A. Note on our notation for sets and functions

We use the following notation for sets and functions:

2^X	the power-set (set of all subsets) of X
Y^X	the set of all functions from X to Y
$f : X \to Y$	f is a function from X to Y

Functions from some set X to some set Y are a special type of relations between X and Y. Thus a function $f : X \to Y$ is a subset of $X \times Y$, namely

$$f = \{(x, y) \in X \times Y \mid y = f(x)\}$$

If $h : X \to Y$ is a (not necessarily binary) classifier, and P is a probability distribution over $X \times Y$, then the probability of misclassification is $P(\mathrm{err}_h)$, where err_h is the complement of h in $X \times Y$, that is

$$\mathrm{err}_h = \{(x, z) \in X \times Y \mid z \neq f(x)\} = (X \times Y) \setminus h$$

If $Y = \{0, 1\}$ is a binary label space, then it is also common to identify classifiers $h : X \to \{0, 1\}$ with a subset of the domain, namely the set $h^{-1}(1)$, that is the set of points that is mapped to label 1 under h:

$$h^{-1}(1) = \{x \in X \mid h(x) = 1\}$$

We switch between identifying h with $h^{-1}(1)$ and viewing h as a subset of $X \times Y$, depending on which view aids the simplicity of argument in a given context.

We defined the margin areas of a classifier (with respect to a perturbation type) again as subsets of $X \times Y$.

$$\mathrm{mar}_h^{\mathcal{U}} = \{(x, y) \in X \times Y \mid \exists z \in \mathcal{U}(x) : h(x) \neq h(z)\}$$

Note, that here, if for a given domain point x, we have $(x, y) \in \mathrm{mar}_h^{\mathcal{U}}$ for some $y \in Y$, then $(x, y') \in \mathrm{mar}_h^{\mathcal{U}}$ for all $y' \in Y$. Thus, the sets $\mathrm{mar}_h^{\mathcal{U}} \subseteq X \times Y$ are not functions. Rather, they can naturally be identified with their projection on X, and we again do so if convenient in the context.

The given definitions of err_h and $\mathrm{mar}_h^{\mathcal{U}}$, naturally let us express the robust loss as the probably measure of a subset of $X \times Y$:

$$\mathcal{L}_P^{\mathcal{U}}(h) = P(\mathrm{err}_h \cup \mathrm{mar}_h^{\mathcal{U}}).$$

B. Note on measurability

Here, we note that allowing the perturbation type $\mathcal{U}$ to be an arbitrary mapping from the domain X to 2^X can easily lead to the adversarial loss being not measurable, even if $\mathcal{U}(x)$ is a measurable set for every x. Consider the case $X = \mathbb{R}$, and a distribution P with P_X uniform on the interval $[0, 2]$. Consider a subset $M \subseteq (0, 1)$ that is not Borel-measurable. Consider a simple threshold function

$$f : \mathbb{R} \to \{0, 1\}, \quad f(x) = \mathbb{1}[x < 1]$$

and a the following perturbation type:

$$\mathcal{U}(x) = \begin{cases} \emptyset & \text{if} \quad x \notin M \\ \{x + 1\} & \text{if} \quad x \in M \end{cases}$$

Clearly, f is a measurable function, and every set $\mathcal{U}(x)$ is measurable. However, we get $\mathrm{mar}_f^{\mathcal{U}} = M$, that is, the margin area of f under these perturbations is not measurable, and therefore the adversarial loss with respect to $\mathcal{U}$ is not measurable. Note that the same phenomenon can occur for sets $\mathcal{U}$ that are always open intervals containing the point x. With the same function f, for perturbation sets

$$\mathcal{U}(x) = \begin{cases} \mathcal{B}_r(x) \cap (0, 1) & \text{if} \quad x < 1, x \notin M \\ \mathcal{B}_r(x) \cap (1, 2) & \text{if} \quad x > 1 \\ (0, 2) & \text{if} \quad x \in M \text{ or } x = 1 \end{cases}$$

we get $\mathrm{mar}_f^{\mathcal{U}} = M \cup \{1\}$, which again is not measurable.

We may thus make the following implicit assumptions on the sets $\mathcal{U}(x)$:

- $x \in \mathcal{U}(x)$ for all $x \in X$
- if X is an uncountable domain, we assume X is equipped with a separable metric and $\mathcal{U}(x) = \mathcal{B}_r(x)$ is an open ball around x

Note that the latter assumption implies that $\mathrm{mar}_h^{\mathcal{U}}$ is measurable for a measurable predictor h. This can be seen as follows: It h is a (Borel-)measurable function, then both $h^{-1}(1) = \{x \in X \mid h(x) = 1\}$ and $h^{-1}(0) = \{x \in X \mid h(x) = 0\}$ are measurable sets by definition. Now, if we consider "blowing up" these sets by adding open balls around each of their members, we obtain open (as a union of open sets), and thus measurable sets:

$$\mathcal{M}_r^1 := \bigcup_{x \in h^{-1}(1)} \mathcal{B}_r(x)$$

and

$$\mathcal{M}_r^0 := \bigcup_{x \in h^{-1}(0)} \mathcal{B}_r(x).$$

Now the margin area can be expressed as a simple union of intersections, and is therefore also measurable:

$$\mathrm{mar}_h^{\mathcal{U}} = (\mathcal{M}_r^1 \cap h^{-1}(0)) \cup (\mathcal{M}_r^0 \cap h^{-1}(1))$$

Note that this equality depends on the balls as perturbation sets inducing a symmetric relation, that is $x \in \mathcal{U}(z)$ if and only if $z \in \mathcal{U}(x)$. This condition does not hold in the above counterexample construction. However, this argument shows it is sufficient (together with openness) for measurability of the sets $\mathrm{mar}_h^{\mathcal{U}}$.

C. Proofs and additional results to Section 3

C.1. Some background

We first briefly recall the notions of ϵ-nets and ϵ-approximations and their role in learning binary hypothesis classes of finite VC-dimension. We will frequently use these concepts in our proofs in this section.

ϵ-nets and ϵ-approximations (Haussler & Welzl, 1987) Let Z be some domain set and let $\mathcal{G} \subseteq 2^Z$ be a collection of (measurable) subsets of Z and let D be a probability distribution over Z. Let $\epsilon \in (0,1)$. A finite set $S \subseteq Z$ is an *ϵ-net* for $\mathcal{G}$ with respect to D if

$$S \cap G \neq \emptyset$$

for all $G \in \mathcal{G}$ with $P(G) \geq \epsilon$. That is, an ϵ-net "hits" every set in the collection $\mathcal{G}$ that has probability weight at least ϵ. A finite set $S \subseteq Z$ is an *ϵ-approximation* for $\mathcal{G}$ with respect to D if

$$\left| P(G) - \frac{|G \cap S|}{|S|} \right| \leq \epsilon$$

for all $G \in \mathcal{G}$. It is well known that, given also $\delta \in (0,1)$, if $\mathcal{G}$ has finite VC-dimension, then an iid sample S of size at least $\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathcal{G})+\log(1/\delta)}{\epsilon}\right)$ from distribution D is an ϵ-net for $\mathcal{G}$ with probability at least $(1-\delta)$ (see, eg, Theorem 28.3 in (Shalev-Shwartz & Ben-David, 2014)); and an iid sample S of size at least $\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathcal{G})+\log(1/\delta)}{\epsilon^2}\right)$ from distribution D is an ϵ-approximation for $\mathcal{G}$ with probability at least $(1-\delta)$ (we are omitting logarithmic factors here).

Learning VC-classes ((Vapnik & Chervonenkis, 1971; Valiant, 1984; Blumer et al., 1989) If X is a domain, $Y = \{0,1\}$ is a binary label space, and $\mathcal{H} \subseteq Y^X \subseteq 2^{(X \times Y)}$ is a hypothesis class of finite VC-dimension, then the class of error sets $\mathrm{err}_{\mathcal{H}} = \{\mathrm{err}_h \mid h \in \mathcal{H}\}$, that is the class of complements of $\mathcal{H}$, has finite VC-dimension $\mathrm{VC}(\mathrm{err}_{\mathcal{H}}) = \mathrm{VC}(\mathcal{H})$. For distributions P over $X \times Y$, we get that sufficiently large samples (as indicated above) are ϵ-nets of $\mathrm{err}_{\mathcal{H}}$. Now, if a sample S is an ϵ-net of the class $\mathrm{err}_{\mathcal{H}}$ with respect to P, then every function in the *version space* $\mathcal{V}_S(\mathcal{H})$ of S with respect to $\mathcal{H}$ has error less than ϵ. Recall the version space is defined as those functions in $\mathcal{H}$ that have zero error on the points in S, that is

$$\mathcal{V}_S(\mathcal{H}) = \{h \in \mathcal{H} \mid \mathcal{L}_S^{0/1}(h) = 0\}.$$

If P is realizable by $\mathcal{H}$, an empirical risk minimizing (ERM) learner, will output a hypothesis from the version space (the version space is non-empty under the realizability assumption) and therefore output a predictor of binary loss at most ϵ (with high probability).

For general (not necessarily realizable) learning, note that large enough samples S are ϵ-approximation of $\mathrm{err}_{\mathcal{H}}$ (with high probability at least $1-\delta$ as above). This is also referred to as *uniform convergence* for the hypothesis class $\mathcal{H}$. Thus, every function $h \in \mathcal{H}$ has true loss that is ϵ-close to its empirical loss on h, and any empirical risk minimizer is a successful learner for $\mathcal{H}$ even in the agnostic case.

With these preparations, we proceed to the proofs of Theorem 7, Theorem 10 and Theorem 30.

C.2. Proofs

Proof of Theorem 7. We recall that the robust loss of a classifier h with respect to distribution P over $X \times Y$ is given by

$$\mathcal{L}_P^{\mathcal{U}}(h) \;=\; P(\mathrm{err}_h \cup \mathrm{mar}_h^{\mathcal{U}})$$

Thus, to show that empirical risk minimization with respect to the robust loss is a successful learner, we need to guarantee that large enough samples are ϵ-approximations for the class $\mathcal{G} = \{(\mathrm{err}_h \cup \mathrm{mar}_h^{\mathcal{U}}) \subseteq X \times Y \;\mid\; h \in \mathcal{H}\}$ of point-wise unions error and margin regions.

A simple counting argument involving Sauer's Lemma (see Chapter 6 in (Shalev-Shwartz & Ben-David, 2014), and exercises therein) shows that $\mathrm{VC}(\mathcal{G}) \leq 2D\log(D)$, where $D = \mathrm{VC}(\mathcal{H}) + \mathrm{VC}(\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}})$. Thus, a sample of size $\tilde{\Theta}\left(\frac{D\log D+\log(1/\delta)}{\epsilon^2}\right)$ will be an ϵ-approximation of $\mathcal{G}$ with respect to P with probability at least $1-\delta$ over the sample. Thus any empirical risk minimizer with respect to $\ell^{\mathcal{U}}$ is a successful proper and agnostic robust learner for $\mathcal{H}$. □

Proof of Theorem 10. Note that robust realizability means there exists a $h^* \in \mathcal{H}$ with $\mathcal{L}_P^{\mathcal{U}}(h^*) = 0$ and this implies $\mathcal{L}_P^{0/1}(h^*) = 0$. That is, the distribution is (standard) realizable by $\mathcal{H}$. The above outlined VC-theory tells us that for an iid sample S of size $\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathcal{H})+\log(\frac{1}{\delta})}{\epsilon}\right)$ guarantees that all functions in the *version space* of S (that is all $h \in \mathcal{H}$ with $\mathcal{L}_S(h) = 0$) have true binary loss at most ϵ (with probability at least $1-\delta$). Now, with access to P_X a learner can remove all hypotheses with $P(\mathrm{mar}_h^{\mathcal{U}}) > 0$ from the version space and return any remaining hypothesis. Note that, since h^* is assumed to satisfy $\mathcal{L}_P^{\mathcal{U}}(h^*) = 0$, we have $P(\mathrm{err}_{h^*}) = 0$ and $P(\mathrm{mar}_h^{\mathcal{U}}) = 0$, therefore, the pruned version will contain at least one function. Now, for any function h_p in the the pruned version space, we obtain

$$\begin{aligned}\mathcal{L}_P^{\mathcal{U}}(h_p) \;&=\; P(\mathrm{err}_{h_p} \,\cup\, \mathrm{mar}_{h_p}^{\mathcal{U}}) \\ &\leq\; P(\mathrm{err}_{h_p}) + P(\mathrm{mar}_{h_p}^{\mathcal{U}}) \\ &\leq\; \epsilon + 0 \;=\; \epsilon.\end{aligned}$$

Thus, access to the marginal allows for a successful learner in the robust-realizable case. □

Proof of Theorem 12. We will modify the lower bound construction of Theorem 9 as follows: we add an additional

point x_8 to the domain set, which has zero probability mass under both P^1 and P^2. We set $\mathcal{U}(x_8) = \mathcal{U}(x_7) = \{x_7, x_8\}$. We modify the probability weights of points $x_1, \ldots, x_6$ under P^1 and P^2 by dividing them by 2 (i.e., all respective denominators in the proof of Theorem 9 become 12, and we add weight accordingly to x_7, so that $P^i(x_7) = 1/2 + 1/12$ under both distributions. Functions h_1 and h_2 are extended to the new point by setting $h_1(x_8) = h_2(x_8) = 0$. Thus, the indistinguishability phenomenon of the construction remains the same.

Now we add a function $h_r = \mathbb{1}[x = x_8]$ to the class $\mathcal{H}$. This yields $P^i(\mathrm{err}_{h_r}) = P^i(x_8) = 0$, for $i \in \{1, 2\}$, thus both distributions are realizable with respect to the $0/1$-loss now. However $P^i(\mathrm{mar}^{\mathcal{U}}_{h_r}) = 1/2 + 1/12$, for $i \in \{1, 2\}$, thus h_r has adversarial loss $1/2 + 1/12$ on both distribution and the construction thus remains otherwise analogous. We now have $\mathcal{L}^{\mathcal{U}}_{P^i}(h_i) = 2/12$, thus h_r is does not affect the optimal robust classifier in $\mathcal{H}$.

Additionally, we add the constant 1 function h_c to the class $\mathcal{H}$. For this function (as for any constant classifier) the margin area is empty, thus the distributions are "margin realizable" by $\mathcal{H}$. However, we have $P^i(\mathrm{err}_{h_c}) = 1$, for $i \in \{1, 2\}$, thus h_c also has adversarial loss 1 on both distribution and the construction still remains otherwise unchanged. □

C.3. Additional results

C.3.1. 0/1-Realizability
$\exists h^* \in \mathcal{H}$ with $\mathcal{L}^{0/1}_P(h^*) = 0$

Theorem 12 shows that $0/1$-realizability does not suffice for semi-supervised learning with a margin oracle for $\mathcal{H}$. However, here we show that the following *extended margin oracle* does suffice: we assume that the learner has oracle access to the weights of the sets $\mathrm{mar}^{\mathcal{U}}_h$, $h\Delta h'$, and $\mathrm{mar}^{\mathcal{U}}_h \cap (h\Delta h')$, for all $h, h' \in \mathcal{H}$, where the sets $h\Delta h' \subseteq X$ are defined as follows:

$$h\Delta h' = \{x \in X \mid h(x) \neq h'(x)\}.$$

Theorem 29. *Let X be some domain, $\mathcal{H}$ a hypothesis class with finite VC-dimension and $\mathcal{U} : X \to 2^X$ any perturbation type. If a learner is given additional access to an extended margin oracle for $\mathcal{H}$, then $\mathcal{H}$ is properly learnable with respect to the robust loss $\ell^{\mathcal{U}}$ and the class of distributions P that are $0/1$-realizable by $\mathcal{H}$, that is we have $\mathcal{L}^{0/1}_P(\mathcal{H}) = 0$, with labeled sample complexity $\tilde{O}(\frac{\mathrm{VC}(\mathcal{H}) + \log(1/\delta)}{\epsilon})$.*

Proof. As in the proof of Theorem 10, since we assume the distribution to be $0/1$-realizable by $\mathcal{H}$, the version space of a labeled sample of the given size will include only functions with (true) binary loss at most ϵ. The learner can choose a function h_e from this version space. Now, given the extended margin oracle, the learner can choose a function h_r that minimizes the robust loss with respect to labeling function h_e. That is, the extended margin oracle allows to find the minimizer in $\mathcal{H}$ of the robust loss on a distribution (P_X, h_e), that shares the marginal with the data generating distribution P, but labels domain points according to h_e.

Let $h^* \in \mathcal{H}$ be a function with $\mathcal{L}^{0/1}_P(h^*) = 0$. Thus, we can identify the distribution P with (P_X, h^*). Now we first show that for any classifier h, the difference between its robust loss with respect to $P = (P_X, h^*)$ and with respect to (P_X, h_e) is bounded by ϵ.

Let $h \in \mathcal{H}$ be given. Then we have

$$\begin{aligned}\mathcal{L}^{\mathcal{U}}_P(h) &= \mathcal{L}^{\mathcal{U}}_{(P_X, h^*)}(h) \\ &= P_X(\mathrm{mar}^{\mathcal{U}}_h \cup (h^*\Delta h)) \\ &= P_X(\mathrm{mar}^{\mathcal{U}}_h) + P_X((h^*\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h)\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}^{\mathcal{U}}_{P_X, h_e}(h) &= P_X(\mathrm{mar}^{\mathcal{U}}_h \cup (h_e\Delta h)) \\ &= P_X(\mathrm{mar}^{\mathcal{U}}_h) + P_X((h_e\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h).\end{aligned}$$

Thus, we get

$$\begin{aligned}&|\mathcal{L}^{\mathcal{U}}_P(h) - \mathcal{L}^{\mathcal{U}}_{P, h_e}(h)| \\ \leq\ &|P((h^*\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h) - P_X((h_e\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h)| \\ \leq\ &|P((h^*\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h) - (P_X((h_e\Delta h^*) \setminus \mathrm{mar}^{\mathcal{U}}_h) \\ &\quad + P_X((h^*\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h))| \\ \leq\ &|P((h_e\Delta h^*) \setminus \mathrm{mar}^{\mathcal{U}}_h)| \\ \leq\ &P((h_e\Delta h^*) \leq \epsilon.\end{aligned}$$

where the second inequality follows from

$$(h_e\Delta h) \subseteq (h_e\Delta h^*) \cup (h^*\Delta h),$$

and thus

$$(h_e\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h \subseteq ((h_e\Delta h^*) \setminus \mathrm{mar}^{\mathcal{U}}_h) \cup ((h^*\Delta h) \setminus \mathrm{mar}^{\mathcal{U}}_h).$$

Note that $|\mathcal{L}^{\mathcal{U}}_P(h) - \mathcal{L}^{\mathcal{U}}_{P, h_e}(h)| \leq \epsilon$ for all $h \in \mathcal{H}$ implies that we also have:

$$|\inf_{h \in \mathcal{H}} \mathcal{L}^{\mathcal{U}}_P(h) - \inf_{h \in \mathcal{H}} \mathcal{L}^{\mathcal{U}}_{P_X, h_e}(h)| \leq \epsilon$$

Thus, for the output h_r of the above procedure, we get

$$\begin{aligned}\mathcal{L}^{\mathcal{U}}_P(h_r) &\leq \mathcal{L}^{\mathcal{U}}_{P_X, h_e}(h_r) + \epsilon \\ &= \inf_{h \in \mathcal{H}} \mathcal{L}^{\mathcal{U}}_{P_X, h_e}(h) + \epsilon \\ &\leq \inf_{h \in \mathcal{H}} \mathcal{L}^{\mathcal{U}}_P(h) + 2\epsilon\end{aligned}$$

Substituting $\epsilon/2$ for ϵ in this argument completes the proof. □

C.3.2. $0/1$-REALIZABILITY ON A $\mathcal{U}$-CLUSTERABLE TASK: $\exists h^* \in \mathcal{H}$ WITH $\mathcal{L}_P^{0/1}(h^*) = 0$ AND $\exists f^* \in \mathcal{F}$ WITH $\mathcal{L}_P^{\mathcal{U}}(f^*) = 0$

We start by observing that the existence of an $f^* \in \mathcal{F}$ with $\mathcal{L}_P^{\mathcal{U}}(f^*) = 0$ implies that the support of P_X is sitting on $\mathcal{U}$-separated clusters. Note that we do not assume that the perturbation type $\mathcal{U}$ induces a symmetric relation; we can nevertheless consider the clusters as connected components of a directed graph where we place a directed edge between two domain instances x and x' if and only if x is in the support of P_X and $x' \in \mathcal{U}(x)$. The assumption $\mathcal{L}_P^{\mathcal{U}}(f^*) = 0$ then implies that these clusters are label-homogeneous. This observation leads to a simple, yet improper learning scheme for the robust loss.

We show that, if the distribution is also $0/1$-realizable by $\mathcal{H}$, a leaner that knows that marginal, can return a hypothesis with robust loss at most ϵ. We note that here, the learner does not return a hypothesis from the class $\mathcal{H}$. In return, the guarantee is stronger in the sense that the robust loss of the returned classifier is close to the overall (among all binary predictors, rather than just those in $\mathcal{H}$) best achievable robust loss.

Theorem 30. *Let X be some domain, $\mathcal{H}$ a hypothesis class with finite VC-dimension and $\mathcal{U} : X \to 2^X$ any perturbation type. If a learner has access to a labeled sample of size*

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

and, additionall has access to P_X, then the class $\mathcal{F}$ of all binary predictors is learnable with respect to the robust loss $\ell^{\mathcal{U}}$ and the class of distributions P that are realizable by $\mathcal{H}$ (that is, $\mathcal{L}_P^{0/1}(\mathcal{H}) = 0$) and robust realizable with respect to $\mathcal{F}$ (that is, $\mathcal{L}_P^{\mathcal{U}}(\mathcal{F}) = 0$).

Proof. Recall that, to avoid measurability issues, we either assume a countable domain, or, in case of an uncountable domain, that the perturbation sets are open balls with respect to some separable metric. The arguments below hold for both cases.

We now start by observing that the existence of an $f^* \in \mathcal{F}$ with $\mathcal{L}_P^{\mathcal{U}}(f^*) = 0$ implies that the support of P_X is sitting on $\mathcal{U}$-separated clusters. Note that (in the case of a countable domain) we do not assume that the perturbation type $\mathcal{U}$ induces a symmetric relation. We derive the clusters as follows: we define a (directed) graph on X, where we place an edge between from domain elements x to x' if and only if x is in the support of P_X and $x' \in \mathcal{U}(x)$. We now let $\mathcal{C} \subseteq 2^X$ be the collection of connected components of the induced undirected graph. Since $\mathcal{L}_P^{\mathcal{U}}(f^*) = 0$, thus $P(\mathrm{mar}_{f^*}^{\mathcal{U}}) = 0$, the function f^* is label homogeneous on these clusters (except, potentially, for subsets of P_X-measure 0, and we may then identify f^* with a function that is label homogeneous on the clusters).

Now, since P is $\mathcal{H}$-realizable, there is an $h^* \in \mathcal{H}$ with $\mathcal{L}_P^{0/1}(h^*) = 0$. Note that h^* is not necessarily label homogeneous on the clusters (since h^* may have a positive robust loss, that is it may be the case that $P(\mathrm{mar}_{h^*}^{\mathcal{U}}) > 0$). However, h^* agrees with f^* on the support of P_X

(except on a set with measure 0), since both functions have zero binary loss, $\mathcal{L}_P^{0/1}(h^*) = \mathcal{L}_P^{0/1}(f^*) = 0$. Let $\mathrm{supp}(P_X)$ denote the support of P_X. That is, for any cluster $C \in \mathcal{C}$, h^* is label-homogeneous (and in agreement with f^*) on the subset $C \cap \mathrm{supp}(P_X)$.

Note that, since we assume knowledge of the marginal, we may assume that a learner knows the collection of clusters $\mathcal{C}$ and the support of P_X. We now define a learning scheme as follows.

As in the proof of Theorem 10, due to the $\mathcal{H}$-realizability ($\mathcal{L}_P^{0/1}(\mathcal{H}) = 0$), we know that with high probability over a large enough sample S, all functions $h \in \mathcal{V}_S(\mathcal{H})$ in the version space satisfy $\mathcal{L}_P^{0/1}(h) \leq \epsilon$. Moreover, due to the $\mathcal{H}$-realizability, there will exist functions (for example h^*) in the version space that label the intersections $C \cap \mathrm{supp}(P_X)$ of the clusters in with the support of P_X homogeneously. Thus, employing the knowledge of P_X, the learner can prune the version space by removing all functions from the version space that don't label all sets $C \cap \mathrm{supp}(P_X)$ homogeneously, and pick a function h_p from this pruned version space.

Now the learner can construct a new classifier f_p, that agrees with h_p on the sets $C \cap \mathrm{supp}(P_X)$ and labels the full clusters homogeneously, that is, if $x \in C \cap \mathrm{supp}(P_X)$ for some cluster $C \in \mathcal{C}$, then we set $f_p(x') = h_p(x)$ for all $x' \in C$. Now, by construction of f_p (recall the definition of the clusters), we get $P(\mathrm{mar}_{f_p}^{\mathcal{U}}) = 0$. Moreover, we have $P(\mathrm{err}_{f_p}) \leq \epsilon$ (inherited from h_p since h_p and f_p agree on the support of P_X). Thus

$$\mathcal{L}_P^{\mathcal{U}}(f_p) \;\leq\; \epsilon \;\leq\; \mathcal{L}_P^{\mathcal{U}}(\mathcal{H}) + \epsilon,$$

which is what we needed to show. □

D. Proof from Section 4

Proof of Observation 17. We prove this statement for the case when the certifier is restricted to be deterministic, and leave the proof of the probabilistic case to future work. Suppose the entire data distribution is concentrated on one point, and wlog suppose the point is the origin and has label 1. Let B be the unit ball centred at the origin. Thus the certifier's task is to determine if h passes through B or not. We construct a scheme for answering the certifier's queries in a way so that no matter what sequence of queries

it chooses to ask, once it commits to a verdict, we can find a halfspace that is consistent with the answers we provided to the queries, but inconsistent with the certifier's verdict.

It is easier to work in a dual space using a standard duality argument, where the dual of a point (a, b) is the line $ax + by + 1 = 0$ and vice versa. This duality transform has the following two useful properties: 1) a point is to the left of a line if and only if the dual of the point is to the left of the dual of the line, and 2) a point is inside the unit ball if and only if its dual does not intersect the unit ball. Thus in the dual space, the certifier picks a line and asks whether the hidden point is to its left or right, and needs to determine if the hidden point is inside the unit ball or not. Our strategy, then, is to consider the arrangement of lines created by the certifier's queries thus far, and locate a cell that contains a part of B's circumference. We answer the certifier's query as if the point was inside this cell. This cell will have a non-zero volume whenever the certifier stops, and we can select a point inside the cell that is inside or outside B depending on the certifier's answer. That we can always find such a cell can be seen with an argument using induction. For the base case, there are no queries and hence no lines. Thus the entire place is such a cell. Suppose we have identified such a cell after seeing m lines. If the next line does not pass through the cell it still satisfies the property in question. If the next line does pass through the cell, it divides the cell into two smaller cells one of which will satisfy the property. □

E. Proof of Theorem 28

We start by providing the definition of proper sample compression for adversarially robust learning.

Definition 31 (Adversarially Robust Proper Compression)**.** *We say $(\mathcal{H},\mathcal{U})$ admits robust proper compression of size k if there exist a (decoder) function $\phi : (X \times Y)^k \to \mathcal{H}$ such that the following holds: for every $h \in \mathcal{H}$ and every $S_X \subset X$, there exist $K_X \subset S_X$ such that*

$$\forall x \in S_X, \ell^{\mathcal{U}}(h, x, h(x)) = \ell^{\mathcal{U}}(\phi(K), x, h(x))$$

where K is the labeled version of K_X (labeled by h).

Note that in the above definition, $k = |K|$ can potentially depend on the size of the set, $m = |S_X|$. However, this dependence should be sub-linear (e.g., logarithmic) to later result in a non-vacuous sample complexity upper bound. The following theorem draws the connection between compression and robust learning.

Theorem 32. *If $(\mathcal{H},\mathcal{U})$ admits an adversarially robust proper compression of size k, then the sample complexity of robust learning of $(\mathcal{H},\mathcal{U})$ in the robustly realizable setting is $O(k \log(k/\epsilon)/\epsilon^2)$.*

Proof. This theorem can be proved in a similar way to that of classical (non-robust) sample compression proposed by (Littlestone & Warmuth, 1986). For the proof in the context of robust compression we refer the reader to Lemma 11 in (Montasser et al., 2019). Note that the hypothesis returned by the decoder of the compression scheme has to have zero robust loss on all of the samples (due to robust realizability). □

In order to proceed, we need to show that for properly compressible classes, the existence of a perfect proper efficient adversary means that a small-sized robust proper compression scheme exists.

Theorem 33. *Let $\mathcal{H}$ be any properly (non-robustly) compressible class. Assume $(\mathcal{H},\mathcal{U})$ has a perfect proper adversary with query complexity $O(m)$. Then $(\mathcal{H},\mathcal{U})$ admits a robust proper compression of size $O(VC(\mathcal{H}) \log(m))$.*

Let us postpone the proof of Theorem 33 for now and complete the proof of Theorem 28.

Proof of Theorem 28. Assume that $(\mathcal{H},\mathcal{U})$ has a perfect, proper, and efficient adversary. Based on Theorem 33, we conclude that $(\mathcal{H},\mathcal{U})$ admits a robust proper compression scheme of size $O(VC(\mathcal{H}) \log(m))$. We can now use Theorem 32 to bound the sample complexity of learning. In particular, it will be enough to have $m > \Omega(k \log(k/\epsilon)/\epsilon^2)$ where $k = \Theta(VC(\mathcal{H}) \log(m))$. Therefore, it will suffice to have $m = \Omega(VC(\mathcal{H}) \log^2(VC(\mathcal{H})/\epsilon)/\epsilon^2)$. □

Therefore, it only remains to construct a robust proper compression scheme and prove Theorem 33. We denote by S_X the unlabeled portion of the sample S.

Proof of Theorem 33. Recall that we want to show that there exists $K_X \subset S_X$ such that

$$\forall x \in S_X, \ell^{\mathcal{U}}(h, x, h(x)) = \ell^{\mathcal{U}}(\phi(K), x, h(x))$$

where K is the labeled version of K_X (labeled by h). We know that $(\mathcal{H},\mathcal{U})$ has a perfect adversary with query complexity $O(m)$. Let Q_S be the set of queries that the adversary asks on S to find the adversarial points (so $|Q_S| = O(m)$). Let Q be the labeled version of Q_S (i.e., each query with its answer from h). We claim that for a proper compression to succeed it will be enough to have

$$\forall z \in S_X \cup Q_S, \phi(K)\big|_z = h\big|_z \tag{1}$$

The reason is that if the two hypotheses from $\mathcal{H}$ have the same behaviour on $T = S_X \cup Q_S$ then they should have the same robust loss on S_X as well (otherwise the adversary

would not be perfect). The final step is to come up with a proper compression scheme that satisfies (1).

Let T_Y be the labeled version of T that is labeled by h. Recall that $\mathcal{H}$ is a properly (non-robustly) compressible class. Therefore, T_Y can be properly (non-robustly) compressed into a set $I \subseteq T_Y$ such that $|T_Y| = O(VC(\mathcal{H}) \log(|T_Y|))$. The catch is that I may contain points that are outside of S, and therefore we cannot simply use I for robust proper compression. We can modify the compression scheme by adding some additional bits of information so that its output contains only points from S. For any $x \in S$, let $Q_x \subseteq \mathcal{U}(x)$ be the set of points that the adversary queries to attack x. Note that $|Q_x| = O(1)$ due to the efficiency of the adversary. We replace any $(x, y) \in S \setminus I$ with (x_0, y_0) where $(x_0, y_0) \in S$ and $x \in \mathcal{U}(x_0)$. Also, we use a constant number of bits to encode the labels of Q_{x_0} and also the subset of Q_{x_0} that was chosen by the non-robust compression scheme. The decoder works as follows. Given (x_0, y_0), it can simulate the adversary on x_0 (using the bits that represent the labels) to recover Q_{x_0}. It can use the other part of the bits to recover the subset of Q_{x_0} that was present in I (let us call this set G_x). Finally, it would run the decoder of the proper non-robust compression scheme on $\cup_{x \in S} G_x$.

□