the realizable setting simply by returning any $h \in H$. On the other hand, we will show it is only $O(c)$-agnostically learnable. First, notice that for any labelling $f : X \to Y$, there exists a hypothesis $h \in H$ which matches f on the second bit, and therefore for any marginal D over X:

$$OPT \leq err_{D,f}(h) \leq 1.$$

As a result, it suffices to show that for every $m \in \mathbb{N}$ and (randomized) algorithm $\mathcal{A}$ using m samples there exists a labeling $f : X \to Y$ and marginal distribution D_X such that

$$\mathbb{E}_{S \sim D_X^m} \mathbb{E}_{x \sim D_X} [\ell(\mathcal{A}(S, f(S))(x), f(x))] \geq c/12. \tag{1}$$

As long as this holds Markov's inequality gives that every algorithm must have error at least $\Omega(c)$ with constant probability.

For simplicity, we will restrict our attention in the rest of the proof to the marginal distribution D_X which is uniform over the set $[k]$ for some natural number k we will fix later. To prove Equation (1), by Yao's minimax principle it is enough to prove there is a distribution μ over functions $f : [k] \to Y$ such that any *deterministic* algorithm $\mathcal{A}$ has expected loss at least $c/12$ over μ:

$$\mathbb{E}_{f \sim \mu} \mathbb{E}_{S \sim D_X^m} \mathbb{E}_{x \sim D_X} [\ell(\mathcal{A}(S, f(S))(x), f(x))] > c/12.$$

We now show that the above holds for μ being uniform over all functions from $[k]$ to Y for any $k > 2m$. Here, we have that

$$\mathbb{E}_{x \sim D_X} \mathbb{E}_{f \sim \mu} \mathbb{E}_{S \sim D_X^m} [\ell(\mathcal{A}(S, f(S))(x), f(x)))] \geq \mathbb{E}_{x \sim D_X} \left[\Pr_{S \sim D_X^m}[x \notin S] \cdot c/4 \right],$$

where the last step follows from noting that for any value (a, b) that $\mathcal{A}(S, f(S))$ assigns to $x \notin S$, $f(x)$ will be $(1 - a, 1 - b)$ with probability $1/4$ incurring a loss of c. The result then follows by noting that for every $x \in [k]$:

$$\Pr_{S \sim D_X^m}[x \notin S] = \left(1 - \frac{1}{k}\right)^m \geq 1/e$$

24

since $1 - x \geq \exp\{-x/(1-x)\}$ and $k > 2m$. Therefore, we get that

$$\mathbb{E}_{x \sim D_X} \mathbb{E}_{f \sim \mu} \mathbb{E}_{S \sim D_X^m} [\ell(\mathcal{A}(S, f(S))(x), f(x))] \geq c/4e,$$

which completes the proof. ■

Note that this bound holds even if $\mathcal{A}$ is allowed to be improper. It is also worth noting that if we are willing to increase the size of Y, the learner's error in this bound can actually be increased all the way to c, the maximum possible (see Proposition 20). This is in fact tight, as we will show that any loss function like the above satisfying a c-approximate triangle inequality can be c-agnostically learned (that is learned to within $c \cdot OPT + \varepsilon$ error).

Appendix C. Four Modification Archetypes

C.1. Discretization: Infinite Label Classes

In the previous section, we showed that our base reduction characterizes the equivalence of realizable and agnostic learning for loss functions satisfying the identity of indiscernibles for all *finite* label classes. In this section, we discuss a technique called *discretization* that allows our reduction to extend this result to infinite label classes. Normally, it's clear that when Y is infinite our standard reduction will fail: since the total number of possible labelings of a finite sample may be infinite, LEARNINGTOCOVER may output an infinite set. In fact, this is more than a technical barrier: realizable and agnostic learning simply aren't equivalent for infinite label classes.

Proposition 20 *Let ℓ be any continuous loss function (in the first variable) over $\mathbb{R}$ satisfying the identity of indiscernibles. Then there exists a class $(\mathscr{D}, X, H, \ell)$ which is realizably learnable but not agnostically learnable.*

Proof Let $X = \mathbb{N}$ and $Y = [0, 2]$. To construct our class, we first consider the set of all boolean functions over X with finite support. Each function f in this class may equivalently be thought of as a binary string in $\{0, 1\}^*$. Denote the corresponding decimal value of this string in $[0, 1]$ by s_f. To construct H, for every boolean function $f : \mathbb{N} \to \{0, 1\}$ with finite support, include in H the function $h_f(x) = f(x) + s_f$. Note that (X, H) is clearly realizably learnable under any distribution family $\mathscr{D}$ and any loss function, since a single sample always uniquely determines h_f. On the other hand, adding even the smallest amount of noise erases this unique identification, making the class impossible to learn.

More formally, let $\mathscr{D}$ be the family of all distributions. By the continuity of ℓ and the fact that $\ell(0, 0) = \ell(1, 1) = 0$, for all $\varepsilon > 0$ notice that there exists $\gamma = \gamma(\varepsilon) > 0$ such that $\max_{0 \leq \gamma' \leq \gamma}\{\ell(\gamma', 0), \ell(1 + \gamma', 1)\} < \varepsilon$. Let $n_\gamma \in \mathbb{N}$ be the index of the first non-zero digit in the binary representation of γ. The idea is to note that beyond these first n_γ coordinates, our class if within ε of an arbitrary boolean function. More formally, notice that for any distribution D and boolean function f which is 0 on $[n_\gamma]$, we have that $\text{OPT}_H(f) := \min_{h \in H}\{\text{err}_{D \times f, \ell}(h)\} \leq \varepsilon$. The bound then follows from the fact that such arbitrary functions are not learnable.

In more detail, Yao's minimax principle states that it is sufficient to show that for any potential sample complexity $m(\varepsilon, \delta)$, there exists a randomized strategy for the adversary such that no deterministic learner can achieve $OPT + c$ accuracy with constant probability for some constant $c > 0$. To this end, consider the following strategy: the adversary chooses the uniform distribution over

$[n_\gamma, n_\gamma + 2m(\varepsilon, \delta)]$, and a binary function on that interval uniformly at random (recall OPT is at most ε for every such function). Since the learner can only see $1/2$ of the mass, any strategy must be incorrect on half of the remaining points in expectation. In particular conditioned on any sample, the expected loss of any predicted labeling of an unseen point is at least $\ell_{\text{min-err}} = \min\limits_{y \in [0,2]} \{\frac{\ell(y,0)+\ell(y,1)}{2}\}$ (since each unseen label appears with $1/2$ probability conditioned on the learner's sample). The total expected loss of any strategy is then at least $\ell_{\text{min-err}}/2$, which is bounded away from 0. Setting ε and c sufficiently small then gives the desired result. ∎

Proposition 20 relies crucially on the fact that the adversary can erase a significant amount of information with a very small label perturbation. In the rest of this section, we'll discuss a technique for modifying our reduction that shows this is essentially *the only barrier* between realizable and agnostic learning (at least for a broad class of loss functions). The key is to require a slightly stronger notion of learnability based upon *discretization*.

Definition 21 (Discretization) *We say $(\mathscr{D}, X, H', \ell)$ is an ε-discretization of $(\mathscr{D}, X, H, \ell)$ if the following three conditions hold:*

1. *H' is **probably bounded**. That is for all $n \in \mathbb{N}$, $\delta > 0$, and $D \in \mathscr{D}$ there exists a bound $m(n, \delta) \in \mathbb{N}$ such that:*
$$\Pr_{S \sim D^n}[|Im(H'|_S)| \leq m(n, \delta)] \geq 1 - \delta.$$

2. *H' **point-wise**[17] ε-**covers** H with respect to ℓ. That is for all for all $h \in H$, there exists $h' \in H'$ satisfying:*
$$\forall x \in X : \ell(h'(x), h(x)) \leq \varepsilon.$$

3. *H' is **always useful**. That is for all $h' \in H'$, there exists $h \in H$ such that:*
$$\forall x \in X : \ell(h'(x), h(x)) \leq \varepsilon.$$

Note that most realistic settings have reasonable discretizations (e.g. it is enough to have some Lipshitz-like condition and a weak tail-bound on the loss). We now define a basic notion of learnability based on discretization which essentially serves to rule out adversarial constructions in the vein of Proposition 20.

Definition 22 (Discretely-learnable) *We say $(\mathscr{D}, X, H, \ell)$ is discretely-learnable with sample complexity $n(\varepsilon, \delta)$ if there is some constant $c_1 > 0$ such that for all $\varepsilon, \delta > 0$ there exists an ε-discretization H_ε which is $(c_1\varepsilon, \delta)$-PAC-learnable in at most $n(\varepsilon, \delta)$ samples. We call the learner proper if it outputs hypotheses in H.*

We'll prove that discrete and agnostic learnability are equivalent as long as the loss satisfies an approximate triangle inequality.

---

17. We note that in most cases this condition can generally be weakened to the expected guarantee $\mathbb{E}_{x \sim D_X}[\ell(h'(x), h(x))] \leq \varepsilon$, but the stronger notion is needed for some applications such as adversarial robustness.

Definition 23 (Approximate pseudometric) *We call a loss function $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ a c-approximate pseudometric if for all triples $y_1, y_2, y_3 \in Y$:*

$$\ell(y_1, y_3) \leq c(\ell(y_1, y_2) + \ell(y_2, y_3)).$$

Approximate pseudometrics are natural choices for loss functions in practice and capture a broad set of scenarios including finite-range losses and standard setups such as ℓ_p-regression, and have seen some previous study in the literature (Crammer et al., 2008). By modifying the first step of our reduction to take discretization into account and leveraging the approximate triangle inequality in the second, we prove that discrete learnability and c-agnostic learnability are equivalent under c-approximate pseudometrics.

Theorem 24 *Let $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ be a bounded c-approximate pseudometric. Then the following are equivalent for all $(\mathscr{D}, X, H, \ell)$:*

1. *$(\mathscr{D}, X, H, \ell)$ is discretely-learnable.*

2. *$(\mathscr{D}, X, H, \ell)$ is c-agnostically learnable.*

Proof The proof is similar to Theorem 18. We first show the forward direction. Assume $(\mathscr{D}, X, H, \ell)$ is discretely-learnable. Fix $\varepsilon' = \frac{\varepsilon}{4c^2 c_1}$ (where c_1 is the constant from Definition 22), and let $H_{\varepsilon'}$ be a learnable ε'-discretization of H. We argue that running LEARNINGTOCOVER on $H_{\varepsilon'}$ gives the desired agnostic learner. Since ℓ is bounded, it is sufficient to prove that $C(S_U)$ contains a hypothesis h' such that $\mathrm{err}_{D,\ell}(h') \leq c \cdot OPT + \varepsilon/2$. Empirical risk minimization then works as in the finite case.

Let $h_{OPT} \in H$ be an optimal hypothesis. Since $H_{\varepsilon'}$ is a discretization of H, there exists $h_{OPT}^{\varepsilon'} \in H_{\varepsilon'}$ such that:

$$\forall x \in X : \ell(h_{OPT}(x), h_{OPT}^{\varepsilon'}(x)) < \varepsilon'.$$

Further, by the guarantees of discrete learnability, with probability at least $1 - \delta/2$ there exists $h' \in C(S_U)$ such that close to $h_{OPT}^{\varepsilon'}$ in the following sense:

$$\mathop{\mathbb{E}}_{x \sim D_X}[\ell(h'(x), h_{OPT}^{\varepsilon'}(x))] \leq c_1 \varepsilon' = \frac{\varepsilon}{4c^2}$$

Plugging in the previous observation and applying our approximate triangle inequality, we get that h' is close to h_{OPT} in the following sense:

$$\mathop{\mathbb{E}}_{x \sim D_X}[\ell(h'(x), h_{OPT}(x))] \leq c \left(\mathop{\mathbb{E}}_{x \sim D_X}[\ell(h'(x), h_{OPT}^{\varepsilon'}(x))] + \mathop{\mathbb{E}}_{x \sim D_X}[\ell(h_{OPT}^{\varepsilon'}(x), h_{OPT}(x))] \right)$$
$$\leq \frac{\varepsilon}{2c}$$

The final step is to transfer from the marginal D_X to the full joint distribution of the adversary, which follows immediately from a similar application of the approximate triangle inequality. This is the only step that loses a factor in the OPT term:

$$\mathrm{err}_{D,\ell}(h') = \mathop{\mathbb{E}}_{(x,y) \sim D}[\ell(h'(x), y)]$$
$$\leq c \left(\mathop{\mathbb{E}}_{(x,y) \sim D}[\ell(h'(x), h_{OPT}(x))] + \mathop{\mathbb{E}}_{(x,y) \sim D}[\ell(h_{OPT}(x), y)] \right)$$
$$\leq c \cdot OPT + \varepsilon/2$$

as desired.

We now prove the reverse direction, which is essentially immediate. Assume the existence of a c-agnostic learner for $(\mathscr{D}, X, H, \ell)$. Given a discretization H_ε, we want to show $(\mathscr{D}, X, H_\varepsilon, \ell)$ is learnable to within $c_1 \varepsilon$ error for some $c_1 > 0$. This is achieved simply by running the agnostic learner on $(\mathscr{D}, X, H_\varepsilon, \ell)$. Since H_ε is "always useful", every $h \in H_\varepsilon$ is ε-close to some $h' \in H$ in the sense that:

$$\forall x \in X : \ell(h'(x), h(x)) \leq \varepsilon.$$

In particular, this means that for any choice of h by the adversary there exists $h' \in H$ with low error:

$$\mathrm{err}_{D,\ell}(h') = \mathbb{E}[\ell(h'(x), h(x))] \leq \varepsilon.$$

As a result, running the c-agnostic learner for $(\mathscr{D}, X, H, \ell)$ returns a hypothesis of at most $(c+1)\varepsilon$ error with high probability. ∎

It is worth noting that bounded loss is not really necessary for Theorem 24. More generally we can require that $(\mathscr{D}, X, H, \ell)$ is "finitely learnable" in the sense that for all finite subsets $H' \subset H$, $(\mathscr{D}, X, H', \ell)$ is agnostically learnable. When ℓ is bounded, this is true for any finite class by empirical risk minimization.

It is also worth noting that various modifications to the definition of loss (e.g. defining loss between hypotheses rather than on Y directly) will continue to work with the above. Similarly, there are various cases when one can get better than $c \cdot OPT$ accuracy for a c-approximate pseudometric, generally by instead optimizing over some surrogate loss function. For instance, if a simple transformation of the loss gives a c'-approximate pseudometric for $c' < c$, then one can generally learn up to $c' \cdot OPT$.[18] As an example, note that while square loss $\ell_2(x, y) = (x - y)^2$ is a 2-approximate pseudometric, taking $\sqrt{\mathrm{err}_{D,\ell_2}}$ gives a true metric between hypotheses. As a result, as long as OPT is bounded, we can get truly agnostic learning by optimizing $\sqrt{\mathrm{err}_{D,\ell_2}}$ instead. This strategy works for any polynomial loss, such as $\ell_p(x, y) = |x - y|^p$.

On the other hand, outside of these special cases, Theorem 24 is tight: there exist c-approximate pseudometric loss functions which cannot be c'-agnostically learned for any $c' < c$. The argument is similar to Proposition 20, but requires a bit more care.

Proposition 25 *There exists a discretely-learnable class over a c-approximate pseudometric that is not c'-agnostically learnable for any $c' < c$.*

Proof The proof is similar to Proposition 19. We consider the same instance space $X = \mathbb{N}$ and hypothesis class H:

$$H = \{h : h(x) = (0, \cdot) \,\forall\, x \in X\}.$$

The loss function $\ell : Y \times Y \to \{0, 1, c\}$ is also the same, but extended to the larger domain $Y = \mathbb{N}^2$:

$$\ell((b_1, r_1), (b_2, r_2)) = \begin{cases} 0 & b_1 = b_2 \\ 1 & b_1 \neq b_2 \text{ and } r_1 = r_2 \\ c & \text{otherwise.} \end{cases}$$

---

18. This does require that OPT is bounded.

As before, note that (X, H, ℓ) is trivially realizably learnable by always returning any $h \in H$, ℓ is a c-psuedometric by definition, and for any labeling $f : X \to Y$ there exists $h \in H$ such that for all distributions D:

$$OPT \leq err_{D,f}(h) \leq 1.$$

We now show that the class (X, H, ℓ) is only c-agnostically learnable. Since $OPT \leq 1$, it suffices to show that for every $m \in \mathbb{N}$, large enough $n \in \mathbb{N}$, and randomized algorithm $\mathcal{A}$ on m samples, there exists a labeling $f : X \to Y$ and a marginal distribution D_X such that:

$$\mathbb{E}_{S \sim D_X^m} \mathbb{E}_{x \sim D_X} [\ell(\mathcal{A}(S, f(S))(x), f(x))] \geq \left(1 - \frac{1}{n}\right)^3 c. \tag{2}$$

For $n \geq \frac{1}{1 - (1 - (c - c')/c)^{1/3}}$, applying Markov's inequality to Equation (2) implies that $\mathcal{A}$ has error at least c' with constant probability.

For simplicity, we now restrict our attention to the marginal D_X which is uniform over the set $[k]$ for some $k \in \mathbb{N}$ to be fixed. By Yao's minimax principle, its enough to prove that there exists a *distribution μ over functions $f : [k] \to [n]^2$* such that any *deterministic* algorithm $\mathcal{A}$ the following holds

$$\mathbb{E}_{f \sim \mu} \mathbb{E}_{S \sim D_X^m} \mathbb{E}_{x \sim D_X} [\ell(\mathcal{A}(S, f(S))(x), f(x))] \geq \left(1 - \frac{1}{n}\right)^3 c.$$

We now show that the above holds for μ being uniform over all functions from $[k]$ to $[n]^2$ when $k > \frac{2m}{\ln(n/(n-1))}$. Similar to Proposition 19, we have that

$$\mathbb{E}_{x \sim D_X} \mathbb{E}_{f \sim \mu} \mathbb{E}_{S \sim D_X^m} [\ell(\mathcal{A}(S, f(S))(x), f(x))] \geq \mathbb{E}_{x \sim D_X} \left[\Pr_{S \sim D_X^m} [x \notin S] \cdot \left(1 - \frac{1}{n}\right)^2 \cdot c \right],$$

since no matter the assignment $\mathcal{A}$ gives to $x \notin S$, it will be wrong on both coordinates with probability $(1 - 1/n)^2$ over the randomness of μ. The result follows by noting that for every $x \in [k]$

$$\Pr_{S \sim D_{X \times f(X)}^m} [(x, \cdot) \notin S] = \left(1 - \frac{1}{k}\right)^m \geq 1 - \frac{1}{n}$$

since $1 - x \geq \exp\{-x/(1 - x)\}$, and we have assumed $k > \frac{2m}{\ln(n/(n-1))}$ and $n > 1$. Therefore, we get that

$$\mathbb{E}_{x \sim D_X} \mathbb{E}_{f \sim \mu} \mathbb{E}_{S \sim D_X^m} [\ell(\mathcal{A}(S, f(S))(x), f(x))] \geq \left(1 - \frac{1}{n}\right)^3 c$$

which completes the proof. ■

C.2. Sub-sampling: Malicious Noise

Now that we've seen how to handle practical problems like regression over infinite label spaces, we'll discuss a technique that helps handle data corruption and data-dependent assumptions: sub-sampling. The main idea is as follows. Say that the original unlabeled sample we draw is, in some sense, partially corrupted: perhaps an adversary has changed some fraction of examples (malicious noise), or some portion of the sample is un-realizable for a concept in the class (robust and partial

learning). In either case, there generally exists a core subset of "clean" samples that we can use to recover the guarantees of LEARNINGTOCOVER. Since we cannot necessarily *identify* these, the idea is to run LEARNINGTOCOVER over enough subsets of the unlabeled sample that we find a clean subsample with high probability. In this section we'll discuss the application of this technique in detail to Kearns and Li (1993)'s well-studied malicious noise model. In the appendix, we discuss applications to recently popular adversarially robust (Appendix E) and partial learning (Appendix F) models.

To start, let's recall the standard malicious noise model. In this variant of PAC learning, instead of having access to the standard sample oracle from the adversary's distribution D over $X \times Y$, we have access to a *malicious oracle* $\mathcal{O}_M(\cdot)$ which, with probability η, outputs an adversarially chosen pair (x, y), and otherwise samples from D as usual.

Definition 26 (PAC-learning with Malicious Noise) *We call $(X, H, \mathcal{D}, \ell)$ (agnostically) (ε, δ)-learnable with malicious noise at rate $\eta = \eta(\varepsilon)$ if there exists an algorithm $\mathcal{A}$ and function $m = m_{mal}(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$, and distributions D over $X \times Y$ satisfying:*

1. The marginal $D_X \in \mathcal{D}$,

$\mathcal{A}$ outputs a good hypothesis with high probability over samples drawn from the malicious oracle of size $n(\varepsilon, \delta)$:

$$\Pr_{S \sim \mathcal{O}_M(\cdot)^m}[err_{D,\ell}(\mathcal{A}(S)) > OPT + \varepsilon] \leq \delta.$$

where $OPT = \min_{h \in H}\{err_{D,\ell}(h)\}$.

In other words, malicious noise essentially gives a worst-case formalization of the idea that an η-fraction of the learner's data is (adversarial) garbage.

Let's now formalize the argument above: modifying LEARNINGTOCOVER to run over subsamples gives a sample-efficient algorithm for learning with malicious noise. For readability, we'll (somewhat informally) restate the algorithm with this change.

Algorithm 3: Malicious to Realizable Reduction

Input: Realizable PAC-Learner $\mathcal{A}$, Accuracy Parameter $\varepsilon < 1/2$, Noise Parameter $\eta < \frac{\varepsilon}{1+\varepsilon}$, Unlabeled Sample Oracle $\mathcal{O}_U$, Labeled Sample Oracle $\mathcal{O}_L$

Algorithm:

1. Draw an unlabeled sample $S_U \sim \mathcal{O}_U$, and labeled sample $S_L \sim \mathcal{O}_L$.

2. Run LEARNINGTOCOVER over all $S \in (S_U)_{\eta'} := \{S \subseteq S_U : |S| = \lfloor(1 - \eta')|S_U|\rfloor\}$, where:

$$\eta' = \frac{3\eta + \varepsilon/(1 - \varepsilon)}{4}.$$

3. **Return** the hypothesis in $C(S_U) := \bigcup_{S \in (S_U)_{\eta'}} C(S)$ with lowest empirical error over S_L.

We now prove that Algorithm 3 gives an (agnostic) learner that is tolerant to malicious noise.

Theorem 27 *Let $(\mathscr{D}, X, H)$ be realizably PAC-learnable with sample complexity $n(\varepsilon, \delta)$. Then for any $\eta < \frac{\varepsilon}{1+\varepsilon}$, Algorithm 3 is an agnostic learner for $(\mathscr{D}, X, H)$ tolerant to η malicious noise. Furthermore letting $\Delta = \frac{\varepsilon}{1+\varepsilon} - \eta$ and $\beta = \left(1 + \frac{\eta}{\Delta}\right) \log\left(\frac{1}{\eta}\right)$, its sample complexity is at most:*

$$m_{mal}(\varepsilon, \delta) \leq O\left(\beta \frac{n(\Delta/4, \delta/4)}{\Delta} + \frac{\log\left(\Pi_H\left(O\left(n(\Delta/2, \delta/2) + \frac{\eta^2 \log(1/\delta)}{\Delta^2}\right)\right)\right) + \log(1/\delta)}{\Delta^2} + \beta \frac{\eta^2 \log(1/\delta)}{\Delta^3}\right)$$

where we've assumed $\varepsilon < 1/2$ for simplicity.

Proof To start, we'll review for completeness a fairly standard analysis of empirical risk minimization under malicious noise. Assume for the moment that the output of LEARNINGTOCOVER, $C(S_U)$, contains a hypothesis h' satisfying $err(h') \leq OPT + \beta_1$. Say we draw M labeled samples for the ERM step, and an $\eta' = \eta + \beta_2$ fraction are corrupted by the adversary. For large enough M, we can assume by a Chernoff and union bound that the empirical loss of every hypothesis returned by LEARNINGTOCOVER is at most some β_3 away from its true loss on the un-corrupted portion of M (we will make all these assumptions formal in a moment). Given these facts, notice that the empirical loss of h' on M is at most:

$$err_M(h') \leq (1 - \eta')(OPT + \beta_1 + \beta_3) + \eta'.$$

On the other hand, the empirical error of any $h_\varepsilon \in H$ whose true error is greater than $OPT + \varepsilon$ is at least:

$$err_M(h) > (1 - \eta')(OPT + \varepsilon - \beta_3).$$

To ensure that our ERM works, it is enough to show that for any such h, $err_M(h) > err_M(h')$. A simple calculation shows that this is satisfied as long as $\beta_1 + 2\beta_3 \leq \varepsilon$ and $\beta_1 + \beta_2 + 2\beta_3 \leq \Delta$. Setting $\beta_1 = \beta_2 = \beta_3 = \Delta/4$ gives the desired result.

It is left to argue that our assumptions above hold with high probability. First, note that by a Chernoff bound, the probability that $\eta' \geq \eta + \Delta/4$ on a set of M samples is at most $e^{-c\left(\frac{\Delta}{\eta}\right)^2 M}$ for some $c > 0$, so this occurs with high probability as long as $M \geq \Omega(\log(1/\delta)\eta^2/\Delta^2)$. Similarly, the empirical error of every hypothesis in $C(S_U)$ on the remaining clean samples will be within $\Delta/4$ of its true error as long as $M \geq \Omega(\log(|C(S_U)|/\delta)\eta^2/\Delta^2)$.

It is then left to show that $C(S_U)$ contains a hypothesis of error at most $OPT + \Delta/4$. To show this, it is enough to ensure that we run LEARNINGTOCOVER over a clean subsample of size at least $n(\Delta/4, \delta/4)$ with high probability. If we draw $|S_U| = O\left(\frac{n(\Delta/4, \delta/4)}{1 - \eta'} + \log(1/\delta)\frac{\eta^2}{\Delta^2}\right)$ unlabeled samples, a similar Chernoff bound to the above promises that at most an η' fraction are corrupted with high probability, and therefore that at least $n(\Delta/4, \delta/4)$ remain un-corrupted. Running LEARNINGTOCOVER over all subsets of size $(1 - \eta')|S_u|$ then gives the desired result. The sample complexity bound follows from choosing M large enough to satisfy the above conditions along with the fact that $|C(S_U)| \leq \binom{n}{(1-\eta')n}\Pi_H(n)$. ∎

It's worth noting that the error tolerance of Theorem 27 is tight. In their original introduction of malicious noise, Kearns and Li (1993) proved that for most non-trivial concept classes, no PAC-learner can be tolerate $\frac{\varepsilon}{1+\varepsilon}$ malicious noise. Theorem 27 also extends to other scenarios we've seen

so far such as arbitrary loss over finite label classes and approximate pseudometrics. The proof remains mostly the same, though the optimal error tolerance may differ.

Since our agnostic model restricts the adversary to choosing a distribution whose marginal lies in the original family, Theorem 27 provides the first insight on robustness against an adversary who can corrupt the *underlying data* as well as the labels. One might wonder whether this result can be pushed further: is it possible to be robust against an adversary who can corrupt the marginal over X in some stronger sense? Unfortunately, the answer is no: malicious noise is necessarily the most distributional corruption we can handle. Let's look at two basic lower bounds to see why. First, we'll consider an adversary who can *remove* a portion of the learner's sample.

Proposition 28 *For any $\delta > 0$, there exists a class $(\mathscr{D}, X, H)$ which is realizably PAC-learnable, but not learnable under an adversary who can remove a δ fraction of the learners sample.*

Proof This follows from a result of Dudley, Kulkarni, Richardson, and Zeitouni (Dudley et al., 1994) that there exists an unlearnable class $(\mathscr{D}, X, H)$ such that for some $n(\varepsilon, \delta)$, (D, X, H) is learnable in $n(\varepsilon, \delta)$ samples for every $D \in \mathscr{D}$. The lower bound then follows simply from adding an extra unique identifying point x_D to X for every distribution D, and modifying each $D \in \mathscr{D}$ to have $\Theta(\delta)$ support on x_D. This modified class is clearly learnable, since after drawing $O(1/\delta)$ samples, the learner will draw x_D and identify the distribution D with good probability. However, the class is not learnable under an adversary who removes points, since with high probability the adversary can completely remove any mention of x_D from the learner's sample, reducing to the original unlearnable class $(\mathscr{D}, X, H)$. ■

An adversary who can *add* samples is similarly powerful. In the realizable setting, if the adversary is allowed to add an arbitrary number of correctly labeled points to the learners sample, basic classifiers such as halfspaces become unlearnable (Blum et al., 2021). On the other hand if the adversary is limited to adding only a few additional samples, realizable learning may remain possible, but even trivial concept classes cannot be agnostically learnable.

Proposition 29 *There exists a class (X, H) which for any $\gamma > 0$ is realizably but not agnostically learnable under an adversary who can add a γ fraction of correctly labeled[19] points to the learner's sample.*

Proof Let $X = \{x, x_1, x_2\}$ and $H = \{h_1, h_2\}$ be any class such that $h_1(x) = h_2(x)$, but $h_1(x_i) \neq h_2(x_i)$ for $i = 1, 2$. In the realizable setting, note that a single labeled example on x_1 or x_2 exactly determines the hypothesis. As long as there is less than $1 - \varepsilon$ mass on x, the learner will draw such a sample after $O(1/\varepsilon)$ samples with good probability. Further, if the mass on x is at least $1 - \varepsilon$, then h_1 and h_2 are both valid outputs. As a result, any ERM is a valid PAC-learner. Since adding correctly labeled examples can only help this learner, the class remains realizably learnable under an adversary who can add an arbitrary number of clean samples.

In the agnostic setting, consider an adversary who chooses a labeling f such that $f(x) = h_1(x) = h_2(x)$, $f(x_1) = h_1(x_1)$, and $f(x_2) = h_2(x_2)$. The optimal hypothesis h_{OPT} is then decided by the amount of mass on x_1 and x_2 in the marginal distribution. Namely if the adversary

---

19. For simplicity, we'll assume the adversary is restricted to picking a marginal distribution and labeling rather than any joint distribution over $X \times Y$.

chooses a distribution D over $\{x, x_1, x_2\}$, the optimal error is $\min\{D(x_1), D(x_2)\}$. The idea is then to note that for $\gamma' \leq \gamma/4$, the learner cannot distinguish between the two following distributions:

$$D_1(x_1) = c_1\gamma', D_1(x_2) = (1 - c_1)\gamma'$$

and

$$D_2(x_1) = (1 - c_1)\gamma', D_2(x_2) = c_1\gamma'$$

where $1/2 > c_1 > 0$ is some small constant. Informally, if the two distributions are indistinguishable, any learner will always incur error of around $\frac{(1-c_1)\gamma'}{2}$, whereas OPT is $c_1\gamma'$ for both distributions.

Let's now give the formal argument. By Yao's Minimax Principle it is enough to prove there is a strategy over distributions such that any deterministic learner has high error. In particular, if we can prove that the expected error is at least $3 \cdot OPT$, then $\Pr[\text{error} \geq 2 \cdot OPT] \geq OPT$. Since OPT is just some constant $c_1\gamma'$ (dependent only on γ), this is sufficient to prove the result. Moving on, consider the strategy in which the adversary chooses the labeling described above, and chooses each marginal (D_1 or D_2) with probability $1/2$. We'll break our analysis into two cases dependent on the sample complexity of the learner. If the learner uses $O(1/\gamma')$ examples, then there is a constant probability of drawing a sample only consisting of the point x. Let f' be the hypothesis returned by the deterministic learner on this sample. By construction, f' must disagree with either h_1 or h_2 on x_1 or x_2. Assume f' differs on $h_1(x_1)$ (the other cases will follow similarly). When the distribution is D_2, f' has error at least $(1 - c_1)\gamma'$. Since this occurs with constant probability *independent* of the choice of c_1, choosing c_1 sufficiently small leads to an expected error of at least $3c_1\gamma'$ as desired.

On the other hand, when there are $n = \Omega(1/\gamma')$ samples, we claim that the adversary can force the following sample to occur with constant probability: $2\gamma'n$ instances of x_1 and x_2, and $(1 + \gamma)n - 4\gamma'n$ instances of x. This follows from the fact that for the appropriate choice of constant for n, a Chernoff bound gives that both x_1 and x_2 occur at most $2\gamma'n$ times with constant probability. Since the adversary is allowed to add $\gamma n \geq 4\gamma'n$ arbitrary examples, they can add instances of x_1, x_2, and x until the above sample is achieved. The remainder of the argument is then the same as the previous case, as any learner response on this sample will incur similarly high expected error. ■

It is also reasonable to consider distributional corruption in the semi-supervised setting, where the unlabeled and labeled data-sets might have different underlying distributions. We discuss this model in Appendix .

C.3. Replacing ERM: Semi-Private Learning

So far we have focused on property generalization for two forms of noise-tolerance—agnostic learning and learning with malicious noise. In this section, we'll show how to use Algorithm 1 to generalize a broader spectrum of finitely-satisfiable properties through replacing the ERM process with a generic finite learner with the desired property. Our prototypical example will be *privacy*, which is well known to be finitely-satisfiable via McSherry and Talwar (2007)'s *exponential mechanism*. To start, we'll cover a few basic privacy definitions.

Definition 30 (Differential Privacy) *A learning algorithm $\mathcal{A}$ is said to be α-differentially private if for all neighboring inputs S, S' which differ on a single example:*

$$\Pr[\mathcal{A}(S) \in T] \leq e^\alpha \Pr[\mathcal{A}(S') \in T],$$

for all measurable events T in the range of $\mathcal{A}$.

The exponential mechanism is one of the most widely used techniques in privacy. Informally, the algorithm allows for differentially private selection of a "good" choice from a finite set of objects (potential hypotheses in our case). More formally, let $s : (X \times Y)^* \times H \to \mathbb{R}$ be a "score" function, and define "sensitivity" Δ_s to be

$$\Delta_s = \max_{h \in H} \max_{S,S'} |s(S,h) - s(S',h)|$$

where S, S' are two neighbouring datasets. The exponential mechanism selects an item with a good score with high probability, while maintaining privacy.

Definition 31 (Exponential Mechanism McSherry and Talwar (2007)) *The exponential mechanism M_E on inputs S, H, s with privacy parameter α selects and outputs $h \in H$ with probability*

$$\frac{\exp(\frac{\alpha s(S,h)}{2\Delta})}{\sum_{h' \in H} \exp(\frac{\alpha s(S,h')}{2\Delta})}.$$

It is well known that the exponential mechanism leads to a private learner for finite hypothesis classes under bounded loss.

Theorem 32 (Theorem 3.4 Kasiviswanathan et al. (2011)) *Let $(\mathscr{D}, X, H, \ell)$ be a finite class with a bounded loss function. Then the sample complexity of α-differentially private learning $(\mathscr{D}, X, H, \ell)$ is at most:*

$$n_{pri}(\alpha, \varepsilon, \delta) \leq O\left(\log\left(\frac{|H|}{\delta}\right) \max\{\varepsilon^{-2}, \varepsilon^{-1}\alpha^{-1}\}\right).$$

We note that (Kasiviswanathan et al., 2011, Theorem 3.4) only considers classification loss, but the extension to bounded loss is immediate. Unfortunately, even with the power of the exponential mechanism, privacy is a very restrictive condition in the general PAC framework, since we're most often interested in *infinite* hypothesis sets. Indeed even *improper* private learning requires finiteness of a highly restrictive measure known as representation dimension (Beimel et al., 2013a), which can be infinite for classes of VC dimension 1. As a result, the past decade has seen the introduction of a number of weaker, more practical definitions of privacy. In this section we'll focus on a model introduced in 2013 by Beimel, Nissim, and Stemmer (Beimel et al., 2013b) called *semi-private learning*.

Definition 33 (Semi-Private Learning) *We call a class $(\mathscr{D}, X, H, \ell)$ semi-private PAC-Learnable if there exists an algorithm $\mathcal{A}$ and two functions $n_{pub} = n_{pub}(\varepsilon, \delta, \alpha)$ and $n_{pri} = n_{pri}(\varepsilon, \delta, \alpha)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ whose marginal D_X is in $\mathscr{D}$, $\mathcal{A}$ satisfies the following:*

1. *$\mathcal{A}$ outputs a good hypothesis with high probability:*

$$\Pr_{S_U \sim D_X^{n_{pub}}, S_L \sim D^{n_{pri}}} [err_{D,\ell}(\mathcal{A}(S_U, S_L)) > OPT + \varepsilon] \leq \delta.$$

2. *$\mathcal{A}$ is semi-private. That is for all $S_U \in X^{n_{pub}}$:*

$$\mathcal{A}(S_U, \cdot) \text{ is } \alpha\text{-differentially private.}$$

In other words, semi-private learning offers a model for applications where *labeled data* is sensitive, but some (perhaps opt-in) users might not care about their participation itself being released. Unlike standard private learning, distribution-free semi-private classification is known to be characterized by VC dimension, just like realizable PAC-learning (Beimel et al., 2013b). The best sample complexity bounds are due to Alon, Bassily, and Moran (ABM) (Alon et al., 2019a), who use uniform convergence to build a uniform cover for H from unlabeled data, and then apply the exponential mechanism to the resulting cover.

Due to their reliance on uniform convergence, ABM's techniques fail in the more general settings we consider. Further, their use of uniform covers results in sub-optimal public sample complexity even for distribution-free classification. We prove in Appendix J that these objects require asymptotically more samples than non-uniform covers (at least in the distribution-family model), and therefore cannot be used to achieve optimal semi-private learning. We circumvent both of these issues by appealing directly to a realizable learner to build a weaker non-uniform cover. For readability, we first restate the algorithm here.

Algorithm 4: Semi-Private to Realizable Reduction

Input: Realizable PAC-Learner $\mathcal{A}$, Unlabeled Sample Oracle $\mathcal{O}_U$, Labeled Sample Oracle $\mathcal{O}_L$
Algorithm:

1. Draw an unlabeled sample $S_U \sim \mathcal{O}_U$, and labeled sample $S_L \sim \mathcal{O}_L$.

2. Run LEARNINGTOCOVER over S_U to get $C(S_U)$.

3. **Return** the hypothesis in $C(S_U)$ given by applying the exponential mechanism with respect to S_L.

We prove that Algorithm 4 gives a semi-private agnostic learner in the distribution-family setting.

Theorem 34 (PAC-learning implies Semi-Private Learning) *Let $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ be a bounded c-approximate pseudometric. Then the following are equivalent for all triples $(\mathscr{D}, X, H, \ell)$:*

1. *$(\mathscr{D}, X, H, \ell)$ is discretely-learnable*

2. *$(\mathscr{D}, X, H, \ell)$ is c-agnostically, semi-private learnable.*

Proof The proof is essentially the same as Theorem 24. The only difference in the argument is to replace the generic ERM learner over the output of LEARNINGTOCOVER with the exponential mechanism (Kasiviswanathan et al., 2011). ∎

Let's now take a look at what Theorem 34 implies about the special case of distribution-free classification.

Corollary 35 *Let (X, H) be a class of VC-dimension d with sample complexity $n(\varepsilon, \delta)$. The sample complexity of (agnostic) semi-private learning (X, H) is at most:*

$$n_{pub}(\varepsilon, \delta, \alpha) \leq n(\varepsilon/2, \delta/2)$$

35

and

$$n_{pri}(\varepsilon, \delta, \alpha) \leq O\left(\left(d\log\left(\frac{n(\varepsilon/2, \delta/2)}{d}\right) + \log\left(\frac{1}{\delta}\right)\right)\max\{\varepsilon^{-2}, \varepsilon^{-1}\alpha^{-1}\}\right).$$

Further, the sample complexity of improperly (agnostic) semi-private learning (X, H) is

$$n_{pub}(\varepsilon, \delta, \alpha) \leq O\left(\frac{d + \log(1/\delta)}{\varepsilon}\right)$$

and

$$n_{pri}(\varepsilon, \delta, \alpha) \leq O\left(\left(d\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\max\{\varepsilon^{-2}, \varepsilon^{-1}\alpha^{-1}\}\right).$$

Proof LEARNINGTOCOVER uses $n(\varepsilon/2, \delta/2)$ samples by definition. In the improper case, Hanneke showed that $n(\varepsilon/2, \delta/2) \leq O\left(\frac{d + \log(1/\delta)}{\varepsilon}\right)$. Since the class has VC dimension d, the size of the resulting cover is at most $(e \cdot n(\varepsilon/2, \delta/2)/d)^d$, and the private sample complexity bound then follows from Theorem 32. ∎

This improves over the recent upper bound of ABM who showed that

$$n_{pub}(\varepsilon, \delta, \alpha) \leq O\left(\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

In fact, for constant d and δ, Corollary 35 completely resolves the unlabeled sample complexity of semi-private learning, as ABM (Alon et al., 2019a) prove a matching lower bound.

Theorem 36 (Public Lower Bound Alon et al. (2019a)) *Every class that is not privately learnable requires at least $\Omega(1/\varepsilon)$ unlabeled samples to learn in the semi-private model under classification error.*

On the other hand, we note that the *private* sample complexity remains off by a log factor from the best known lower bounds of Chaudhuri and Hsu (2011).

Theorem 37 (Private Lower Bound Chaudhuri and Hsu (2011)) *There exist classes of VC dimension $O(1)$ which require at least:*

$$n_{pri}(\varepsilon, \delta, \alpha) \geq \Omega\left(\max\{\varepsilon^{-2}, \varepsilon^{-1}\alpha^{-1}\}\right)$$

private samples to learn.

While we have now resolved the public sample complexity of improper learning, it remains an interesting open problem for the proper regime where certain adversarial examples are known to require an extra $\log(1/\varepsilon)$ factor in the standard PAC sample complexity (Daniely and Shalev-Shwartz, 2014; Hanneke, 2019). We conjecture that Theorem 34 should still be tight in this setting: namely that the unlabeled semi-private sample complexity should always be at least the realizable PAC sample complexity.

Conjecture 38 *Let (X, H) be a hypothesis class which is not privately learnable. Then the realizable sample complexity of (X, H) lower bounds the unlabeled sample complexity of proper semi-private learning:*

$$n_{pub}(\varepsilon, 1/2) \geq \Omega(n(\varepsilon, 1/2)).$$

C.4. Changing the Base Model: Covariate Shift

One issue with semi-supervised models like semi-private learning is that, in practice, the distribution over unlabeled data probably won't match the labeled data exactly. In this section, we'll talk about a final modification to our reduction to tackle such scenarios and more generally to extend property generalization beyond the realizable PAC setting: replacing the base learner. In fact, we already saw this strategy used to a lesser extent in Appendix C.1, where we replaced our standard realizable base learner with a discrete learner. Here we'll look at an application in which we assume our initial learner is robust to *covariate shift* (Shimodaira, 2000), meaning that even if the distribution underlying the data shifts between train and test time, the algorithm will continue to perform well. This stronger assumption will allow us to build semi-private learners that can handle corruption between the public and private databases. To start, let's formalize covariate shift in the distribution-family model.

Definition 39 (Covariate Shift) *Let $(\mathscr{D}, X, H, \ell)$ be any class, and for every $\varepsilon > 0$ let C_ε be a "covariate-shift" function that maps every $D \in \mathscr{D}$ to some family of distributions over X. Given any distribution $D \in \mathscr{D}$ and any $h \in H$, let the error of a potential labeling be given by its worst-case error over $C_\varepsilon(D)$, that is:*

$$CS\text{-}err_{D \times h, \ell, \varepsilon}(c) = \max_{D' \in C_\varepsilon(D)} \left\{ \mathbb{E}_{x \sim D'}[\ell(c(x), h(x))] \right\}.$$

We say that $(\mathscr{D}, X, H, \ell)$ is realizably learnable under covariate shift $C = \{C_\varepsilon\}$ if there exists an algorithm $\mathcal{A}$ and function $n = n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$, $D \in \mathscr{D}$, and $h \in H$:

$$\Pr_{S \sim D^n}[CS\text{-}err_{D \times h, \ell, \varepsilon}(\mathcal{A}(S, h(S))) > \varepsilon] \leq \delta.$$

We call such a learner robust to covariate shift.

We emphasize that in the above definition, the covariate shift family *scales with the error parameter ε*. This is a bit different than Shimodaira (2000)'s original definition, but is a natural choice in our context since we consider algorithms which only use access to the original source distribution (sometimes called "conservative domain adaptation" (David et al., 2010)). In this setting, we'd expect that as we demand higher accuracy, the amount of covariate shift we can tolerate will decrease. Indeed in the agnostic model, it's clear this scaling is necessary by a similar argument to Proposition 29.

   The key observation to apply learning under covariate shift in our reduction is simply to notice that the non-uniform cover output by LEARNINGTOCOVER must contain a hypothesis close to optimal under any shifted distribution in $C_\varepsilon(D)$. This can then be used to analyze any semi-supervised model where the marginal of the labeled distribution may be corrupted from D to any distribution in $C_\varepsilon(D)$. In this section, we'll again focus on the setting of semi-private learning. First, let's formalize what it means to be semi-private learnable under covariate shift.

Definition 40 (Semi-Private Learning under Covariate Shift) *We call a class $(\mathscr{D}, X, H, \ell)$ semi-private (agnostically) PAC-Learnable under covariate shift $C = \{C_\varepsilon\}$ if there exists an algorithm $\mathcal{A}$ and two functions $n_{pub} = n_{pub}(\varepsilon, \delta, \alpha)$ and $n_{pri} = n_{pri}(\varepsilon, \delta, \alpha)$ such that for all $\varepsilon, \delta > 0$ and distributions $D_X \in \mathscr{D}$ and D' over $X \times Y$ whose marginal $D'_X \in C_\varepsilon(D_X)$, $\mathcal{A}$ satisfies the following:*

1. $\mathcal{A}$ outputs a good hypothesis over D' with high probability:

$$\Pr_{S_U \sim D_X^{n_{pub}}, S_p \sim D'^{n_{pri}}}[err_{D',\ell}(\mathcal{A}(S_U, S_p)) > OPT' + \varepsilon] \le \delta,$$

where $OPT' = \min_{h \in H}[err_{D',\ell}(h)]$ is the minimum error over the **shifted** distribution.

2. $\mathcal{A}$ is semi-private. That is for all $S_U \in X^{n_{pub}}$:

$$\mathcal{A}(S_U, \cdot) \text{ is } \alpha\text{-differentially private}$$

In other words, we'd like to recover a near-optimal hypothesis even when the marginal distribution over private data is shifted from the public data. This is a realistic scenario in practice, since the distribution of "opt-in" users is likely different from the marginal over the total population. We'll show that this issue is solvable in the semi-private setting as long as the analogous issue in the non-private setting (distribution shift between train and test time) can be resolved.

Theorem 41 *Let $(\mathscr{D}, X, H, \ell)$ be any class on a finite label space with loss function satisfying the identity of indiscernibles which is realizably learnable under covariate shift $C = \{C_\varepsilon\}$. Then $(\mathscr{D}, X, H, \ell)$ is semi-private, agnostically learnable under covariate shift C with sample complexity:*

$$m(\varepsilon, \delta) \le n(\eta_\ell \varepsilon, \delta/2) + O\left(\frac{\log\left(\frac{\Pi_H(n(\eta_\ell \varepsilon, \delta/2))}{\delta}\right)}{\varepsilon^2}\right)$$

where we recall $\eta_\ell \ge \Omega\left(\frac{\min_{a \ne b}(\ell(a,b))}{\max_{a \ne b}(\ell(a,b))}\right)$ is a constant depending only on ℓ.

Proof The proof is essentially the same as Theorem 18. The only difference is to note that for all marginal distributions $D_X \in \mathscr{D}$ and choices of shift $D'_X \in C_\varepsilon(D_X)$, the output of LEARNINGTO-COVER has some h' satisfying

$$\mathbb{E}_{x \sim D'_X}[\ell(h'(x), h_{OPT'}(x))] \le \eta_\ell \varepsilon,$$

where $h_{OPT'}$ is some optimal hypothesis over the shifted distribution D'. As in the standard analysis, this is guaranteed by including the output of the realizable learner on $h_{OPT'}$ (robustness to covariate shift promises this output is close under D'_X with high probability). The remainder of the argument is exactly as in Theorem 24, with the exception of working over the shifted distribution D'_X instead of D_X. ∎

We note that this result can also be extended to the more general loss functions discussed in Appendix C.1 without too much difficulty with the appropriate definition of discrete learnability under covariate shift.

Since Theorem 41 is a bit abstract, let's take a look at one concrete application. Given a class (X, H), the class-dependent total variation distance is a metric on distributions measuring the worst case distance across elements of $H \Delta H := \{h \Delta h' : h, h' \in H\}$:

$$TV_{H\Delta H}(D, D') := \max_{h \in H\Delta H}\{|D(h) - D'(h)|\}.$$

It is not hard to see that any realizable learner is robust to $O(\varepsilon)$ covariate shift in $TV_{H\Delta H}$ distance. We can then apply Theorem 41 to build a robust semi-private learner.

Corollary 42 *Let (X, H) be a class of VC-dimension d, and for every $\varepsilon > 0$ and distribution D over X, define a coviarate shift function:*

$$C_\varepsilon(D) := \left\{ D' : TV_{H \Delta H}(D, D') \leq \varepsilon/2 \right\}.$$

Then (X, H) is semi-private learnable under covariate shift $C = \{C_\varepsilon\}$ in only

$$n_{pub}(\varepsilon, \delta, \alpha) \leq O\left(\frac{d + \log(1/\delta)}{\varepsilon} \right)$$

unlabeled samples and

$$n_{pri}(\varepsilon, \delta, \alpha) \leq O\left(\left(d \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right) \max\{\varepsilon^{-2}, \varepsilon^{-1}\alpha^{-1}\} \right)$$

labeled samples.

Finally, we note again that our original learner in these results is robust to covariate shift despite having no access to samples from the new distribution. Unfortunately, this model does come with fairly strong lower bounds regarding the type of covariate shifts to which it is possible to be robust (David et al., 2010). One solution to this problem is to consider a relaxed variant called (non-conservative) domain adaptation, where the learner additionally has access to a small number of unlabeled samples from the test-time distribution. It is certainly possible to define an analog in the semi-private setting, but naively the use of unlabeled data from the private distribution breaks our reduction since privacy won't be preserved. We leave as an open question whether any sort of PAC-learner in the *non*-conservative model could imply semi-private learners with stronger robustness to covariate shift. Some progress has been made in this direction recently by Bassily, Moran, and Nandi (Bassily et al., 2020b) for distribution-free classification of halfspaces.

Appendix D. Doubly Bounded Loss

In this section we discuss a natural generalization of loss functions over finite label classes we call doubly bounded loss: for all distinct $y, y' \in Y$, we require $\ell(y, y') \in [a, b]$ for some $b \geq a > 0$. This is trivially satisfied by any loss function on a finite label class that satisfies the identity of indiscernibles.

Definition 43 (Doubly Bounded Loss) *We say $\ell : Y \to Y$ is (a, b)-bounded if there exist $b \geq a > 0$ for any distinct $y, y' \in Y$:*

$$\ell(y, y') \in [a, b].$$

As discussed in Appendix C.1, since we now allow Y to be infinite, we need to work with discrete learnability instead of realizable learnability. We can use a slight modification to the discretization technique in Theorem 24 to prove the equivalence of discrete and agnostic learnability for doubly-bounded loss functions. Note that this is stronger than our guarantee for c-approximate pseudometrics, which only gives c-agnostic learnability.

Theorem 44 *Let $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$ be an (a, b)-bounded loss function. Then for any class $(\mathscr{D}, X, H, \ell)$ the following are equivalent:*

1. $(X, H, \mathscr{D}, \ell)$ *is (properly) discretely-learnable*

2. $(X, H, \mathscr{D}, \ell)$ *is (properly) agnostically learnable.*

Proof The proof that agnostic learnability implies discrete learnability is the same as in Theorem 24, so we focus only the forward direction. Assume $(\mathscr{D}, X, H, \ell)$ is discretely-learnable. Fix $\varepsilon' = \frac{a\varepsilon}{4b}$, and let $H_{\varepsilon'}$ be a learnable ε'-discretization of H. We argue that running LEARNINGTOCOVER on $H_{\varepsilon'}$ (using the promised discrete learner) gives the desired agnostic learner. As before, it is sufficient to prove that $C(S_U)$ contains a hypothesis h' such that $\text{err}_{D,\ell}(h') \leq OPT + \varepsilon/2$. Since ℓ is upper bounded, standard empirical risk minimization arguments then give the desired result.

To see why $C(S_U)$ has this property, recall from Lemma 17 that for any $h \in H_{\varepsilon'}$, with probability at least $1 - \delta/2$ there exists $h' \in C(S_U)$ that is ε'-close to h in the following sense:

$$\mathbb{E}_{x \sim D_X}[\ell(h'(x), h(x))] \leq \frac{a\varepsilon}{4b}.$$

Because ℓ is a-lower bounded, this implies that h' must be close to h in classification error:

$$\Pr_{x \sim D_X}[h(x) \neq h'(x)] \leq \frac{\varepsilon}{4b},$$

and since the loss is bounded by b, the risk of h' cannot be much more than of h:

$$\text{err}_{D,\ell}(h') \leq \text{err}_{D,\ell}(h) + \varepsilon/4.$$

Let $h_{OPT} \in H$ be an optimal hypothesis. Since $H_{\varepsilon'}$ ε'-covers H, there exists $h_{OPT}^{\varepsilon'} \in H_{\varepsilon'}$ such that:

$$\text{err}_{D,\ell}(h_{OPT}^{\varepsilon'}) - \text{err}_{D,\ell}(h_{OPT}) \leq \varepsilon'.$$

Let $h_{OPT}'^{\varepsilon'} \in H_{\varepsilon'}$ denote the output of the base learner $\mathcal{A}$ on the labeling given by $h_{OPT}^{\varepsilon'}$. Then by the above, we have that:

$$\begin{aligned}
\text{err}_{D,\ell}(h_{OPT}'^{\varepsilon'}) &= \mathop{\mathbb{E}}_{(x,y) \sim D}[\ell(h_{OPT}'^{\varepsilon'}(x), y)] \\
&\leq \mathop{\mathbb{E}}_{(x,y) \sim D}[\ell(h_{OPT}^{\varepsilon'}(x), y)] + \varepsilon/4 \\
&\leq \mathop{\mathbb{E}}_{(x,y) \sim D}[\ell(h_{OPT}^{\varepsilon'}(x), y) - \ell(h_{OPT}(x), y) + \ell(h_{OPT}(x), y)] + \varepsilon/4 \\
&= \text{err}_{D,\ell}(h_{OPT}^{\varepsilon'}) - \text{err}_{D,\ell}(h_{OPT}) + OPT + \varepsilon/4 \\
&\leq OPT + \varepsilon/2
\end{aligned}$$

as desired. ∎

It is worth noting that the upper bound on the loss can be removed if the adversary is restricted to choosing a marginal over Y which is weakly concentrated.

Appendix E. Robust Learning

Robust learning is an extension to the PAC setting that models an adversary with the power to *perturb* examples at test time. In practice, this corresponds to the fact that we'd like our predictors to be stable to small amounts of adversarial noise—this could range anywhere from a sticker on a stop-sign tricking a self-driving car, to completely imperceptible perturbations that totally fool standard classifiers. The latter was famously demonstrated by Athalye, Engstrom, Ilyas, and Kwok (Athalye et al., 2018), who showed how to generate such perturbations and provided the classic example of tricking a standard ImageNet classifier into thinking a turtle was a gun. Their seminal work caused an explosion of both practical and theoretical research in the area.[20]

Formally, adversarial robustness is modeled simply by changing the error function to be the maximum error over some pre-defined set of neighboring perturbations.

Definition 45 (Robust Loss) *Let X be an instance space and $\mathcal{U} : X \to P(X)$ a "perturbation function" mapping elements to a set of possible perturbations. Given a loss function $\ell : Y \times Y \to \mathbb{R}_{\geq 0}$, the robust loss of a concept $h : X \to Y$ with respect to a distribution D over $X \times Y$ is:*

$$R\text{-}err_{\mathcal{U},D}(h) = \mathop{\mathbb{E}}_{(x,y)\sim D} \left[\max_{x'\in\mathcal{U}(x)} \ell(h(x'), y) \right].$$

In other words, a hypothesis with low robust loss performs well even against an adversary who can perturb x to any "nearby point" (i.e. any $x' \in \mathcal{U}(x)$). Standard realizable and agnostic Robust PAC-learning are then simply defined by replacing the standard error function with the robust error function. Robust learning in the distribution-family model does require one extra twist: we need to make sure that each hypothesis in the class actually has a corresponding distribution over which it is realizable. To this end, we introduce a basic notion of closure for distribution families.

Definition 46 (Robust Closure) *Let $\mathscr{D}$ be a set of distributions over an instance space X and H a concept class. Given any concept h, let X_h denote the set of points in X on which h has 0 robust loss with respect to itself, that is:*

$$X_h \coloneqq \{x \in X : \forall x' \in \mathcal{U}(x), \ell(h(x'), h(x)) = 0\}.$$

For notational simplicity, let $D|_h$ denote the restriction $D|_{X_h}$. The robust closure of $\mathscr{D}$ under H is:

$$\mathscr{D}_H \coloneqq \mathscr{D} \cup \bigcup_{D\in\mathscr{D},h\in H} D|_h.$$

In the robust distribution-family model, it only really makes sense to define realizable learnability over the robust closure of $\mathscr{D}$, since otherwise there may be hypotheses in the class that are not realizable with respect to any distribution in $\mathscr{D}$ and cannot be chosen by the adversary at all. With this in mind, let's formalize this model.

Definition 47 ((Realizable) Distribution-Family Robust PAC Learning) *A class $(\mathscr{D}, X, H, \ell)$ is Robustly PAC-learnable in the realizable setting with respect to perturbation function $\mathcal{U}$ if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ satisfying:*

---

20. Their work has over 900 citations despite being only four years old.

1. *The marginal $D_X \in \mathscr{D}_H$,*

2. $\min_{h \in H} R\text{-}err_{\mathcal{U},D}(h) = 0,$

then:

$$\Pr_{S \sim D^{n(\varepsilon,\delta)}}[R\text{-}err_{\mathcal{U},D}(\mathcal{A}(S)) > \varepsilon] \leq \delta.$$

Agnostic learnability is defined similarly, but since the adversary is unrestricted, there is no longer any need to take the robust closure.

Definition 48 ((Agnostic) Distribution-Family Robust PAC Learning) *A class $(\mathscr{D}, X, H, \ell)$ is Robustly PAC-learnable in the agnostic setting with respect to perturbation function $\mathcal{U}$ if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ satisfying $D_X \in \mathscr{D}$, then:*

$$\Pr_{S \sim D^{n(\varepsilon,\delta)}}[R\text{-}err_{\mathcal{U},D}(\mathcal{A}(S)) > OPT + \varepsilon] \leq \delta,$$

where $OPT = \min_{h \in H}\{R\text{-}err_{\mathcal{U},D}(h)\}$.

We note that different works consider different models of access to the perturbation set $\mathcal{U}$ as well (e.g. assuming $\mathcal{U}$ is known to the learner (Montasser et al., 2019), or has some type of oracle access (Montasser et al., 2020, 2021)). Our reduction requires fairly weak access to $\mathcal{U}$—it is enough to be able to estimate the empirical robust loss of a hypothesis h over any finite sample $S \subset X$. With this in mind, let's now prove realizable and agnostic robust learning are equivalent in the distribution-family model. We'll focus on the special case of (multi-class) classification, and start by re-stating our modified algorithm for simplicity of presentation.

Algorithm 5: Agnostic to Realizable Reduction Robust Setting

Input: Realizable Robust PAC-Learner $\mathcal{A}$

Algorithm:

1. Draw an unlabeled sample S_U, and labeled sample S_L.

2. Run $\mathcal{A}$ over all possible subsets and labelings of S_U to get:

$$C(S) := \{\mathcal{A}(S, h(S)) \mid S \subseteq S_U, h \in H|_S\}.$$

3. **Return** the hypothesis in $C(S)$ with lowest empirical robust error over S_L.

Theorem 49 *If $(\mathscr{D}, X, H)$ is robustly PAC-learnable in the realizable setting with sample complexity $n(\varepsilon, \delta)$, then Algorithm 5 robustly learns $(\mathscr{D}, X, H)$ in the agnostic setting in:*

$$m_U(\varepsilon, \delta) \leq O\left(\max_{\mu \in [0, 1-\varepsilon]}\left\{\frac{n(\varepsilon/(2(1-\mu)), \delta/3)}{1-\mu}\right\}\right)$$

unlabeled samples, and

$$m_L(\varepsilon, \delta) \leq O\left(\frac{m_U(\varepsilon, \delta) + \log(1/\delta)}{\varepsilon^2}\right)$$

labeled samples.

Proof The proof is similar to Theorem 18, but like malicious noise (Theorem 27), requires the use of subsampling. The key issue is that for a distribution D over $X \times Y$, the optimal hypothesis h_{OPT} may not be realizable with respect to D_X, that is we may have:

$$\text{R-err}_{U, D_X \times h_{OPT}}(h_{OPT}) > 0.$$

As a result, our realizable learner (and therefore LEARNINGTOCOVER) has no guarantees over this distribution. On the other hand, our realizable learner does have good guarantees over the *restricted marginal* $D_X|_{h_{OPT}}$. We can then fix the above issue by running LEARNINGTOCOVER of all *subsets* of S including its restriction to $X_{h_{OPT}}$. We will see that this essentially simulates running the realizable learner on the realizable restriction $D_X|_{h_{OPT}}$ and recovers the desired guarantees.

Let's take a look at this more formally. As in our previous arguments it is enough to prove that $C(S)$ contains a hypothesis h' with robust error at most $OPT + \varepsilon/2$, since a standard Chernoff bound tells us that $O(\log(|C(S)|/\delta)/\varepsilon^2)$ labeled examples are enough to estimate the robust loss of every hypothesis in $C(S)$ with high probability. We note that this is the only step in our reduction which requires access to the perturbation set $\mathcal{U}$.

It is left to show that $C(S)$ satisfies this property. Let $D|_{h_{OPT}}$ denote the restriction of D to $X_{h_{OPT}}$, the points in X on which h_{OPT} has 0 robust loss with respect to itself. Let $\bar{D}|_{h_{OPT}}$ be the restriction to the complement, that is $X \setminus X_{h_{OPT}}$. The idea is to decompose our analysis into two separate parts over $D|_{h_{OPT}}$ and $\bar{D}|_{h_{OPT}}$. With this in mind, let μ^* denote the mass of D_X on $X_{h_{OPT}}$, and let OPT' denote the robust error of h_{OPT} over $D|_{h_{OPT}}$. Since we are restricting our attention to classification error, notice that we can decompose OPT as:

$$
\begin{aligned}
OPT &= \text{R-err}_{\mathcal{U}, D}(h_{OPT}) \\
&= \Pr_{(x,y) \sim D} \left[\exists x' \in \mathcal{U}(x) : h_{OPT}(x') \neq y \right] \\
&= \mu^* \Pr_{(x,y) \sim \bar{D}|_{h_{OPT}}} \left[\exists x' \in \mathcal{U}(x) : h_{OPT}(x') \neq y \right] + (1 - \mu^*) \Pr_{(x,y) \sim D|_{h_{OPT}}} \left[\exists x' \in \mathcal{U}(x) : h_{OPT}(x') \neq y \right] \\
&= \mu^* + (1 - \mu^*) OPT',
\end{aligned}
$$

where the last step follows from noting that by definition for all x in the support of $\bar{D}|_{h_{OPT}}$, h_{OPT} is not constant on $\mathcal{U}(x)$. To get a function within $\varepsilon/2$ robust loss of OPT, we claim it is sufficient to prove $C(S)$ contains some h within robust error $\varepsilon/(2(1 - \mu^*))$ of h_{OPT} over $D|_{h_{OPT}}$, that is some h satisfying:

$$\Pr_{x \sim D_X|_{h_{OPT}}} \left[\exists x' \in \mathcal{U}(x) : h(x') \neq h_{OPT}(x') \right] \leq \varepsilon/(2(1 - \mu^*)). \tag{3}$$

This follows from a similar analysis to the above. Namely, letting $R(h, x, y)$ denote the event

$$R(h, x, y) := \exists x' \in \mathcal{U}(x) : h(x') \neq y$$

for notational simplicity, we can break down R-err$_{\mathcal{U},D}(h)$ as:

$$
\begin{aligned}
\text{R-err}_{\mathcal{U},D}(h) = \mu^* \Pr_{(x,y)\sim \tilde{D}|_{h_{OPT}}}[R(h,x,y)] &+ (1-\mu^*) \Pr_{(x,y)\sim D|_{h_{OPT}}}[R(h,x,y)] \\
\leq \mu^* &+ (1-\mu^*) \Pr_{(x,y)\sim D|_{h_{OPT}}}[R(h,x,y)] \\
\leq \mu^* + (1-\mu^*) &\left(\Pr_{(x,y)\sim D|_{h_{OPT}}}[R(h_{OPT},x,y)] + \Pr_{(x,y)\sim D|_{h_{OPT}}}\left[\exists x' \in \mathcal{U}(x) : h(x') \neq h_{OPT}(x')\right] \right) \\
\leq \mu^* + (1-\mu^*) &\left(OPT' + \frac{\varepsilon}{2(1-\mu^*)} \right) \\
= OPT + \varepsilon/2. &
\end{aligned}
$$

It remains to prove that $C(S)$ contains a hypothesis satisfying Equation (3) with high probability. To see this, note that by definition of realizable robust learning, on a labeled sample $(S, h_{OPT}(S)) \sim D|_{h_{OPT}} \times h_{OPT}$ of size $n(\varepsilon/(2(1-\mu^*)), \delta/3)$ our learner will output h satisfying:

$$
\Pr_{x \sim D_X|_{h_{OPT}}}\left[\exists x' \in \mathcal{U}(x) : h(x') \neq h_{OPT}(x)\right] \leq \varepsilon/(2(1-\mu^*))
$$

with probability at least $1-\delta/3$. To get Equation (3), it is then enough to note that h_{OPT} is constant on $\mathcal{U}(x)$ for all x in the support of $D_X|_{h_{OPT}}$ by definition.

The idea is now to draw a large enough unlabeled sample such that with probability at least $1-\delta/3$, the restriction to $X_{h_{OPT}}$ is at least this size. By a Chernoff bound, it is enough to draw $c_1 \frac{n(\varepsilon/(1-\mu^*), \delta/3)}{1-\mu^*}$ points to achieve this for some large enough constant $c_1 > 0$.[21] Since we do not know μ^*, we'll need to draw $c_1 \max_{\mu \in [0, 1-\varepsilon]} \left\{ \frac{n(\varepsilon/(1-\mu), \delta/3)}{1-\mu} \right\}$ points to ensure this property holds (if $\mu^* \geq 1 - \varepsilon$, note that any hypothesis gives a valid solution). By a union bound we have that this overall process succeeds with probability at least $1 - 2\delta/3$, and outputting the hypothesis in $C(S)$ with the lowest empirical robust risk then succeeds with probability $1 - \delta$ as desired. ∎

Theorem 49 can be extended to many of the generic property generalization results in the main body, including approximate pseudometric loss, malicious noise, and semi-private learning, though the exact parameters may be somewhat weaker (e.g. learning over non-binary loss may incur additional factors and lead to c-agnostic rather than truly agnostic learning).

Appendix F. Partial PAC-Learning

Partial PAC-learning is an extension of the standard PAC model to functions that are only defined on a certain portion of the input. Originally introduced by Long (2001) and recently developed in greater depth by Alon, Hanneke, Holzman, and Moran (AHHM) (Alon et al., 2021b), this model allows for the theoretical formalization of popular data-dependent assumptions such as *margin* that have no known analog in the PAC model. Combined with the distribution-family framework, this captures a significant portion of learning assumptions studied in both theory and practice (e.g. learning halfspaces with margin and distributional tail bounds). Let's formalize this model, starting with partial functions.

---

21. We've assumed for simplicity that $n(\varepsilon, \delta) \geq \Omega(\log(1/\delta))$. This assumption can be removed by including an extra additive factor of $\log(1/\delta)$.

Definition 50 (Partial Function) *Let X be an instance space and Y a label space. A partial function is a labeling $f : X \to Y \cup \{*\}$, where elements labeled "$*$" are thought as of undefined. The support of f, denoted $supp(f)$, is the set of elements $x \in X$ s.t. $f(x) \neq *$.*

Standard Partial PAC-learning is defined much like the standard model with the simple modification that "$*$" labels are always considered to be incorrect. As a result, in the realizable case, when the adversary selects a particular partial function f, their marginal distribution over the instance space X must be restricted to lying on $supp(f)$. This makes formalizing data-dependent assumptions easy. If one wanted to consider halfspaces with margin γ for instance, one simply labels every point within γ of the decision boundary as "$*$." Interestingly, much like the distribution-family setting, Partial-PAC learning falls outside both the uniform convergence and the sample compression paradigm (Alon et al., 2021b). AHHM also show a dramatic failure of empirical risk minimization: not only does naively applying an ERM to the partial class fail, it will also fail on any total extension of the class. Despite the lack of these standard tools, both Long and AHHM were able to show that distribution-free classification of partial classes is still controlled by VC dimension, and as a result that the equivalence of realizable and agnostic learnability extends to this setting. In this section, we'll discuss how a variant of our reduction shows that this result extends to the distribution-family model, extended loss function, and to properties beyond agnostic learning.

In the distribution-family model, formalizing realizable learnability requires some slight changes from the standard model, since we need to make sure our hypotheses are actually realizable over some distribution in the family (this is automatic in the distribution-free setting). To this end, we introduce a basic notion of closure for distribution families.

Definition 51 (Partial Closure) *Let $\mathscr{D}$ be a set of distributions over an instance space X and H a concept class. Given any concept h, and distribution D over X, let $D|_h$ denote the restriction $D|_{supp(f)}$. The partial closure of $\mathscr{D}$ under H is:*

$$\mathscr{D}_H := \mathscr{D} \cup \bigcup_{D \in \mathscr{D}, h \in H} D|_h.$$

In the realizable model it makes more sense to work with the closure of $\mathscr{D}$ than $\mathscr{D}$ itself, since otherwise the class H may contain hypotheses that cannot be realized over any distribution, and therefore cannot be accessed by the adversary at all. For simplicity, we'll also restrict our attention to (multi-class) classification where the label space $Y = [m]$, and recall that the loss of any undefined point is always 1.

Definition 52 ((Realizable) Distribution-Family Partial PAC Learning) *A partial class $(\mathscr{D}, X, H)$ is PAC-learnable in the realizable setting if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ satisfying:*

1. The marginal $D_X \in \mathscr{D}_H$,

2. $\min_{h \in H} err_D(h) = 0$,

then:

$$\Pr_{S \sim D^{n(\varepsilon, \delta)}}[err_D(\mathcal{A}(S)) > \varepsilon] \leq \delta,$$

where the error $err_D(h)$ is standard classification error:

$$err_D(h) = \Pr_{(x,y) \sim D}[h(x) \neq y].$$

Agnostic learnability is defined analogously, but since the adversary is unrestricted, there is no need to move to the closure of $\mathscr{D}$.

Definition 53 ((Agnostic) Distribution-Family Partial PAC Learning) *A partial class $(\mathscr{D}, X, H)$ is PAC-learnable in the agnostic setting if there exists an algorithm $\mathcal{A}$ and function $n(\varepsilon, \delta)$ such that for all $\varepsilon, \delta > 0$ and distributions D over $X \times Y$ satisfying $D_X \in \mathscr{D}$, then:*

$$\Pr_{S \sim D^{n(\varepsilon, \delta)}}[err_D(\mathcal{A}(S)) > OPT + \varepsilon] \leq \delta.$$

The issue with our standard reduction strategy for partial functions is that in the agnostic model, the adversary's marginal distribution over X might have support outside of $\text{supp}(h_{OPT})$, which causes LEARNINGTOCOVER to lose its guarantee of outputting a non-uniform cover. This can be dealt with by a variant of our subsampling technique. If we run LEARNINGTOCOVER over all subsamples of the unlabeled sample S_U, one of these subsamples must match the support of h_{OPT}. This is in fact the same strategy used for adversarial robustness in Appendix E, but we will include the algorithm again to make this section self-contained.

Algorithm 6: Agnostic to Realizable Reduction Partial PAC Setting

Input: Realizable Robust PAC-Learner $\mathcal{A}$

Algorithm:

1. Draw an unlabeled sample S_U, and labeled sample S_L.

2. Run $\mathcal{A}$ over all possible subsets and labelings of S_U to get:

$$C(S) \coloneqq \{\mathcal{A}(S, h(S)) \mid S \subseteq S_U, h \in H|_S\}.$$

3. **Return** the hypothesis in $C(S)$ with lowest empirical error over S_L.

Theorem 54 *If $(\mathscr{D}, X, H)$ is a realizably PAC-learnable partial class with sample complexity $n(\varepsilon, \delta)$, then Algorithm 6 agnostically learns $(\mathscr{D}, X, H)$ in*

$$m_U(\varepsilon, \delta) \leq O\left(\max_{\mu \in [0, 1-\varepsilon]} \left\{ \frac{n(\varepsilon/(2(1-\mu)), \delta/3)}{1-\mu} \right\}\right)$$

unlabeled samples, and

$$m_L(\varepsilon, \delta) \leq O\left(\frac{m_U(\varepsilon, \delta) + \log(1/\delta)}{\varepsilon^2}\right)$$

labeled samples.

Proof The proof is essentially the same as for Theorem 49, but we repeat it here for completeness. As always, it is enough to prove that $C(S)$ (from Algorithm 6) contains a hypothesis h' with error at most $OPT + \varepsilon/2$. The key issue with our standard reduction is that the optimal hypothesis h_{OPT} may be undefined on certain examples in the unlabeled sample S_U. By running over all subsamples of S_U, we in essence simulate pulling samples only from the support of h_{OPT}, which is enough to get the desired guarantee.

More formally, let $D|_{h_{OPT}}$ be the restriction of D to supp(h_{OPT}), and $\bar{D}|_{h_{OPT}}$ the restriction to its complement $X \setminus \text{supp}(h_{OPT})$. The idea is to decompose our analysis into two separate parts over $D|_{h_{OPT}}$ and $\bar{D}|_{h_{OPT}}$. With this in mind, let μ^* denote the mass of D_X on the undefined portion of h_{OPT}, and let OPT' denote the error of h_{OPT} over $D|_{h_{OPT}}$. Since we have restricted our attention to classification, notice that we can decompose OPT as:

$$\text{err}_D(h_{OPT}) = \Pr_{(x,y)\sim D}[h_{OPT}(x) \neq y]$$
$$= \mu^* \Pr_{(x,y)\sim \bar{D}|_{h_{OPT}}}[h_{OPT}(x) \neq y] + (1-\mu^*)\Pr_{(x,y)\sim D|_{h_{OPT}}}[h_{OPT}(x) \neq y]$$
$$= \mu^* + (1-\mu^*)OPT'.$$

We'd like to prove that $C(S)$ contains a hypothesis h within $\varepsilon/2$ error of optimal. We claim it is sufficient to show that $C(S)$ contains a hypothesis within $\varepsilon/(2(1-\mu^*))$ classification distance of h_{OPT}, since:

$$\text{err}_D(h) = \mu^* \mathop{\mathbb{E}}_{(x,y)\sim \bar{D}|_{h_{OPT}}}[h(x) \neq y] + (1-\mu^*)\mathop{\mathbb{E}}_{(x,y)\sim D|_{h_{OPT}}}[h(x) \neq y]$$
$$\leq \mu^* + (1-\mu^*)\mathop{\mathbb{E}}_{(x,y)\sim D|_{h_{OPT}}}[h(x) \neq y]$$
$$\leq \mu^* + (1-\mu^*)\left(\mathop{\mathbb{E}}_{(x,y)\sim D|_{h_{OPT}}}[h_{OPT}(x) \neq y] + \frac{\varepsilon}{2(1-\mu^*)}\right)$$
$$= \mu^* + (1-\mu^*)OPT' + \frac{\varepsilon}{2}$$
$$= OPT + \varepsilon/2,$$

where the third line follows from the fact that h' and h_{OPT} only differ on a $\frac{\varepsilon}{2(1-\mu^*)}$ fraction of inputs over $D|_{h_{OPT}}$.

It is left to argue that $C(S)$ contains such a hypothesis h. Recall that on a labeled sample $(S, h(S)) \sim D|_{h_{OPT}} \times h_{OPT}$ of size $n(\varepsilon/(2(1-\mu^*)), \delta/3)$, LEARNINGTOCOVER will contain h that is $\varepsilon/(2(1-\mu^*))$-close to h_{OPT} in classification error over $D|_{h_{OPT}}$ with probability at least $1 - \delta/3$. The idea is then to draw a large enough unlabeled sample such that with probability at least $1 - \delta/3$, the restriction of the sample to $D|_{h_{OPT}}$ is at least this size (since we run over every subsample, we will always hit this restriction). By a Chernoff bound, it is enough to draw $c_1 \frac{n(\varepsilon/(2(1-\mu^*)),\delta/3)}{1-\mu^*}$ points to achieve this for some large enough constant $c_1 > 0$.[22] Since we do not know μ^*, we'll need to draw $c_1 \max_{\mu \in [0, 1-\varepsilon]}\left\{\frac{n(\varepsilon/(1-\mu),\delta/3)}{1-\mu}\right\}$ points to ensure this property holds (if $\mu^* \geq 1 - \varepsilon$, note that any hypothesis gives a valid solution). By a union bound we have that this overall process succeeds with probability at least $1 - 2\delta/3$, and outputting the hypothesis in $C(S)$ with the lowest empirical risk then succeeds with probability $1 - \delta$ as desired. ∎

Like Theorem 49, Theorem 54 can be extended to many of the generic property generalization results in the main body, including approximate pseudometric loss, malicious noise, and semi-private learning, though it may experience some degradation of parameters (e.g. c-agnostic rather than truly agnostic learning) depending on how the loss of "$*$" values are formalized in these settings.

---

22. As in Theorem 49, we've assumed for simplicity that $n(\varepsilon, \delta) \geq \Omega(\log(1/\delta))$. This assumption can be removed by including an extra additive factor of $\log(1/\delta)$.

Appendix G. Uniform Stability

Uniform stability, originally introduced by Bousquet and Elisseeff (2002), is a useful algorithmic property that is closely tied to both generalization and privacy. Informally, an algorithm $\mathcal{A}$ is said to be uniformly stable if for all elements $x \in X$, the probability that $\mathcal{A}$ changes its output on x over neighboring datasets is small.

Definition 55 (Uniform Stability) *A learning algorithm is said to be α-uniformly stable if for all neighboring inputs S, S' which differ on a single example, all $x \in X$, and all $y \in Y$:*

$$\Pr[\mathcal{A}(S)(x) = y] \leq \Pr[\mathcal{A}(S')(x) = y] + \alpha.$$

Uniform stability can also be thought of as a form of *private prediction* (Dwork and Feldman, 2018), which protects against adversaries who have restricted access to a model only through prediction responses on individual points (this is often the case in practice since it is common to release APIs with query access rather than full models). Like semi-privacy, this definition has the benefit of maintaining practicality in a reasonable range of circumstances while weakening the stringent requirements of standard private learning. Indeed, it is well known that in the distribution-free classification setting, uniformly stable learning and private prediction are both possible for any class with finite VC dimension (Shalev-Shwartz et al., 2010; Dwork and Feldman, 2018; Dagan and Feldman, 2020). Unsurprisingly, these previous works (at least those working in the agnostic model), rely on uniform convergence and uniform covers. We'll show these can be replaced with a variant of our standard reduction. The argument is otherwise similar to the proof in Dagan and Feldman (2020).

Theorem 56 *Let $(\mathscr{D}, X, H)$ be a realizably learnable class with sample complexity $n(\varepsilon, \delta)$. Then there exists an α-uniformly stable, α-semi private algorithm that agnostically learns $(\mathscr{D}, X, H)$ in only*

$$m_U(\varepsilon, \delta, \alpha) \leq O\left(\frac{n(\varepsilon/2, \delta/2)}{\alpha}\right)$$

unlabeled samples, and

$$m_L(\varepsilon, \delta, \alpha) \leq O\left(\frac{\log\left(\Pi_H(n(\varepsilon, \delta))\right)}{\min\{\alpha\varepsilon, \varepsilon^2\}}\right)$$

labeled samples.

Proof The proof boils down to a standard subsampling trick first noted by Shalev-Shwartz et al. (2010). Instead of drawing our standard unlabeled sample of size $n(\varepsilon/2, \delta/2)$, we draw a sample of size $\frac{n(\varepsilon/2, \delta/2)}{2\alpha}$ and run LEARNINGTOCOVER over a random $\alpha/2$ fraction of the sample. This ensures that swapping out any individual sample can only effect the result with probability $\alpha/2$. Since this subsample is of size $n(\varepsilon/2, \delta/2)$, LEARNINGTOCOVER keeps its standard guarantees and the output $C(S_U)$ has a hypothesis within $\varepsilon/2$ of optimal with probability $1 - \delta/2$. We can now apply the exponential mechanism with privacy parameter $\alpha/4$, which ensures the algorithm is $\alpha/2$-uniformly stable with respect to the labeled sample as well. The sample complexity bounds come from standard analysis of the exponential mechanism and the size of $C(S_U)$. Semi-privacy comes for free due to our use of the exponential mechanism. ∎

As in previous sections, Theorem 56 can be extended to any of the generic property generalization results in the main body, including for instance c-approximate pseudometric loss, malicious noise, and robustness to covariate shift.

Appendix H. Statistical Query Model

Kearns (1998) statistical query model is a popular modification of PAC learning where the sample oracle is replaced with the ability to ask noisy statistical questions about the data.

Definition 57 (Realizable SQ-learning) *Given a distribution D over X and $h \in H$, let $STAT(D, h)$ be an oracle which upon input of a function $\psi : X \times Y \to [-1, 1]$ and tolerance $\tau \in \mathbb{R}_{\geq 0}$ may output any estimate of the expectation of ψ up to τ error, that is:*

$$STAT(D, h)(\psi, \tau) \in \mathbb{E}_{x \sim D}[\psi(x, h(x))] \pm \tau.$$

We call a class $(\mathscr{D}, X, H, \ell)$ SQ-learnable if for all $\varepsilon > 0$, there exists some tolerance $\tau = \tau(\varepsilon)$, query complexity $n(\varepsilon, \tau)$, and an algorithm $\mathcal{A}$ such that for all $D \in \mathscr{D}$ and $h \in H$, $\mathcal{A}$ achieves ε error in at most $n(\varepsilon, \tau)$ oracle calls to $STAT(D, h)$ with tolerance at worst τ.[23]

Agnostic learning is then defined analogously where D, h is replaced with a generic distribution over $X \times Y$ whose marginal lies in $\mathscr{D}$. We can use a basic form of discretization to prove property generalization in the SQ model.

Theorem 58 *Let ℓ be a c-approximate pseudometric and $(\mathscr{D}, X, H, \ell)$ a realizably SQ-learnable class with query complexity $n(\varepsilon, \tau)$. Then $(\mathscr{D}, X, H, \ell)$ is c-agnostically SQ-learnable up to $\varepsilon + \tau$ error in $(1/\tau)^{n(\varepsilon, \tau)}$ statistical queries of tolerance at worst τ.*

Proof The idea is similar to our discretization in Theorem 24. The realizable SQ-learner $\mathcal{A}$ makes some finite $n(\varepsilon, \tau)$ queries. Let $C_{\mathcal{A}}$ denote the set of outputs of $\mathcal{A}$ when fed every possible combination of responses from the discretized set $\{-1, -1 + 2\tau, \ldots, 1 - 2\tau, 1\}$. For every $D \in \mathscr{D}$ and $h \in H$, one of these combinations must be a valid query response in the realizable model, so $C_{\mathcal{A}}$ covers $(\mathscr{D}, X, H, \ell)$. By the same arguments of Theorem 24, $C_{\mathcal{A}}$ must contains a hypothesis with error at most $c \cdot OPT + \varepsilon$. Since we can directly compute the loss of every element in $C_{\mathcal{A}}$ up to τ error in the SQ model simply by querying the loss function, this gives the desired result in $|C_{\mathcal{A}}| = (1/\tau)^{n(\varepsilon, \tau)}$ queries. $\blacksquare$

We note that while our reduction in this model experiences exponential blowup in the number of queries, this should really be thought of as corresponding to a blow up in *run-time* instead of "sample complexity" in the standard sense (which corresponds more closely to τ).

Appendix I. Fairness

Recent years have seen rising interest in an algorithmic property called *fairness*. Informally, fairness tries to tackle the issue that "well-performing" classifiers in the standard sense may actually be discriminatory against certain individuals or subgroups. We will consider a form of fair leaning introduced by Rothblum and Yona (2018) called Probably Approximately Correct and Fair (PACF) learning. Their definition is based off of a notion of fairness that ensures that similar individuals are treated similarly with respect to a fixed metric.

---

23. While we generally think of τ as being at worst polynomial in ε, this is not strictly necessary for the model.

Definition 59 (Metric Fairness) *Let $d : X \times X \to \mathbb{R}_{\geq 0}$ be a similarity measure on X and D a distribution over X. A classifier $h : X \to Y_{out}$ is called (α, γ)-fair with respect to d and D if h acts similarly on most similar individuals:*

$$\Pr_{x,x' \sim D}[|h(x) - h(x')| > d(x, x') + \gamma] \leq \alpha.$$

We note that the output space Y_{out} may differ from the label space Y in general learning problems.

In fact, this definition only really makes sense when the output classifier h is allowed to be real-valued (as this allows for some flexibility in the $|h(x) - h(x')|$ term). As such, when considering settings such as binary classification where $Y = \{0, 1\}$ is discrete, Rothblum and Yona (2018)'s initial formalization considers returning *probabilistic* classifiers with $Y_{out} = [0, 1]$. Here $h(x) = y \in [0, 1]$ is taken to be the probability of the label being 1. The error of a probabilistic classifier h with respect to any distribution D over $X \times \{0, 1\}$ is then given by its expected ℓ_1 distance:

$$err_D(h) = \mathbb{E}_{(x,y) \sim D}[|h(x) - y|].$$

For simplicity we'll focus in this section on this same regime extended to the distribution-family model.

In broad strokes, the goal of Fair PAC learning is to output a fair classifier satisfying standard PAC guarantees. Practically this requires a few modifications. First, since there may be no fair classifier satisfying these guarantees, we will only require our output to be as good as the best fair classifier. Second, we will actually allow some slack in the fairness parameters, which Rothblum and Yona Rothblum and Yona (2018) show is a practical way to ensure that fair learnability remains possible across a broad range of classes.[24]

Definition 60 (PACF-learning Rothblum and Yona (2018)) *We say $(\mathscr{D}, X, H)$ is (agnostically) (α, γ)-PACF-learnable with respect to a similarity metric $d : X \times X \to Y$ if there exists an algorithm $\mathcal{A}$ and function $n = n(\varepsilon, \varepsilon_\alpha, \varepsilon_\gamma, \delta)$ such that for all $\varepsilon, \varepsilon_\alpha, \varepsilon_\gamma, \delta > 0$, and distributions D over $X \times Y$ such that $D_X \in \mathscr{D}$, $\mathcal{A}(S)$ satisfies the following guarantees with probability $1 - \delta$ over samples S of size n:*

1. $\mathcal{A}(S)$ is accurate:

$$err_{D,\ell}(\mathcal{A}(S)) \leq OPT_{\alpha,\gamma} + \varepsilon$$

2. $\mathcal{A}(S)$ is $(\alpha + \varepsilon_\alpha, \gamma + \varepsilon_\gamma)$-fair.

Here $OPT_{\alpha,\gamma}$ is the optimal error of any (α, γ)-fair classifier, that is:

$$OPT_{a,b} := \min_{h \in H^d_{D_X,a,b}} \{err_{D,\ell}(h)\},$$

and

$$H^d_{D_X,a,b} = \{h \in H : h \text{ is } (a, b)\text{-fair with respect to } d \text{ and } D_X\}$$

---

24. We note that our presentation of this definition differs slightly from Rothblum and Yona (2018). Their (α, γ)-PACF-learnability formally corresponds to $(\alpha - \varepsilon_\alpha, \gamma - \varepsilon_\gamma)$-PACF-learnability in our version.

Realizable learnability is defined similarly, where the adversary is constrained to picking distributions which have 0 error with respect to some (α, γ)-fair classifier in H. We show that property generalization holds for the PACF model.

Theorem 61 (Agnostic → Realizable (PACF Setting)) *Let $(\mathscr{D}, X, H)$ be any class that is realizably (α, γ)-PACF learnable with sample complexity $n(\varepsilon, \varepsilon_\alpha, \varepsilon_\gamma, \delta)$. Then $(\mathscr{D}, X, H)$ is agnostically (α, γ)-fair-PAC learnable in only*

$$m_U(\varepsilon, \varepsilon_\alpha, \varepsilon_\gamma, \delta) \leq n(\varepsilon/2, \varepsilon_\alpha, \varepsilon_\gamma, \delta/2)$$

unlabeled samples, and

$$m_L(\varepsilon, \varepsilon_\alpha, \varepsilon_\gamma, \delta) \leq O\left(\frac{\log(\Pi_H(n(\varepsilon/2, \varepsilon_\alpha, \varepsilon_\gamma, \delta/2))) + \log(1/\delta)}{\varepsilon^2}\right)$$

labeled samples.

Proof The key observation is that the definition of fairness depends only on the classifier h and the *marginal distribution D_X*. Let h_{OPT} be the hypothesis achieving the minimum error over $H_{D_X, \alpha, \gamma}$. By the above observation, with probability $1 - \delta$ the hypothesis set $C(S_U)$ returned by LEARNINGTOCOVER contains an $(\alpha + \varepsilon_\alpha, \gamma + \varepsilon_\gamma)$-*fair* h satisfying:

$$\mathbb{E}_{x \sim D_X}[|h(x) - h_{OPT}(x)|] \leq \varepsilon/2.$$

Since ℓ_1 error is a metric (and therefore satisfies the triangle inequality), we can use our argument for c-pseudometric loss functions from Theorem 24 to argue that choosing the lowest empirical risk $(\alpha + \varepsilon_\alpha, \gamma + \varepsilon_\gamma)$-*fair* classifier in $C(S_U)$ with respect to a sufficiently large labeled sample S_L gives the desired learner. ∎

With care, this result can be extended to a broader range of loss functions as well as to other finitely-satisfiable properties covered in this work.

Appendix J. Notions of Coverability

In this section we discuss the connection between non-uniform covers and several previous notions of coverability used in various learning applications. For simplicity, we'll restrict our attention to covering with respect to standard classification distance, that is given a distribution D and hypotheses h and h' over some instance space X:

$$d_D(h, h') = \Pr_{x \sim D}[h(x) \neq h'(x)].$$

To start, let's recall the basic notion of an ε-cover specified to this measure for simplicity.

Definition 62 (ε-cover) *Let X be an instance space, Y a label space, and let $L_{X,Y}$ denote the set of all labelings $c : X \to Y$. A set $C \subset L_{X,Y}$ is said to form an ε-cover for (D, X, H) if for every hypothesis $h \in H$, there exists $c \in C$ such that*

$$d_D(c, h) \leq \varepsilon.$$

*C is called **proper** if $C \subset H$.*

Finite ε-covers are exceedingly useful in learning theory. As discussed in Section 5, a common strategy in the literature is to use unlabeled samples to construct an ε-cover with high probability (Balcan and Blum, 2010; Hanneke and Yang, 2015; Alon et al., 2019a; Bassily et al., 2020a). This results in a distribution over potential covers we call a uniform (ε, δ)-cover.

Definition 63 (Uniform (ε, δ)-cover) *Let X be an instance space, Y a label space, and let $L_{X,Y}$ denote the set of all labelings $c : X \to Y$. A distribution D_C over the power set $P(L_{X,Y})$ is said to form a **uniform** (ε, δ)-cover for (D, X, H) if:*

$$\Pr_{C \sim D_C}[C \text{ is an } \varepsilon\text{-cover for } (D, X, H)] \geq 1 - \delta.$$

*D_C is called **proper** if its support lies entirely in H.*

In this work, we introduce a weaker non-uniform variant of this notion where each h has an individual guarantee of being covered by the distribution, but it is not necessarily the case that a sample will cover all $h \in H$ simultaneously.

Definition 64 (Non-Uniform (ε, δ)-cover) *Let X be an instance space, Y a label space, and let $L_{X,Y}$ denote the set of all labelings $c : X \to Y$. A distribution D_C over the power set $P(L_{X,Y})$ is said to form a **non-uniform** (ε, δ)-cover for (D, X, H) if for every fixed hypothesis $h \in H$,*

$$\Pr_{C \sim D_C}[C \text{ is an } \varepsilon\text{-cover for } (D, X, \{h\})] \geq 1 - \delta.$$

*D_C is called **proper** if its support lies entirely in H.*

In the context of learning, we are usually interested not just in the existence of these covers, but in the more challenging problem of constructing them from a small number of unlabeled samples. In other words, given a class $(\mathscr{D}, X, H)$, we'd like to know how many unlabeled samples from an adversarially chosen distribution $D \in \mathscr{D}$ are necessary to build a uniform (or non-uniform) (ε, δ)-cover for (D, X, H). In Appendix C.3, we saw that the ability to construct a non-uniform (ε, δ)-cover from $O\left(\frac{\log(1/\delta)}{\varepsilon}\right)$ samples was crucial to give a semi-private learner with optimal public sample complexity. This improved over recent work of Alon, Bassily, and Moran (ABM) (Alon et al., 2019a), who showed that it is possible to build a uniform (ε, δ)-cover in $O\left(\frac{\log(1/\varepsilon)+\log(1/\delta)}{\varepsilon}\right)$ samples.

It is interesting to ask whether non-uniformity is really necessary here, or whether ABM's analysis is simply sub-optimal. We'll show that the former is true, at least in the proper distribution-family setting: the $\log(1/\varepsilon)$ gap between these models is necessary and uniform covers cannot be used to build optimal semi-private learners.

Theorem 65 (Separation of Uniform and Non-Uniform Covers) *There exists an instance space X, hypothesis class H, and family of distributions $\mathscr{D}$ such that for any sufficiently small $\varepsilon > 0$, the following statements holds:*

1. *Any algorithm which returns a finite proper uniform $(\varepsilon, 1/3)$-cover for $(\mathscr{D}, X, H)$ requires at least $\Omega(1/\varepsilon \cdot \log(1/\varepsilon))$ samples.*

2. *There exists an algorithm which returns a finite proper non-uniform (ε, δ)-cover for $(\mathscr{D}, X, H)$ in $O(\log(1/\delta)/\varepsilon)$ samples.*

Proof Let the instance space $X = \mathbb{N}$ and H be the class of indicators along with the all 0's function, that is $H = \{h_i : i \in \mathbb{N}\} \cup \{h_0\}$ where $h_i(x) = \mathbb{1}\{x = i\}$ and h_0 is 0 everywhere. We consider the family of distribution $\mathscr{D} = \{\mathscr{D}_{n,k}\}_{n,k>0}$ given by k-sets of $[n]$ where

$$\mathscr{D}_{n,k} = \{\mathrm{unif}(T) : T \subset [n] \text{ and } |T| = k\},$$

where $\mathrm{unif}(T)$ is a uniform distribution over T.

We start with the first claim, that building a bounded uniform $(\varepsilon, 1/2)$-cover needs at least $\Omega(1/\varepsilon \log(1/\varepsilon))$ samples. More formally, for any error parameter $\varepsilon > 0$ and size bound $m = m(\varepsilon) \in \mathbb{N}$, let $k = \lfloor 1/(2\epsilon) \rfloor$. We will show that for any algorithm $\mathcal{A}$ on $k \log(k)$ samples that outputs at most m hypotheses, $\mathcal{A}$ must fail to output an ε-cover with probability at least $1/2$.

Let $n \gg m, k$ be some natural number to be fixed later and consider the family of distributions $\mathscr{D}_{n,k}$. By Yao's minimax principle, it is sufficient to show that there exists a distribution over the elements in $\mathscr{D}_{n,k}$ such that any deterministic algorithm over $k \log(k)$ samples outputting a set of (at most) m hypotheses fails to give a proper ε-cover with probability $1/2$. We claim that taking the uniform distribution over $\mathscr{D}_{n,k}$ suffices. To formalize this, it is useful to observe the following claim.

Claim 66 *Any subset of hypotheses $C \subset H$ of size m can be a proper ε-cover for H under at most $\binom{m}{k}$ distributions in $\mathscr{D}_{n,k}$.*

Let's prove the result under this assumption. The key observation is that by standard lower bounds on the coupon collector problem, a sample S of $k \log(k)$ points from any $\mathrm{unif}(T) \in \mathscr{D}_{n,k}$ will not include $\mathrm{unif}(T)$'s entire support with probability at least $1/2$. With this in mind, assume that the input sample S contains only $\mathrm{supp}(S) = j < k = \mathrm{supp}(\mathrm{unif}(T))$ elements. As a result, there are $\binom{n-j}{k-j}$ consistent distributions with S, and by Claim 66, $\mathcal{A}(S)$ is a proper ε-cover for at most $\binom{m}{k}$ of them. Since S is equally likely to have been sampled from any of these distributions, the probability that $\mathcal{A}(S)$ is a proper ε-cover is at most:

$$\Pr[\mathcal{A} \text{ fails given } \mathrm{supp}(S) = j < k] \geq \frac{\binom{n-j}{k-j} - \binom{m}{k}}{\binom{n-j}{k-j}}.$$

Taking n sufficiently larger than m and k, we can make this probability as close to 1 as desired for any $0 < j < k$. Finally, since samples of this form occur with probability at least $1/2$, the algorithm fails with probability at least $1/3$ as desired. It is left to prove Claim 66.

Proof [Proof of Claim 66] Notice that for any distribution $\mathrm{unif}(T) \in \mathscr{D}_{n,k}$, any $i \in T$ and any $j \neq i$, $d_{\mathrm{unif}(T)}(h_i, h_j) > 2\varepsilon$. Let C be any proper ε-cover of H under distribution $\mathrm{unif}(T)$. Then, by the above argument, it must contain $\{h_i : i \in T\}$. Since $|T| = k$, C can be a proper ε-cover of H under at most $\binom{|C|}{k}$ distributions in $\mathscr{D}_{n,k}$. ∎

We now move to proving that a proper *non-uniform* (ε, δ)-cover can be built in only $O(\log(1/\delta)/\varepsilon)$ samples. This follows from the fact that for any $n \geq k$ and distribution $\mathrm{unif}(T) \in \mathscr{D}_{n,k}$, each $i \in T$ is in the random sample S with probability $1 - \delta$. Since each h_j for $j \notin T$ is covered by h_0, outputting $\{h_i : i \in S\} \cup \{h_0\}$ generates a proper non-uniform (ε, δ)-cover. ∎

The construction in Theorem 65 can easily be modified to give a class with the same gap which is not privately learnable (say by embedding a single copy of a threshold over $[0, 1]$). Since any such

class requires at least $\Omega(\frac{1}{\varepsilon})$ public samples to semi-privately learn by Theorem 36,[25] Theorem 65 then provides a separation between using uniform and non-uniform covers in semi-private learning: the former provably requires an extra log factor, while the latter matches the lower bound exactly. Unfortunately, our proof of this result only holds in the proper setting, as Claim 66 fails when improper hypotheses are allowed. We conjecture that this is not an inherent barrier: the separation should continue to hold in the improper case, albeit with some different analysis.

We have now seen a weak separation between uniform and non-uniform covers, but one might reasonably wonder whether a much stronger separation is possible. In particular, all previous constructions of uniform covers use uniform convergence, but there exist simple examples of learnable classes in the distribution-family model that fail this property: do such classes provide an example of objects which are non-uniformly coverable but not uniformly coverable? Surprisingly, the answer is no! It turns out that an algorithm for non-uniform covering can always be used to construct a uniform covering without too much overhead. Moreover, we'll see that the $\log(1/\varepsilon)$ gap is tight when (X, H) has finite VC dimension.

To prove this, it will actually be useful to make a brief aside and introduce another closely related notion of covering called *fractional covers*. These objects are essentially a form of non-uniform covering which output a single hypothesis instead of a set of them.

Definition 67 (Fractional cover) *Let X be an instance space, Y a label space, and let $L_{X,Y}$ denote the set of all labelings $c : X \to Y$. A distribution D_C over $L_{X,Y}$ is said to form a fractional (ε, p)-cover for a hypothesis class H for (D, X, H) if for any fixed $h \in H$, a sample from D_C covers h with probability p:*

$$\Pr_{c \sim D_C}[d(c, h) \leq \varepsilon] \geq p.$$

Fractional covers are closely connected to non-uniform covers. In fact, one can easily move between the two by sampling or subsampling.

Proposition 68 (Non-uniform cover $\iff$ Fractional cover) *Let (D, X, H) be any class, C_{frac} a fractional (ε, p)-cover, and $C_{n\text{-}u}$ a non-uniform $(\varepsilon, 1/2)$-cover. Then the following hold:*

1. *Drawing $\log_{1/(1-p)}(1/\delta)$ samples from C_{frac} gives a non-uniform (ε, δ)-cover.*

2. *Choosing a random hypothesis from $C_{n\text{-}u}$ gives a fractional $(\varepsilon, 1/2|C|)$-cover.*

Proof Both statements are essentially immediate from definition. For any fixed $h \in H$, if we draw M samples from C_{frac}, the probability we fail to cover h is $(1-p)^M$, so setting $M = \log_{1/(1-p)}(1/\delta)$ gives the desired non-uniform cover. On the other hand, for any fixed $h \in H$, a sample from $C \sim C_{\text{n-u}}$ contains c ε-close to h with probability $1/2$. Outputting a uniformly random element of C then gives an element within ε of h with probability $1/2|C|$ as desired. ∎

It will also be useful to note a classical relation between covers and fractional covers.

Lemma 69 *If there exists a fractional (ε, p)-cover for (D, X, H), then there exists a 2ε-cover of size $1/p$.*

Proof This follows from classical packing-covering duality. The existence of a fractional (ε, p)-cover implies there cannot exist a 2ε-packing of size greater than $1/p$ (that is, a set of more than $1/p$ hypotheses in H that are pairwise 2ε-separated with respect to D). By packing-covering duality, this implies the existence of a 2ε-cover of size $\lceil 1/p \rceil + 1$. ∎

With this in hand, let's show that uniform covers can be constructed for any realizably learnable class, regardless of whether we have uniform convergence.

Theorem 70 (Realizable learning → Uniform cover) *Let $(\mathscr{D}, X, H)$ be realizably PAC-learnable with sample complexity $n(\varepsilon, \delta)$. Then it is possible to construct a uniform (ε, δ)-cover for $(\mathscr{D}, X, H)$ in $n(\varepsilon/2, \delta')$ samples where $\delta' = O\left(\frac{\delta}{\Pi_H(n(\varepsilon/2, 1/2))}\right)$.*

Proof We'll start by proving a slightly more general fact. If for every $D \in \mathscr{D}$, (D, X, H) has a proper $(\varepsilon/2)$-cover C_D of size at most $C = C(\varepsilon/2)$, then it is possible to construct a uniform (ε, δ)-cover in $n(\varepsilon/2, \delta/C)$ samples. This is essentially immediate from Lemma 17, which states that running LEARNINGTOCOVER over a sample of size $n(\varepsilon/2, \delta/C)$ gives a non-uniform $(\varepsilon/2, \delta/C)$-cover. Union bounding over C_D then gives that a sample from the non-uniform cover $(\varepsilon/2)$-covers C_D with probability at least $1 - \delta$. Since C_D is itself an $(\varepsilon/2)$-cover, this implies that the entire class H ε-covered by the sample with probability at least $1 - \delta$ as desired.

It remains to show that for every $D \in \mathscr{D}$, (D, X, H) has a proper $(\varepsilon/2)$-cover of size $O(\Pi_H(n(\varepsilon/2, 1/2)))$. This follows from combining Lemma 68 and Lemma 69. In particular, Lemma 17 implies that running LEARNINGTOCOVER over a sample of size $n(\varepsilon/2, 1/2)$ produces a non-uniform $(\varepsilon/2, 1/2)$-cover of size at most $\Pi_H(n(\varepsilon/2, 1/2))$. Lemma 68 states that subsampling from this cover gives a fractional $(\varepsilon/2, 1/(2\Pi_H(n(\varepsilon/2, 1/2))))$-cover, which in turn implies the existence of a $(\varepsilon/2)$-cover of size $O(\Pi_H(n(\varepsilon/2, 1/2)))$ as desired. We note that this last argument is similar to an observation made in Benedek and Itai (1991) seminal work on the distribution-dependent model. ∎

When (X, H) has finite VC-dimension d, note that Theorem 70 exactly matches the lower bound exhibited in Theorem 65 as the required number of samples for a uniform (ε, δ)-cover becomes:

$$n(\varepsilon/2, \delta') \leq O\left(\frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

This also matches the bound given by ABM (Alon et al., 2019a) using uniform convergence.