# Simplifying Adversarially Robust PAC Learning with Tolerance

**Hassan Ashtiani** ZOKAEIAM@MCMASTER.CA
*McMaster University, Hamilton, Canada*

**Vinayak Pathak** PATH.VINAYAK@GMAIL.COM
*Timaeus*

**Ruth Urner** RUTH@EECS.YORKU.CA
*York University, EECS Department, Toronto, Canada*

**Editors:** Nika Haghtalab and Ankur Moitra

## Abstract

Adversarially robust PAC learning has proved to be challenging, with the currently best known learners (Montasser et al., 2021a) relying on improper methods based on intricate compression schemes, resulting in sample complexity exponential in the VC-dimension. A series of follow up work considered a slightly relaxed version of the problem called adversarially robust learning *with tolerance* (Ashtiani et al., 2023; Bhattacharjee et al., 2023; Raman et al., 2024) and achieved better sample complexity in terms of the VC-dimension. However, those algorithms were either improper and complex, or required additional assumptions on the hypothesis class $\mathcal{H}$. We prove, for the first time, the existence of a simpler learner that achieves a sample complexity linear in the VC-dimension without requiring additional assumptions on $\mathcal{H}$. Even though our learner is improper, it is "almost proper" in the sense that it outputs a hypothesis that is "similar" to a hypothesis in $\mathcal{H}$.

We also use the ideas from our algorithm to construct a semi-supervised learner in the tolerant setting. This simple algorithm achieves comparable bounds to the previous (non-tolerant) semi-supervised algorithm of Attias et al. (2022a), but avoids the use of intricate subroutines from previous works, and is "almost proper."

**Keywords:** Adversarially robust learning, proper learning, semi-supervised learning

## 1. Introduction

In standard PAC-learning, the user encodes their domain knowledge by specifying a hypothesis class $\mathcal{H}$ that they think achieves a small expected loss on the data distribution. In adversarially robust PAC-learning, in addition to $\mathcal{H}$, the user also has knowledge of some perturbation type $\mathcal{U} : X \to 2^X$ encoding a belief that all points in the close environment $\mathcal{U}(x)$ of point $x$ share the same label as $x$. Imagine, for example, that $X$ is a domain of images, and $\mathcal{U}(x)$ is a small $\ell_0$-ball around $x$. This encodes the domain knowledge that one should not be able to change the label of an image by changing a small number of pixels. Thus, for standard PAC-learning, the goal is to find a hypothesis $h$ that achieves a small expected binary loss $\mathbb{E}_{(x,y)\sim P}\mathbb{1}\left[h(x) \neq y\right]$, whereas for adversarially robust PAC-learning the goal is to get a small expected adversarial loss $\mathbb{E}_{(x,y)\sim P}\mathbb{1}\left[\exists z \in \mathcal{U}(x) : h(z) \neq y\right]$.

Unlike standard PAC learning, adversarially robust PAC-learning of VC classes requires an improper learner in general, and the best known learner (Montasser et al., 2019) uses a potentially exponential number of samples in the VC-dimension of the class. To address this, Ashtiani et al. (2023) defined a mild relaxation of the problem where the expected adversarial loss (with respect to $\mathcal{U}$) of the learner is compared with the best achievable adversarial loss with respect to a slightly larger perturbation type $\mathcal{V}$. It was argued that the user is typically impartial to aim for robustness

with respect to $\mathcal{U}$ versus $\mathcal{V}$ if they are very "similar". A *tolerance* parameter $\gamma$ was then introduced to capture the relationship between $\mathcal{U}$ and $\mathcal{V}$ (see Section 2.1 for a precise definition). It has been shown (Ashtiani et al., 2023; Bhattacharjee et al., 2023) that tolerant adversarial learning can be achieved with a number of samples that is linear in the VC-dimension, $\log(\frac{1}{\gamma})$ and ambient dimension $d$. However, such a bound has so far only been achieved through either a complex compression-based algorithm (Ashtiani et al., 2023), or by adding additional assumptions on $\mathcal{H}$ and $\mathcal{U}$ (Bhattacharjee et al., 2023) (such as the property of "regularity" that has appeared in several other works (Awasthi et al., 2021; Raman et al., 2024) under varying names).

In our work, we show for the first time a *simple* learning algorithm that achieves sample complexity linear in VC, $\log(\frac{1}{\gamma})$, and $d$ and does not require any assumption on $\mathcal{H}$. In the realizable case, the algorithm is easy to state: run Robust ERM (RERM) on the training set to get a hypothesis $h \in \mathcal{H}$ and output a "smoothed" version of $h$: for a (perturbed) test point $z$ output the majority over a small set $\mathcal{W}(z)$ around $z$. For the agnostic case, we present a slight modification using a discrete "cover" of the space. Even though our learners are improper, the improperness only appears in the last smoothing step, meaning the final output is simply a smoothed version of a hypothesis in $\mathcal{H}$.

Next, we extend our ideas to the semi-supervised setting. In an earlier work (Attias et al., 2022a) it was proved that semi-supervised adversarially robust learning has a small labeled sample complexity as long as a large number of unlabeled samples are present. However, their algorithm involved invoking complex subroutines such as the one-inclusion-graph algorithm. We show that if we add tolerance then a simpler semi-supervised learner achieves comparable bounds.

### 1.1. Related work

Adversarially robust PAC-learning was formulated to study an empirical phenomenon, first encountered in image classification: state of the art models were vulnerable to adversarial attacks, namely imperceptible perturbations of an input image that led the otherwise highly accurate model to erroneously change its output (Szegedy et al., 2014). Adversarial robustness has since developed into a fertile area of research in the last decade. However, developing sound and efficient practical methods as well as obtaining a theoretical understanding of the problem remain challenging.

From a theoretical perspective, the sample complexity of adversarial PAC learning (Feige et al., 2015; Montasser et al., 2019) has been widely investigated (Feige et al., 2015; Attias et al., 2022b; Ashtiani et al., 2020; Montasser et al., 2021a). However, the best known upper bounds (Montasser et al., 2019, 2022) are based on rather involved and impractical learning methods, namely on intricate compression schemes (Moran and Yehudayoff, 2016) or one-inclusion-graphs. Also, the known sample complexity bounds are exponential in VC-dimension of the hypothesis class. Variations of the problem such as semi-supervised learning (Ashtiani et al., 2020; Attias et al., 2022a) and learning real-valued functions (Attias and Hanneke, 2023) have also been studied and similar upper bounds have been derived. It can be easily shown that VC-dimension does not provide a lower bound, since any class with infinite VC-dimension is trivially learnable when $\mathcal{U}(x)$ is the entire domain $X$. A dimension characterizing adversarially robust PAC-learning was obtained in (Montasser et al., 2022), but the dimension is based on a global variant of one-inclusion graph (Haussler et al., 1994) with potentially infinite vertices and edges. A simpler characterization remains elusive.

A major drawback of the standard PAC framing of adversarial robustness is the fixation on one perturbation type, which realistically cannot be known by the learner. Various alternatives have been investigated, such as robustness that is adaptive to the underlying distribution (Bhattacharjee and

Chaudhuri, 2021) or robustness with respect to a large collection of perturbation sets (Montasser et al., 2021a; Lechner et al., 2024). For the latter approach, it was shown that under structural assumptions on the class or perturbation types (such as a linear ordering by inclusion) and access to additional oracles (a perfect attack oracle provides witness points to adversarial vulnerability) adversarially robust learning is still possible.

Adversarial learning with tolerance was introduced in Ashtiani et al. (2023) and further studied by Bhattacharjee et al. (2023); Raman et al. (2024). A related notion (corresponding to the special case of tolerance parameter $\gamma = 1$ in our terminology) was considered in additional studies (Montasser et al., 2021b; Blum et al., 2022). Another relaxation of the adversarial problem was studied in Robey et al. (2022); Raman et al. (2024) where the requirement is robustness to the majority of perturbations in a perturbation set (as opposed to all perturbations in the set).

## 2. Notations and Setup

We denote by $X$ the input domain (often $X = \mathbb{R}^d$) and by $Y = \{0, 1\}$ the binary label space. We assume that $X$ is equipped with a metric dist. A hypothesis $h : X \to Y$ is a function that assigns a label to each point in the domain. A hypothesis class $\mathcal{H}$ is a set of hypotheses. For a sample $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, we use the notation $S_X = (x_1, x_2, \ldots, x_n)$ to denote the collection of domain points $x_i$ in $S$. The binary (also called 0-1) loss of $h$ on data point $(x, y) \in X \times Y$ is defined by $\ell^{0/1}(h, x, y) = \mathbb{1}[h(x) \neq y]$, where $\mathbb{1}[.]$ is the indicator function. Let $P$ by a probability distribution over $X \times Y$. Then the *expected binary loss* of $h$ with respect to $P$ is defined by $\mathcal{L}_P^{0/1}(h) = \mathbb{E}_{(x,y)\sim P}[\ell^{0/1}(h, x, y)]$. Similarly, the *empirical binary loss* of $h$ on sample $S = ((x_1, y_1), \ldots, (x_n, y_n)$ is defined as $\mathcal{L}_S^{0/1}(h) = \frac{1}{n}\sum_{i=1}^n \ell^{0/1}(h, x_i, y_i)$. We let the *approximation error* of $\mathcal{H}$ with respect to $P$ be denoted by $\mathcal{L}_P^{0/1}(\mathcal{H}) = \inf_{h\in\mathcal{H}} \mathcal{L}_P^{0/1}(h)$.

A *learner* $\mathcal{A}$ is a function that takes in a finite sequence of labeled instances $S \in (X \times Y)^n$ and outputs a hypothesis $h = \mathcal{A}(S)$. See Appendix Section A to recall the standard PAC learning requirement for binary classification (Vapnik and Chervonenkis, 1971; Valiant, 1984).

Throughout the paper, we use $\tilde{O}(\cdot)$ to hide factors poly-logarithmic in $1/\epsilon$.

### 2.1. Tolerant Adversarial PAC Learning

In robust learning under adversarial perturbations (or simply adversarial learning), it is assumed that an adversary can replace a test point $x$ with any point $z$ in $\mathcal{U}(x)$, where $\mathcal{U}(x) \subseteq X$ is a predefined set of "admissible perturbations" for $x$. We call the function $\mathcal{U} : X \to 2^X$ the *perturbation type*.

**Definition 1 (Adversarial loss )** *The adversarial loss of $h$ with respect to $\mathcal{U}$ on $(x, y) \in X \times Y$ is defined by $\ell^{\mathcal{U}}(h, x, y) = \max_{z\in\mathcal{U}(x)} \ell^{0/1}(h, z, y)$. The* expected adversarial loss *with respect to $P$ is defined by $\mathcal{L}_P^{\mathcal{U}}(h) = \mathbb{E}_{(x,y)\sim P}\ell^{\mathcal{U}}(h, x, y)$. Similarly, the* empirical adversarial loss *of $h$ on sample $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ is defined by $\mathcal{L}_S^{\mathcal{U}}(h) = \frac{1}{n}\sum_{i=1}^n \ell^{\mathcal{U}}(h, x_i, y_i)$. Finally, the* adversarial approximation error *of $\mathcal{H}$ with respect to $\mathcal{U}$ and $P$ is defined by $\mathcal{L}_P^{\mathcal{U}}(\mathcal{H}) = \inf_{h\in\mathcal{H}} \mathcal{L}_P^{\mathcal{U}}(h)$.*

The adversarial loss encompasses a classifier mispredicting on an instance and the instance being too close to the classifier's decision boundary. The following definition isolates the latter component.

**Definition 2 (Margin loss)** *Given $h \in \mathcal{H}, x \in X$, and a perturbation type $\mathcal{U}$, we define margin loss $\ell^{\mathcal{U},\mathrm{mar}}(h, x) = \mathbb{1}[\exists x_1, x_2 \in \mathcal{U}(x) : h(x_1) \neq h(x_2)]$. We define $\mathcal{L}_S^{\mathcal{U},\mathrm{mar}}$ and $\mathcal{L}_P^{\mathcal{U},\mathrm{mar}}$ accordingly.*

Analogously to PAC learning for the binary loss (Definition 16), one can define PAC learning with respect to the adversarial loss. We here define the more general setting of *tolerant* adversarial learning. Consider two perturbation types $\mathcal{U}$ and $\mathcal{V}$. We say $\mathcal{U}$ is *contained in* $\mathcal{V}$ and write it as $\mathcal{U} \prec \mathcal{V}$ if $\mathcal{U}(x) \subseteq \mathcal{V}(x)$ for all $x \in X$. Introducing tolerance relaxes adversarial learning by comparing the robust loss of the algorithm with respect to $\mathcal{U}$ with the approximation error of $\mathcal{H}$ with respect to a larger perturbation type $\mathcal{V}$.

**Definition 3 (Tolerant Adversarial PAC Learner(Ashtiani et al., 2023))** *Let $\mathcal{P}$ be a set of distributions over $X \times Y$, $\mathcal{H}$ a hypothesis class, and $\mathcal{U} \prec \mathcal{V}$ two perturbation types. We say $\mathcal{A}$ $(\mathcal{U}, \mathcal{V})$-*tolerantly *PAC learns $\mathcal{H}$ with respect to $\mathcal{P}$ with $m_{\mathcal{A}} : (0,1)^2 \to \mathbb{N}$ samples if the following holds: for every distribution $P \in \mathcal{P}$ and every $\epsilon, \delta \in (0,1)$, if $S$ is an i.i.d. sample of size at least $m_{\mathcal{A}}(\epsilon, \delta)$ from $P$, then with probability at least $1 - \delta$ (over the randomness of $S$) we have*

$$\mathcal{L}_P^{\mathcal{U}}(\mathcal{A}(S)) \leq \mathcal{L}_P^{\mathcal{V}}(\mathcal{H}) + \epsilon.$$

*We say $\mathcal{A}$ is a tolerant PAC learner in the* agnostic setting *if $\mathcal{P}$ is the set of all distributions over $X \times Y$, and in the* tolerantly realizable setting *if $\mathcal{P} = \{P : \mathcal{L}_P^{\mathcal{V}}(\mathcal{H}) = 0\}$.*

We call $\mathcal{U}$ the *actual perturbation type* and to $\mathcal{V}$ the *reference perturbation type*. The above definition recovers the standard definition of PAC learning under adversarial perturbations (Montasser et al., 2019) when $\mathcal{U}(x) = \mathcal{V}(x)$ for all $x \in X$. The smallest function $m : (0,1)^2 \to \mathbb{N}$ for which there exists a learner $\mathcal{A}$ that satisfies the above definition with $m_{\mathcal{A}} = m$ is referred to as the (realizable or agnostic) *sample complexity* of the problem.

In our work, we often consider a specific form of actual and reference perturbation types $\mathcal{U}$ and $\mathcal{V}$, namely $\mathcal{V}$ resulting from $\mathcal{U}$ by "inflating" it with a third perturbation type $\mathcal{W}$. Let $\mathcal{W}$ and $\mathcal{U}$ be two perturbation types. We now define $\mathcal{V}$ by setting $\mathcal{V}(x) = \{x'' \mid \exists x' \in \mathcal{U}(x) \text{ st } x'' \in \mathcal{W}(x')\}$[1]. Inspired by its role in our methods, we also refer to $\mathcal{W}$ as the *smoothing perturbation type*.

Perturbation types are defined naturally when $X$ is equipped with a metric $\text{dist}(.,.)$. In this case, $\mathcal{U}(x)$ can be defined by a ball of radius $r$ around $x$, i.e., $\mathcal{U}(x) = \mathcal{B}_r(x) = \{z \in X \mid \text{dist}(x,z) \leq r\}$. Now one can inflate $\mathcal{U}(x)$ with $\mathcal{W}(x) = \mathcal{B}_{r\gamma}(x)$ to create $\mathcal{V}(x) = \mathcal{B}_{(1+\gamma)r}(x)$. The perturbation types that we consider in this paper are mostly of this form. We call $\gamma > 0$ the *tolerance parameter* and we will refer to $(\mathcal{U}, \mathcal{V})$-tolerance also as *$\gamma$-tolerance* in this case.

We will further assume that our actual, reference and smoothing perturbation types $\mathcal{U}, \mathcal{V}$ and $\mathcal{W}$ are so that the perturbation sets $\mathcal{U}(x), \mathcal{V}(x)$ and $\mathcal{W}(x)$ admit the definition of a uniform measure over them. We use the notation $\mu_{\mathcal{U}(x)}, \mu_{\mathcal{V}(x)}$ and $\mu_{\mathcal{W}(x)}$ for these measures. We will use the notation $x \sim \mathcal{U}(x)$ etc to denote sampling from these uniform measures of the perturbation sets.

The following complexity measure was introduced by Montasser et al. (2019) and adopted in various works (Attias et al., 2022a; Shao et al., 2022) with different names. We adopt the one used by Attias et al. (2022a). The definition immediately yields $\text{VC}_{\mathcal{U}}(\mathcal{H}) \leq \text{VC}(\mathcal{H})$ for all $\mathcal{H}$ and $\mathcal{U}$.

**Definition 4 ($\text{VC}_{\mathcal{U}}$-dimension)** *Let $\mathcal{U} : X \to 2^X$ be some perturbation type and $\mathcal{H} \subseteq \{0,1\}^X$ a hypothesis class. We say that $\mathcal{H}$ $\mathcal{U}$-shatters a set of points $K \subseteq X$ if for every labeling $y \in \{0,1\}^K$ there exists a function $h \in \mathcal{H}$ with $h(z) = y(x)$ for all $z \in \mathcal{U}(x)$ and all $x \in K$. The $\text{VC}_{\mathcal{U}}$-dimension of the class $\mathcal{H}$ is the supremum over the sizes of sets that $\mathcal{H}$ can $\mathcal{U}$-shatter.*

1. The result of Montasser et al. (2021b) for transductive adversarial learning can be thought of as a result in the tolerant setting for $\mathcal{W}(x) = \mathcal{U}^{-1}(x) = \{x' \mid x \in \mathcal{U}(x')\}$.

### 2.2. Why add tolerance?

As has been argued before (Ashtiani et al., 2023), tolerance is a good way to capture the user's ambivalence over precisely which perturbation type to be robust against. If the user has a specific perturbation type $\mathcal{U}$ in mind, then we aim to find a hypothesis $h$ such that $\mathcal{L}_P^{\mathcal{U}}(h) \leq \mathcal{L}_P^{\mathcal{U}}(\mathcal{H}) + \epsilon$. But if the user is happy with any perturbation type between $\mathcal{U}$ and $\mathcal{V}$, we should aim for $\mathcal{L}_P^{\mathcal{U}}(h) \leq \mathcal{L}_P^{\mathcal{U}}(\mathcal{H}) + \xi + \epsilon$, where $\xi = \mathcal{L}_P^{\mathcal{V}}(\mathcal{H}) - \mathcal{L}_P^{\mathcal{U}}(\mathcal{H})$ denotes the extra slack due to user's ambivalence. This gives us Definition 3. Below we describe another motivation to study tolerance.

**Tolerance and learning with smoothed adversaries** Standard learning theory tells us that learning with respect to iid data is characterized by VC-dimension, but learning when data is generated by a worst-case adversary is characterized by Littlestone dimension (which can be much bigger than VC-dimension) (Shalev-Shwartz and Ben-David, 2014). Inspired by the smoothed analysis of algorithms (Spielman and Teng, 2004), it has been proved that if the worst-case adversary is "smoothed" then learning becomes characterized by VC-dimension again (Block et al., 2022; Haghtalab et al., 2020, 2024). To smooth an adversary, one takes the points generated by the worst-case adversary and perturbs them uniformly at random within a neighbourhood. In adversarially robust learning a worst-case adversary is allowed to generate points anywhere within $\mathcal{U}(x)$. One can consider a smoothed adversary in a similar way, where the point $z$ generated by the worst-case adversary is perturbed uniformly at random within $\mathcal{W}(z)$. Learning with respect to this adversary is then exactly the problem of tolerantly robust learning. Our results essentially show that adversarially robust learning with smoothed adversaries is easier.

### 2.3. Empirical Risk Minimization (ERM) and basic VC theory

It is well known that, for binary classification a class $\mathcal{H}$ is PAC learnable if and only if its *VC-dimension* is finite and that such classes can be learned through *Empirical Risk Minimization (ERM)*. More generally, ERM is a successful PAC learning principle whenever the VC-dimension of the *loss class* $\mathcal{H}_\ell$ of a class $\mathcal{H}$ induced by loss function $\ell : \{0,1\}^X \times X \times Y \to \{0,1\}$ is finite. This induced loss class is a collection of subsets of $X \times Y$ defined by $\mathcal{H}_\ell = \{h_\ell \subseteq X \times Y \ : \ h \in \mathcal{H}\}$, where $h_\ell = \{(x,y) \in X \times Y \ : \ \ell(h,x,y) = 1)\}$. For the adversarial loss with respect to a perturbation type $\mathcal{V}$ we will use the notation $\mathcal{H}_\mathcal{V}$ to denote the loss class of $\mathcal{H}$ with respect to $\ell^\mathcal{V}$.

Our learners often employ empirical risk minimzation with respect to various losses as a subroutine. We thus define the following notation for ERM (and Robust ERM) oracles.

**Definition 5 (ERM and RERM oracles)** *Let $S = ((x_1,y_1),\ldots,(x_m,y_m))$. An ERM oracle $\mathcal{A}_\mathcal{H}$ with respect to the hypothesis class $\mathcal{H}$ outputs any $\hat{h} \in \arg\min_{h\in\mathcal{H}} \mathcal{L}_S^{0/1}(h)$. An RERM (Robust ERM) oracle $\mathcal{A}_\mathcal{H}^\mathcal{V}$ with respect to class $\mathcal{H}$ and perturbation type $\mathcal{V}$ outputs any $\hat{h} \in \arg\min_{h\in\mathcal{H}} \mathcal{L}_S^{\mathcal{V}}(h)$.*

A standard result in VC theory is that a collection of subsets of finite VC-dimension enjoys finite sample uniform convergence (Vapnik and Chervonenkis, 1971). This immediately implies that any learner $\mathcal{A}$ that is an empirical risk minimizer for a class $\mathcal{H}$ with respect to loss $\ell$ (that is, $\mathcal{A}$ always outputs an $h \in \mathcal{H}$ with minimal empirical loss) is a successful PAC learner for $\mathcal{H}$ with respect to $\ell$. See appendix Section A for a reminder of this classic PAC learning result and the definition of the VC dimension. While the VC-dimension of the loss class $\mathcal{H}_\mathcal{U}$ for adversarial robust losses $\ell^\mathcal{U}$ can be arbitrarily larger than $\mathrm{VC}(\mathcal{H})$ (Montasser et al., 2019), it has been shown that the VC-dimension of the loss class for *finite* perturbation types can be bounded, a key compnent in our analysis.

**Lemma 6 ((Attias et al., 2022b) Lemma 1)** *Let $\mathcal{H} \subseteq \{0,1\}^X$ be some hypothesis class and let $\mathcal{C} : X \to 2^X$ be a perturbation type that satisfies $|\mathcal{C}(x)| \leq k$ for all $x \in X$ for some $k \in \mathbb{N}$. Then the VC-dimension of the robust loss class is bounded by $\mathrm{VC}(\mathcal{H}_C) \leq \mathrm{VC}(\mathcal{H}) \log(k)$.*

## 3. Supervised Tolerant Learning

The algorithms we propose for tolerantly robust learning of VC classes are *improper* learners. A common aspect of our methods is a two-stage approach, where the learner first determines an RERM hypothesis based on the training data, and then performs a post-processing step on this RERM hypothesis to obtain the final predictor, which in turn is not from the class $\mathcal{H}$. We show that this aspect of non-properness is *necessary* for any successful tolerantly robust learner, even when the perturbation types are balls in a Euclidean space and the sample complexity can additionally depend on the dimension of the space. The proof of the impossibility result below is in Appendix Section B

**Theorem 7** *For any $r \in \mathbb{R}$, any $d \in \mathbb{N}$ and any $g > 0$, there exist a hypothesis class $\mathcal{H}$ over $X = \mathbb{R}^d$ with $\mathrm{VC}(\mathcal{H}) = 1$ that is not properly tolerantly robustly PAC learnable (even in the tolerantly realizable case) for $\mathcal{U}(x) = \mathcal{B}_r(x)$ and $\mathcal{V}(x) = \mathcal{B}_{(1+\gamma)r}(x)$ for any $\gamma$ with $0 < \gamma \leq g$.*

### 3.1. Supervised tolerant learning in the realizable case

We start by presenting a simple robust learner, outlined in Algorithm 1 below, for the tolerantly realizable setting. It proceeds in two stages. Given data $S$, it first determines an RERM hypothesis $\hat{h}$ with respect to the reference perturbation type $\mathcal{V}$. It then smoothes this hypothesis by assigning each domain point $x$ the majority label of $\hat{h}$ in $\mathcal{W}(x)$ for a smoothing perturbation type $\mathcal{W}$.

**Algorithm 1** RERM-and-Smooth

**Input:** Data $S = ((x_1, y_1), \ldots, (x_m, y_m))$, access to an RERM oracle $\mathcal{A}^{\mathcal{V}}_{\mathcal{H}}$, and smoothing perturbation type $\mathcal{W}$.
Set $\hat{h} = \mathcal{A}^{\mathcal{V}}_{\mathcal{H}}(S)$
**Output:** $\mathrm{sm}_{\mathcal{W}}(\hat{h})$ defined by

$$\mathrm{sm}_{\mathcal{W}}(\hat{h})(x) = \mathbb{1}\left[\mathbb{E}_{x' \sim \mathcal{W}(x)} \hat{h}(x') \geq 1/2\right]$$

The analysis of the method in Algorithm 1 employs the notion of $\eta$-nets for the pre-images of the label classes under hypothesis class $\mathcal{H}$. For a hypothesis $h \in \{0,1\}^X$, we let $h_0 = h^{-1}(0) = \{x \in X : h(x) = 0\}$ denote the pre-image of label 0 under $h$, and analogously we let $h_1 = h^{-1}(1) = \{x \in X : h(x) = 1\}$ denote the pre-image of label 1 under $h$. Using this notation, we define the following version of an $\eta$-net for a binary hypothesis class.

**Definition 8 ($\eta$-net for class $\mathcal{H}$)** *Let $\mathcal{H} \subseteq \{0,1\}^X$ be some hypothesis class, let $D$ be a distribution over $X$ and let $\eta > 0$. A domain subset $C \subseteq X$ is an $\eta$-net for $\mathcal{H}$ with respect to $D$ if whenever $D(h_y) \geq \eta$ for some $y \in \{0,1\}$ and $h \in \mathcal{H}$, then $h_y \cap C \neq \emptyset$.*

An $\eta$-net $C$ for $\mathcal{H}$ ensures that whenever a label-pre-image $h_0 = h^{-1}(0)$ or $h_1 = h^{-1}(1)$ has mass at least $\eta$ under distribution $D$, then the net $C$ contains at least one point in this pre-image. Standard VC-theory guarantees that a sample of size $O\left(\frac{\mathrm{VC}(\mathcal{H})}{\eta} \log \frac{1}{\eta}\right)$ is a $\eta$-net for $\mathcal{H}$ with probability lower

bounded by a constant, say $2/3$. This in particular means that for any distribution $D$ and class $\mathcal{H}$ there exists an $\eta$-net of size $O\left(\frac{\mathrm{VC}(\mathcal{H})}{\eta}\log\frac{1}{\eta}\right)$.

The analysis of Algorithm 1 above will employ a discrete perturbation type $\mathcal{C} \prec \mathcal{V}$, that has finite perturbation sets $\mathcal{C}(x)$, is included in the reference perturbation type $\mathcal{V}$, and is so that each set $\mathcal{C}(x)$ is an $\eta$-net for the uniform distribution $\mu_{\mathcal{V}(x)}$ over the perturbation set $\mathcal{V}(x)$. By the above argument, the discrete type can be chosen so that the sizes of the perturbation sets are uniformly bounded, $|\mathcal{C}(x)| = O\left(\frac{\mathrm{VC}(\mathcal{H})}{\eta}\log\frac{1}{\eta}\right)$.

**Theorem 9** *Let $\mathcal{H}$ be a hypothesis class of finite VC-dimension ($\mathrm{VC}(\mathcal{H}) < \infty$), let $0 < \eta < 1/3$, and let $\mathcal{V}, \mathcal{U}$ and $\mathcal{W}$ be perturbation types that satisfy $\mathcal{U} \prec \mathcal{V}$, $\mathcal{W}(z) \subseteq \mathcal{V}(x)$ for all $x \in X$ and $z \in \mathcal{U}(x)$, and $\mu_{\mathcal{V}(x)}(\mathcal{W}(z)) \geq 3\eta$ for all $x \in X$ and $z \in \mathcal{U}(x)$. Then Algorithm 1 $(\mathcal{U}, \mathcal{V})$-tolerant robustly PAC learns $\mathcal{H}$ in the tolerantly realizable case with sample complexity bounded by*

$$m(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})\log(\mathrm{VC}(\mathcal{H})/\eta) + \log(1/\delta)}{\epsilon}\right).$$

**Proof** Let $\mathcal{U}, \mathcal{V}, \mathcal{W}$ and $\eta$ satisfy the conditions stated in the theorem. We define a new perturbation type $\mathcal{C} \prec \mathcal{V}$ as follows: for $x \in X$ set $\mathcal{C}(x)$ to be an $\eta$-net for $\mathcal{H}$ with respect to $\mu_{\mathcal{V}(x)}$, the uniform distribution over $\mathcal{V}(x)$. As outlined above, since $\mathrm{VC}(\mathcal{H})$ is finite, basic VC-theory guarantees that $\mathcal{C}$ exists and can be chosen so that $|\mathcal{C}(x)| = O(\mathrm{VC}(\mathcal{H}))/\eta$. We let $k \in \mathbb{N}$ denote a uniform upper bound on the sizes of the perturbation sets in $\mathcal{C}$, that is $|\mathcal{C}(x)| \leq k$ for all $x \in X$. This implies that the VC-dimension of the loss class $\mathcal{H}_{\mathcal{C}}$ of $\mathcal{H}$ with respect to $\ell^{\mathcal{C}}$ is bounded by $\mathrm{VC}(\mathcal{H})\log(k) = O(\mathrm{VC}(\mathcal{H})\log(\mathrm{VC}(\mathcal{H})/\eta))$ (Lemma 6).

Now, let $h$ be some hypothesis and let $\mathrm{sm}_{\mathcal{W}}(h)$ be the $\mathcal{W}$-smoothed version of $h$ as defined above. We next prove that for any distribution $P$, we have

$$\mathcal{L}_P^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(h)) \;\leq\; \mathcal{L}_P^{\mathcal{C}}(h) \;\leq\; \mathcal{L}_P^{\mathcal{V}}(h) \tag{1}$$

The second inequality is immediate from $\mathcal{C} \prec \mathcal{V}$. To prove the the first inequality we will show that it holds pointwise, that is $\ell^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(h), x, y) \leq \ell^{\mathcal{C}}(h, x, y)$ for all $(x, y) \in X \times Y$.

Indeed, assume that for some $(x, y) \in X \times Y$ we have $\ell^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(h), x, y) = 1$. This means that there exists some $z \in \mathcal{U}(x)$ with $\mathrm{sm}_{\mathcal{W}}(h)(z) \neq y$. This implies that $\mathbb{1}\left[\mathbb{E}_{x' \sim \mathcal{W}(z)} h(x') \geq 1/2\right] \neq y$, which means that $\mathbb{P}_{x' \sim \mathcal{W}(x)}[h(x') \neq y] > 1/2$. Now $\mu_{\mathcal{V}(x)}(\mathcal{W}(z)) \geq 3\eta$ together with $\mathcal{W}(z) \subseteq \mathcal{V}(x)$ (where $\mu_{\mathcal{V}(x)}$ is a uniform measure over $\mathcal{V}$) implies that $\mathbb{P}_{x' \sim \mathcal{V}(x)}[h(x') \neq y] \geq (3/2)\eta > \eta$. Now, since $\mathcal{C}(x)$ is an $\eta$-net with respect to $\mu_{\mathcal{V}(x)}$ for $\mathcal{H}$, this implies that there exists a $c \in \mathcal{C}(x)$ with $h(c) \neq y$. Thus, we have $\ell^{\mathcal{C}}(h, x, y) = 1$ which is what we needed to show.

Note that, since $\mathcal{C} \prec \mathcal{V}$, any distribution $P$ that is realizable by $\mathcal{H}$ with respect to $\ell^{\mathcal{V}}$ (as assumed in the tolerantly realizable case) is also realizable with respect to $\ell^{\mathcal{C}}$. Further, any RERM hypothesis with respect to $\ell^{\mathcal{V}}$ is also an RERM hypothesis class with respect to $\ell^{\mathcal{C}}$. Now the above bound on the VC-dimension on the loss class with respect to $\ell^{\mathcal{C}}$ implies that any RERM learner is a successful robust PAC learner for $\mathcal{C}$, and thus, with the stated samples sizes, with high probability at least $1 - \delta$ we have $\mathcal{L}_P^{\mathcal{C}}(\hat{h}) \leq \epsilon$. Now Equation 1 implies $\mathcal{L}_P^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(\hat{h})) \leq \epsilon$ as required. ■

**Remark 10** *If $\mathcal{V}$, $\mathcal{U}$ and $\mathcal{W}$ are Euclidian balls of radii $r(1+\gamma)$, $r$ and $r\gamma$ respectively, we have $\mu_{\mathcal{V}(x)}(\mathcal{W}(z)) = \frac{\gamma^d}{(1+\gamma)^d}$. Choosing $\eta = \frac{1}{3(1+1/\gamma)^d}$ yields sample complexity bound $m(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})(\log(\mathrm{VC}(\mathcal{H}))+d\log(1+1/\gamma))+\log(1/\delta)}{\epsilon}\right)$ with Algorithm 1 in the tolerantly realizable case.*

**Remark 11** *We emphasize that the learner in Algorithm 1 is not employing the discrete perturbation type $\mathcal{C}$ and thus does not need knowledge (or ability to construct) $\mathcal{C}$. Rather $\mathcal{C}$ and its properties as providing local $\eta$-nets are solely a tool of the analysis.*

A modification of the method presented here can be shown to also provide tolerantly robust guarantees in the agnostic case (see appendix Section C for details of this modification and analysis). However, the modified learner requires knowledge of the discrete perturbation type $\mathcal{C}$, thus we loose the attractive property from Remark 11. We next present an alternative method, which is slightly more complex, for tolerantly robust learning for which we derive guarantees in both the realizable and agnostic case. For that, we employ a global discretization which, in many natural settings, can be chosen to consist of grid points, and thus the learner can readily access this discretization. This method also achieves a slightly better sample complexity by eliminating the $\log(\mathrm{VC}(\mathcal{H}))$ factor.

### 3.2. Supervised agnostic tolerant learning using a global discretization

In this section we present a tolerant learner that uses a global discretization of the space. More specifically, the method is based on a countable domain subset $C \subseteq X$ as a discretization of the space. We will show that if $C$ is a $r\gamma$-cover of $X = \mathbb{R}^d$, that is, the discretization $C$ is such that for all $x \in X$ there exists a point $c \in C$ with $\mathrm{dist}(x, c) \leq r\gamma$, our algorithm is a successful tolerant robust PAC learner for perturbation types $\mathcal{U}(x) = \mathcal{B}_r(x)$ and $\mathcal{V}(x) = \mathcal{B}_{(1+\gamma)r}(x)$. Given a discretization $C \subseteq X$, we let $\mathcal{C}$ denote the induced discretization of perturbation type $\mathcal{V}$, that is $\mathcal{C}(x) = \mathcal{V}(x) \cap C$.

For concreteness, we may assume that $C$ consists of evenly spaced grid points of a grid with side-length $2r\gamma/d$. This will be an $r\gamma$-cover with respect to any $\ell_p$-norm with $p \geq 1$, and the sizes of the sets $\mathcal{C}(x)$ will be uniformly bounded by $k := \mathcal{C}(x) \leq \left(\frac{(1+\gamma)d}{\gamma}\right)^d = \Theta\left((1+\frac{1}{\gamma})^d d^d\right)$.

Our proposed method, Algorithm 2 below, acts in two stages. Given a training dataset $S$ and discretization $C$, it first determines an RERM hypothesis $\hat{h}$ with respect to the discretized perturbation type $\mathcal{C}$. Given $\hat{h}$, it then produces a discretized version of this predictor by assigning every domain point $x$ the $\hat{h}$ label of its nearest neighbor in $C$ (breaking ties arbitrarily). More precisely, for a set $C \subseteq X$ and hypothesis $h \in \{0,1\}^X$ we define the $C$-nearest neighbor discretized hypothesis $\mathrm{nn}_C(h)$ by

$$\mathrm{nn}_C(h)(x) = h(z) \text{ for some } z \in \mathrm{argmin}_{c \in C}\mathrm{dist}(x, c).$$

We note that, in general, $\mathrm{nn}_C(h) \notin \mathcal{H}$ even for hypotheses $h \in \mathcal{H}$. However the predictor $\mathrm{nn}_C(h)$ can be viewed as "close to being from $\mathcal{H}$" in the sense that $\mathrm{nn}_C(h)$ is a discretized version of $h$, its decision boundary being moved slightly to pass along grid lines (or more generally voronoi cells) induced by $C$. In that sense, informally, our $\gamma$-tolerant learner is "$\gamma$-close to being proper".

**Algorithm 2** RERM-and-Discretize

**Input:** Data $S = ((x_1, y_1), \ldots, (x_m, y_m))$, discretization $C \subseteq X$, access to RERM oracle $\mathcal{A}^{\mathcal{C}}_{\mathcal{H}}$.
Set $\hat{h} = \mathcal{A}^{\mathcal{C}}_{\mathcal{H}}(S)$.
**Output:** $h = \mathrm{nn}_{\mathcal{C}}(\hat{h})$ defined by
$\mathrm{nn}_C(\hat{h})(x) = \hat{h}(z)$ for some $z \in \mathrm{argmin}_{c \in C}\mathrm{dist}(x, c)$

**Theorem 12** *Let $\mathcal{H}$ be a hypothesis class of finite VC-dimension $\mathrm{VC}(\mathcal{H}) < \infty$ and let $\mathcal{V}$ and $\mathcal{U}$ be perturbation types that are balls of radii $r(1+\gamma)$ and $r$ respectively for some $r > 0$. Let discretization $C$ be an $r\gamma$-cover of $X$, let $\mathcal{C}$ be the induced perturbation type, where $\mathcal{C}(x) = \mathcal{V}(x) \cap C$ for all $x \in X$ and let $k$ be a uniform upper bound on the perturbations sets in $\mathcal{C}$, that is $|\mathcal{C}(x)| \leq k$ for all $x \in X$. Then Algorithm 2 $\gamma$-tolerant robustly PAC learns $\mathcal{H}$. Moreover the sample complexity in the tolerantly realizable case is bounded by $m(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})\log(k)+\log(1/\delta)}{\epsilon}\right)$ and in the agnostic case by $n(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})\log(k)+\log(1/\delta)}{\epsilon^2}\right)$.*

**Proof** Let $h$ be some hypothesis and let $\mathrm{nn}_C(h)$ be the $C$-discretized version of $h$ as defined above. We start by proving that for any distribution $P$, we have

$$\mathcal{L}_P^{\mathcal{U}}(\mathrm{nn}_C(h)) \;\leq\; \mathcal{L}_P^{\mathcal{C}}(h) \;\leq\; \mathcal{L}_P^{\mathcal{V}}(h) \tag{2}$$

We show that the first inequality holds pointwise, $\ell^{\mathcal{U}}(\mathrm{nn}_C(h), x, y) \leq \ell^{\mathcal{C}}(h, x, y)$ for all $(x, y) \in X \times Y$ (the second follows from $\mathcal{C} \prec \mathcal{V}$). Indeed, assume we have $\ell^{\mathcal{U}}(\mathrm{nn}_C(h), x, y) = 1$ for some $(x, y)$. Then there exists some $z \in \mathcal{U}(x)$ with $\mathrm{nn}_C(h)(z) \neq y$, and thus for some $z' \in \mathrm{argmin}_{c \in C}\mathrm{dist}(z, c)$ we have $h(z') \neq y$. Since $C$ is a $r\gamma$-cover of $X$ we know $\mathrm{dist}(z, z') \leq r\gamma$. Further, since $z \in \mathcal{U}(x)$, we know $\mathrm{dist}(x, z) \leq r$. Now the triangle inquality implies $\mathrm{dist}(x, z') \leq \mathrm{dist}(x, z) + \mathrm{dist}(z, z') \leq r + r\gamma = r(1+\gamma)$. Thus $z' \in C \cap \mathcal{V}(x) = \mathcal{C}(x)$, and now $h(z') \neq y$ means $\ell^{\mathcal{C}}(h, x, y) = 1$. Thus we have $\ell^{\mathcal{U}}(\mathrm{nn}_C(h), x, y) \leq \ell^{\mathcal{C}}(h, x, y)$ for all $(x, y) \in X \times Y$, which implies Equation 2 above.

Since $\mathcal{C}(x) \leq k$ for all $x \in X$, we have $\mathrm{VC}(\mathcal{H}_{\mathcal{C}}) \leq \mathrm{VC}(\mathcal{H})\log(k)$ (Lemma 6). For the realizable case, the result now follows exactly as in the proof of Theorem 9. For the agnostic case, note that Equation 2 also implies that for any distribution $P$, $\mathcal{L}_P^{\mathcal{C}}(\mathcal{H}) \leq \mathcal{L}_P^{\mathcal{V}}(\mathcal{H})$. Due to $\mathrm{VC}(\mathcal{H}_{\mathcal{C}}) \leq \mathrm{VC}(\mathcal{H})\log(k)$, any empirical risk minimizing learner $\mathcal{A}_{\mathcal{H}}^{\mathcal{C}}$, PAC learns $\mathcal{H}$ with respect to $\ell^C$, yielding $\mathcal{L}_{\mathcal{C}}^P(\hat{h}) \leq \mathcal{L}_P^{\mathcal{C}}(\mathcal{H}) + \epsilon \leq \mathcal{L}_{\mathcal{V}}^P(\mathcal{H}) + \epsilon$ with high probability at least $1 - \delta$ over the training samples for the stated agnostic sample sizes. Combining this expression with Equation 2, we obtain $\mathcal{L}_P^{\mathcal{U}}(\mathrm{nn}_C(\hat{h})) \leq \mathcal{L}_P^{\mathcal{C}}(\hat{h}) \leq \mathcal{L}_P^{\mathcal{V}}(\mathcal{H}) + \epsilon$ as required. ■

**Remark 13** *As outlined above, in case of $X = \mathbb{R}^d$ equipped with any $\ell_p$ norm, we can choose the $r\gamma$-cover $C$ so that $k := \mathcal{C}(x) \leq \left(\frac{(1+\gamma)\sqrt{d}}{2\gamma}\right)^d = \Theta\left((1+\frac{1}{\gamma})^d d^{d/2}\right)$, thus we obtain sample complexity bounds $m(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})d(\log(1+\frac{1}{\gamma})+\log d)+\log(1/\delta)}{\epsilon}\right)$ in the tolerantly realizable case and $n(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})d(\log(1+\frac{1}{\gamma})+\log d)+\log(1/\delta)}{\epsilon^2}\right)$ in the agnostic case.*

## 4. Semi-supervised learning

In semi-supervised learning, in addition to the $m$ labeled examples $S_l = ((x_1, y_1), \ldots, (x_m, y_m))$, the learner $\mathcal{A}$ is also given $n$ unlabeled examples $S_u = (x_1, \ldots, x_n)$ drawn iid from the marginal $P_X$ of the data distribution. Analogous to the supervised setting, we can define agnostic and realizable, as well as their tolerant versions for the semi-supervised setting. The sample complexity of semi-supervised learning can be quantified by two functions: the number of labeled samples (defined by function $m_l(\epsilon, \delta)$) and the number of unlabeled samples (denoted by $m_u(\epsilon, \delta)$).

The sample complexity of semi-supervised adversarially robust learning was studied in (Attias et al., 2022a) where it was shown that in the realizable case, the labeled sample complexity with respect to a perturbation set $\mathcal{U}$ is characterized by $\mathrm{VC}_{\mathcal{U}}$, see Definition 4, which can be significantly smaller than the standard VC-dimension of a class. They also showed that in the agnostic case bounding the labeled sample complexity with $\mathrm{VC}_{\mathcal{U}}$ is impossible. On the other hand, they proposed a "factor-$\alpha$ agnostic learner" (a.k.a. a semi-agnostic learner) that guarantees $\mathcal{L}_P^{\mathcal{U}}(\mathcal{A}(S_l, S_u)) \leq \alpha \cdot \mathcal{L}_P^{\mathcal{U}}(\mathcal{H}) + \epsilon$ for $\alpha = 3$ and whose labeled sample complexity is characterized by $\mathrm{VC}_{\mathcal{U}}$. Specifically, they showed that given a supervised agnostic learner with sample complexity $m(\epsilon, \delta)$, the realizable semi-supervised case can be solved with $m(\epsilon/3, \delta/2)$ unlabeled and $O\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log 1/\delta}{\epsilon}\right)$ labeled samples, and the factor-3 agnostic semi-supervised case can be solved with $m(\epsilon/3, \delta/2)$ unlabeled and $O\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{\log 1/\delta}{\epsilon^2}\right)$ labeled samples.

Their algorithm has two steps. First, they transform $\mathcal{H}$ into a "partial" concept class (Alon et al., 2022), and use the 1-inclusion graph based learner from (Alon et al., 2022) to learn a partial concept $h$ using the labeled set $S_l$. Next, they label $S_u$ using $h$ and invoke the compression-based adversarially robust agnostic learner from (Montasser et al., 2021a) on this set. While the algorithm is simple to describe, both of its steps invoke other algorithms that are complex and rely on improper learning. For a tolerantly robust learning guarantee, we can use in their second step the simpler supervised learner with a tolerant robustness guarantee, thus achieving an unlabeled sample complexity of $\tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})d(\log(1+\frac{1}{\gamma})+\log d)+\log(1/\delta)}{\epsilon^2}\right)$.

However, we can show that with tolerance, their algorithms can be further simplified if we allow for an extra $\log \mathrm{VC}(\mathcal{H})$ factor in the labeled sample complexity. For both realizable and agnostic settings, we now present algorithms that are easy to state, do not require using complex subroutines, and achieve a labeled sample complexity that depends logarithmically on $\mathrm{VC}(\mathcal{H})$. Moreover, like the supervised case, our algorithms are "almost proper."

#### 4.0.1. REALIZABLE

Our main insight is that using just the unlabeled set $S_u$, we can identify a small, finite set $\mathcal{H}'$ of candidates from $\mathcal{H}$ such that robust learning $\mathcal{H}'$ suffices. To create $\mathcal{H}'$, we simply iterate through all robustly realizable labelings $y = (y_1, \ldots, y_m)$ of $S_u$ and call RERM on $((x_1, y_1), \ldots, (x_m, y_m))$.

As in the result of Theorem 12 for Algorithm 2, our SSL method in Algorithm 3 employs an $r\gamma$-cover $C$ as a global discretization of the space, and uses the induced perturbation type $\mathcal{C}$ with $\mathcal{C}(x) = \mathcal{V}(x) \cap C$ for all $x$. As discussed in Section 3.2, defining $C$ to be a grid with appropriate side-length is a simple way to obtain such a cover.

**Theorem 14** *Algorithm 3 is a $\gamma$-tolerant adversarially robust learner in the realizable setting with unlabeled sample complexity $m_u = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})(d\log(1+1/\gamma)+\log d)+\log 1/\delta}{\epsilon}\right)$ and labeled sample complexity $m_l = \tilde{O}\left(\frac{\mathrm{VC}_{\mathcal{V}}(\mathcal{H})\log m_u+\log 1/\delta}{\epsilon^2}\right)$.*

**Proof** The high level idea of the proof is as follows. Since this is the realizable setting, let $h^* \in \mathcal{H}$ be a hypothesis such that $\mathcal{L}_P^{\mathcal{V}}(h^*) = 0$. In the main for loop (Line 3) of the algorithm, when the label $y = (h^*(x_1), \ldots, h^*(x_n))$ is picked, we show that the output $h'$ will satisfy $\mathcal{L}_P^{\mathcal{C}}(h') \leq \frac{\epsilon}{2}$ (with high probability over the randomness of $S_u$) as long as $S_u$ is of an appropriate size (to be bound later). Thus, the approximation error of the finite class $\mathcal{H}'$ with respect to $\ell^{\mathcal{C}}$ is bounded by $\frac{\epsilon}{2}$. Next, recall

**Algorithm 3** Semi-supervised learner (realizable)

**Input:** Labeled data $S_l = ((x_1, y_1), \ldots, (x_m, y_m))$, unlabeled data $S_u = (x_1, \ldots, x_n)$, access to RERM oracle $\mathcal{A}^{\mathcal{V}}_{\mathcal{H}}$, $r\gamma$-cover $C \subseteq X$
Set $\mathcal{H}' = \{\}$
**for** each labeling $y = (y_1, \ldots, y_n) \in \{0,1\}^n$ of $S_u$ **do**
  Create labeled set $S^y_u = ((x_1, y_1), \ldots, (x_n, y_n))$.
  Let $h' = \mathcal{A}^{\mathcal{V}}_{\mathcal{H}}(S^y_u)$
  **if** $\mathcal{L}^{\mathcal{V}}_{S^y_u}(h') = 0$ **then**
    Add $h'$ to $\mathcal{H}'$
  **end if**
**end for**
Define RERM oracle $\mathcal{A}^{\mathcal{C}}_{\mathcal{H}'}$ for induced discrete perturbation type $\mathcal{C}(x) = C \cap \mathcal{V}(x)$
Let $\hat{h} = \mathcal{A}^{\mathcal{C}}_{\mathcal{H}'}(S_l)$.
**Output:** $\mathrm{nn}_C(\hat{h})$ defined by:
  $\mathrm{nn}_C(\hat{h})(x) = \hat{h}(x')$ where $x'$ is the nearest neighbour of $x$ in $C$.

that for finite perturbation types such as $\mathcal{C}$, the loss class $\mathcal{H}_{\mathcal{C}}$ of a class $\mathcal{H}$ of finite VC-dimension, has finite VC-dimension as well (see Lemma 6). Thus RERM learner $\mathcal{A}^{\mathcal{C}}_{\mathcal{H}'}$ is an agnostic learner with respect to $\ell^{\mathcal{C}}$ for $\mathcal{H}'$. This implies that $\mathcal{L}^{\mathcal{C}}_P(\hat{h}) \leq \min_{h \in \mathcal{H}'} \mathcal{L}^{\mathcal{C}}_P(h) + \frac{\epsilon}{2}$ (with high probability over the randomness of $S_l$) as long as $S_l$ is of an appropriate size. Thus overall, $\mathcal{L}^{\mathcal{C}}_P(\hat{h}) \leq \epsilon$ and this implies $\mathcal{L}^{\mathcal{U}}_P(\mathrm{nn}_C(\hat{h})) \leq \epsilon$ (as in proof of Theorem 12, see Equation 2 therein).

Now it only remains to argue that the stated bounds on the size of $S_l$ and $S_u$ suffice for the above. Since we use $S_u$ to obtain an $h'$ that is $\frac{\epsilon}{2}$-close to $h^*$ in terms of $\mathcal{L}^{\mathcal{C}}_P$, an unlabeled data set size $|S_u| \geq O\left(\frac{\mathrm{VC}(\mathcal{H}) \log k + \log 1/\delta}{\epsilon}\right)$ where $k = \max_x |\mathcal{C}(x)|$ suffices (see Lemma 6). Using the bound on $k$, we get $O\left(\frac{\mathrm{VC}(\mathcal{H}) d(\log(1+1/\gamma) + \log d) + \log 1/\delta}{\epsilon}\right)$. Next, to bound the size of $S_l$, note that $\mathcal{H}'$ is a finite set and thus $\mathrm{VC}(\mathcal{H}') \leq O(\log |\mathcal{H}'|)$. Moreover, $\mathcal{H}'$ contains exactly one hypothesis per robustly realizable (with respect to $\mathcal{V}$) labeling of $S_u$. From a simple application of Sauer lemma, we get that $|\mathcal{H}'| \leq O\left(|S_u|^{\mathrm{VC}_{\mathcal{V}}(\mathcal{H})}\right)$ and thus $\mathrm{VC}(\mathcal{H}') \leq O(\mathrm{VC}_{\mathcal{V}}(\mathcal{H}) \log |S_u|)$. Thus number of labeled samples required is $|S_l| \geq O\left(\frac{\mathrm{VC}_{\mathcal{V}}(\mathcal{H}) \log |S_u| d(\log(1+1/\gamma) + \log d) + \log 1/\delta}{\epsilon}\right)$. ■

#### 4.0.2. AGNOSTIC

For the agnostic case, the set $\mathcal{H}'$ formed by calling RERM for all $2^{|S_u|}$ labelings of $S_u$ can be quite large: unlike the realizable case, since the outputs of RERM no longer robustly label all points of $S_u$, the size of the resulting $\mathcal{H}'$ cannot be bounded by anything better than $O\left(|S_u|^{\mathrm{VC}(\mathcal{H})}\right)$. However, we show that we can prune $\mathcal{H}'$ to get a new smaller set $\mathcal{H}''$ such that learning with respect to $\mathcal{H}''$ on the labeled set gives a factor-3 agnostic learner. The purpose of pruning is to ensure that if two hypotheses $h_1, h_2 \in \mathcal{H}'$ have the property that whenever both robustly label a point in $S_u$, they give it the same label, then we keep only one of them. We still need to decide which one to keep. Our algorithm keeps the one that is robust on a larger number of points.

Similar to the realizable case, we assume we are given an $r\gamma$-cover $C$ and the corresponding perturbation type $\mathcal{C}$. In the agnostic case, we need to call RERM$^{\mathcal{C}}$ instead of RERM$^{\mathcal{V}}$ for creating

$\mathcal{H}'$. As a result, the bound on $|\mathcal{H}''|$ is $O\left(|S_u|^{\mathrm{VC}_{\mathcal{C}}(\mathcal{H})}\right)$ and thus the labeled sample complexity depends on $\mathrm{VC}_{\mathcal{C}}(\mathcal{H})$, which can be bigger than $\mathrm{VC}_{\mathcal{U}}(\mathcal{H})$, but still smaller than $\mathrm{VC}(\mathcal{H})$. We provide a proof sketch for the guarantees of our algorithm. The detailed proof can be found in Appendix D.

**Algorithm 4** Semi-supervised learner (agnostic)

**Input:** Labeled data $S_l = ((x_1, y_1), \ldots, (x_m, y_m))$, unlabeled data $S_u = (x_1, \ldots, x_n)$, $r\gamma$-cover $C \subseteq X$, access to RERM oracle $\mathcal{A}^{\mathcal{C}}_{\mathcal{H}}$.
Set $\mathcal{H}' = \{\}$
**for** each labeling $y = (y_1, \ldots, y_n) \in \{0,1\}^n$ of $S_u$ **do**
  Create labeled set $S^y_u = ((x_1, y_1), \ldots, (x_n, y_n))$.
  Add $h' = \mathcal{A}^{\mathcal{C}}_{\mathcal{H}}(S^y_u)$ to $\mathcal{H}'$.
**end for**
$\mathcal{H}'' = \{\}$
**for** s = 0 to $|S_u|$ **do**
  **for** $h \in \mathcal{H}'$ such that $\mathcal{L}^{\mathcal{C},\mathrm{mar}}_{S_u}(h) = s/|S_u|$ **do**
    **if** $\forall h' \in \mathcal{H}', \exists x \in S_u$ such that $\ell^{\mathcal{C},\mathrm{mar}}(h, x) = \ell^{\mathcal{C},\mathrm{mar}}(h', x) = 0$, but $h'(x) \neq h(x)$ **then**
      Add $h$ to $\mathcal{H}''$
    **end if**
  **end for**
**end for**
Define RERM oracle $\mathcal{A}^{\mathcal{C}}_{\mathcal{H}''}$ for induced discrete perturbation type $\mathcal{C}(x) = C \cap \mathcal{V}(x)$
Let $\hat{h} = \mathrm{RERM}^{\mathcal{C}}_{\mathcal{H}''}(S_l)$.
**Output:** $\mathrm{nn}_C(\hat{h})$ defined by:
  $\mathrm{nn}_C(\hat{h})(x) = \hat{h}(x')$ where $x'$ is the nearest neighbour of $x$ in $C$.

**Theorem 15** *Algorithm 4 is a factor-3 agnostic learner in the semi-supervised setting with tolerance parameter $\gamma$ with unlabeled sample complexity $m_u = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})d(\log(1+1/\gamma)+\log d)+\log 1/\delta}{\epsilon^2}\right)$ and labeled sample complexity $m_l = \tilde{O}\left(\frac{\mathrm{VC}_{\mathcal{C}}(\mathcal{H})\log m_u+\log 1/\delta}{\epsilon^2}\right)$.*

**Proof** [sketch] The proof involves two steps. First we show $|\mathcal{H}''| \leq O\left(|S_u|^{\mathrm{VC}_{\mathcal{C}}(\mathcal{H})\log|S_u|}\right)$, and then we show there exists $h'' \in \mathcal{H}''$ such that $\mathcal{L}^{\mathcal{C}}_P(h'') \leq 3 \cdot \mathcal{L}^{\mathcal{C}}_P(\mathcal{H}) + \epsilon$.

To bound the size, we use a more general version of Sauer lemma from (Alon et al. (2022) Theorem 12). Recall that the standard Sauer lemma shows that the number of different labelings induced by a hypothesis class $\mathcal{H}$ on $m$ points is bounded by $O\left(m^{\mathrm{VC}(\mathcal{H})}\right)$. Here, two labelings are considered different if they are different on at least one of the $m$ points. The way we have defined $\mathcal{H}''$, any two $h_1, h_2 \in \mathcal{H}''$ are *robustly* different on at least one point, i.e., there exists $x \in S_u$ such that $h_1(x) \neq h_2(x)$ and $\ell^{\mathcal{C},\mathrm{mar}}(h_1, x) = \ell^{\mathcal{C},\mathrm{mar}}(h_2, x) = 0$. The generalized Sauer lemma can be used to show that the number of such hypotheses depends on the size of the largest set they can *robustly* shatter, in particular, that $|\mathcal{H}''| \leq O\left(|S_u|^{\mathrm{VC}_{\mathcal{C}}(\mathcal{H})\log|S_u|}\right)$.

To show that $\mathcal{H}''$ contains a good hypothesis, a similar argument as in the realizable case shows that there exists $h' \in \mathcal{H}'$ such that $\mathcal{L}^{\mathcal{C}}_P(h') \leq \mathcal{L}^{\mathcal{C}}_P(\mathcal{H}) + \epsilon$, implying $\mathcal{H}'$ contains a good hypothesis. But the risk is that it might get pruned out in the pruning step. However, if it gets pruned, it is because of another hypothesis $h''$ that robustly labels more points of $S_u$ than $h'$ and is consistent with $h''$ on

every $x \in S_u$ that is robustly labeled by both. For any $h$, let $S_u^h = \{x \in S_u \mid \ell^{\mathcal{C},\mathrm{mar}}(h,x) = 0\}$. Then $h'$ and $h''$ must robustly agree on at least $\frac{(1-|S_u^{h'}|+1-|S_u^{h''}|)}{|S_u|} \leq 2 \cdot \frac{(1-|S_u^{h'}|)}{|S_u|} \approx 2 \cdot \mathcal{L}_P^{\mathcal{C}}(\mathcal{H})$ fraction of the points. Since $\mathcal{L}_P^{\mathcal{C}}(h') \approx \mathcal{L}_P^{\mathcal{C}}(\mathcal{H})$, and $h', h''$ robustly agree on $2 \cdot \mathcal{L}_P^{\mathcal{C}}(\mathcal{H})$ fraction of points, we get that $\mathcal{L}_P^{\mathcal{C}}(h'') \approx 3 \cdot \mathcal{L}_P^{\mathcal{C}}(\mathcal{H})$. Here the last few steps rely on some carefully constructed generalization arguments that have been laid out in Appendix D. ■

## 5. Discussion

The main message of this work is that adding tolerance vastly simplifies the task of adversarially robust learning, and is perhaps a more natural formulation of the robust learning problem in the first place. Recent developments in statistical learning theory often involved establishing novel sample complexity bounds or characterizations of learnability through rather impractical learners (Montasser et al., 2022; Brukhim et al., 2022). While these provide important insights into learnability, we view our work also as promoting more PAC type analysis of methods that are closer to what is used in applications, as well as explorations into how the frameworks of analysis we choose may affect the feasibility of developing such guarantees for natural learners. Our work shows how the slight shift to tolerance for the adversarial robustness task enables PAC type guarantees for simpler learners. One interesting open question on a technical level is to eliminiate the dependence on $d$ for $\gamma = 1$. Montasser et al. (2021b) prove the existence of a *transductive* learner for $\gamma = 1$ whose sample complexity is independent of $d$. It will be nice to explore if our techniques can be used to obtain similar bounds in the inductive case. Our algorithms depend on $d$ even for $\gamma = 1$ essentially because we need to cover the set $\mathcal{V}(x)$ with $\mathcal{W}$ thus giving us a cover of size $(1 + 1/\gamma)^d$. If, instead, we only had to cover $\mathcal{U}(x)$, we would get $1/\gamma^d$, which, for $\gamma = 1$ becomes independent of $d$.

## Acknowledgments

Vinayak thanks Open Philanthropy for supporting part of the research. Hassan and Ruth are both also affiliate faculty members at Toronto's Vector Institute and their research is supported by (separate) NSERC Discovery grants.

## References

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of pac learnability of partial concept classes. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021*, pages 658–671. IEEE Computer Society, 2022.

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Black-box certification and learning under adversarial perturbations. In *International Conference on Machine Learning*, pages 388–398. PMLR, 2020.

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Adversarially robust learning with tolerance. In *International Conference on Algorithmic Learning Theory*, pages 115–135. PMLR, 2023.

Idan Attias and Steve Hanneke. Adversarially robust pac learnability of real-valued functions. In *International Conference on Machine Learning*, pages 1172–1199. PMLR, 2023.

Idan Attias, Steve Hanneke, and Yishay Mansour. A characterization of semi-supervised adversarially robust pac learnability. *Advances in Neural Information Processing Systems*, 35:23646–23659, 2022a.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for adversarially robust learning. *Journal of Machine Learning Research*, 23(175):1–31, 2022b.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the existence of the adversarial bayes classifier. *Advances in Neural Information Processing Systems*, 34:2978–2990, 2021.

Robi Bhattacharjee and Kamalika Chaudhuri. Consistent non-parametric methods for maximizing robustness. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 9036–9048, 2021.

Robi Bhattacharjee, Max Hopkins, Akash Kumar, Hantao Yu, and Kamalika Chaudhuri. Robust empirical risk minimization with tolerance. In *International Conference on Algorithmic Learning Theory*, pages 182–203. PMLR, 2023.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *Conference on Learning Theory*, pages 1716–1786. PMLR, 2022.

Avrim Blum, Omar Montasser, Greg Shakhnarovich, and Hongyang Zhang. Boosting barely robust learners: A new perspective on adversarial robustness. *Advances in Neural Information Processing Systems*, 35:1307–1319, 2022.

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.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 943–955. IEEE, 2022.

Uriel Feige, Yishay Mansour, and Robert Schapire. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory, COLT*, pages 637–657, 2015.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. *Advances in Neural Information Processing Systems*, 33:9203–9215, 2020.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. *Journal of the ACM*, 71(3):1–34, 2024.

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

David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Tosca Lechner, Vinayak Pathak, and Ruth Urner. Adversarially robust learning with uncertain perturbation sets. *Advances in Neural Information Processing Systems*, 36, 2024.

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

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

Omar Montasser, Steve Hanneke, and Nathan Srebro. Transductive robust learning guarantees. *arXiv preprint arXiv:2110.10602*, 2021b.

Omar Montasser, Steve Hanneke, and Nati Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. *Advances in Neural Information Processing Systems*, 35:37458–37470, 2022.

Shay Moran and Amir Yehudayoff. Sample compression schemes for vc classes. *Journal of the ACM (JACM)*, 63(3):1–10, 2016.

Vinod Raman, Unique Subedi, and Ambuj Tewari. On proper learnability between average-and worst-case robustness. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexander Robey, Luiz Chamon, George J Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average and worst-case performance. In *International Conference on Machine Learning*, pages 18667–18686. PMLR, 2022.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Han Shao, Omar Montasser, and Avrim Blum. A theory of pac learnability under transformation invariances. *Advances in Neural Information Processing Systems*, 35:13989–14001, 2022.

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR*, 2014.

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

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

## Appendix A. Basic VC theory

The following definition abstracts the notion of PAC learning (Vapnik and Chervonenkis, 1971; Valiant, 1984).

**Definition 16 (PAC Learner)** *Let $\mathcal{P}$ be a set of distributions over $X \times Y$ and $\mathcal{H}$ a hypothesis class. We say $\mathcal{A}$ PAC learns $\mathcal{H}$ with respect to $\mathcal{P}$ with $m_{\mathcal{A}} : (0,1)^2 \to \mathbb{N}$ samples if the following holds: for every distribution $P \in \mathcal{P}$ over $X \times Y$, and every $\epsilon, \delta \in (0,1)$, if $S$ is an i.i.d. sample of size at least $m_{\mathcal{A}}(\epsilon, \delta)$ from $P$, then with probability at least $1 - \delta$ (over the randomness of $S$) we have*

$$\mathcal{L}_P(\mathcal{A}(S)) \leq \mathcal{L}_P(\mathcal{H}) + \epsilon.$$

*$\mathcal{A}$ is called an* agnostic learner *if $\mathcal{P}$ is the set of all distributions*[2] *on $X \times Y$, and a* realizable learner *if $\mathcal{P} = \{P : \mathcal{L}_P(\mathcal{H}) = 0\}$.*

The following is a classic result of PAC learnability for finite VC classes.

**Theorem 17 ((Vapnik and Chervonenkis, 1971; Blumer et al., 1989; Haussler, 1992))** *Let $\mathcal{H} \subseteq \{0,1\}^X$ be a hypothesis class and let $\ell$ be a loss function such that $\mathrm{VC}(\mathcal{H}_\ell) < \infty$ is finite. Then, any empirical risk minizing learner $\mathcal{A}$ PAC learns $\mathcal{H}$ with respect to loss $\ell$ with sample complexity $O\left(\frac{\mathrm{VC}(\mathcal{H}_\ell)+\log(1/\delta)}{\epsilon^2}\right)$ in the agnostic case and sample complexity $\tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H}_\ell)+\log(1/\delta)}{\epsilon}\right)$ in the realizable case.*

For completeness, we here also provide the definition of the VC-dimension for collection of subsets. Note that any binary hypothesis class $\mathcal{H} \subseteq \{0,1\}^X$ is also a collection of subsets of $X \times \{0,1\}$.

**Definition 18 (VC-dimension)** *Let $Z$ be some domain set and $\mathcal{C} \subseteq 2^Z$ be a collection of subsets of $Z$. We say that some set $K \subseteq Z$ is* shattered *by $\mathcal{C}$ if $|\{c \cap K \ : \ c \in \mathcal{C}\}| = 2^{|K|}$. The* VC-dimension *of $\mathcal{C}$ is the supremum over the sizes of domain subsets that are shattered by $\mathcal{C}$.*

## Appendix B. Proof of Theorem 7

As a reminder, we restate the Theorem here:

**Theorem 7** *For any $r \in \mathbb{R}$, any $d \in \mathbb{N}$ and any $g > 0$, there exist a hypothesis class $\mathcal{H}$ over $X = \mathbb{R}^d$ with $\mathrm{VC}(\mathcal{H}) = 1$ that is not properly tolerantly robustly PAC learnable (even in the tolerantly realizable case) for $\mathcal{U}(x) = \mathcal{B}_r(x)$ and $\mathcal{V}(x) = \mathcal{B}_{(1+\gamma)r}(x)$ for any $\gamma$ with $0 < \gamma \leq g$.*

**Proof** Let $r \in \mathbb{R}$, $d \in \mathbb{N}$ and $g > 0$ be given. We will provide a construction with $X = \mathbb{R}^1$, which can readily be embedded into any higher dimensional space. Our definition of a hypothesis class $\mathcal{H}$ will closely follow the construction of the hardness of proper learning in the case of standard adversarial robustness without tolerance (Theorem 1 by Montasser et al. (2019)).

We construct a hypothesis class $\mathcal{H}$ with VC-dimension 1 for which the VC-dimension of the adversarial loss class $\mathcal{H}_{\mathcal{V}}$ is arbitrarily large. Moreover, we will define $\mathcal{H}$ in such a way that the

2. Subject to mild measurability conditions, namely the loss sets being measurable for all $h \in \mathcal{H}$.

adversarial losses with respect to $\mathcal{V}$ and $\mathcal{U}$ are identical, and thus tolerance will not alleviate the difficulty (as it would in the original construction).

Let $n \in \mathbb{N}$ be given. We now first construct a class $\mathcal{H}_n$ with VC-dimension 1 and loss class VC-dimension $n$. Consider $n$ points $x_1, x_2, \ldots, x_n \in \mathbb{R}$ spaced apart with distances $|x_i - x_j| > 2r(1+g)$ for all $i \neq j$. That is, the points are positioned so that balls of radius $(1+\gamma)r$ around any two $x_i \neq x_j$ do not intersect for any $0 < \gamma < g$. Let $(p_i)_{i \in \mathbb{N}}$ be an enumeration of prime numbers (that is $p_1 = 2, p_2 = 3, p_3 = 5$ etc). Define a one to one mapping $f$ between subsets of $[n] = \{1, 2, 3, \ldots, n\}$ and the first $2^n$ prime numbers.

For each subset $Z \subseteq [n]$, we define a hypothesis $h_Z$ as follows:

$$h_Z(x) = \begin{cases} 1 & \text{if } x = x_j + \frac{r}{(p_{f(Z)})^m} \text{ for } j \in Z, m \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

That is, each hypothesis $h_Z$ labels most of $X = \mathbb{R}$, and in particular all points $x_1, x_2, \ldots, x_n$ with 0. The only points $h_Z$ labels with 1 are the above defined sequences approaching the points $x_j$ where $j$ is an index in $Z$. These sequences are the points $x_j + \frac{r}{(p_{f(Z)})^m}$ for $m \in \mathbb{N}$, that is defined by a prime number that is uniquely associated with the set $Z$.

Now we set $\mathcal{H}_n$ be the set of all these hypotheses, that is $\mathcal{H}_n = \{h_Z \ : \ Z \subseteq [n]\}$. Note that in this construction, every point in $X = \mathbb{R}$ gets labeled with 1 by at most one hypothesis in $\mathcal{H}_n$, and thus $\mathrm{VC}(\mathcal{H}_n) = 1$.

Now note that for a subset $Z \subseteq [n]$, and an $j \in Z$, the point $x_j$ is arbitrarily close to points that are labeled with 1 by $h_Z$, and thus the robust loss $\ell^{\mathcal{B}}(h_Z, x_j, 0) = 1$ for any ball perturbation type $\mathcal{B}(x) = \mathcal{B}_\rho(x)$ for any (arbitrarily small) radius $\rho > 0$. In particular $\ell^{\mathcal{U}}(h_Z, x_j, 0) = \ell^{\mathcal{V}}(h_Z, x_j, 0) = 1$ if (and only if) $j \in Z$. This shows that the VC dimension of the robust loss class $\mathrm{VC}(\mathcal{H}_{\mathcal{V}}) = \mathrm{VC}(\mathcal{H}_{\mathcal{U}}) = n$.

From here the argument for impossibility of proper learning in the $(\mathcal{U}, \mathcal{V})$-tolerantly realizable setting proceeds as in the proof of Theorem 1 by Montasser et al. (2019). We restrict the class $\mathcal{H}_n$ to only contain functions corresponding to subsets $n/2$:

$$\mathcal{H}'_n = \{h_Z \ : \ Z \subseteq [n], |Z| = n/2\}$$

. Now, we consider the set of distributions $\mathcal{P}_n$ that ditribute their mass uniformly over $n/2$ domain points among the $x_1, x_2, \ldots x_n$ with label 0. These distributions are $\mathcal{V}$-robustly realizable by $\mathcal{H}'_n$, in particular the function $h_Z$ for $Z$ that is the complement of the distribution's support in $\{x_1, x_2, \ldots x_n\}$ has robust loss 0 with respect to $\mathcal{V}$. However, standard arguments show that any *proper* learner that sees only samples of sizes $n/4$ cannot correctly identify this required complement set and thus has high probabitiy of outputting a function $h_Z \in \mathcal{H}'_n$ where $Z$ intersects significantly with the support of the distribution and thus suffers high robust loss. In particular it suffers this high robust loss with respect to $\mathcal{V}$ and equally with respect to $\mathcal{U}$ or balls of *any* smaller radius. Thus, the relaxation to the tolerant setting does not facilitate proper learning in this case.

Finally, by copying the above constrution into disjoint intervals of $X = \mathbb{R}$ for all $m \in \mathbb{N}$ the resulting class $\mathcal{H} = \bigcup_{m \in \mathbb{N}} \mathcal{H}'_m$ has VC dimension $\mathrm{VC}(\mathcal{H}) = 1$, but cannot be learned by any proper learner in the tolerantly realazble setting.

■

## Appendix C. Local discretization for both realizable and agnostic setting

A slightly modified version of our first method Algorihm 1 can be shown to also work in the agnostic case. For completeness, we state the modified algorithm and complete and unified analysis for both realizabe and agnostic settings here. Algorithm 5 below is a variant where the initial RERM hypothesis is chosen with respect to a discretized type $\mathcal{C} \prec \mathcal{V}$.

**Algorithm 5** RERM-and-Smooth

**Input:** Data $S = ((x_1, y_1), \ldots, (x_m, y_m))$, access to an RERM oracle $\mathcal{A}^{\mathcal{V}}_{\mathcal{H}}$, and smoothing perturbation type $\mathcal{W}$.
Set $\hat{h} = \mathcal{A}^{\mathcal{V}}_{\mathcal{H}}(S)$
**Output:** $\mathrm{sm}_{\mathcal{W}}(\hat{h})$ defined by

$$\mathrm{sm}_{\mathcal{W}}(\hat{h})(x) = \mathbb{1}\left[\mathbb{E}_{x' \sim \mathcal{W}(x)} \hat{h}(x') \geq 1/2\right]$$

**Theorem 19** *Let $\mathcal{H}$ be a hypothesis class of finite VC-dimension ($\mathrm{VC}(\mathcal{H}) < \infty$), let $0 < \eta < 1/3$, and let $\mathcal{V}, \mathcal{U}, \mathcal{W}$ and $\mathcal{C}$ be perturbation types that satisfy*
- *$\mathcal{U} \prec \mathcal{V}$ and $\mathcal{C} \prec \mathcal{V}$,*
- *$\mathcal{W}(z) \in \mathcal{V}(x)$ for all $x \in X$ and $z \in \mathcal{U}(x)$,*
- *$\mathcal{C}(x)$ is finite and an $\eta$-net of $\mathcal{H}$ with respect the uniform measure $\mu_{\mathcal{V}(x)}$ over $\mathcal{V}(x)$ for all $x \in X$,*
- *$\mu_{\mathcal{V}(x)}(\mathcal{W}(z)) \geq 3\eta$ for all $x \in X$ and $z \in \mathcal{U}(x)$.*

*Then Algorithm 1 $(\mathcal{U}, \mathcal{V})$-tolerant robustly PAC learns $\mathcal{H}$. Moreover, the perturbation type $\mathcal{C}$ can be chosen so that the resulting sample complexity in the tolerantly realizable case is bounded by*

$$m(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H}) \log(\mathrm{VC}(\mathcal{H})/\eta) + \log(1/\delta)}{\epsilon}\right)$$

*and in the agnostic case by*

$$n(\epsilon, \delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H}) \log(\mathrm{VC}(\mathcal{H})/\eta) + \log(1/\delta)}{\epsilon^2}\right)$$

**Proof** Let $h$ be some hypothesis and let $\mathrm{sm}_{\mathcal{W}}(h)$ be the $\mathcal{W}$-smoothed version of $h$ as defined above. We start by proving that for any distribution $P$, we have

$$\mathcal{L}^{\mathcal{U}}_P(\mathrm{sm}_{\mathcal{W}}(h)) \;\leq\; \mathcal{L}^{\mathcal{C}}_P(h) \;\leq\; \mathcal{L}^{\mathcal{V}}_P(h) \tag{3}$$

The second inequality is immediate by observing that $\mathcal{C}(x) \subseteq \mathcal{V}(x)$ for all $x \in X$. To prove the the first inequality we will show that it holds pointwise, that is $\ell^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(h), x, y) \leq \ell^{\mathcal{C}}(h, x, y)$ for all $(x, y) \in X \times Y$.

Indeed, assume that for some $(x, y) \in X \times Y$ we have $\ell^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(h), x, y) = 1$. This means that there exists some $z \in \mathcal{U}(x)$ with $\mathrm{sm}_{\mathcal{W}}(h)(z) \neq y$. This implies that $\mathbb{1}\left[\mathbb{E}_{x' \sim \mathcal{W}(z)} \hat{h}(x') \geq 1/2\right] \neq y$, which means that $\mathbb{P}_{x' \sim \mathcal{W}(x)}[\hat{h}(x') \neq y] > 1/2$. Now $\mu_{\mathcal{V}(x)}(\mathcal{W}(z)) \geq 3\eta$ together with $\mathcal{W}(z) \subseteq \mathcal{V}(x)$ (where $\mu_{\mathcal{V}(x)}$ is a uniform measure over $\mathcal{V}$) implies that $\mathbb{P}_{x' \sim \mathcal{V}(x)}[\hat{h}(x') \neq y] \geq (3/2)\eta > \eta$. Now, since $\mathcal{C}(x)$ is an $\eta$-net with respect to $\mu_{\mathcal{V}(x)}$ for $\mathcal{H}$, this implies that there exists a $c \in \mathcal{C}(x)$ with $\hat{h}(c) \neq y$. Thus, we have $\ell^{\mathcal{C}}(h, x, y) = 1$ which is what we needed to show.

Since $\mathrm{VC}(\mathcal{H})$ is finite, $\mathcal{C}$ can be chosen so that $|\mathcal{C}(x)| \leq 3\mathrm{VC}(\mathcal{H})/\eta$ and satisfy $\mathcal{C}(x)$ being an $\eta$-net for $\mathcal{H}$ with respect to $\mu_{\mathcal{V}(x)}$ for each $\mathcal{V}(x)$. Note that since the perturbation sets of type $\mathcal{C}$ are finite and their sizes are uniformly upper bounded by $3\mathrm{VC}(\mathcal{H})/\eta$, the VC-dimension of the loss class of $\mathcal{H}$ with respect to $\ell^{\mathcal{C}}$ is bounded by $\mathrm{VC}(\mathcal{H})\log(3\mathrm{VC}(\mathcal{H})/\eta)$ (see Lemma 6).

**Realizable case.** Note that, since $\mathcal{C} \prec \mathcal{V}$, any distribution $P$ that is realizable by $\mathcal{H}$ with respect to $\ell^{\mathcal{V}}$ (as assumed in the tolerantly realizable case) is also realizable with respect to $\ell^{\mathcal{C}}$. Now the above bound on the VC-dimension on the loss class with respect to $\ell^{\mathcal{C}}$ implies that any RERM learner is a successful robust PAC learner for perturbation type $\mathcal{C}$, and thus, with the stated samples sizes, with high probability at least $1-\delta$ we have $\mathcal{L}_P^{\mathcal{C}}(\hat{h}) \leq \epsilon$. Now Equation 1 implies $\mathcal{L}_P^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(\hat{h})) \leq \epsilon$ as required.

**Agnostic case.** Note that Equation 1 also implies that for any distribution $P$, the approximation error of $\mathcal{H}$ with respect to $\ell^{\mathcal{C}}$ is upper bounded by the approximation error of $\mathcal{H}$ with respect to $\ell^{\mathcal{V}}$

$$\mathcal{L}_P^{\mathcal{C}}(\mathcal{H}) \;\leq\; \mathcal{L}_P^{\mathcal{V}}(\mathcal{H})$$

By the bound on the VC-dimension of the loss class of $\mathcal{H}$ with respect to $\ell^{\mathcal{C}}$, any empirical risk minimizing learner for $\mathcal{H}$ with respect to $\ell^{\mathcal{C}}$ outputting $\hat{h}$, is a successful robust PAC learner with respect to $\ell^{C}$, yielding

$$\mathcal{L}_P^{\mathcal{C}}(\hat{h}) \leq \mathcal{L}_P^{\mathcal{C}}(\mathcal{H}) + \epsilon \;\leq\; \mathcal{L}_P^{\mathcal{V}}(\mathcal{H}) + \epsilon$$

with high probability at least $1-\delta$ over the training samples for the stated sample sizes in the agnostic case. Combining this expression with Equation 1, we obtain

$$\mathcal{L}_P^{\mathcal{U}}(\mathrm{sm}_{\mathcal{W}}(\hat{h})) \;\leq\; \mathcal{L}_P^{\mathcal{C}}(\hat{h}) \;\leq\; \mathcal{L}_P^{\mathcal{V}}(\mathcal{H}) + \epsilon$$

as required for tolerantly robust learning. This completes the proof of the theorem.

■

**Remark 20** *For the case that $\mathcal{V}$, $\mathcal{U}$ and $\mathcal{W}$ are Euclidian balls of radii $r(1+\gamma)$, $r$ and $r\gamma$ respectively in $\mathbb{R}^d$, have $\mu_{\mathcal{V}(x)}(\mathcal{W}(z)) = \frac{\gamma^d}{(1+\gamma)^d}$. Thus we can chose $\eta = \frac{1}{3(1+1/\gamma)^d}$ and obtain sample complexity bounds*

$$m(\epsilon,\delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})(\log(\mathrm{VC}(\mathcal{H})) + d\log(1+1/\gamma)) + \log(1/\delta)}{\epsilon}\right)$$

*in the tolerantly realizable case and*

$$n(\epsilon,\delta) = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})(\log(\mathrm{VC}(\mathcal{H})) + d\log(1+1/\gamma)) + \log(1/\delta)}{\epsilon^2}\right)$$

*in the agnostic case.*

## Appendix D. Proofs for agnostic semi-supervised learning

**Theorem 15** *Algorithm 4 is a factor-3 agnostic learner in the semi-supervised setting with tolerance parameter $\gamma$ with unlabeled sample complexity $m_u = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})d(\log(1+1/\gamma)+\log d)+\log 1/\delta}{\epsilon^2}\right)$ and labeled sample complexity $m_l = \tilde{O}\left(\frac{\mathrm{VC}_{\mathcal{C}}(\mathcal{H})\log m_u+\log 1/\delta}{\epsilon^2}\right)$.*

**Proof** We divide the proof into two parts. First, we show that $\mathcal{H}''$ has a small size, and then we show that $\mathcal{H}''$ contains a hypothesis whose robust loss with respect to $\mathcal{C}$ is not much worse compared to the optimal hypothesis $h^*$. These two facts combined show that $\hat{h}$ gets a small robust loss with respect to $\mathcal{C}$. This means $\mathrm{nn}_C(\hat{h})$ gets a small robust loss with respect to $\mathcal{V}$.

To show that $\mathcal{H}''$ has a small size we use a result from (Alon et al., 2022). For any hypothesis $h \in \mathcal{H}$, define a "partial" hypothesis $\mathrm{p}(h)$ as follows. We say $\mathrm{p}(h)(x) = 0$ if $h(z) = 0$ for every $z \in \mathcal{C}(x)$, $\mathrm{p}(h)(x) = 1$ if $h(z) = 1$ for every $z \in \mathcal{C}(x)$, and $\mathrm{p}(h)(x) = \star$ otherwise. Transforming every hypothesis of $\mathcal{H}$ in this way gives us a new hypothesis class $\mathrm{p}(\mathcal{H})$. We call such a hypothesis class a partial hypothesis class, and a class that does not contain any partial hypotheses a total hypothesis class. Given any set $S = \{x_1, \ldots, x_m\} \in X^m$, we say that $\mathrm{p}(\mathcal{H})$ shatters $S$ if for every labeling $\{y_1, \ldots, y_m\} \in \{0,1\}^m$, there exists some $\mathrm{p}(h) \in \mathrm{p}(\mathcal{H})$ such that $\mathrm{p}(h)(x_i) = y_i$ for every $x_i \in S$. Note that even though $\mathrm{p}(h)$ is allowed to output $\star$, it is supposed to shatter a set only using the labels 0 and 1. The VC-dimension of the class $\mathrm{p}(\mathcal{H})$ is defined as the size of the biggest set that is shattered by it. It is easy to see that this is equal to $\mathrm{VC}_{\mathcal{C}}$. We restate the following lemma for our context, which was originally shown in (Alon et al., 2022).

**Lemma 21 ( Alon et al. (2022) Theorem 12)** *Given the set $S_u$ and the class $p(\mathcal{H})$ of partial hypotheses, there exists a class $\tilde{\mathcal{H}}$ of total hypotheses such that $|\tilde{\mathcal{H}}| \leq |S_u|^{O(\mathrm{VC}_{\mathcal{C}}(\mathcal{H})\log|S_u|)}$, and for every $p(h) \in p(\mathcal{H})$, there exists $\tilde{h} \in \tilde{\mathcal{H}}$ such that $\tilde{h}(x) = p(h)(x)$ for every $x \in S_u$ satisfying $p(h)(x) \neq \star$.*

To see that $\mathcal{H}''$ is small, consider the corresponding partial version $\mathrm{p}(\mathcal{H}'')$. Using Lemma 21, we can construct a total hypothesis class $\tilde{\mathcal{H}}''$ that has a hypothesis $\tilde{h}$ for every $h \in \mathcal{H}''$ such that $h(x) = \tilde{h}(x)$ for every $x \in S_u$ that satisfies $\ell^{\mathcal{C},\mathrm{mar}}(h,x) = 0$ (i.e., $h$ is robust on $x$). Moreover, we can see that for each $h$, the corresponding $\tilde{h}$ is unique. This is because due to the way $\mathcal{H}''$ is constructed, for any two $h_1, h_2 \in \mathcal{H}''$, there exists $x \in S_u$ such that $h_1(x) \neq h_2(x)$ and $\ell^{\mathcal{C},\mathrm{mar}}(h_1,x) = \ell^{\mathcal{C},\mathrm{mar}}(h_2,x) = 0$. This means the corresponding $\tilde{h_1}$ and $\tilde{h_2}$ must be different and $x$ is the witness. Since there is a unique hypothesis in $\tilde{\mathcal{H}}''$ for every hypothesis in $\mathcal{H}''$, we get $|\mathcal{H}''| \leq |\tilde{\mathcal{H}}''| \leq |S_u|^{O(\mathrm{VC}_{\mathcal{C}}(\mathcal{H})\log|S_u|)}$.

Next, we show that $\mathcal{H}''$ contains a "good" hypothesis. Imagine that $S_u$ was generated by first sampling a labeled set of size $|S_u|$ from distribution $P$ and then removing the labels. Let $y = (y_1, \ldots, y_m)$ be the labels that were removed. Then, at some point $y$ will be considered by the first for loop (Line 3) in Algorithm 4 and a corresponding hypothesis $h'$ will be added to $\mathcal{H}'$. Since $h'$ is the result of running RERM$^{\mathcal{C}}$, using Lemma 6, we know that with high probability over the randomness of $S_u$, we have $\mathcal{L}_P^{\mathcal{C}}(h') \leq \mathcal{L}_P^{\mathcal{C}}(\mathcal{H}) + \epsilon$. Thus $\mathcal{H}'$ does contain a good hypothesis. But the risk is that it might get pruned out in the pruning step.

We can show that because of the way we do the pruning, there will be another hypothesis $h'' \in \mathcal{H}''$ that is not much worse than $h'$ on $S_u$. But we want a guarantee wrt $P$, and thus we need a uniform convergence result that links the closeness of two hypotheses on $S_u$ with their closeness on $P$.

We use the following lemmas.

**Definition 22** *For any two hypotheses $h_1, h_2 \in \mathcal{H}$ and $x \in X$, define:*

$$\ell^{\mathcal{C},par}(h_1, h_2, x) = \mathbb{1}\left[p(h_1)(x) \neq p(h_2)(x)\right].$$

*Also, define $\mathcal{L}_{S_u}^{\mathcal{C},par}(h_1, h_2)$ and $\mathcal{L}_{P}^{\mathcal{C},par}(h_1, h_2)$ appropriately.*

**Lemma 23** *For $S_u \geq O\left(\frac{\mathrm{VC}(\mathcal{H})d(\log(1+1/\gamma)+\log d)+\log 1/\delta}{\epsilon^2}\right)$, with probability at least $1-\delta$ over the randomness of $S_u$, we have that $|\mathcal{L}_{S_u}^{\mathcal{C},par}(h_1, h_2) - \mathcal{L}_{P}^{\mathcal{C},par}(h_1, h_2)| \leq \epsilon$*

**Proof** Let $\max_{x\in X} |\mathcal{C}(x)| \leq k$ and consider the hypothesis class $\mathcal{H} \times \mathcal{H}$ consisting of pairs of hypotheses from $\mathcal{H}$. For any $h_1, h_2 \in \mathcal{H}$ define the function $(h_1, h_2) : X \to Y$ as $(h_1, h_2)(x) = \ell^{\mathcal{C},\mathrm{par}}(h_1, h_2, x)$. We show that $\mathrm{VC}(\mathcal{H} \times \mathcal{H}) \leq O(\mathrm{VC}(\mathcal{H}) \log k)$. To see this, consider a sequence $S = (x_1, \ldots, x_m)$ of $m$ points from $X$ and count the number of patterns induced on it by members of $\mathcal{H} \times \mathcal{H}$. It is easy to see that for any $x \in X$, the value of $(h_1, h_2)(x)$ is uniquely determined once we specify the values of $h_1(x), h_2(x), \ell^{\mathcal{C},\mathrm{mar}}(h_1, x)$, and $\ell^{\mathcal{C},\mathrm{mar}}(h_2, x)$. Thus the total number of patterns is at most the product of the number of patterns induced by each and thus can be bounded by $m^{O(\mathrm{VC}(\mathcal{H}) \log k)}$). Standard uniform convergence results for VC classes concludes the proof. ■

**Lemma 24** *For any two hypotheses $h_1, h_2$, $|\mathcal{L}_P^{\mathcal{C}}(h_1) - \mathcal{L}_P^{\mathcal{C}}(h_2)| \leq \mathcal{L}_P^{\mathcal{C},par}(h_1, h_2)$.*

**Proof** In fact, it is easy to see that for any $x$, if $\ell^{\mathcal{C},\mathrm{par}}(h_1, h_2, x) = 0$, then for all $y \in \{0, 1\}$, $\ell^{\mathcal{C}}(h_1, x, y) = \ell^{\mathcal{C}}(h_2, x, y)$. Thus the lemma follows. ■

Now, we are ready to complete the proof. If $h' \in \mathcal{H}''$, then $\mathcal{H}''$ contains a good hypothesis. Otherwise, if $h'$ gets pruned out, that can only be because there was another hypothesis $h''$ such that $|S_u^{h''}| \geq |S_u^{h'}|$ and $h'(x) = h''(x)$ for all $x \in S_u^{h'} \cap S_u^{h''}$. We have:

$$\begin{aligned}
\mathcal{L}_{S_u}^{\mathcal{C},\mathrm{par}}(h', h'') &\leq \frac{(1 - |S_u^{h'}|) + (1 - |S_u^{h''}|)}{|S_u|} \\
&\leq 2 \cdot \frac{1 - |S_u^{h'}|}{|S_u|} \\
&= 2 \cdot \mathcal{L}_{S_u}^{\mathcal{C},\mathrm{mar}}(h') \\
&\leq 2 \cdot \mathcal{L}_{P}^{\mathcal{C},\mathrm{mar}}(h') + 2\epsilon \\
&\leq 2 \cdot \mathcal{L}_{P}^{\mathcal{C}}(h') + 2\epsilon \\
&= 2 \cdot \mathcal{L}_{P}^{\mathcal{C}}(\mathcal{H}) + 2\epsilon \qquad (4)
\end{aligned}$$

Finally, when we run RERM$^{\mathcal{C}}$ on $S_l$ with respect to $\mathcal{H}''$, we get $\hat{h}$, that satisfies:

$$\begin{aligned}
\mathcal{L}_P^{\mathcal{C}}(\hat{h}) &\leq \mathcal{L}_P^{\mathcal{C}}(\mathcal{H}'') + \epsilon \\
&\leq \mathcal{L}_P^{\mathcal{C}}(h'') + \epsilon \\
&\leq \mathcal{L}_P^{\mathcal{C}}(h') + \mathcal{L}_P^{\mathcal{C},\mathrm{par}}(h'', h') + 2\epsilon \qquad (5) \\
&\leq \mathcal{L}_P^{\mathcal{C}}(\mathcal{H}) + \mathcal{L}_{S_u}^{\mathcal{C},\mathrm{par}}(h'', h') + 3\epsilon \qquad (6) \\
&\leq 3 \cdot \mathcal{L}_P^{\mathcal{C}}(\mathcal{H}) + 5\epsilon \qquad (7)
\end{aligned}$$

Here, (5) follows from Lemma 24, (6) follows from Lemma 23 and (7) follows from (4). Thus overall, $\mathcal{L}_P^{\mathcal{C}}(\hat{h}) \leq \epsilon$ and this implies $\mathcal{L}_P^{\mathcal{U}}(\mathrm{nn}_C(\hat{h})) \leq \epsilon$ (as in proof of Theorem 12, Equation 2). ■