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 $Y$
- $f : X \to Y$ if $f$ 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 $P(\text{err}_h)$, where $\text{err}_h$ is the complement of $h$ in $X \times Y$, that is

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

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 sets 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 $X \times Y$.

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

Note, that here, if for a given domain point $x$, we have $(x, y) \in \text{mar}^U_h$ for some $y \in Y$, then $(x, y') \in \text{mar}^U_h$ for all $y' \in Y$. Thus, the sets $\text{mar}^U_h \subseteq X \times Y$ 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 $\text{err}_h$ and $\text{mar}^U_h$, naturally let us express the robust loss as the probably measure of a subset of $X \times Y$:

$$\mathcal{L}^U_P(h) = P(\text{err}_h \cup \text{mar}^U_h)$$

B. Note on measurability

Here, we note that allowing the perturbation type $U$ to be an arbitrary mapping from the domain $X$ to $2^X$ can easily lead to the adversarial loss being not measurable, even if $U(x)$ 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

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

and a the following perturbation type:

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

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

With the same function $f$, for perturbation sets

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

we get $\text{mar}^U_f = M \cup \{1\}$, which again is not measurable.

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

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

Note that the latter assumption implies that $\text{mar}^U_h$ 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:

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

and

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

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

$$\text{mar}^U_h = (\mathcal{M}_r^+ \cap h^{-1}(0)) \cup (\mathcal{M}_r^0 \cap h^{-1}(1))$$

Note that this equality depends on the balls as perturbation sets inducing a symmetric relation, that is $x \in U(z)$ if and only if $z \in U(x)$. 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 $\text{mar}^U_h$. 
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 \( G \subseteq 2^Z \) 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 \( G \) with respect to D if

\[
S \cap G \neq \emptyset
\]

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

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

for all \( G \in G \). It is well known that, given also \( \delta \in (0, 1) \), if \( G \) has finite VC-dimension, then an iid sample \( S \) of size at least \( \tilde{\Theta} \left( \frac{\text{VC}(G) + \log(1/\delta)}{\epsilon^2} \right) \) from distribution D is an ε-net for \( G \) 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 \( \tilde{\Theta} \left( \frac{\text{VC}(G) + \log(1/\delta)}{\epsilon^2} \right) \) from distribution D is an ε-approximation for \( G \) 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 \( \mathcal{H} \subseteq Y^X \subseteq \mathbb{2}^{(X \times Y)} \) is a hypothesis class of finite VC-dimension, then the class of error sets \( \text{err}_\mathcal{H} = \{ \text{err}_h \mid h \in \mathcal{H} \} \), that is the class of complements of \( \mathcal{H} \), has finite VC-dimension \( \text{VC} (\text{err}_\mathcal{H}) = \text{VC}(\mathcal{H}) \). For distributions \( P \) over \( X \times Y \), we get that sufficiently large samples (as indicated above) are ε-nets of \( \text{err}_\mathcal{H} \). Now, if a sample \( S \) is an ε-net of the class \( \text{err}_\mathcal{H} \) with respect to \( P \), then every function in the version space \( V_S(\mathcal{H}) \) of \( S \) with respect to \( \mathcal{H} \) has error less than \( \epsilon \). Recall the version space is defined as those functions in \( \mathcal{H} \) that have zero error on the points in \( S \), that is

\[
V_S(\mathcal{H}) = \{ h \in \mathcal{H} \mid \mathcal{L}_S^{err}(h) = 0 \}.
\]

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 \( \epsilon \) (with high probability).

For general (not necessarily realizable) learning, note that large enough samples \( S \) are ε-approximation of \( \text{err}_\mathcal{H} \) (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

\[
\mathcal{L}_P^{\ell}(h) = P(\text{err}_h \cup \text{mar}_h^{\ell})
\]

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 \( \mathcal{G} = \{ (\text{err}_h \cup \text{mar}_h^{\ell}) \subseteq X \times Y \mid h \in \mathcal{H} \} \) 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 \( \text{VC}(\mathcal{G}) \leq 2D \log(D) \), where \( D = \text{VC}(\mathcal{H}) + \text{VC}(\mathcal{H}^{\ell}_{\text{mar}}) \). 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 \( \mathcal{L}^{\ell} \) is a successful proper and agnostic robust learner for \( \mathcal{H} \). \( \square \)

Proof of Theorem 10. Note that robust realizability means there exists a \( h^* \in \mathcal{H} \) with \( \mathcal{L}_P^{\ell}(h^*) = 0 \) and this implies \( \mathcal{L}_P^{\ell}(h^*) = 0 \). 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 \( \tilde{\Theta} \left( \frac{\text{VC}(\mathcal{H}) + \log(1/\epsilon)}{\epsilon^2} \right) \) 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 \( \epsilon \) (with probability at least \( 1 - \delta \)). Now, with access to \( P_X \) a learner can remove all hypotheses with \( P(\text{mar}_h^{\ell}) > 0 \) from the version space and return any remaining hypothesis. Note that, since \( h^* \) is assumed to satisfy \( \mathcal{L}_P^{\ell}(h^*) = 0 \), we have \( P(\text{err}_h^{\ell}) = 0 \) and \( P(\text{mar}_h^{\ell}) = 0 \), therefore, the pruned version will contain at least one function. Now, for any function \( h_p \) in the the pruned version space, we obtain

\[
\mathcal{L}_P^{\ell}(h_p) = P(\text{err}_{h_p} \cup \text{mar}_{h_p}^{\ell}) \leq P(\text{err}_{h_p}) + P(\text{mar}_{h_p}^{\ell}) \leq \epsilon + 0 = \epsilon.
\]

Thus, access to the marginal allows for a successful learner in the robust-realizable case. \( \square \)

Proof of Theorem 12. We will modify the lower bound construction of Theorem 9 as follows: we add an additional
We modify the probability weights of points with (true) binary loss at most $\epsilon$. Additionally, we add the constant $1$.

Theorem 29.

Theorem 12 shows that $0$-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 $\text{mar}_h^R, h \Delta h$, and $\text{mar}_h^L \cap (h \Delta h^*)$, 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 $\mathcal{L}_P^U$ and the class of distributions $P$ that are $0/1$-realizable by $\mathcal{H}$, that is we have $\mathcal{L}_P^U(\mathcal{H}) = 0$, with labeled sample complexity $O\left(\frac{\text{VC}(\mathcal{H}) + \log(1/\delta)}{\epsilon}\right)$.

**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 $\epsilon$. The learner can choose a function $h_\epsilon$ from this version space. Now, given the extended margin oracle, the learner can choose a function $h_\epsilon$ that minimizes the robust loss with respect to labeling function $h_\epsilon$. That is, the extended margin oracle allows to find the minimizer in $H$ of the robust loss on a distribution $(P_X, h_\epsilon)$, that shares the marginal with the data generating distribution $P$, but labels domain points according to $h_\epsilon$.

Let $h^* \in \mathcal{H}$ be a function with $\mathcal{L}_P^U(h^*) = 0$. 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 $(P_X, h_\epsilon)$ is bounded by $\epsilon$.

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

$$\mathcal{L}_{P_X, h_\epsilon}^U(h) = P_X(\text{mar}_h^L \cup (h^* \Delta h)) = P_X(\text{mar}_h^L) + P_X((h^* \Delta h) \setminus \text{mar}_h^L).$$

and

$$\mathcal{L}_{P_X, h}^U(h) = P_X(\text{mar}_h^L \cup (h_\epsilon \Delta h)) = P_X(\text{mar}_h^L) + P_X((h_\epsilon \Delta h) \setminus \text{mar}_h^L).$$

Thus, we get

$$|\mathcal{L}_{P_X, h}^U(h) - \mathcal{L}_{P_X, h_\epsilon}^U(h)| \leq |P((h^* \Delta h) \setminus \text{mar}_h^L) - P_X((h_\epsilon \Delta h) \setminus \text{mar}_h^L)|$$

$$\leq |P((h^* \Delta h) \setminus \text{mar}_h^L) - P_X((h^* \Delta h^*) \setminus \text{mar}_h^L)|$$

$$+ P_X((h_\epsilon \Delta h) \setminus \text{mar}_h^L))$$

$$\leq |P((h_\epsilon \Delta h^*) \setminus \text{mar}_h^L)|$$

$$\leq P((h_\epsilon \Delta h^*) \setminus \text{mar}_h^L).$$

where the second inequality follows from

$$(h_\epsilon \Delta h) \subseteq (h_\epsilon \Delta h^*) \cup (h^* \Delta h),$$

and thus

$$(h_\epsilon \Delta h) \setminus \text{mar}_h^L \subseteq ((h_\epsilon \Delta h^*) \setminus \text{mar}_h^L) \cup ((h^* \Delta h) \setminus \text{mar}_h^L).$$

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

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

Thus, for the output $h_\epsilon$ of the above procedure, we get

$$\mathcal{L}_{P_X, h_\epsilon}^U(h_\epsilon) \leq \inf_{h \in \mathcal{H}} \mathcal{L}_{P_X, h}^U(h) + \epsilon.$$

Substituting $\epsilon/2$ for $\epsilon$ in this argument completes the proof.
We now start by observing that the existence of an \( f^* \in \mathcal{F} \) with \( \mathcal{L}^0_{\mathcal{P}}(f^*) = 0 \) implies that the support of \( P_X \) is sitting on \( \mathcal{U} \)-separated clusters. Note that we do not necessarily 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 \( x, x' \) if and only if \( x \) is in the support of \( P_X \) and \( x' \in \mathcal{U}(x) \). The assumption \( \mathcal{L}^0_{\mathcal{P}}(f^*) = 0 \) 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 learner that knows that marginal, can return a hypothesis with robust loss at most \( \epsilon \). 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 \( \mathcal{U} : X \rightarrow 2^X \) any perturbation type. If a learner has access to a labeled sample of size

\[
\hat{O}\left(\frac{\text{VC}(\mathcal{H}) + \log 1/\delta}{\epsilon}\right)
\]

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

**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 \( \mathcal{L}^\mathcal{U}(f^*) = 0 \) 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 necessarily 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 \( x, x' \) if and only if \( x \) is in the support of \( P_X \) and \( x' \in \mathcal{U}(x) \). We now let \( C \subseteq 2^X \) be the collection of connected components of the induced undirected graph. Since \( \mathcal{L}^\mathcal{U}(f^*) = 0 \), 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 \( \mathcal{L}^0_{\mathcal{P}}(h^*) = 0 \). 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 \( P(\text{mar}_{h^*}) > 0 \)). 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, \( \mathcal{L}^0_{\mathcal{P}}(h^*) = \mathcal{L}^0_{\mathcal{P}}(f^*) = 0 \). Let \( \text{supp}(P_X) \) denote the support of \( P_X \). That is, for any cluster \( C \subseteq \mathcal{U} \), \( h^* \) is label-homogeneous (and in agreement with \( f^* \)) on the subset \( C \cap \text{supp}(P_X) \).

Note that, since we assume knowledge of the marginal, we may assume that a learner knows the collection of clusters \( C \) 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 (\( \mathcal{L}^0_{\mathcal{P}}(\mathcal{H}) = 0 \)), 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 \( \mathcal{L}^0_{\mathcal{P}}(h) \leq \epsilon \). Moreover, due to the \( \mathcal{H} \)-realizability, there will exist functions (for example \( h^* \)) in the version space that label the intersections \( C \cap \text{supp}(P_X) \) 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 \( C \cap \text{supp}(P_X) \) 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 \( C \cap \text{supp}(P_X) \) and labels the full clusters homogeneously, that is, if \( x \in C \cap \text{supp}(P_X) \) for some cluster \( C \subseteq \mathcal{U} \), 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 \( P(\text{mar}_{h_p}) = 0 \). Moreover, we have \( P(\text{err}_{f_p}) \leq \epsilon \) (inherited from \( h_p \) since \( h_p \) and \( f_p \) agree on the support of \( P_X \)). Thus

\[
\mathcal{L}^\mathcal{U}(f_p) \leq \epsilon \leq \mathcal{L}^\mathcal{U}(\mathcal{H}) + \epsilon,
\]

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.

\[\tag{32} \]

**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).

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

**Theorem 33.** Assume \((\mathcal{H}, \mathcal{U})\) has a perfect proper adversary with query complexity \(f(m)\). Then \((\mathcal{H}, \mathcal{U})\) admits a compression of size \(O(VC(\mathcal{H}) \log(m + f(m)))\).

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 compression scheme of size \(O(VC(\mathcal{H}) \log(m + f(m)))\). Furthermore, because the adversary is efficient we have \(f(m) = \text{poly}(m)\) and therefore we conclude that the size of the compression scheme is \(O(VC(\mathcal{H}) \log(m))\). 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 compression scheme and prove Theorem 33. For this, we will need to first define the notion of weak learners in the context of adversarially robust learning.

Consider the robust empirical risk minimizer (RERM) algorithm for \((\mathcal{H}, \mathcal{U})\) which given a labeled sample \(S\) outputs \(h \in \mathcal{H}\) with minimum robust loss on \(S\).

\[
RERM(S) = \arg \min_{h} L_S^\mathcal{U}_S
\]

(Note that in case of a tie, the RERM algorithm is allowed to output any minimizer of the robust loss.)

We denote by \(S_X\) the unlabeled portion of the sample \(S\). The inflated version of \(S_X\) with respect to \(\mathcal{U}\) is defined by

\[
S_X^\mathcal{U} = \bigcup_{x \in S_X} \mathcal{U}(x)
\]

**Theorem 34 (Weak Learner for Robust Learning).** For every \(S_X \subseteq X\), every probability measure \(\mathcal{P}_X\) over \(S_X^\mathcal{U}\), and every \(h \in \mathcal{H}\), there exists \(W \subseteq S\) such that \(|W| = O(VC(\mathcal{H}))\) and \(L_V^\mathcal{P}(RERM(W)) < 1/3\). Here \(S\) is the labeled version of \(S_X\) by \(h\), and \(\mathcal{P}\) is a distribution over \(X \times Y\) with marginal \(\mathcal{P}_X\) and labeled (deterministically) by \(h\).
Note that in the above theorem, $W$ is not only a subset of the inflated set, but also a subset of the original set $S$. This will become handy when designing a compression scheme, as we would be allowed to only use the given sample $S$ in the scheme, not the inflated set.

**Proof of Theorem 34.** Consider an arbitrary probability measure $P_X$ over the inflated set $S'_X$. By standard VC-theory we know that for a large enough constant $C$, a sample $S' \subset S'_X \times Y$ of size $C \cdot VC(H)$ generated from $P$ will be a $1/3$-approximation of $P$ with respect to $H$ with probability at least 0.99. Therefore, there exist $S' \subset S'_X \times Y$ such that if given to ERM, the robust loss of the resulting hypothesis, $L_{p,1}^{0/1}(ERM(S'))$, will be smaller than $1/3$. Now define

$$W = \{(\mu(z), y) \mid (z, y) \in S'\}$$

where $\mu(z) = \arg\min_{x \in S_X} \{||x - z|| \mid (x, y) \in S, z \in \mathcal{U}(x)\}$ (and we break the ties arbitrarily). By definition of the RERM algorithm, running RERM on $W$ will have the effect of running an ERM on $S'$. Therefore, we can conclude that $L_{p,1}^{0/1}(RERM(W)) = L_{p,1}^{0/1}(ERM(S')) \leq 1/3$. \hfill \qed

Now we can use these weak learners along with a boosting scheme to construct the compression scheme.

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

$$\forall x \in S_X, \forall z \in \mathcal{U}(x), \phi(K)|_z = h|_z$$

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

$$\forall z \in S_X \cup Q_S, \phi(K)|_z = h|_z$$

The reason is that if the two hypotheses have the same behaviour on $T = S_X \cup Q_S$ then they should have the same behaviour on the $S'_{X}$ as well (otherwise the adversary would not be perfect). Furthermore, because the adversary is proper, we know that $Q_S \subset S'_X$. In other words, we just need to compress the subset $T$ instead of $S'_X$. Also, we know that $|T| \leq m + f(m)$.

We will use the idea of boosting to do the sample compression for $T$. Based on Theorem 34, we know that we have weak learners that can reach error $1/3$ on for any marginal over $T$, and we can encode each of these weak classifiers using a subset $W \subset S$ of size $O(VC(H))$. We can now use AdaBoost (Schapire & Freund, 2013, Theorem 3.1) to combine $O(\log(1/\epsilon))$ of these classifiers to reach error $\epsilon$. However, since the labeling of $h$ is deterministic, the error of the combined classifier will be zero if it is smaller than $1/(m + f(m))$. As a result, we only need to combine $\log(m + f(m))$ weak classifiers each of which using $O(VC(H))$ samples. Therefore, the size of the compression scheme is $O(\log(m + f(m))VC(H))$. \hfill \qed