A Notation and Definitions

We first overview standard definitions about sample complexity in PAC learning.

PAC Sample Complexity. The realizable PAC sample complexity $\mathcal{M}(\mathcal{H};\varepsilon,\delta)$ of $\mathcal{H}$ is defined as

$$\mathcal{M}(\mathcal{H};\varepsilon,\delta) = \inf_{A\in\mathcal{A}} \mathcal{M}_A(\mathcal{H};\varepsilon,\delta)\,, \tag{1}$$

where the infimum is over all possible learning algorithms and $\mathcal{M}_A(\mathcal{H};\varepsilon,\delta)$ is the minimal integer such that for any $m \geq \mathcal{M}_A(\mathcal{H};\varepsilon,\delta)$, every distribution $\mathcal{D}_\mathcal{X}$ on $\mathcal{X}$, and, true target $h^\star \in \mathcal{H}$, the expected loss $\mathbf{E}_{x\sim\mathcal{D}_\mathcal{X}}[\ell(A(T)(x), h^\star(x))]$ of A is at most ε with probability $1-\delta$ over the training set $T = \{(x, h^\star(x)) : x \in S\}$, $S \sim \mathcal{D}_\mathcal{X}^m$.

PAC Cut-Off Sample Complexity. We slightly overload the notation of the sample complexity and we define

$$\mathcal{M}(\mathcal{H};\varepsilon,\delta,\gamma) = \inf_{A\in\mathcal{A}} \mathcal{M}_A(\mathcal{H};\varepsilon,\delta,\gamma)\,, \tag{2}$$

where the infimum is over all possible learning algorithms and $\mathcal{M}_A(\mathcal{H};\varepsilon,\delta,\gamma)$ is the minimal integer such that for any $m \geq \mathcal{M}_A(\mathcal{H};\varepsilon,\delta,\gamma)$, every distribution $\mathcal{D}_\mathcal{X}$ on $\mathcal{X}$, and, true target $h^\star \in \mathcal{H}$, the expected cut-off loss $\mathbf{Pr}_{x\sim\mathcal{D}_\mathcal{X}}[\ell(A(T)(x), h^\star(x)) > \gamma]$ of A is at most ε with probability $1-\delta$ over the training set $T = \{(x, h^\star(x)) : x \in S\}$, $S \sim \mathcal{D}_\mathcal{X}^m$.

Lemma 1 (Equivalence Between Sample Complexities)**.** *For every $\varepsilon,\delta \in (0,1)^2$ and every $\mathcal{H} \subseteq [0,1]^\mathcal{X}$, where $\mathcal{X}$ is the input domain, it holds that*

$$\mathcal{M}(\mathcal{H};\sqrt{\varepsilon},\delta,\sqrt{\varepsilon}) \leq \mathcal{M}(\mathcal{H};\varepsilon,\delta) \leq \mathcal{M}(\mathcal{H};\varepsilon/2,\delta,\varepsilon/2)$$

Proof. Let A be a learning algorithm. We will prove the statement for each fixed A and for each data-generating distribution $\mathcal{D}$, so the result follows by taking the infimum over the learning algorithms.

Assume that the cut-off sample complexity of A is $\mathcal{M}_A(\mathcal{H};\varepsilon/2,\delta,\varepsilon/2)$. Then, with probability $1-\delta$ over the training sample $S \sim \mathcal{D}$, for its expected loss it holds that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\ell(A(S;x), h^\star(x))] \leq \frac{\varepsilon}{2} + \left(1-\frac{\varepsilon}{2}\right)\cdot\frac{\varepsilon}{2} \leq \varepsilon\,,$$

thus, $\mathcal{M}_A(\mathcal{H};\varepsilon,\delta) \leq \mathcal{M}_A(\mathcal{H};\varepsilon/2,\delta,\varepsilon/2)$.

The other direction follows by using Markov's inequality. In particular, if we have that with probability at least $1-\delta$ over $S \sim \mathcal{D}$ it holds that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\ell(A(S;x), h^\star(x))] \leq \varepsilon\,,$$

then Markov's inequality gives us that

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_\mathcal{X}}[\ell(A(S;x), h^\star(x)) \geq \sqrt{\varepsilon}] \leq \sqrt{\varepsilon}\,,$$

which shows that $\mathcal{M}_A(\mathcal{H};\varepsilon,\delta) \geq \mathcal{M}_A(\mathcal{H};\sqrt{\varepsilon},\delta,\sqrt{\varepsilon})$. □

ERM Sample Complexity. In the special case where $\mathcal{A}$ is the class ERM of all possible ERM algorithms, i.e., algorithms that return a hypothesis whose sample error is exactly 0, we define the ERM sample complexity as the number of samples required by the worst-case ERM algorithm, i.e.,

$$\mathcal{M}_{\mathrm{ERM}}(\mathcal{H};\varepsilon,\delta) = \sup_{A\in\mathrm{ERM}} \mathcal{M}_A(\mathcal{H};\varepsilon,\delta)\,, \tag{3}$$

and its cut-off analogue as

$$\mathcal{M}_{\mathrm{ERM}}(\mathcal{H};\varepsilon,\delta,\gamma) = \sup_{A\in\mathrm{ERM}} \mathcal{M}_A(\mathcal{H};\varepsilon,\delta,\gamma)\,, \tag{4}$$

B γ-Graph Dimension and ERM Learnability

In this section, we show that γ-graph dimension determines the learnability of $\mathcal{H} \subseteq [0,1]^\mathcal{X}$ using *any* ERM learner. We first revisit the notion of partial concept classes which will be useful for deriving our algorithms.

B.1 Partial Concept Classes and A Naive Approach that Fails

[AHHM22] proposed an extension of the binary PAC model to handle *partial concept classes*, where $\mathcal{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$, for some input domain $\mathcal{X}$, where $h(x) = \star$ should be thought of as h not knowing the label of $x \in \mathcal{X}$. The main motivation behind their work is that partial classes allow one to conveniently express *data-dependent* assumptions. As an intuitive example, a halfspace with margin is a partial function that is undefined inside the forbidden margin and is a well-defined halfspace outside the margin boundaries. Instead of dealing with concept classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ where each concept $h \in \mathcal{H}$ is a **total function** $h : \mathcal{X} \to \mathcal{Y}$, we study **partial concept classes** $\mathcal{H} \subseteq (\mathcal{Y} \cup \{\star\})^{\mathcal{X}}$, where each concept h is now a **partial function** and $h(x) = \star$ means that the function h is **undefined** at x. We define the support of h as the set $\mathrm{supp}(h) = \{x \in \mathcal{X} : h(x) \neq \star\}$. Similarly as in the case of total classes, we say that a finite sequence $S = (x_1, y_1, \ldots, x_n, y_n)$ is realizable with respect to $\mathcal{H}$ if there exists some $h^* \in \mathcal{H}$ such that $h^*(x_i) = y_i, \forall i \in [n]$.

An important notion related to partial concept classes is that of *disambiguation*.

Definition 14 (Disambiguation of Partial Concept Class [AHHM22])**.** *Let $\mathcal{X}$ be an input domain. A total concept class $\overline{\mathcal{H}} \subseteq \{0,1\}^{\mathcal{X}}$ is a special type of a partial concept. Given some partial concept class $\mathcal{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ we say that $\overline{H}$ is a disambiguation of $\mathcal{H}$ if for any finite sequence $S \in (\mathcal{X} \times \{0,1\})^*$ if S is realizable with respect to $\mathcal{H}$, then S is realizable with respect to $\overline{\mathcal{H}}$.*

Intuitively, by disambiguating a partial concept class we convert it to a total concept class without reducing its "expressivity".

Let us first describe an approach to prove the upper bound, i.e., that if the scaled-graph dimension is finite for all scales then the class is ERM learnable, that does not work. We could perform the following transformation, inspired by the multiclass setting [DSS14]: for any $h \in \mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, let us consider the function $\widetilde{h} : \mathcal{X} \times [0,1] \to \{0,1\}$ with $\widetilde{h}(x,y) = 1$ if and only if $h(x) = y$, $\widetilde{h}(x,y) = 0$ if and only if $\ell(h(x),y) > \varepsilon$ and $\widetilde{h}(x,y) = \star$ otherwise. This induces a new binary *partial* hypothesis class $\widetilde{\mathcal{H}} = \{\widetilde{h} : h \in \mathcal{H}\} \subseteq \{0,1,\star\}^{\mathcal{X}}$. We note that $\mathbb{D}^{\mathrm{G}}_{\varepsilon}(\mathcal{H}) = \mathrm{VC}(\widetilde{\mathcal{H}})$. However, we cannot use ERM for the partial concept class since in general this approach fails. In particular, a sufficient condition for applying ERM is that $\mathrm{VC}(\{\mathrm{supp}(\widetilde{h}) : \widetilde{h} \in \widetilde{\mathcal{H}}\}) < \infty$.

Remark 1. *Predicting $\star$ in [AHHM22] implies a mistake for the setting of partial concept classes. However, in our regression setting, $\star$ is interpreted differently and corresponds to loss at most γ which is desirable. In particular, the hard instance for proper learners in the partial concepts paper (see Proposition 4 in [AHHM22]) is good in settings where predicting $\star$ does not count as a mistake, as in our regression case.*

B.2 Main Result

We are now ready to state the main result of this section. We will prove the next statement.

Theorem 1. *Let ℓ be the absolute loss function. For every class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and for any $\varepsilon, \delta, \gamma \in (0,1)^3$, the sample complexity bound for realizable PAC regression by any ERM satisfies*

$$\Omega\left(\frac{\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}) + \log(1/\delta)}{\varepsilon}\right) \leq \mathcal{M}_{\mathrm{ERM}}(\mathcal{H};\varepsilon,\delta,\gamma) \leq O\left(\frac{\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

In particular, any ERM algorithm $\mathbb{A}$ achieves

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\ell(A(S;x), h^\star(x)] \leq \inf_{\gamma\in[0,1]} \gamma + \widetilde{\Theta}\left(\frac{\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}) + \log(1/\delta)}{n}\right),$$

with probability at least $1-\delta$ over S of size n.

Proof. We prove the upper bound and the lower bound of the statement separately.

Upper Bound for the ERM learner. We deal with the cut-off loss problem with parameters $(\varepsilon,\delta,\gamma) \in (0,1)^3$. Our proof is based on a technique that uses a "ghost" sample to establish generalization guarantees of the algorithm. Let us denote

$$\mathrm{er}_{\mathcal{D},\gamma}(h) \triangleq \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\ell(h(x),h^\star(x)) > \gamma], \tag{1}$$

and for a dataset $z \in (\mathcal{X} \times [0,1])^n$,

$$\widehat{\mathrm{er}}_{z,\gamma}(h) \triangleq \frac{1}{|z|} \sum_{(x,y)\in z} \mathbb{I}\{\ell(h(x),y) > \gamma\}\,. \tag{2}$$

We will start by showing the next symmetrization lemma in our setting. Essentially, it bounds the probability that there exists a bad ERM learner by the probability that there exists an ERM learner on a sample r whose performance on a hidden sample s is bad. For a similar result, see Lemma 4.4 in [AB99].

Lemma 2 (Symmetrization)**.** *Let $\varepsilon, \gamma \in (0,1)^2, n > 0$. Fix $Z = \mathcal{X} \times [0,1]$. Let*

$$Q_{\varepsilon,\gamma} = \{z \in Z^n : \exists h \in \mathcal{H} : \widehat{\mathrm{er}}_{z,\gamma}(h) = 0,\ \mathrm{er}_{\mathcal{D},\gamma}(h) > \varepsilon\} \tag{3}$$

and

$$R_{\varepsilon,\gamma} = \{(r,s) \in Z^n \times Z^n : \exists h \in \mathcal{H} : \mathrm{er}_{\mathcal{D},\gamma}(h) > \varepsilon,\ \widehat{\mathrm{er}}_{r,\gamma}(h) = 0,\ \widehat{\mathrm{er}}_{s,\gamma}(h) \geq \varepsilon/2\}\,. \tag{4}$$

Then, for $n \geq c/\varepsilon$, where c is some absolute constant, we have that

$$\mathcal{D}^n(Q_{\varepsilon,\gamma}) \leq 2\mathcal{D}^{2n}(R_{\varepsilon,\gamma})\,.$$

Proof. We will show that $\mathcal{D}^{2n}(R_{\varepsilon,\gamma}) \geq \frac{\mathcal{D}^n(Q_{\varepsilon,\gamma})}{2}$. By the definition of $R_{\varepsilon,\gamma}$ we can write

$$\mathcal{D}^{2n}(R_{\varepsilon,\gamma}) = \int_{Q_{\varepsilon,\gamma}} \mathcal{D}^n(s : \exists h \in \mathcal{H}, \mathrm{er}_{\mathcal{D},\gamma}(h) > \varepsilon, \widehat{\mathrm{er}}_{r,\gamma}(h) = 0, \widehat{\mathrm{er}}_{s,\gamma}(h) \geq \varepsilon/2) d\mathcal{D}^n(r)\,.$$

For $r \in Q_{\varepsilon,\gamma}$, fix $h_r \in \mathcal{H}$ that satisfies $\widehat{\mathrm{er}}_{r,\gamma}(h_r) = 0,\ \mathrm{er}_{\mathcal{D},\gamma}(h) > \varepsilon$. It suffices to show that for h_r

$$\mathcal{D}^n(s : \widehat{\mathrm{er}}_{s,\gamma}(h_r) \geq \varepsilon/2) \geq 1/2\,.$$

Then, the proof of the lemma follows immediately. Since $\mathrm{er}_{\mathcal{D},\gamma}(h_r) > \varepsilon$, we know that $n \cdot \widehat{\mathrm{er}}_{s,\gamma}(h)$ follows a binomial distribution with probability of success on every try at least ε and n number of tries. Thus, the multiplicative version of Chernoff's bound gives us

$$\mathcal{D}^n(s : \widehat{\mathrm{er}}_{s,\gamma}(h_r) < \varepsilon/2) \leq e^{-\frac{n\cdot\varepsilon}{8}}\,.$$

Thus, if $n = c/\varepsilon$, for some appropriate absolute constant c we see that

$$\mathcal{D}^n(s : \widehat{\mathrm{er}}_{s,\gamma}(h_r) < \varepsilon/2) < 1/2\,,$$

which concludes the proof. □

Next, we can use a random swap argument to upper bound $\mathcal{D}^{2n}(R_{\varepsilon,\gamma})$ with a quantity that involves a set of permutations over the sample of length $2n$. The main idea behind the proof is to try to leverage the fact that each of the labeled examples is as likely to occur among the first n examples or the last n examples.

Following [AB99], we denote by Γ_n the set of all permutations on $\{1,\ldots,2n\}$ that swap i and $n+i$, for all i that belongs to $\{1,\ldots,n\}$. In other words, for all $\sigma \in \Gamma_n$ and $i \in \{1,\ldots,n\}$ either $\sigma(i) = i, \sigma(n+i) = n+i$ or $\sigma(i) = n+i, \sigma(n+i) = i$. Thus, we can think of σ as acting on coordinates where it (potentially) swaps one element from the first half of the sample with the corresponding element on the second half of the sample. For some $z \in Z^{2n}$ we overload the notation and denote $\sigma(z)$ the effect of applying σ to the sample z.

We are now ready to state the bound. Importantly, it shows that by (uniformly) randomly choosing a permutation $\sigma \in \Gamma_n$ we can bound the probability that a sample falls into the bad set $R_{\varepsilon,\gamma}$ by a quantity that does not depend on the distribution $\mathcal{D}$.

Lemma 3 (Random Swaps; Adaptation of Lemma 4.5 in [AB99])**.** *Fix $Z = \mathcal{X} \times [0,1]$. Let $R_{\varepsilon,\gamma}$ be any subset of Z^{2n} and $\mathcal{D}$ any probability distribution on Z. Then*

$$\mathcal{D}^{2n}(R_{\varepsilon,\gamma}) = \mathop{\mathbf{E}}_{z\sim\mathcal{D}^{2n}} \mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z) \in R_{\varepsilon,\gamma}] \leq \max_{z\in Z^{2n}} \mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z) \in R_{\varepsilon,\gamma}]\,,$$

where $\mathbb{U}(\Gamma_n)$ is the uniform distribution over the set of swapping permutations Γ_n.

Proof. First, notice that the bound

$$\mathop{\mathbf{E}}_{z\sim\mathcal{D}^{2n}}\mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z)\in R_{\varepsilon,\gamma}]\leq\max_{z\in Z^{2n}}\mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z)\in R_{\varepsilon,\gamma}]\,,$$

follows trivially and the maximum exists since $\mathbf{Pr}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z)\in R_{\varepsilon,\gamma}]$ takes finitely many values for any finite n and all $z\in Z^{2n}$. Thus, the bulk of the proof is to show that

$$\mathcal{D}^{2n}(R_{\varepsilon,\gamma})=\mathop{\mathbf{E}}_{z\sim\mathcal{D}^{2n}}\mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z)\in R_{\varepsilon,\gamma}]\,.$$

First, notice that since example is drawn i.i.d., for any swapping permutation $\sigma\in\Gamma_n$ we have that

$$\mathcal{D}^{2n}(R_{\varepsilon,\gamma})=\mathcal{D}^{2n}\left(\left\{z\in Z^{2n}:\sigma(z)\in R_{\varepsilon,\gamma}\right\}\right)\,. \tag{5}$$

Thus, the following holds

$$\begin{aligned}
\mathcal{D}^{2n}(R_{\varepsilon,\gamma})&=\int_{Z^{2n}}\mathbb{I}\{z\in R_{\varepsilon,\gamma}\}\,d\mathcal{D}^{2n}(z)\\
&=\frac{1}{|\Gamma_n|}\sum_{\sigma\in\Gamma_n}\int_{Z^{2n}}\mathbb{I}\{\sigma(z)\in R_{\varepsilon,\gamma}\}\,d\mathcal{D}^{2n}(z)\\
&=\int_{Z^{2n}}\left(\frac{1}{|\Gamma_n|}\sum_{\sigma\in\Gamma_n}\mathbb{I}\{\sigma(z)\in R_{\varepsilon,\gamma}\}\right)d\mathcal{D}^{2n}(z)\\
&=\mathop{\mathbf{E}}_{z\sim\mathcal{D}^{2n}}\mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z)\in R_{\varepsilon,\gamma}]\,,
\end{aligned}$$

where the first equation follows by definition, the second by Equation (5), the third because the number of terms in the summation is finite, and the last one by definition. □

As a last step we can bound the above RHS by using all possible patterns when $\mathcal{H}$ is (roughly speaking) projected in the sample $rs\in Z^{2n}$.

Lemma 4 (Bounding the Bad Event)**.** *Fix $Z=\mathcal{X}\times[0,1]$. Let $R_{\varepsilon,\gamma}\subseteq Z^{2n}$ be the set*

$$R_{\varepsilon,\gamma}=\{(r,s)\in Z^n\times Z^n:\exists h\in\mathcal{H}:\widehat{\mathrm{er}}_{r,\gamma}(h)=0,\ \widehat{\mathrm{er}}_{s,\gamma}(h)\geq\varepsilon/2\}\,.$$

Then

$$\mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z)\in R_{\varepsilon,\gamma}]\leq(2n)^{O(\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})\log(2n))}2^{-n\varepsilon/2}\,.$$

Proof. Throughout the proof we fix $z=(z_1,...,z_{2n})\in Z^{2n}$, where $z_i=(x_i,y_i)=(x_i,h^\star(x_i))$ and let $S=\{x_1,...,x_{2n}\}$. Consider the projection set $\mathcal{H}|_S$. We define a *partial binary* concept class $\mathcal{H}'\subseteq\{0,1,\star\}^{2n}$ as follows:

$$\mathcal{H}':=\left\{h'\in\{0,1,\star\}^{2n}:\exists h\in\mathcal{H}|_S:\forall i\in[2n]\begin{Bmatrix}h(x_i)=y_i, & h'(i)=0\\ \ell(h(x_i),y_i)>\gamma, & h'(i)=1\\ 0<\ell(h(x_i),y_i)\leq\gamma, & h'(i)=\star\end{Bmatrix}\right\}\,.$$

Importantly, we note that, by definition, $\mathrm{VC}(\mathcal{H}')\leq\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})$.

Currently, we have a partial binary concept class $\mathcal{H}'$. As a next step, we would like to replace the $\star$ symbols and essentially reduce the problem to a total concept class. This procedure is called disambiguation (cf. Definition 14). The next key lemma shows that there exists a compact (in terms of cardinality) disambiguation of a VC partial concept class for finite instance domains.

Lemma 5 (Compact Disambiguations, see [AHHM22])**.** *Let $\mathcal{H}$ be a partial concept class on a finite instance domain $\mathcal{X}$ with $\mathrm{VC}(\mathcal{H})=d$. Then there exists a disambiguation $\overline{\mathcal{H}}$ of $\mathcal{H}$ with size $|\overline{\mathcal{H}}|=|\mathcal{X}|^{O(d\log|\mathcal{X}|)}$.*

This means that there exists a disambiguation $\overline{\mathcal{H}'}$ of $\mathcal{H}'$ of size at most

$$(2n)^{O(\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})\log(2n))}\,.$$

Since this (total) binary concept class is finite, we can apply the following union bound argument. We have that $\sigma(z) \in R$ if and only if some $h \in \mathcal{H}$ satisfies

$$\frac{\sum_{i=1}^{n} \mathbb{I}\{\ell(h(x_{\sigma(i)}), y_{\sigma(i)}) > \gamma\}}{n} = 0, \quad \frac{\sum_{i=1}^{n} \mathbb{I}\{\ell(h(x_{\sigma(n+i)}), y_{\sigma(n+i)}) > \gamma\}}{n} \geq \varepsilon/2\,.$$

We can relate this event with an event about the disambiguated partial concept class $\overline{\mathcal{H}'}$ since the number of 1's can only increase. In particular, for any swapping permutation σ of the $2n$ points, if there exists a function in $\mathcal{H}$ that is correct on the first n points and is off by at least γ on at least $\varepsilon n/2$ of the remaining n points, then there is a function in the disambiguation $\overline{\mathcal{H}'}$ that is 0 on the first n points and is 1 on those same $\epsilon n/2$ of the remaining points.

If we fix some $\sigma \in \Gamma_n$, and some $\overline{h'} \in \overline{\mathcal{H}'}$ then $\overline{h'}$ is a witness that $\sigma(z) \in R_{\varepsilon,\gamma}$ only if $\forall i \in [n]$ we do not have that $\overline{h'}(i) = 1, \overline{h'}(i+n) = 1$. Thus at least one of $\overline{h'}(i), \overline{h'}(n+i)$ must be zero. Moreover, at least $n\varepsilon/2$ entries must be non-zero. Thus, when we draw random swapping permutation the probability that all the non-zero entries land on the second half of the sample sample is at most $2^{-n\varepsilon/2}$.

Crucially since the number of possible functions is at most $|\overline{\mathcal{H}'}| \leq (2n)^{O(\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})\log(2n))}$ a union bound gives us that

$$\mathop{\mathbf{Pr}}_{\sigma\sim\mathbb{U}(\Gamma_n)}[\sigma(z) \in R_{\varepsilon,\gamma}] \leq (2n)^{O(\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})\log(2n))} \cdot 2^{-n\varepsilon/2}\,.$$

Thus, since $z \in Z^{2n}$ was arbitrary we have that

$$\max_{z\in Z^{2n}} \mathop{\mathbf{Pr}}_{\sigma\mathbb{U}(\Gamma_n)}[\sigma(z) \in R_{\varepsilon,\gamma}] \leq (2n)^{O(\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})\log(2n))} \cdot 2^{-n\varepsilon/2}\,.$$

This concludes the proof. □

Lower Bound for the ERM learner. Our next goal is to show that

$$\mathcal{M}_{\mathrm{ERM}}(\mathcal{H};\varepsilon,\delta,\gamma) \geq C_0 \cdot \frac{\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}) + \log(1/\delta)}{\varepsilon}\,.$$

To this end, we will show that there exists an ERM learner satisfying this lower bound. This will establish that the finiteness of γ-graph dimension for any $\gamma \in (0,1)$, is necessary for PAC learnability using *a worst-case* ERM. It suffices to show that there exists a bad ERM algorithm that requires at least $C_0 \frac{\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})+\log(1/\delta)}{\varepsilon}$ samples to cut-off PAC learn $\mathcal{H}$. First let us consider the case where $d = \mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}) < \infty$ and let $S = \{x_1, ..., x_d\}$ be a γ-graph-shattered set by $\mathcal{H}$ with witness f_0. Consider the ERM learner $\mathcal{A}$ that works as follows: Upon seeing a sample $T \subseteq S$ consistent with f_0, $\mathcal{A}$ returns a function $\mathcal{A}(T)$ that is equal to f_0 on elements of T and γ-far from f_0 on $S \setminus T$. Such a function exists since S is γ-graph-shattered with witness f_0. Let us take $\delta < 1/100$ and $\varepsilon < 1/12$. Define a distribution over $S \subseteq \mathcal{X}$ such that

$$\mathbf{Pr}[x_1] = 1 - 2\varepsilon, \quad \mathbf{Pr}[x_i] = 2\varepsilon/(d-1),\ \forall i \in \{2, ..., d\}\,.$$

Let us set $h^\star = f_0$ and consider m samples $\{(z_i, f_0(z_i)\}_{i\in[m]}$. Since we work in the scaled PAC model, $\mathcal{A}$ will make a γ-error on all examples from S which are not in the sample (since in that case the output will be γ-far from the true label). Let us take $m \leq \frac{d-1}{6\varepsilon}$. Then, the sample will include at most $(d-1)/2$ examples which are not x_1 with probability $1/100$, using Chernoff's bound. Conditioned on that event, this implies that the ERM learner will make a γ-error with probability at least $\frac{2\varepsilon}{d-1}\cdot(d-1-\frac{d-1}{2}) = \varepsilon$, over the random draw of the test point. Thus, $\mathcal{M}_{\mathcal{A}}(\mathcal{H};\varepsilon,\delta,\gamma) = \Omega(\frac{d-1}{\varepsilon})$. Moreover, the probability that the sample will only contain x_1 is $(1-2\varepsilon)^m \geq e^{-4\varepsilon m}$ which is greater that δ whenever $m \leq \log(1/\delta)/(4\varepsilon)$. This implies that the γ-cut-off ERM sample complexity is lower bounded by

$$\max\left\{\frac{d-1}{6\varepsilon}, \frac{\log(1/\delta)}{2\varepsilon}\right\} = C_0 \cdot \frac{\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}) + \log(1/\delta)}{\varepsilon}\,.$$

Thus $\mathcal{M}_{\mathrm{ERM}}(\mathcal{H};\varepsilon,\delta,\gamma)$, satisfies the desired bound when the dimension is finite. Finally, it remains to claim about the case where $\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}) = \infty$ for the given γ. We consider a sequence of γ-graph-shattered sets S_n with $|S_n| = n$ and repeat the claim for the finite case. This will yield that for any n the cut-off ERM sample complexity is lower bounded by $\Omega((n + \log(1/\delta))/\varepsilon)$ and this yields that $\mathcal{M}_{\mathrm{ERM}}(\mathcal{H};\varepsilon,\delta,\gamma) = \infty$. □

However, as Example 4 shows, the optimal learner cannot be proper and as a result, this dimension does not characterize PAC learnability for real-valued regression (there exist classes whose γ-graph dimension is infinite but are PAC learnable in the realizable regression setting).

C γ-OIG Dimension and Learnability

In this section we identify a dimension characterizing qualitatively and quantitatively what classes of predictors $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ are PAC learnable and we provide PAC learners that achieve (almost) optimal sample complexity. In particular, we show the following result.

Theorem 2. *Let ℓ be the absolute loss function. For every class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and for any $\varepsilon, \delta, \gamma \in (0,1)^3$, the sample complexity bound for realizable PAC regression satisfies*

$$\Omega\left(\frac{\mathbb{D}^{\mathrm{OIG}}_{2\gamma}(\mathcal{H})}{\varepsilon}\right) \leq \mathcal{M}(\mathcal{H};\varepsilon,\delta,\gamma) \leq O\left(\frac{\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})}{\varepsilon}\log^2\frac{\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right).$$

In particular, there exists an algorithm A such that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\ell(A(S;x),h^\star(x)] \leq \inf_{\gamma\in[0,1]}\gamma + \widetilde{\Theta}\left(\frac{\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H}) + \log(1/\delta)}{n}\right),$$

with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$.

A finite fat-shattering dimension implies a finite OIG dimension. Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ be a function class with finite γ-fat shattering dimension for any $\gamma > 0$. We show that $\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{F})$ is upper bounded (up to constants and log factors) by $\mathbb{D}^{\mathrm{fat}}_{c\gamma}(\mathcal{F})$, for some $c > 0$, where the OIG dimension is defined with respect to the ℓ_1 loss. Note that the opposite direction does not hold. Example 1 exhibits a function class with an infinite fat-shattering dimension that can be learned with a single example, and as a result, the OIG dimension has to be finite. On the one hand, we have an upper bound on the sample complexity of $O\left(\frac{1}{\epsilon}\left(\mathbb{D}^{\mathrm{fat}}_{\tilde{c}\varepsilon}(\mathcal{F})\log^2\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$, for any $\varepsilon, \delta \in (0,1)$. See sections 19.6 and 20.4 about the restricted model in [AB99]. On the other hand, we prove in Lemma 6 a lower bound on the sample complexity of $\Omega\left(\frac{\mathbb{D}^{\mathrm{OIG}}_{2\varepsilon}(\mathcal{F})}{\epsilon}\right)$, for any $\varepsilon \in (0,1)$, and so $\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{F})$ is upper bounded by $\mathbb{D}^{\mathrm{fat}}_{c\gamma}(\mathcal{F})$ up to constants and log factors.

C.1 Proof of the Lower Bound

Lemma 6. *[Lower Bound of PAC Regression] Let A be any learning algorithm and $\varepsilon, \delta, \gamma \in (0,1)^3$ such that $\delta < \varepsilon$. Then,*

$$\mathcal{M}_A(\mathcal{H};\varepsilon,\delta,\gamma) \geq \Omega\left(\frac{\mathbb{D}^{\mathrm{OIG}}_{2\gamma}(\mathcal{H})}{\varepsilon}\right).$$

Proof. Let $n_0 = \mathbb{D}^{\mathrm{OIG}}_{2\gamma}(\mathcal{H})$. Let $n \in \mathbb{N}, 1 < n \leq n_0$. We know that for each such n there exists some $S \in \mathcal{X}^n$ such that the one-inclusion graph of $\mathcal{H}|_S$ has the property that: there exists a finite subgraph $G = (V,E)$ of $G^{\mathrm{OIG}}_{\mathcal{H}|_S}$ such that for any orientation $\sigma : E \to V$ of the subgraph, there exists a vertex $v \in V$ with $\mathrm{outdeg}(v;\sigma,2\gamma) > \mathbb{D}^{\mathrm{OIG}}_{2\gamma}(\mathcal{H})/3$.

Given the learning algorithm $A : (\mathcal{X}\times[0,1])^\star \times \mathcal{X} \to [0,1]$, we can describe an orientation σ_A of the edges in E. For any vertex $v = (v_1,\ldots,v_n) \in V$ let P_v be the distribution over $(x_1,v_1),....,(x_n,v_n)$ defined as

$$P_v((x_1,v_1)) = 1-\varepsilon,\ P_v((x_t,v_t)) = \frac{\varepsilon}{n-1},\ t \in \{2,\ldots,n\}.$$

Let $m = n/(2\varepsilon)$. For each vertex $v \in V$ and direction $t \in [n]$, consider the hyperedge $e_{t,v}$. For each $u \in e_{t,v}$ we define

$$p_t(u) = \Pr_{S\sim P_u^m}[\ell(A(S;x_t),u_t) > \gamma | (x_t,u_t) \notin S],$$

and let $C_{e_{t,v}} = \{u \in e_{t,v} : p_t(u) < 1/2\}$. If $C_{e_{t,v}} = \emptyset$, we orient the edge $e_{t,v}$ arbitrarily. Since for all $u, v \in e_{t,v}$ the distributions P_u^m, P_v^m conditioned on the event that $(x_t,u_t),(x_t,v_t)$ respectively

are not in S are the same, we can see that $\forall u, v \in C_{e_{t,v}}$ it holds that $\ell(u_t, v_t) \leq 2\gamma$. We orient the edge $e_{t,v}$ using an arbitrary element of $C_{e_{t,v}}$.

Because of the previous discussion, we can bound from above the out-degree of all vertices $v \in V$ with respect to the orientation σ_A as follows:

$$\operatorname{outdeg}(v; \sigma_A, 2\gamma) \leq \sum_t \mathbb{I}\{p_t(v) \geq 1/2\} \leq 1 + 2\sum_{t=2}^{n} \mathop{\mathbf{Pr}}_{S \sim P_v^m}[\ell(A(S, x_t), y_t) > \gamma | (x_t, y_t) \notin S] .$$

Notice that

$$\mathop{\mathbf{Pr}}_{S \sim P_v^m}[\ell(A(S, x_t), y_t) > \gamma | (x_t, y_t) \notin S] = \frac{\mathbf{Pr}_{S \sim P_v^m}[\{\ell(A(S, x_t), y_t) > \gamma\} \wedge \{(x_t, y_t) \notin S\}]}{\mathbf{Pr}_{S \sim P_v^m}[(x_t, y_t) \notin S]} ,$$

and by the definition of P_v, we have that

$$\mathop{\mathbf{Pr}}_{S \sim P_v^m}[(x_t, y_t) \notin S] = \left(1 - \frac{\varepsilon}{n-1}\right)^m \geq 1 - \frac{n}{2(n-1)} ,$$

since $m = n/(2\varepsilon)$. Combining the above, we get that

$$\operatorname{outdeg}(v; \sigma_A, 2\gamma) \leq 1 + 2\left(1 - \frac{n}{2(n-1)}\right) \sum_{t=2}^{n} \mathop{\mathbf{E}}_{S \sim P_v^m}[\mathbb{I}\{\ell(A(S, x_t), y_t) > \gamma\} \cdot \mathbb{I}\{(x_t, y_t) \notin S\}] ,$$

and so

$$\begin{aligned}
\operatorname{outdeg}(v; \sigma_A, 2\gamma) &\leq 1 + 2\left(1 - \frac{n}{2(n-1)}\right) \mathop{\mathbf{E}}_{S \sim P_v^m}\left[\sum_{t=2}^{n} \mathbb{I}\{\ell(A(S, x_t), y_t) > \gamma\}\right] \\
&= 1 + 2\left(1 - \frac{n}{2(n-1)}\right) \frac{n-1}{\varepsilon} \mathop{\mathbf{E}}_{S \sim P_v^m}\left[\frac{\varepsilon}{n-1} \sum_{t=2}^{n} \mathbb{I}\{\ell(A(S, x_t), y_t) > \gamma\}\right] \\
&\leq 1 + 2\left(1 - \frac{n}{2(n-1)}\right) \frac{n-1}{\varepsilon} \mathop{\mathbf{E}}_{S \sim P_v^m}\left[\mathop{\mathbf{E}}_{(x,y) \sim P_v}[\mathbb{I}\{\ell(A(S; x), y) > \gamma\}]\right] \\
&= 1 + 2\left(1 - \frac{n}{2(n-1)}\right) \frac{n-1}{\varepsilon} \mathop{\mathbf{E}}_{S \sim P_v^m}\left[\mathop{\mathbf{Pr}}_{(x,y) \sim P_v}[\ell(A(S; x), y) > \gamma]\right] \\
&\leq 1 + \frac{n-2}{\varepsilon} \mathop{\mathbf{E}}_{S \sim P_v^m}\left[\mathop{\mathbf{Pr}}_{(x,y) \sim P_v}[\ell(A(S; x), y) > \gamma]\right]
\end{aligned}$$

By picking "hard" distribution $\mathcal{D} = P_{v^\star}$, where $v^\star \in \arg\max_{v' \in V} \operatorname{outdeg}(v'; \sigma_A, 2\gamma)$ we get that that

$$\begin{aligned}
\mathop{\mathbf{E}}_{S \sim P_{v^\star}^m}\left[\mathop{\mathbf{Pr}}_{(x,y) \sim P_{v^\star}}[\ell(A(S; x), y) > \gamma]\right] &\geq (\operatorname{outdeg}(v^\star; \sigma_A, 2\gamma) - 1) \cdot \frac{\varepsilon}{n-2} \\
&\geq \frac{\varepsilon}{6} ,
\end{aligned}$$

since $\operatorname{outdeg}(v^\star; \sigma_A, 2\gamma) > n/3$. By picking $n = n_0$ we see that when the learner uses $m = n_0/\varepsilon$ samples then its expected error is at least $\varepsilon/6$. Notice that when the learner uses $m' = \mathcal{M}_A(\mathcal{H}; \varepsilon, \delta, \gamma)$ samples we have that

$$\mathop{\mathbf{E}}_{S \sim P_{v^\star}^{m'}}\left[\mathop{\mathbf{Pr}}_{(x,y) \sim P_{v^\star}}[\ell(A(S; x), y) > \gamma]\right] \leq \delta + (1 - \delta)\varepsilon \leq \delta + \varepsilon \leq 2\varepsilon .$$

Thus, we see that for any algorithm A

$$\mathcal{M}_A(\mathcal{H}; \varepsilon, \delta, \gamma) \geq \Omega\left(\frac{\mathbb{D}_{2\gamma}^{\mathrm{OIG}}(\mathcal{H})}{\varepsilon}\right) ,$$

hence

$$\mathcal{M}(\mathcal{H}; \varepsilon, \delta, \gamma) \geq \Omega\left(\frac{\mathbb{D}_{2\gamma}^{\mathrm{OIG}}(\mathcal{H})}{\varepsilon}\right) .$$

□

C.2 Proof of the Upper Bound

Let us present the upper bound. For this proof, we need three tools: we will provide a weak learner based on the scaled one-inclusion graph, a boosting algorithm for real-valued functions, and consistent sample compression schemes for real-valued functions.

To this end, we introduce the one-inclusion graph (OIG) algorithm $\mathbb{A}^{\mathrm{OIG}}_{\gamma}$ for realizable regression at scale γ.

C.2.1 Scaled One-Inclusion Graph Algorithm and Weak Learning

First, we show that every scaled orientation σ of the one-inclusion graph gives rise to a learner $\mathbb{A}_\sigma$ whose expected absolute loss is upper bounded by the maximum out-degree induced by σ.

Lemma 7 (From Orientations to Learners). *Let $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$ and $h^\star \in \mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, let $n \in \mathbb{N}$ and $\gamma \in (0,1)$. Then, for any orientation $\sigma : E_n \to V_n$ of the scaled-one-inclusion graph $G^{\mathrm{OIG}}_{\mathcal{H}} = (V_n, E_n)$, there exists a learner $\mathbb{A}_\sigma : (\mathcal{X} \times [0,1])^{n-1} \to [0,1]^{\mathcal{X}}$, such that*

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}_{\mathcal{X}}^{n-1}}\left[\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\ell(\mathbb{A}_\sigma(x), h^\star(x)) > \gamma\right]\right] \le \frac{\max_{v\in V_n}\mathrm{outdeg}(v;\sigma,\gamma)}{n}\,,$$

where $\mathbb{A}_\sigma$ is trained using a sample S of size $n-1$ realized by $h^\star$.

Algorithm 1 From orientation σ to learner $\mathbb{A}_\sigma$

Input: An $\mathcal{H}$-realizable sample $\{(x_i, y_i)\}_{i=1}^{n-1}$ and a test point $x \in \mathcal{X}$, $\gamma \in (0,1)$.
Output: A prediction $\mathbb{A}_\sigma(x)$.

1. Create the one-inclusion graph $G^{\mathrm{OIG}}_{\mathcal{H}|_{(x_1,\ldots,x_{n-1},x)}}$.
2. Consider the edge in direction n defined by the realizable sample $\{(x_i, y_i)\}_{i=1}^{n-1}$; let
$$e = \{h \in \mathcal{H}|_{(x_1,\ldots,x_{n-1},x)} : \forall i \in [n-1]\ h(i) = y_i\}\,.$$
3. Return $\mathbb{A}_\sigma(x) = \sigma(e)(n)$.

Proof. By the classical leave-one-out argument, we have that

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}_{\mathcal{X}}^{n-1}}\left[\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\ell(\mathbb{A}_\sigma(x), h^\star(x)) > \gamma\right]\right] = \mathop{\mathbf{E}}_{(S,(x,y))\sim\mathcal{D}^n}\left[\mathbb{I}\{\ell(h_S(x), y) > \gamma\}\right] = \mathop{\mathbf{E}}_{S'\sim\mathcal{D}^n, I\sim\mathbb{U}([n])}\left[\mathbb{I}\{\ell(h_{S'_{-I}}(x'_I), y'_I) > \gamma\}\right],$$

where h_S is the predictor $\mathbb{A}_\sigma$ using the examples S, and $\mathbb{U}([n])$ is the uniform distribution on $\{1,\ldots,n\}$. Now for every fixed S' we have that

$$\mathop{\mathbf{E}}_{I\sim\mathbb{U}([n])}\left[\mathbb{I}\{\ell(h_{S'_{-I}}(x'_I), y'_I) > \gamma\}\right] = \frac{1}{n}\sum_{i\in[n]}\mathbb{I}\{\ell(\sigma(e_i)(i), y'_i) > \gamma\} = \frac{\mathrm{outdeg}(y';\sigma,\gamma)}{n}\,,$$

where y' is the node of the scaled OIG that corresponds to the true labeling of S'. By taking expectation over $S' \sim \mathcal{D}^n$ we get that

$$\mathop{\mathbf{E}}_{S'\sim\mathcal{D}^n, I\sim\mathbb{U}([n])}\left[\mathbb{I}\{\ell(h_{S'_{-I}}(x'_I), y'_I) > \gamma\}\right] \le \mathop{\mathbf{E}}_{S'\sim\mathcal{D}^n}\left[\frac{\mathrm{outdeg}(y';\sigma,\gamma)}{n}\right] \le \frac{\max_{v\in V_n}\mathrm{outdeg}(v;\sigma,\gamma)}{n}\,.$$

□

Equipped with the previous result, we are now ready to show that when the learner gets at least $\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})$ samples as its training set, then its expected γ-cutoff loss is bounded away from $1/2$.

Lemma 8 (Scaled OIG Guarantee (Weak Learner)). *Let $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$ and $h^\star \in \mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\gamma \in (0,1)$. Then, for all $n > \mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})$ there exists an orientation $\sigma^\star$ such that for the prediction error of the one-inclusion graph algorithm $\mathbb{A}^{\mathrm{OIG}}_{\sigma^\star} : (\mathcal{X}\times[0,1])^{n-1}\times\mathcal{X} \to [0,1]$, it holds that*

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}_{\mathcal{X}}^{n-1}}\left[\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\ell(\mathbb{A}^{\mathrm{OIG}}_{\sigma^\star}(x), h^\star(x)) > \gamma\right]\right] \le 1/3\,.$$

Proof. Fix $\gamma \in (0,1)$. Assume that $n > \mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H})$ and let $G^{\mathrm{OIG}}_{\mathcal{H}|_{(S,x)}} = (V_n, E_n)$ be the possibly infinite scaled one-inclusion graph. By the definition of the γ-OIG dimension (see Definition 9), for every finite subgraph $G = (V, E)$ of $G^{\mathrm{OIG}}_{\mathcal{H}|_{(S,x)}}$ there exists an orientation $\sigma : E \to V$ such that for every vertex in G the out-degree is at most $n/3$, i.e.,

$$\forall S \in \mathcal{X}^n, \forall \text{ finite } G = (V, E) \text{ of } G^{\mathrm{OIG}}_{\mathcal{H}|_S}, \exists \text{ orientation } \sigma_E \text{ s.t. } \forall v \in V, \text{ it holds } \mathrm{outdeg}(v; \sigma_E, \gamma) \leq n/3\,.$$

First, we need to create an orientation of the whole (potentially infinite) one-inclusion graph.

We will create this orientation using the compactness theorem of first-order logic which states that a set of formulas Φ is satisfiable if and only if it is finitely satisfiable, i.e., every finite subset $\Phi' \subseteq \Phi$ is satisfiable. Let $G^{\mathrm{OIG}}_{\mathcal{H}|_S} = (V_n, E_n)$ be the (potentially infinite) one-inclusion graph of $\mathcal{H}|_S$. Let $\mathcal{Z}$ be the set of pairs $z = (v, e) \in V_n \times E_n$ so that $v \in e$. Our goal is to assign binary values to each $z \in \mathcal{Z}$. We define the following sets of formulas:

- For each $e \in E_n$ we let $\Phi_e :=$ $\exists$ exactly one $v \in e : z(v, e) = 1$.
- For each $v \in V_n$ we let $\Phi_v :=$ $\exists$ at most $n/3$ different $e_{i,f} \in E_n : v \in e_{i,f} \wedge (\exists v' \in e_{i,f} : (z(v', e) = 1 \wedge \ell(v'_i, v_i) > \gamma))$

It is not hard to see that each Φ_e, Φ_v can be expressed in first-order logic. Then, we define

$$\Phi := \left(\bigcap_{e \in E_n} \Phi_e \right) \cap \left(\bigcap_{v \in V_n} \Phi_v \right) .$$

Notice that an orientation of the edges of $G^{\mathrm{OIG}}_{\mathcal{H}|_S}$ is equivalent to picking an assignment of the elements of $\mathcal{Z}$ that satisfies all the Φ_e. Moreover, notice that for such an assignment, if all the Φ_v are satisfied then then maximum γ-scaled out-degree of $G^{\mathrm{OIG}}_{\mathcal{H}|_S}$ is at most $n/3$.

We will now show that Φ is finitely satisfiable. Let Φ' be a finite subset of Φ and let $E' \subseteq E_n, V' \subseteq V_n$, be the set of edges, vertices that appear in Φ', respectively. If $V' = \emptyset$, then we can orient the edges in E' arbitrarily and satisfy Φ'. Similarly, if $E' = \emptyset$ we can let all the $z(e, v) = 0$ and satisfy all the $\Phi_v, v \in V'$. Thus, assume that both sets are non-empty. Consider the finite subgraph of $G^{\mathrm{OIG}}_{\mathcal{H}|_S}$ that is induced by V' and let E'' be the set of edges of this subgraph. For every edge $e \in E' \setminus E''$[4], pick an arbitrary orientation, i.e, for exactly one $v \in e$ set $z(e, v) = 1$ and for the remaining $v' \in e$ set $z(e, v') = 0$. By the definition of $\mathbb{D}_{\gamma}^{\mathrm{OIG}}(\mathcal{H})$ there is an orientation $\sigma_{E''}$ of the edges in E'' such that $\forall v \in V' \mathrm{outdeg}(v; \sigma_{E''}, \gamma) \leq n/3$. For every $e \in E''$ pick the assignment of all the $z(v, e), v \in e$, according to the orientation $\sigma_{E''}$. Thus, because of the maximum out-degree property of $\sigma_{E''}$ we described before, we can also see that all the $\Phi_v, v \in V'$, are satisfied. Hence, we have shown that Φ is finitely satisfiable, so it is satisfiable. This assignment on $z(v, e)$ induces an orientation $\sigma^\star$ under which all the vertices of the one-inclusion graph have out-degree at most $n/3$.

We will next use the orientation $\sigma^\star$ of $G^{\mathrm{OIG}}_{\mathcal{H}|_S} = (V_n, E_n)$ to design a learner $\mathbb{A}^{\mathrm{OIG}}_{\sigma^\star} : (\mathcal{X} \times [0,1])^{n-1} \times \mathcal{X} \to [0,1]$, invoking Lemma 7. In particular, we get that, from Lemma 7 with the chosen orientation,

$$\mathop{\mathbf{E}}_{S \sim \mathcal{D}_{\mathcal{X}}^{n-1}} \left[\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\ell(\mathbb{A}^{\mathrm{OIG}}_{\sigma^\star}(x), h^\star(x)) > \gamma \right] \right] \leq \frac{\max_{v \in V_n} \mathrm{outdeg}(v; \sigma^\star, \gamma)}{n} \leq 1/3\,,$$

which concludes the proof. □

C.2.2 Boosting Real-Valued Functions

Definition 15 (Weak Real-Valued Learner)**.** *Let ℓ be a loss function. Let $\gamma \in [0,1]$, $\beta \in (0, \frac{1}{2})$, and $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$. For a distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$ and true target function $h^\star \in \mathcal{H}$, we say that $f : \mathcal{X} \to [0,1]$ is (γ, β)-weak learner with respect to $\mathcal{D}_{\mathcal{X}}$ and $h^\star$, if*

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_{\mathcal{X}}} \left[\ell(f(x), h^\star(x)) > \gamma \right] < \frac{1}{2} - \beta.$$

[4]Since the edges in E'' are of finite length, we first need to map them to the appropriate edges in E'.

Following [HKS19], we define the weighted median as

$$\text{Median}(y_1,\ldots,y_T;\alpha_1,\ldots,\alpha_T) = \min\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} \right\},$$

and the weighted quantiles, for $\theta \in [0, 1/2]$, as

$$Q_\theta^+(y_1,\ldots,y_T;\alpha_1,\ldots,\alpha_T) = \min\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \theta \right\}$$

$$Q_\theta^-(y_1,\ldots,y_T;\alpha_1,\ldots,\alpha_T) = \max\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j > y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \theta \right\},$$

and we let $Q_\theta^+(x) = Q_\theta^+(h_1(x),\ldots,h_T(x);\alpha_1,\ldots,\alpha_T), Q_\theta^-(x) = Q_\theta^-(h_1(x),\ldots,h_T(x);\alpha_1,\ldots,\alpha_T)$, where $h_1,\ldots,h_T,\alpha_1,\ldots,\alpha_T$ are the values returned by Algorithm 2. The following guarantee holds for this procedure.

Lemma 9 (MedBoost guarantee [Kég03])**.** *Let ℓ be the absolute loss and $S = \{(x_i,y_i)\}_{i=1}^m$, $T = O\left(\frac{1}{\theta^2}\log(m)\right)$. Let $h_1,\ldots,h_T$ and $\alpha_1,\ldots,\alpha_T$ be the functions and coefficients returned from* MedBoost*. For any $i \in \{1,\ldots,m\}$ it holds that*

$$\max\left\{ \ell\left(Q_{\theta/2}^+(x_i), y_i\right), \ell\left(Q_{\theta/2}^-(x_i), y_i\right) \right\} \le \gamma.$$

Algorithm 2 MedBoost [Kég03]

Input: $S = \{(x_i,y_i)\}_{i=1}^m$.
Parameters: γ, β, T.
Initialize $\mathcal{P}_1$ = Uniform(S).
For $t = 1,\ldots,T$:

1. Find a (γ,β)-weak learner h_t with respect to $(x_i,y_i) \sim \mathcal{P}_t$, using a subset $S_t \subseteq S$.
2. For $i = 1,\ldots,m$:
   (a) Set $w_i^{(t)} = 1 - 2\mathbb{I}\left[\ell(h_t(x_i),y_i) > \gamma\right]$.
   (b) Set $\alpha_t = \frac{1}{2}\log\left(\frac{(1-\gamma)\sum_{i=1}^n \mathcal{P}_t(x_i,y_i)\mathbb{I}\left[w_i^{(t)}=1\right]}{(1+\gamma)\sum_{i=1}^n \mathcal{P}_t(x_i,y_i)\mathbb{I}\left[w_i^{(t)}=-1\right]}\right)$.
   (c)
   - If $\alpha_t = \infty$: return T copies of h_t, $(\alpha_1 = 1,\ldots,\alpha_T = 1)$, and S_t.
   - Else: $P_{t+1}(x_i,y_i) = P_t(x_i,y_i)\exp(-\alpha_t w_i^t)/Z_t$, where $Z_t = \sum_{j=1}^n \mathcal{P}_t(x_j,y_j)\exp\left(-\alpha_t w_j^t\right)$.

Output: Functions $h_1,\ldots,h_T$, coefficients $\alpha_1,\ldots,\alpha_T$ and sets $S_1,\ldots,S_T$.

C.2.3 Generalization via Sample Compression Schemes

Sample compression scheme is a classic technique for proving generalization bounds, introduced by [LW86, FW95]. These bounds proved to be useful in numerous learning settings, such as binary classification [GHST05, MY16, BHMZ20], multiclass classification [DSBDSS15, DSS14, DMY16, BCD+22], regression [HKS18, HKS19], active learning [WHEY15], density estimation [ABDH+20], adversarially robust learning [MHS19, MHS20, MHS21, MHS22, AHM22, AH22], learning with partial concepts [AHHM22], and showing Bayes-consistency for nearest-neighbor methods [GKN14, KSW17]. As a matter of fact, compressibility and learnability are known to be equivalent for general learning problems [DMY16]. Another remarkable result by [MY16] showed that VC classes enjoy a sample compression that is independent of the sample size.

We start with a formal definition of a sample compression scheme.

Definition 16 (Sample compression scheme)**.** *A pair of functions (κ,ρ) is a sample compression scheme of size ℓ for class $\mathcal{H}$ if for any $n \in \mathbb{N}$, $h \in \mathcal{H}$ and sample $S = \{(x_i,h(x_i))\}_{i=1}^n$, it holds for the compression function that $\kappa(S) \subseteq S$ and $|\kappa(S)| \le \ell$, and the reconstruction function $\rho(\kappa(S)) = \hat{h}$ satisfies $\hat{h}(x_i) = h(x_i)$ for any $i \in [n]$.*

We show a generalization bound that scales with the sample compression size. The proof follows from [LW86].

Lemma 10 (Sample compression scheme generalization bound)**.** *Fix a margin $\gamma \in [0,1]$. For any $k \in \mathbb{N}$ and fixed function $\phi : (\mathcal{X} \times [0,1])^k \to [0,1]^{\mathcal{X}}$, for any distribution $\mathcal{D}$ over $\mathcal{X} \times [0,1]$ and any $m \in \mathbb{N}$, for $S = \{(x_i, y_i)\}_{i \in [m]}$ i.i.d. $\mathcal{D}$-distributed random variables, if there exist indices $i_1, ..., i_k \in [m]$ such that $\sum_{(x,y) \in S} \mathbb{I}\{\ell(\phi((x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k}))(x), y) > \gamma\} = 0$, then*

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}} [\mathbb{I}\{\ell(\phi((x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k}))(x), y) > \gamma\}] \leq \frac{1}{m-k}(k \log m + \log(1/\delta)) .$$

with probability at least $1 - \delta$ over S.

Proof. Let us define $\widehat{\ell}_\gamma(h; S) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{I}\{\ell(h(x), y) > \gamma\}$ and $\ell_\gamma(h; \mathcal{D}) = \mathbf{E}_{(x,y)\sim\mathcal{D}} [\mathbb{I}\{\ell(h(x), y) > \gamma\}]$. For any indices $i_1, ..., i_k \in [m]$, the probability of the bad event

$$\mathop{\mathbf{Pr}}_{S\sim\mathcal{D}^m} [\widehat{\ell}_\gamma(\phi((x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k})); S) = 0 \wedge \ell_\gamma(\phi((x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k})); \mathcal{D}) > \varepsilon]$$

is at most

$$\mathbf{E}\left[\mathbb{I}\{\ell_\gamma(\phi(\{(x_{i_j}, y_{i_j})\}_{j\in[k]}); \mathcal{D}) > \varepsilon\} \mathbf{Pr}[\widehat{\ell}_\gamma(\phi(\{(x_{i_j}, y_{i_j})\}_{j\in[k]}); S \setminus \{(x_{i_j}, y_{i_j})\}_{j\in[k]}) = 0 | \{(x_{i_j}, y_{i_j})\}_{j\in[k]}]\right] < (1-\varepsilon)^{m-k}$$

where the expectation is over $(x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k})$ and the inner probability is over $S \setminus (x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k})$. Taking a union bound over all m^k possible choices for the k indices, we get that the bad event occurs with probability at most

$$m^k \exp(-\varepsilon(m-k)) \leq \delta \Rightarrow \varepsilon = \frac{1}{m-k}(k \log m + \log(1/\delta)) .$$

□

C.3 Putting it Together

We now have all the necessary ingredients in place to prove the upper bound of Theorem 2. First, we use Lemma 8 on a sample of size $n_0 = \mathbb{D}^{\mathrm{OIG}}_\gamma(\mathcal{H})$ to obtain a learner which makes γ-errors with probability at most $1/3$[5]. Then, we use the boosting algorithm we described (see Algorithm 2) to obtain a learner that does not make any γ-mistakes on the training set. Notice that the boosting algorithm on its own does not provide any guarantees about the generalization error of the procedure. This is obtained through the sample compression result we described in Appendix C.2.3. Since we run the boosting algorithm for a few rounds on a sample whose size is small, we can provide a sample compression scheme following the approach of [DMY16, HKS19].

Lemma 11 (Upper Bound of PAC Regression)**.** *Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and $\varepsilon, \delta, \gamma \in (0,1)^3$. Then,*

$$\mathcal{M}(\mathcal{H}; \varepsilon, \delta, \gamma) \leq O\left(\frac{\mathbb{D}^{\mathrm{OIG}}_\gamma(\mathcal{H})}{\varepsilon} \log^2 \frac{\mathbb{D}^{\mathrm{OIG}}_\gamma(\mathcal{H})}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta}\right) .$$

Proof. Let n be the number of samples $S = ((x_1, y_1), \ldots, (x_n, y_n))$ that are available to the learner, $n_0 = \mathbb{D}^{\mathrm{OIG}}_\gamma(\mathcal{H})$ and let A be the algorithm obtained from Lemma 8. We have that

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}_{\mathcal{X}}^{n_0-1}} \left[\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_{\mathcal{X}}} [\ell(A(S; x), h^\star(x)) > \gamma]\right] \leq 1/3 .$$

This means that, for any distribution $\mathcal{D}_{\mathcal{X}}$ and any labeling function $h^\star \in \mathcal{H}$ we can draw a sample $S^\star = ((x_1, y_1), \ldots, (x_{n_0-1}, y_{n_0-1}))$ with non-zero probability such that

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_{\mathcal{X}}} [\ell(A(S^\star; x), h^\star(x)) > \gamma] \leq \frac{1}{3} .$$

[5] In expectation over the training set.

Notice that such a classifier is a $(\gamma, 1/6)$-weak learner (see Definition 15). Thus, by executing the MedBoost algorithm (see Algorithm 2) for $T = O(\log n)$ rounds we obtain a classifier $\hat{h} : \mathcal{X} \to \mathbb{R}$ such that, $\ell(\hat{h}(x_i), y_i) \leq \gamma, \forall i \in [n]$. We underline that the subset S_t that is used in line 1 of Algorithm 2 has size at most n_0, for all rounds $t \in [T]$. Thus, the total number of samples that is used by MedBoost is at most $O(n_0 \log n)$. Hence, following the approach of [MY16] we can encode the classifiers produced by MedBoost as a compression set that consists of $k = O(n_0 \log n)$ samples that were used to train the classifiers along with $k \log k$ extra bits that indicate their order. Thus, using generalization based on sample compression scheme as in Lemma 10, we have that with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\Big\{\ell(\hat{h}(x), y) > \gamma\Big\}\right] \leq \frac{C}{n - n_0 \log(n)} \left(n_0 \log^2 n + \log(1/\delta)\right) ,$$

which means that for large enough n,

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\Big\{\ell(\hat{h}(x), y) > \gamma\Big\}\right] \leq O\left(\frac{n_0 \log^2 n}{n} + \frac{\log(1/\delta)}{n}\right) .$$

Thus,

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\mathcal{D}}\left[\ell(\hat{h}(x), y) > \gamma\right] \leq O\left(\frac{n_0 \log^2 n}{n} + \frac{\log(1/\delta)}{n}\right) .$$

Hence, we can see that

$$\mathcal{M}(\mathcal{H}; \varepsilon, \delta, \gamma) \leq O\left(\frac{\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})}{\varepsilon} \log^2 \frac{\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta}\right) .$$

□

D γ-DS Dimension and Learnability

In this section, we will show that finiteness of γ-DS dimension is necessary for PAC learning in the realizable case.

Theorem 3. *Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}, \varepsilon, \delta, \gamma \in (0,1)^3$. Then,*

$$\mathcal{M}(\mathcal{H}; \varepsilon, \delta, \gamma) \geq \Omega\left(\frac{\mathbb{D}^{\mathrm{DS}}_{2\gamma}(\mathcal{H}) + \log(1/\delta)}{\varepsilon}\right) .$$

Proof. Let $d = \mathbb{D}^{\mathrm{OIG}}_{2\gamma}(\mathcal{H})$. Then, there exists some $S = (x_1, \ldots, x_d) \in \mathcal{X}^d$ such that $\mathcal{H}|_S$ contains a 2γ-pseudo-cube, which we call $\mathcal{H}'$. By the definition of the scaled pseudo-cube, $\forall h \in \mathcal{H}', i \in [d]$, there is exactly one $h' \in \mathcal{H}'$ such that $h(x_j) = h'(x_j), j \neq i$, and $\ell(h(x_i), h'(x_i)) > 2\gamma$. We pick the target function $h^\star$ uniformly at random among the hypotheses of $\mathcal{H}'$ and we set the marginal distribution $\mathcal{D}_{\mathcal{X}}$ of $\mathcal{D}$ as follows

$$\mathbf{Pr}[x_1] = 1 - 2\varepsilon, \quad \mathbf{Pr}[x_i] = 2\varepsilon/(d-1), \ \forall i \in \{2, ..., d\} .$$

Consider m samples $\{(z_i, h^\star(z_i)\}_{i\in[m]}$ drawn i.i.d. from $\mathcal{D}$. Let us take $m \leq \frac{d-1}{6\varepsilon}$. Then, the sample will include at most $(d-1)/2$ examples which are not x_1 with probability $1/100$, using Chernoff's bound. Let us call this event E. Conditioned on E, the posterior distribution of the unobserved points is uniform among the vertices of the $d/2$-dimensional 2γ-pseudo-cube. Thus, if the test point x falls among the unobserved points, the learner will make a γ-mistake with probability at least $1/2$. To see that, let $\widehat{y}$ be the prediction of the learner on x. Since every hyperedge has size at least 2 and all the vertices that are on the hyperedge differ by at least 2γ in the direction of x, no matter what $\widehat{y}$ is the correct label $y^\star$ is at least γ-far from it. Since $\mathbf{Pr}[E] \geq 1/100$, we can see that $\mathcal{M}_{\mathcal{A}}(\mathcal{H}; \varepsilon, \delta, \gamma) = \Omega(\frac{d}{\varepsilon})$. Moreover, by the law of total probability there must exist a deterministic choice of the target function $h^\star$, that could depend on $\mathcal{A}$, which satisfies the lower bound. For the other part of the lower bound, notices the probability that the sample will only contain x_1 is

$(1-2\varepsilon)^m \geq e^{-4\varepsilon m}$ which is greater that δ whenever $m \leq \log(1/\delta)/(4\varepsilon)$. This implies that the γ-cut-off sample complexity is lower bounded by

$$\max\left\{C_1 \cdot \frac{d}{\varepsilon}, C_2 \cdot \frac{\log(1/\delta)}{\varepsilon}\right\} = C_0 \cdot \frac{\mathbb{D}^{\mathrm{DS}}_{2\gamma}(\mathcal{H}) + \log(1/\delta)}{\varepsilon}\,.$$

Thus $\mathcal{M}_{\mathcal{A}}(\mathcal{H};\varepsilon,\delta,\gamma)$, satisfies the desired bound when the dimension is finite. Finally, it remains to claim about the case where $\mathbb{D}^{\mathrm{DS}}_{2\gamma}(\mathcal{H}) = \infty$ for the given γ. We consider a sequence of 2γ-DS shattered sets S_n with $|S_n| = n$ and repeat the claim for the finite case. This will yield that for any n the γ-cut-off sample complexity is lower bounded by $\Omega((n+\log(1/\delta))/\varepsilon)$ and this yields that $\mathcal{M}(\mathcal{H};\varepsilon,\delta,\gamma) = \infty$. □

We further conjecture that this dimension is also sufficient for PAC learning.

Conjecture 2. *A class $\mathcal{H} \subseteq (0,1)^{\mathcal{X}}$ is PAC learnable in the realizable regression setting with respect to the absolute loss function if and only if $\mathbb{D}^{\mathrm{DS}}_{\gamma}(\mathcal{H}) < \infty$ for any $\gamma \in (0,1)$.*

We believe that there must exist a modification of the approach of [BCD+22] that will be helpful in settling the above conjecture.

Conjecture 3. *There exists $\mathcal{H} \subseteq (0,1)^{\mathcal{X}}$ for which $\mathbb{D}^{\mathrm{Nat}}_{\gamma}(\mathcal{H}) = 1$ for all $\gamma \in (0,1)$ but $\mathbb{D}^{\mathrm{DS}}_{\gamma}(\mathcal{H}) < \infty$ for some $\gamma \in (0,1)$.*

In particular, we believe that one can extend the construction of [BCD+22] (which uses various tools from algebraic topology as a black-box) and obtain a hypothesis class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ that has γ-Natarajan dimension 1 but is not PAC learnable (it will have infinite γ-DS dimension). This construction though is not immediate and requires new ideas related to the works of [JŚ03, Osa13]

E Online Realizable Regression

In this section, we present our results regarding online realizable regression. The next result resolves an open question of [DG22]. It provides an online learner with optimal (off by a factor of 2) cumulative loss in realizable regression.

Theorem 4 (Optimal Cumulative Loss)**.** *Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and $\varepsilon > 0$. Then, there exists a deterministic algorithm (Algorithm 3) whose cumulative loss in the realizable setting is bounded by $\mathbb{D}^{\mathrm{onl}}(\mathcal{H}) + \varepsilon$. Conversely, for any $\varepsilon > 0$, every deterministic algorithm in the realizable setting incurs loss at least $\mathbb{D}^{\mathrm{onl}}(\mathcal{H})/2 - \varepsilon$.*

Algorithm 3 Scaled SOA

Parameters: $\{\varepsilon_t\}_{t\in\mathbb{N}}$.
Initialize $V^{(1)} = \mathcal{H}$.
For $t = 1, \ldots$:

1. Receive $x_t \in \mathcal{X}$.
2. For every $y \in [0,1]$, let $V^{(t)}_{(x_t,y)} = \left\{h \in V^{(t)} : h(x_t) = y\right\}$.
3. Let $\widehat{y}_t$ be an arbitrary label such that
$$\mathbb{D}^{\mathrm{onl}}\left(V^{(t)}_{(x_t,\widehat{y}_t)}\right) \geq \sup_{y'} \mathbb{D}^{\mathrm{onl}}\left(V^{(t)}_{(x_t,y')}\right) - \varepsilon_t\,.$$
4. Predict $\widehat{y}_t$.
5. Receive the true label $y^\star_t$ and incur loss $\ell(\widehat{y}_t, y^\star_t)$.
6. Update $V^{(t+1)} = \left\{h \in V^{(t)} : h(x_t) = y^\star_t\right\}$.

Proof. Let us begin with the upper bound. Assume that $\mathbb{D}^{\mathrm{onl}}(\mathcal{H}) < \infty$. Suppose we are predicting on the t-th point in the sequence and let $V^{(t)}$ be the version space so far, i.e., $V^{(t)} =$

$\{h \in \mathcal{H} : \forall \tau \in [t-1], h(x_\tau) = y_\tau\}$. Let x_t be the next point to predict on. For each label $y \in \mathbb{R}$, let $V^{(t)}_{(x_t,y)} = \{h \in V^{(t)} : h(x_t) = y\}$. From the definition of the dimension $\mathbb{D}^{\text{onl}}$, we know that for all $y, y' \in \mathbb{R}$ such that $V^{(t)}_{(x_t,y)}, V^{(t)}_{(x_t,y')} \neq \emptyset$,

$$\mathbb{D}^{\text{onl}}(V^t) \geq \ell(y, y') + \min\left\{\mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y)}\right), \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y')}\right)\right\}.$$

Let $\widehat{y}_t$ be an arbitrary label with $\mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,\widehat{y}_t)}\right) \geq \sup_{y'} \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y')}\right) - \varepsilon_t$, where ε_t is some sequence shrinking arbitrarily quickly in the number of rounds t. The learner predicts $\widehat{y}_t$. Assume that the adversary picks $y^\star_t$ as the true label and, so, the learner incurs loss $\ell(\widehat{y}_t, y^\star_t)$ at round t. Then, the updated version space $V^{(t)}_{(x_t,y^\star_t)}$ has

$$\mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y^\star_t)}\right) \leq \sup_{y'} \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y')}\right) \leq \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,\widehat{y}_t)}\right) + \varepsilon_t\,,$$

which implies

$$\min\left\{\mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,\widehat{y}_t)}\right), \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y^\star_t)}\right)\right\} \geq \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x,y^\star_t)}\right) - \varepsilon_t\,.$$

This gives that

$$\begin{aligned}\mathbb{D}^{\text{onl}}(V^{(t)}) &\geq \ell(\widehat{y}_t, y^*_t) + \min\left\{\mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,\widehat{y}_t)}\right), \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y^\star_t)}\right)\right\} \\ &\geq \ell(\widehat{y}_t, y^*_t) + \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y^\star_t)}\right) - \varepsilon_t\,,\end{aligned}$$

and, by re-arranging,

$$\mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y^\star_t)}\right) \leq \mathbb{D}^{\text{onl}}(V^{(t)}) - \ell(\widehat{y}_t, y^*_t) + \varepsilon_t\,. \tag{1}$$

So every round reduces the dimension by at least the magnitude of the loss (minus ε_t). Notice that $\mathbb{D}^{\text{onl}}(V^{(t+1)}) = \mathbb{D}^{\text{onl}}\left(V^{(t)}_{(x_t,y^\star_t)}\right)$. Thus, by choosing the $\{\varepsilon_t\}_{t\in\mathbb{N}}$ sequence such that

$$\sum_t \varepsilon_t \leq \varepsilon'\,,$$

and summing up Equation (1) over all $t \in \mathbb{N}$, we get a cumulative loss bound

$$\sum_t \ell(\widehat{y}_t, y^\star_t) \leq \mathbb{D}^{\text{onl}}(\mathcal{H}) + \varepsilon'\,.$$

Hence, we see that by taking the limit as ε' goes to 0 shows that the cumulative loss is upper bounded by $\mathbb{D}^{\text{onl}}(\mathcal{H})$. This analysis shows that Algorithm 3 achieves the cumulative loss bound $\mathbb{D}^{\text{onl}}(\mathcal{H}) + \varepsilon'$, for arbitrarily small $\varepsilon' > 0$.

Let us continue with the lower bound. For any $\varepsilon > 0$, we are going to prove that any deterministic learner must incur cumulative loss at least $\mathbb{D}^{\text{onl}}(\mathcal{H})/2 - \varepsilon$. By the definition of $\mathbb{D}^{\text{onl}}(\mathcal{H})$, for any $\varepsilon > 0$, there exists a tree T_ε such that, for every path $\boldsymbol{y}$,

$$\sum_{i=1}^{\infty} \gamma_{\boldsymbol{y}_{\leq i}} \geq \mathbb{D}^{\text{onl}}(\mathcal{H}) - 2\varepsilon\,,$$

i.e., the sum of the gaps across the path is at least $\mathbb{D}^{\text{onl}}(\mathcal{H}) - 2\varepsilon$. The strategy of the adversary is the following: in the first round, she presents the learner with the instance $x_1 = x_\emptyset$. Then, no matter what label $\hat{y}_1$ the learner picks, the adversary can choose the label $y^\star_1$ so that $|\hat{y}_1 - y^\star_1| \geq \gamma_\emptyset/2$. The adversary can keep picking the instances x_t based on the induced path of the choices of the true labels $\{y^\star_\tau\}_{\tau<t}$ and the loss of the learner in every round t is at least $\gamma_{\boldsymbol{y}_{\leq t}}/2$. Thus, summing up over all the rounds as $T \to \infty$, we see that the total loss of the learner is at least

$$\frac{\mathbb{D}^{\text{onl}}(\mathcal{H})}{2} - \varepsilon\,.$$

□

Remark 2 (Randomized Online Learners). *We highlight that, unlike the setting of realizable online classification, in the case of realizable online regression randomization does not seem to help the learner (see also [FHMM23]). In particular, the lower bound of* $\frac{\mathbb{D}^{\mathrm{onl}}(\mathcal{H})}{2} - \varepsilon$ *holds even for randomized learners. To see it, notice that for all distributions over* $\mathcal{D}$ *over* $[0,1]$ *it holds that*

$$\max_{c_1,c_2} \left\{ \mathop{\mathbf{E}}_{X\sim\mathcal{D}}[\ell(X,c_1)], \mathop{\mathbf{E}}_{X\sim\mathcal{D}}[\ell(X,c_2)] \right\} \geq \ell(c_1,c_2)/2\,.$$

Example 2 (Sequential Complexity Measures). *Sequential fat-shattering dimension and sequential covering numbers are two standard combinatorial measures for regression in online settings [RST15b, RST15a]. Note that Example 1 can be learned with 1 sample even in the online realizable setting. Hence, Example 1 shows that sequential fat-shattering dimension fails to characterize online realizable regression (since this dimension is at least as large as fat-shattering dimension which is infinite in this example). Moreover, we know that sequential covering numbers and sequential fat-shattering dimension are of the same order of magnitude and so they are also infinite in the case of Example 1.*

F Dimension and Finite Character Property

[BDHM+19] gave a formal definition of the notion of "dimension" or "complexity measure", that all previously proposed dimensions in statistical learning theory comply with. In addition to characterizing learnability, a dimension should satisfy the finite character property:

Definition 17 (Finite Character [BDHM+19]). *A dimension characterizing learnability can be abstracted as a function* F *that maps a class* $\mathcal{H}$ *to* $\mathbb{N}\cup\{\infty\}$ *and satisfies the finite character property: for every* $d\in\mathbb{N}$ *and* $\mathcal{H}$*, the statement "*$F(\mathcal{H})\geq d$*" can be demonstrated by a finite set* $X\subseteq\mathcal{X}$ *of domain points, and a finite set of hypotheses* $H\subseteq\mathcal{H}$*. That is, "*$F(\mathcal{H})\geq d$*" is equivalent to the existence of a bounded first order formula* $\phi(\mathcal{X},\mathcal{H})$ *in which all the quantifiers are of the form:* $\exists x\in\mathcal{X},\forall x\in\mathcal{X}$ *or* $\exists h\in\mathcal{H},\forall h\in\mathcal{H}$*.*

Claim 1. *The scaled one-inclusion graph dimension* $\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})$ *satisfies the finite character property.*

Proof. To demonstrate that $\mathbb{D}^{\mathrm{OIG}}(\mathcal{H})\geq d$, it suffices to find a set S of n domain points and present a finite subgraph $G=(V,E)$ of the one-inclusion hypergraph induced by S where every orientation $\sigma:E\to V$ has out-degree at least $n/3$. Note that V is, by definition, a finite collection of datasets realizable by $\mathcal{H}$ and so this means that we can demonstrate that $\mathbb{D}^{\mathrm{OIG}}(\mathcal{H})\geq d$ with a finite set of domain points and a finite set of hypotheses. □

G Examples for Scaled Graph Dimension

These examples are adaptations from [DSS14, DSBDSS15].

Example 3 (Large Gap Between ERM Learners). *For every* $d\in\mathbb{N}$*, consider a domain* $\mathcal{X}_d$ *such that* $|\mathcal{X}_d|=d$ *and* $\mathcal{X}_d,\mathcal{X}_{d'}$ *are disjoint for* $d\neq d'$*. For all* $d\in\mathbb{N}$*, let* $P(\mathcal{X}_d)$ *denote the collection of all finite and co-finite*[6] *subsets of* $\mathcal{X}_d$*. Let us fix* $\gamma\in(0,1)$*. Consider a mapping* $f:\cup_{d\in\mathbb{N}}P(\mathcal{X}_d)\to[0,1]$ *such that* $f(A_d)\in(\gamma,1)$ *for all* $d\in\mathbb{N}, A_d\in P(\mathcal{X}_d)$*, and* $f(A_d)\neq f(A'_{d'})$ *for all* $A_d\neq A'_{d'}, A_d\in P(\mathcal{X}_d), A_{d'}\in P(\mathcal{X}_{d'})$*. Such a mapping exists due to the density of the reals. For any* $d\in\mathbb{N}, A_d\subseteq\mathcal{X}_d$*, let* $h_{A_d}(x)=f(A_d)\cdot\mathbb{1}\{x\in A_d\}$ *and consider the scaled first Cantor class* $\mathcal{H}_{\mathcal{X}_d,\gamma}=\{h_{A_d}:A_d\in P(\mathcal{X}_d)\}$*. We claim that* $\mathbb{D}^{\mathrm{Nat}}_{\gamma}(\mathcal{H}_{\mathcal{X}_d,\gamma})=1$ *and that* $\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}_{\mathcal{X}_d,\gamma})=|\mathcal{X}_d|=d$ *since one can use* $f_{\emptyset}$ *for the* γ*-graph shattering. Consider the following two ERM learners for the scaled first Cantor class* $\mathcal{H}_{\mathcal{X}_d,\gamma}$*:*

1. *Whenever a sample of the form* $S=\{(x_i,0)\}_{i\in[n]}$ *is observed, the first algorithm outputs* $h_{\cup\{x_i\}^c_{i\in[n]}}$ *which minimizes the empirical error. If the sample contains a non-zero element, the ERM learner identifies the correct hypothesis. The sample complexity of PAC learning is* $\Omega(d)$*.*
2. *The second algorithm either returns the all-zero function or identifies the correct hypothesis if the sample contains a non-zero label. This is a good ERM learner* $\mathcal{A}^{ERM}_{good}$ *with sample complexity* $m(\varepsilon,\delta)=\frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)$*.*

[6] A set $S\subseteq\mathcal{X}_d$ is co-finite if its complement S^c is finite.

The construction that illustrates the poor performance of the first learner is exactly the same as in the proof of the lower bound of Theorem 1. The second part of the example is formally shown in Claim 2, which follows.

Claim 2 (Good ERM Learner)**.** *Let $\varepsilon, \delta \in (0,1)^2$. Then, the good ERM learner of Example 3 has sample complexity $\mathcal{M}(\varepsilon, \delta) = \frac{1}{\varepsilon} \log\left(\frac{1}{\delta}\right)$.*

Proof. Let $d \in \mathbb{N}$, $\mathcal{D}_{\mathcal{X}_d}$ be a distribution over $\mathcal{X}_d$ and $h_{A_d^\star}$ be the labeling function. Consider a sample S of length m. If the learner observes a value that is different from 0 among the labels in S, then it will be able to infer $h_{A_d^\star}$ and incur 0 error. On the other hand, if the learner returns the all zero function its error can be bounded as

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}_{\mathcal{X}_d}} \left[\ell(h_\emptyset(x), h_{A_d^\star}(x))\right] \leq \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_{\mathcal{X}_d}} \left[x \in A_d^\star\right].$$

Since in all the $m = \frac{1}{\varepsilon} \log\left(\frac{1}{\delta}\right)$ draws of the training set S there were no elements from $A_d^\star$ we can see that, with probability at least $1 - \delta$ over the draws of S it holds that

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_{\mathcal{X}_d}} \left[x \in A_d^\star\right] \leq \varepsilon\,.$$

Thus, the algorithm satisfies the desired guarantees. □

The next example shows that no proper algorithm can be optimal in the realizable regression setting.

Example 4 (No Optimal PAC Learner Can be Proper)**.** *Let $\mathcal{X}_d$ contain d elements and let $\gamma \in (0,1)$. Consider the subclass of the scaled first Cantor class (see Example 3) with $\mathcal{H}'_{d,\gamma} = \{h_A : A \in P(\mathcal{X}_d), |A| = \lfloor d/2 \rfloor\}$. First, since this class is contained in the scaled first Cantor class, we can employ the good ERM and learn it. However, this learner is improper since $h_\emptyset \notin \mathcal{H}'_{d,\gamma}$. Then, no proper algorithm is able to PAC learn $\mathcal{H}'_{d,\gamma}$ using $o(d)$ examples.*

Proof. Suppose that an adversary chooses $h_A \in \mathcal{H}'_{d,\gamma}$ uniformly at random and consider the distribution on $\mathcal{X}_d$ which is uniform on the complement of A, where $|A| = O(d)$. Note that the error of every hypothesis $h_B \in \mathcal{H}'_{d,\gamma}$ is at least $\gamma |B \setminus A|/d$. Therefore, to return a hypothesis with small error, the algorithm must recover a set that is almost disjoint from A and so recover A. However the size of A implies that it cannot be done with $o(d)$ examples.

Formally, fix $x_0 \in \mathcal{X}_d$ and $\varepsilon \in (0,1)$. Let $A \subseteq \mathcal{X}_d \setminus \{x_0\}$ of size $d/2$. Let $\mathcal{D}_A$ be a distribution with mass $\mathcal{D}_A((x_0, h_A(x_0))) = 1 - 16\varepsilon$ and is uniform on the points $\{(x, h_A(x)) : x \in A^c\}$, where A^c is the complement of A (without x_0).

Consider a proper learning algorithm $\mathcal{A}$. We will show that there is some algorithm-dependent set A, so that when $\mathcal{A}$ is run on $\mathcal{D}_A$ with $m = O(d/\varepsilon)$, it outputs a hypothesis with error at least γ with constant probability.

Pick A uniformly at random from all sets of size $d/2$ of $\mathcal{X}_d \setminus \{x_0\}$. Let Z be the random variable that counts the number of samples in the m draws from $\mathcal{D}_A$ that are not $(x_0, h_A(x_0))$. Standard concentration bounds imply that with probability at least $1/2$, the number of points from $(\mathcal{X}_d \setminus \{x_0\}) \setminus A$ is at most $d/4$. Conditioning on this event, A is a uniformly chosen random set of size $d/2$ that is chosen uniformly from all subsets of a set $\mathcal{X}' \subset \mathcal{X}_d$ with $|\mathcal{X}'| \geq 3d/4$ (these points are not present in the sample). Now assume that the learner returns a hypothesis h_B, where B is a subset of size $d/2$. Note that $\mathbf{E}[|B \setminus A|] \geq d/6$. Hence there exists a set A such that with probability $1/2$, it holds that $|B \setminus A| \geq d/6$. This means that $\mathcal{A}$ incurs a loss of at least γ on all points in $B \setminus A$ and the mass of each such point is $\Omega(\varepsilon/d)$. Hence, in total, the learner will incur a loss of order $\gamma \cdot \varepsilon$. □

H Extension to More General Loss Functions

Our results can be extended to loss functions that satisfy approximate pseudo-metric axioms (see e.g., [HKLM22, CKW08]). The main difference from metric losses is that we allow an approximate triangle inequality instead of a strict inequality. Many natural loss functions are captured by this definition, such as the well-studied ℓ_p losses for the regression setting. Abstractly, in this context, the

label space[7] is an abstract non-empty set $\mathcal{Y}$, equipped with a general loss function $\ell : \mathcal{Y}^2 \to \mathbb{R}_{\geq 0}$ satisfying the following property.

Definition 18 (Approximate Pseudo-Metric)**.** *For $c \geq 1$, a loss function $\ell : \mathcal{Y}^2 \to \mathbb{R}_{\geq 0}$ is c-approximate pseudo-metric if (i) $\ell(x, x) = 0$ for any $x \in \mathcal{Y}$, (ii) $\ell(x, y) = \ell(y, x)$ for any $x, y \in \mathcal{Y}$, and, (iii) ℓ satisfies a c-approximate triangle inequality; for any $y_1, y_2, y_3 \in \mathcal{Y}$, it holds that $\ell(y_1, y_2) \leq c(\ell(y_1, y_3) + \ell(y_2, y_3))$.*

Furthermore, note that all dimensions for $\mathcal{H}$, $\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H})$, $\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H})$, $\mathbb{D}^{\mathrm{DS}}_{\gamma}(\mathcal{H})$, and $\mathbb{D}^{\mathrm{onl}}_{\gamma}(\mathcal{H})$ are defined for loss functions satisfying Definition 18.

Next, we provide extensions of our main results for approximate pseudo-metric losses and provide proof sketches for the extensions.

ERM Learnability for Approximate Pseudo-Metrics. For ERM learnability and losses satisfying Definition 18, we can obtain the next result.

Theorem 5. *Let ℓ be a loss function satisfying Definition 18. Then for every class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $\mathcal{H}$ is learnable by any ERM in the realizable PAC regression setting under ℓ if and only if $\mathbb{D}^{\mathrm{G}}_{\gamma}(\mathcal{H}) < \infty$ for all $\gamma \in (0, 1)$.*

The proof of the upper bound and the lower bound follow in the exact same way as with the absolute loss.

PAC Learnability for Approximate Pseudo-Metrics. As for PAC learning with approximate pseudo-metric losses, we can derive the next statement.

Theorem 6. *Let ℓ be a loss function satisfying Definition 18. Then every class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is PAC learnable in the realizable PAC regression setting under ℓ if and only if $\mathbb{D}^{\mathrm{OIG}}_{\gamma}(\mathcal{H}) < \infty$ for any $\gamma \in (0, 1)$.*

Proof Sketch. We can generalize the upper bound in Theorem 2 for the scaled OIG dimension as follows. One of the ingredients of the proof for the absolute loss is to construct a sample compression scheme through the median boosting algorithm (cf. Algorithm 2). While the multiplicative update rule is defined for any loss function, the median aggregation is no longer the right aggregation for arbitrary (approximate) pseudo-metrics. However, for each such loss function, there exists an aggregation such that the output value of the ensemble is within some cutoff value from the true label for each example in the training set, which means that we have a sample compression scheme for some cutoff loss. In particular, we show that by using weak learners with cutoff parameter $\gamma/(2c)$, where c is the approximation level of the triangle inequality, the aggregation of the base learners can be expressed as a sample compression scheme for cutoff loss with parameter γ.

Indeed, running the boosting algorithm with $(\gamma/(2c), 1/6)$-weak learners yields a set $h_1, \ldots, h_N$ of weak predictors, with the property that for each training example (x, y), at least $2/3$ of the functions h_i (as weighted by coefficients α_i), $1 \leq i \leq N$, satisfy $\ell(h_i(x), y) \leq \gamma/(2c)$. For any x, let $\hat{h}(x)$ be a value in $\mathcal{Y}$ such that at least $2/3$ of h_i (as weighted by α_i), $1 \leq i \leq N$, satisfy $\ell(h_i(x), \hat{h}(x)) \leq \gamma/(2c)$, if such a value exists, and otherwise $\hat{h}(x)$ is an arbitrary value in $\mathcal{Y}$. In particular, note that on the training examples (x, y), the label y satisfies this property, and hence $\hat{h}(x)$ is defined by the first case. Thus, for any training example, there exists h_i (indeed, at least $2/3$ of them) such that *both* $\ell(h_i(x), y) \leq \gamma/(2c)$ and $\ell(h_i(x), \hat{h}(x)) \leq \gamma/(2c)$ are satisfied, and therefore we have

$$\ell(\hat{h}(x), y) \leq c(\ell(\hat{h}(x), h_i(x)) + \ell(h_i(x), y)) \leq \gamma.$$

This function $\hat{h}$ can be expressed as a sample compression scheme of size $O\left(\mathbb{D}^{\mathrm{OIG}}_{\gamma/(2c)}(\mathcal{H}) \log(m)\right)$ for cutoff loss with parameter γ: namely, it is purely defined by the h_i functions, where each h_i is

[7] We would like to mention that, in general, we do not require that the label space is bounded. In contrast, we have to assume that the loss function takes values in a bounded space. This is actually necessary since having an unbounded loss in the regression task would potentially make the learning task impossible. For instance, having some fixed accuracy goal, one could construct a learning instance (distribution over labeled examples) that would make estimation with that level of accuracy trivially impossible.

specified by $O\left(\mathbb{D}^{\mathrm{OIG}}_{\gamma/(2c)}(\mathcal{H})\right)$ training examples, and we have $N = O(\log(m))$ such functions, and $\hat{h}$ satisfies $\ell(\hat{h}(x), y) \leq \gamma$ for all m training examples (x, y). Thus, by standard generalization bounds for sample compression, we get an upper bound that scales with $\tilde{O}\left(\mathbb{D}^{\mathrm{OIG}}_{\gamma/(2c)}(\mathcal{H})\frac{1}{m}\right)$ for the cutoff loss with parameter γ, and hence by Markov's inequality, an upper bound

$$\mathbf{E}[\ell(\hat{h}(x), y)] = \tilde{O}\left(\mathbb{D}^{\mathrm{OIG}}_{\gamma/(2c)}(\mathcal{H})\frac{1}{m\gamma}\right).$$

We next deal with the lower bound. For the absolute loss, we scale the dimension by 2γ instead of γ since for any two possible labels y_1, y_2 the learner can predict some intermediate point, and we want to make sure that the prediction will be either γ far from y_1 or y_2. For an approximate pseudo-metric, we should take instead $2c\gamma$ in order to ensure that the prediction is γ far, which means that the lower bounds in Theorem 2 hold with a scale of $2c\gamma$. □

Online Learnability for Approximate Pseudo-Metrics. Finally, we present the more general statement for online learning.

Theorem 7. *Let ℓ be a loss function satisfying Definition 18 with parameter $c \geq 1$. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $\varepsilon > 0$. Then, there exists a deterministic algorithm whose cumulative loss in the realizable setting is bounded by $\mathbb{D}^{\mathrm{onl}}(\mathcal{H}) + \varepsilon$. Conversely, for any $\varepsilon > 0$, every deterministic algorithm in the realizable setting incurs loss at least $\mathbb{D}^{\mathrm{onl}}(\mathcal{H})/(2c) - \varepsilon$.*

Proof Sketch. The upper bound of Theorem 4 works for any loss function. Recall the proof idea; in every round t there is some $\widehat{y}_t \in \mathcal{Y}$ the learner can predict such that no matter what the adversary picks as the true label $y^\star_t$, the online dimension of the version space at round t, i.e, $V = \{h \in \mathcal{H} : h(x_\tau) = y^\star_\tau, 1 \leq \tau \leq t\}$, decreases by $\ell(y^\star_t, \widehat{y}_t)$, minus some shrinking number ϵ_t that we can choose as a parameter. Therefore we get that the sum of losses is bounded by the online dimension and the sum of ϵ_t that we can choose to be arbitrarily small.

The lower bound for online learning in Theorem 4 would be $\mathbb{D}^{\mathrm{onl}}(\mathcal{H})/(2c) - \varepsilon$, for any $\epsilon > 0$, since the adversary can force a loss of $\gamma_{\boldsymbol{y}\leq t}/(2c)$ in every round t, where $\gamma_{\boldsymbol{y}\leq t}$ is the sum of the gaps across the path $\boldsymbol{y}$.

□