Adversarially Robust PAC Learnability of Real-Valued Functions

Idan Attias [1] **Steve Hanneke** [2]

Abstract

We study robustness to test-time adversarial attacks in the regression setting with ℓ_p losses and arbitrary perturbation sets. We address the question of which function classes are PAC learnable in this setting. We show that classes of finite fat-shattering dimension are learnable in both realizable and agnostic settings. Moreover, for convex function classes, they are even properly learnable. In contrast, some non-convex function classes provably require improper learning algorithms. Our main technique is based on a construction of an adversarially robust sample compression scheme of a size determined by the fat-shattering dimension. Along the way, we introduce a novel agnostic sample compression scheme for real-valued functions, which may be of independent interest.

1. Introduction

Learning a predictor that is resilient to test-time adversarial attacks is a fundamental problem in contemporary machine learning. A long line of research has studied the vulnerability of deep learning-based models to small perturbations of their inputs (e.g., Szegedy et al. (2013); Biggio et al. (2013); Goodfellow et al. (2014); Madry et al. (2017)). From the theoretical standpoint, there has been a lot of effort to provide provable guarantees of such methods (e.g., Feige et al. (2015); Schmidt et al. (2018); Khim & Loh (2018); Yin et al. (2019); Cullina et al. (2018); Attias et al. (2019; 2022); Montasser et al. (2021b; 2020a;b; 2021a); Ashtiani et al. (2020); Dan et al. (2020); Awasthi et al. (2020; 2021b;a; 2022a;b; 2023); Bhattacharjee et al. (2021); Xing et al. (2021); Mao et al. (2023)), which is the focus of this work.

In the robust PAC learning framework, the problem of learning binary function classes was studied by Montasser et al. (2019). They showed that uniform convergence does not hold in this setting, and as a result, robust empirical risk minimization is not sufficient to ensure learnability. Yet, they showed that VC classes are learnable, by considering an improper learning rule; the learning algorithm outputs a function that is not in the function class that we aim to learn.

In this work, we provide a theoretical understanding of the robustness of real-valued predictors in the PAC learning model, with arbitrary perturbation sets. The work of Attias et al. (2019) considered this question for finite perturbation sets, they obtained sample complexity guarantees based on uniform convergence, which is no longer true for arbitrary perturbation sets. We address the fundamental question, *which real-valued function classes are robustly learnable?*

Furthermore, we study the robust learnability of convex classes, a natural and commonly studied subcategory for regression. We address the question, *are real-valued convex classes properly robustly learnable?* On the one hand, some non-convex function classes provably require improper learning due to Montasser et al. (2019). On the other hand, Mendelson (2019) showed that non-robust regression with the mean squared error is properly learnable.

We study the following learning models for real-valued functions. An adversarial attack is formalized by a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, where $\mathcal{U}(x)$ is the set of possible perturbations (attacks) on x. In practice, we usually consider $\mathcal{U}(x)$ to be the ℓ_1 ball centered at x. In this work, we have no restriction on $\mathcal{U}$, besides $x \in \mathcal{U}(x)$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X} \times [0,1]$ and let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ be a concept class. In our first model, the robust error of concept h is defined as

$$\mathrm{Err}_{\ell_p}(h;\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{z\in\mathcal{U}(x)} |h(z)-y|^p\right], 1 \le p < \infty.$$

The learner gets an i.i.d. sample from $\mathcal{D}$, and would like to output function $\hat{h}$, such that with high probability,

$$\mathrm{Err}_{\ell_p}(\hat{h};\mathcal{D}) \le \inf_{h\in\mathcal{H}} \mathrm{Err}_{\ell_p}(h;\mathcal{D}) + \epsilon. \tag{1}$$

The sample complexity for learning $\mathcal{H}$ is the minimal i.i.d. sample from $\mathcal{D}$ such that there exists a learning algorithm

[1]Department of Computer Science, Ben-Gurion University, Israel [2]Department of Computer Science, Purdue University, USA. Correspondence to: Idan Attias <idanatti@post.bgu.ac.il>, Steve Hanneke <steve.hanneke@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

with output as in Equation (1). We refer to this model as *Robust Regression* with ℓ_p robust losses. This is a robust formulation of the classic nonparametric regression setting.

In the second model, the robust error of concept h is defined as

$$\mathrm{Err}_\eta(h;\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\left\{\sup_{z\in\mathcal{U}(x)} |h(z)-y| \geq \eta\right\}\right].$$

We refer to the loss function in this model as *cutoff loss*, where $\eta \geq 0$ is a predefined cutoff parameter. The learner gets an i.i.d. sample from $\mathcal{D}$, and would like to output function $\hat{h}$, such that with high probability,

$$\mathrm{Err}_{\eta+\beta}(\hat{h};\mathcal{D}) \leq \inf_{h\in\mathcal{H}} \mathrm{Err}_\eta(h;\mathcal{D}) + \epsilon,$$

where $\beta > 0$ is a predefined parameter. The sample complexity is defined similarly to the previous model. We refer to this model as *Robust* (η,β)*-Regression*. The non-robust formulation of this setting was studied by, e.g., Anthony & Bartlett (2000); Simon (1997). See also Anthony et al. (1999, section 21.4) and references therein.

Main results and roadmap. Denote the γ-fat-shattering dimension of $\mathcal{H}$ by $\mathrm{fat}(\mathcal{H},\gamma)$, and the dual γ-fat-shattering dimension by $\mathrm{fat}^*(\mathcal{H},\gamma)$, which is the dimension of the dual class. The dimension of the dual class is finite as long as the γ-fat-shattering of the primal class is finite (see Kleer & Simon (2021), and Equation (7)).

- In Section 3 we provide a learning algorithm for robust regression with ℓ_p losses, with sample complexity

  $$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}^3(\mathcal{H},\epsilon/p)\,\mathrm{fat}^*(\mathcal{H},\epsilon/p)}{\epsilon^5}\right).$$

  Moreover, this algorithm is *proper* for convex function classes. We circumvent a negative result regarding non-convex function classes, for which proper learning is impossible, even for binary-valued functions (Montasser et al., 2019).

- In Section 4 we provide a learning algorithm with a substantial sample complexity improvement for the ℓ_1 loss,

  $$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H},\epsilon)\,\mathrm{fat}^*(\mathcal{H},\epsilon)}{\epsilon^2}\right).$$

- In Section 5, we provide learning algorithms for the (η,β)-robust regression setting in the realizable and agnostic settings. Our sample complexity for the realizable case is

  $$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H},\beta)\,\mathrm{fat}^*(\mathcal{H},\beta)}{\epsilon}\right),$$

  and

  $$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H},\beta)\,\mathrm{fat}^*(\mathcal{H},\beta)}{\epsilon^2}\right)$$

  for the agnostic case.

Technical contributions and related work. The setting of agnostic adversarially robust regression with finite perturbation sets was studied by Attias et al. (2019). Subsequently, improved bounds appeared in Kontorovich & Attias (2021). Adversarially robust PAC learnability of binary-valued function classes with arbitrary perturbation sets was studied by Montasser et al. (2019). They showed that uniform convergence does not hold in this setting, which means that some classes provably require improper learning. Their main technique is constructing a sample compression scheme from a boosting-style algorithm, where the generalization follows from sample compression bounds.

First, we explain our new technical ideas behind the algorithms for robust (η,β)-regression, and compare it to the ones of Montasser et al. (2019) in the classification setting. We then explain why the approach for learning these models fails in the general robust regression setting and introduce the new ingredients behind the proofs for this setting.

Robust (η,β)*-regression.* We construct an adversarially robust sample compression scheme of a size determined by the fat-shattering dimension of the function class. The following steps are different from the binary-valued case. First, we use a modified *boosting algorithm for real-valued functions*. In the non-robust setting, Hanneke et al. (2019) showed how to convert a boosting algorithm (originally introduced by Kégl (2003)), into a sample compression scheme. In order to find weak learners (and prove their existence), we rely on *generalization from approximate interpolation* (see Anthony & Bartlett (2000) and Anthony & Bartlett (2000, section 21.4)). The idea is that any function $f \in \mathcal{F}$ that approximately interpolates a sample $S \sim \mathcal{D}^m$, that is, $|f(x)-y| \leq \eta$ for $(x,y) \in S$, also satisfies that $\mathcal{P}\{(x,y) : |f(x)-y| \leq \eta+\beta\} > 1-\epsilon$ with high probability, as long as $\tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{F},\beta)/\epsilon) \leq |S|$. Crucially, this result relies on uniform convergence and does not apply to the robust loss function. Another difference is in the discretization step. In the classification setting, we inflate the data set to include all possible perturbations (potentially infinite set). We then define a function class $\hat{\mathcal{H}}$ by running a robust empirical minimizer on every subset of size $\mathrm{VC}(\mathcal{H})$ from the training set, where $\mathcal{H}$ is the class we want to learn. $\hat{\mathcal{H}}$ induces a finite partition on the inflated set into regions, such that any $h \in \hat{\mathcal{H}}$ has a constant error in each region. This is no longer true in the real-valued case. Instead, we discretize the inflated set by taking a *uniform cover* using the supremum metric, and controlling the errors that arise from the cover.

Robust regression. We first explain which natural techniques fail. We cannot run boosting for the ℓ_p loss as explained by Hanneke et al. (2019): "Duffy & Helmbold (2002, Remark 2.1) spell out a central technical challenge: no boosting algorithm can always force the base regressor to output a useful function by simply modifying the distribution over the sample. This is because unlike a binary classifier, which localizes errors on specific examples, a real-valued hypothesis can spread its error evenly over the entire sample and it will not be affected by reweighting".

As a first attempt, we could try to learn with respect to the cutoff loss (with a fixed cutoff parameter), and conclude learnability in the general regression setting. However, the ℓ_p loss can spread over different values for different points, which means that this approach fails. In another possible attempt, we could try to solve the realizable case first and try to reduce agnostic to realizable learning as in Montasser et al. (2019) for binary-valued functions, as we prove the agnostic setting for robust (η, β)-regression. However, this attempt fails for the same reasons we mentioned above.

Therefore, we introduce a novel technique for working with *changing* cutoffs. We establish generalization from approximate interpolation with *different* cutoff parameters, and thereby, we find a learner that approximates the loss of the target function on different points. Utilizing this idea, we provide a learning algorithm for ℓ_p robust losses that constructs an ensemble and predict with the average. Further, we show that this algorithm is *proper* for convex function classes. In contrast, some non-convex function classes provably require improper learning (Montasser et al., 2019). Moreover, we show how to reduce the sample complexity substantially for the ℓ_1 robust loss with a different algorithm, by constructing an ensemble of weak learners and predicting with the median. Both algorithms can be represented as an agnostic sample compression scheme for the robust loss. This is a new result since constructing a sample compression scheme for real-valued functions is known only for the realizable setting (Hanneke et al., 2019). We believe that this technique may be of independent interest.

2. Problem Setup and Preliminaries

Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ be a concept class. We implicitly assume that all concept classes are satisfying mild measure-theoretic conditions (see e.g., Dudley (1984, section 10.3.1) and Pollard (2012, appendix C)). Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = [0,1]$. Define a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ that maps an input to an arbitrary set $\mathcal{U}(x) \subseteq \mathcal{X}$, such that $x \in \mathcal{U}(x)$.

We consider the following loss functions. For $1 \leq p < \infty$, define the ℓ_p *robust loss* function of h on (x, y) with respect to a perturbation function $\mathcal{U}$,

$$\ell_{p,\mathcal{U}}(h;(x,y)) = \sup_{z \in \mathcal{U}(x)} |h(z) - y|^p \,. \tag{2}$$

We define also the η*-ball robust loss* function of h on (x, y) with respect to a perturbation function $\mathcal{U}$,

$$\ell^{\eta}_{\mathcal{U}}(h;(x,y)) = \mathbb{I}\left\{ \sup_{z \in \mathcal{U}(x)} |h(z) - y| \geq \eta \right\}. \tag{3}$$

The non-robust version of this loss function is also known as η-ball or η-tube loss (see for example Anthony et al. (1999, Section 21.4)).

Define the error of a function h on distribution $\mathcal{D}$, with respect to the ℓ_p robust loss,

$$\mathrm{Err}_{\ell_p}(h;\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{z \in \mathcal{U}(x)} |h(z) - y|^p \right],$$

and the error with respect to the η-ball robust loss

$$\mathrm{Err}_{\eta}(h;\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\left\{ \sup_{z \in \mathcal{U}(x)} |h(z) - y| \geq \eta \right\} \right].$$

Note that in our model the learner is tested on the original label y while observing only the perturbed example z. There are formulations of robustness where the learner is compared to the value of the optimal function in the class on the perturbed example, i.e., if the optimal function in the class is $h^\star$, then the ℓ_1 robust loss would be $\sup_{z \in \mathcal{U}(x)} |h(z) - h^\star(z)|$. For more details and comparisons of the two models, see Gourdeau et al. (2021); Diochnos et al. (2018); Bubeck et al. (2019).

Learning models. We precisely define the models for robustly learning real-valued functions. Our first model is learning with the ℓ_p robust losses (see Equation (2)), we refer to this model as *Robust Regression*.

Definition 2.1 (Robust regression)**.** For any $\epsilon, \delta \in (0,1)$, the sample complexity robust (ϵ, δ)-PAC learning a concept class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ with ℓ_p robust losses, denoted by $\mathcal{M}(\epsilon, \delta, \mathcal{H}, \mathcal{U}, \ell_p)$, is the smallest integer m such that the following holds: there exists a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \to [0,1]^{\mathcal{X}}$, such that for any distribution $\mathcal{D}$ over $\mathcal{X} \times [0,1]$, for an i.i.d. random sample $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$ over S, it holds that

$$\mathrm{Err}_{\ell_p}(\mathcal{A}(S);\mathcal{D}) \leq \inf_{h \in \mathcal{H}} \mathrm{Err}_{\ell_p}(h;\mathcal{D}) + \epsilon.$$

If no such m exists, define $\mathcal{M}(\epsilon, \delta, \mathcal{H}, \mathcal{U}, \ell_p) = \infty$, and $\mathcal{H}$ is not robustly (ϵ, δ)-PAC learnable. We use the shorthand $\mathcal{M} = \mathcal{M}(\epsilon, \delta, \mathcal{H}, \mathcal{U}, \ell_p)$ for notational simplicity.

Our second model is learning with the η-ball robust loss (see Equation (3)) in the realizable and agnostic settings, we refer to this model by *Robust* (η, β)*-regression*. We say that a distribution $\mathcal{D}$ is η*-uniformly realizable* with respect to $\mathcal{H}$ and $\mathcal{U}$, if there exists $h^\star \in \mathcal{H}$ such that

$$\mathrm{Err}_\eta(h^\star; \mathcal{D}) = 0. \tag{4}$$

Definition 2.2 (Robust (η, β)-regression)**.** For any $\beta, \epsilon, \delta \in (0, 1)$ and $\eta \in [0, 1]$, the sample complexity of *realizable* robust $(\eta, \beta, \epsilon, \delta)$-PAC learning a concept class $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$, denoted by $\mathcal{M}_{\mathrm{RE}}(\eta, \beta, \epsilon, \delta, \mathcal{H}, \mathcal{U})$, is the smallest integer m such that the following holds: there exists a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \to [0, 1]^{\mathcal{X}}$, such that for any distribution $\mathcal{D}$ over $\mathcal{X} \times [0, 1]$ that is η-uniformly realizable w.r.t. $\mathcal{H}$ and $\mathcal{U}$ (see Equation (4)), for an i.i.d. random sample $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$ over S, it holds that

$$\mathrm{Err}_{\eta+\beta}(\mathcal{A}(S); \mathcal{D}) \leq \epsilon.$$

If no such m exists, define $\mathcal{M}_{\mathrm{RE}}(\eta, \epsilon, \delta, \mathcal{H}, \mathcal{U}) = \infty$, and $\mathcal{H}$ is not robustly $(\eta, \beta, \epsilon, \delta)$-PAC learnable.

The *agnostic* sample complexity, denoted by $\mathcal{M}_{\mathrm{AG}}(\eta, \beta, \epsilon, \delta, \mathcal{H}, \mathcal{U})$, is defined similarly with the following difference. We require the learning algorithm to output a function, such that with probability at least $1 - \delta$,

$$\mathrm{Err}_{\eta+\beta}(\mathcal{A}(S); \mathcal{D}) \leq \inf_{h \in \mathcal{H}} \mathrm{Err}_\eta(h; \mathcal{D}) + \epsilon.$$

We use the shorthand $\mathcal{M}_{\mathrm{RE}}^{\eta,\beta} = \mathcal{M}_{\mathrm{RE}}(\eta, \beta, \epsilon, \delta, \mathcal{H}, \mathcal{U})$ and $\mathcal{M}_{\mathrm{AG}}^{\eta,\beta} = \mathcal{M}_{\mathrm{AG}}(\eta, \beta, \epsilon, \delta, \mathcal{H}, \mathcal{U})$ for notational simplicity.

We do not define the setting of robust regression in the realizable setting since it coincides with the realizable setting of robust (η, β)-regression, by taking $\eta = 0, \beta = \epsilon/2$, and rescaling ϵ to $\epsilon/2$. Moreover, we could define the ℓ_p variant of the η-ball loss in robust (η, β)-regression, however, results for our definition translate immediately by taking $\eta^{1/p}$.

Note that there is a fundamental difference between the models. In the robust (η, β)-regression, we demand from the learning algorithm to find a function that is almost everywhere within $\eta + \beta$ from the target function in class. That is, on $1 - \epsilon$ mass of elements in the support of $\mathcal{D}$, we find an approximation up to $\eta + \beta$. On the other hand, in the robust regression model, we aim to be close to the target function on average, and the error can possibly spread across all elements in the support.

Proper and improper learning algorithms. The learning algorithm is not limited to returning a function that is inside the concept class that we aim to learn. When learning a class $\mathcal{H}$, whenever the learning algorithm returns a function inside the class, that is, $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$, we say that the algorithm is proper and the class in properly learnable. Otherwise, we say that the algorithm is improper. Improper algorithms are extremely powerful and using them often circumvents computational issues and sample complexity barriers (Srebro et al., 2004; Candes & Recht, 2012; Anava et al., 2013; Hazan et al., 2015; Hanneke, 2016; Hazan & Ma, 2016; Hazan et al., 2012; Agarwal et al., 2019; Daniely et al., 2011; Daniely & Shalev-Shwartz, 2014; Angluin, 1988; Montasser et al., 2019).

Oracles. We make use of the following robust empirical risk minimizers. Let a set $S = \{(x_i, y_i)\}_{i=1}^m$. Define an η-robust empirical risk minimizer $\eta\text{-RERM}_{\mathcal{H}} : (\mathcal{X} \times \mathcal{Y})^m \times [0, 1]^m \times \mathbb{N} \to \mathcal{H}$,

$$\eta\text{-RERM}_{\mathcal{H}}(S, \boldsymbol{\eta}_S, p) := \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{m} \sum_{(x,y) \in S} \mathbb{I}\left[\sup_{z \in \mathcal{U}(x)} |h(z) - y|^p \geq \eta(x, y)\right], \tag{5}$$

where $\boldsymbol{\eta}_S = (\eta(x_1, y_1), \ldots, \eta(x_m, y_m))$. We refer to $\eta(x, y)$ as a *cutoff* parameter. Note that η is a function of (x, y) and not necessarily a constant.

Define a robust empirical risk minimizer for the ℓ_p robust loss, $\ell_p\text{-RERM}_{\mathcal{H}} : (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$,

$$\ell_p\text{-RERM}_{\mathcal{H}}(S) := \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{m} \sum_{(x,y) \in S} \sup_{z \in \mathcal{U}(x)} |h(z) - y|^p. \tag{6}$$

Complexity measures. *Fat-shattering dimension.* Let $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$ and $\gamma > 0$. We say that $S = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ is γ-shattered by $\mathcal{F}$ if there exists a witness $r = (r_1, \ldots, r_m) \in [0, 1]^m$ such that for each $\sigma = (\sigma_1, \ldots, \sigma_m) \in \{-1, 1\}^m$ there is a function $f_\sigma \in \mathcal{F}$ such that

$$\forall i \in [m] \begin{cases} f_\sigma(x_i) \geq r_i + \gamma, & \text{if } \sigma_i = 1 \\ f_\sigma(x_i) \leq r_i - \gamma, & \text{if } \sigma_i = -1. \end{cases}$$

The fat-shattering dimension of $\mathcal{F}$ at scale γ, denoted by $\mathrm{fat}(\mathcal{F}, \gamma)$, is the cardinality of the largest set of points in $\mathcal{X}$ that can be γ-shattered by $\mathcal{F}$. This parametrized variant of the Pseudo-dimension (Alon et al., 1997) was first proposed by Kearns & Schapire (1994). Its key role in learning theory lies in characterizing the PAC learnability of real-valued function classes (Alon et al., 1997; Bartlett & Long, 1998).

Dual fat-shattering dimension. Define the dual class $\mathcal{F}^* \subseteq [0, 1]^{\mathcal{H}}$ of $\mathcal{F}$ as the set of all functions $g_w : \mathcal{F} \to [0, 1]$ defined by $g_w(f) = f(w)$. If we think of a function class as a matrix whose rows and columns are indexed by functions and points, respectively, then the dual class is given by the transpose of the matrix. The dual fat-shattering at scale γ, is

defined as the fat-shattering at scale γ of the dual-class and denoted by $\mathrm{fat}^*(\mathcal{F},\gamma)$. We have the following bound due to Kleer & Simon (2021, Corollary 3.8 and inequality 3.1),

$$\mathrm{fat}^*(\mathcal{F},\gamma) \lesssim \frac{1}{\gamma} 2^{\mathrm{fat}(\mathcal{F},\gamma/2)+1}. \tag{7}$$

Covering numbers. We say that $\mathcal{G} \subseteq [0,1]^\Omega$ is ϵ-cover for $\mathcal{F} \subseteq [0,1]^\Omega$ in d_∞ norm, if for any $f \in \mathcal{F}$ the exists $g \in \mathcal{G}$ such that for any $x \in \Omega$, $|f(x) - g(x)| \le \epsilon$. The ϵ-covering number of $\mathcal{F}$ is the minimal cardinality of any ϵ-cover, and denoted by $\mathcal{N}(\epsilon, \mathcal{F}, d_\infty)$.

Sample compression scheme. We say that a pair of functions (κ, ρ) is *uniform α-approximate sample compression scheme* of size k for a class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and the ℓ_p loss, if for any $m \in \mathbb{N}$, $h^\star \in \mathcal{H}$, and sample $S = \{(x_i, y_i)\}_{i=1}^m$, it holds for the compression function that $\kappa(S) \subseteq S$, $|\kappa(S)| \le k$, and the reconstruction function $\rho(\kappa(S)) = \hat{h}$ satisfies $\forall i \in [m]$

$$\left|\hat{h}(x_i) - y_i\right|^p \le \left|h^\star(x_i) - y_i\right|^p + \alpha. \tag{8}$$

Similarly, we define an *adversarially robust uniform α-approximate sample compression scheme* if $\forall i \in [m]$

$$\sup_{z_i \in \mathcal{U}(x_i)} \left|\hat{h}(z_i) - y_i\right|^p \le \sup_{z_i \in \mathcal{U}(x_i)} \left|h^\star(z_i) - y_i\right|^p + \alpha. \tag{9}$$

Notation. We use the notation $\tilde{\mathcal{O}}(\cdot)$ for omitting poly-logarithmic factors of $(\mathrm{fat}(\mathcal{H},\gamma), \mathrm{fat}^*(\mathcal{H},\gamma), 1/\epsilon, 1/\delta, 1/\eta, 1/\beta)$. We denote $[n] = \{1,\dots,n\}$, and $\exp(\cdot) = e^{(\cdot)}$. $\lesssim$ and $\gtrsim$ denote inequalities up to a constant factor, and $\approx$ denotes equality up to a constant factor. Vectors are written using bold symbols.

3. Robust Regression for ℓ_p Losses

In this section, we provide an algorithm and prove its sample complexity for robust regression with ℓ_p losses. Moreover, our learning algorithm is *proper* for convex function classes, arguably the most commonly studied subcategory of real-valued function classes for regression. This result circumvents a negative result from Montasser et al. (2019); there exist, non-convex function classes, where proper learning is impossible.

Theorem 3.1. *The sample complexity of Algorithm 1 for robust (ϵ,δ)-PAC learning a concept class $\mathcal{H}$ with the ℓ_p robust loss is*

$$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}^3(\mathcal{H}, c\epsilon/p)\,\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}{\epsilon^5} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

for some numerical constant $c \in (0,\infty)$.

Recall that $\mathrm{fat}^*(\mathcal{F},\epsilon) \lesssim \frac{1}{\epsilon} 2^{\mathrm{fat}(\mathcal{F},\epsilon/2)+1}$ by Equation (7).

Remark 3.2. The output of Algorithm 1 is a convex combination of the functions from the concept class, which is a proper predictor, assuming convexity of the function class.

Remark 3.3. Similar to non-robust regression, our results generalize to loss functions with bounded codomain $[0,M]$. The generalization bound should be multiplied by pM^p and the scaling of the fat-shattering dimension should be ϵ/pM^p.

In the following result, we establish generalization from approximate interpolation for *changing* cutoff parameters for different points. This generalizes a result by (Anthony & Bartlett, 2000), where the cutoff parameter is fixed for all points. The proof is in Appendix A.

Theorem 3.4 (Generalization from approximate interpolation with changing cutoffs)**.** *Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ with a finite fat-shattering dimension (at any scale). For any $\beta,\epsilon,\delta \in (0,1)$, any function $\eta : \mathcal{X}\times\mathcal{Y} \to [0,1]$, any distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, for a random sample $S \sim \mathcal{D}^m$, if*

$$m = \mathcal{O}\left(\frac{1}{\epsilon}\left(\mathrm{fat}(\mathcal{H},\beta/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H},\beta/8)}{\beta\epsilon}\right) + \log\frac{1}{\delta}\right)\right),$$

then with probability at least $1-\delta$ over S, for any $h \in \mathcal{H}$ satisfying $|h(x) - y| \le \eta(x,y)$, $\forall (x,y) \in S$, it holds that $\mathbb{P}_{(x,y)\sim\mathcal{D}}\{(x,y) : |h(x) - y| \le \eta(x,y) + \beta\} \ge 1-\epsilon$.

Algorithm 1 Improper Robust Regressor with High-Vote

Input: $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, $S = \{(x_i,y_i)\}_{i=1}^m$.
Parameters: ϵ.
Algorithms used: ℓ_p-RERM$_{\mathcal{H}}$ (Equation (6)), η-RERM$_{\mathcal{H}}$ (Equation (5)), a variant of Multiplicative Weights (Algorithm 4).

1. Compute $h^\star \leftarrow \ell_p\text{-RERM}_{\mathcal{H}}(S)$.
   Denote $\eta(x,y) = \sup_{z\in\mathcal{U}(x)} |h^\star(z) - y|, \forall (x,y) \in S$.
2. Inflate S to $S_{\mathcal{U}}$ to include all perturbed points.
3. Discretize $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$: (i) Construct a function class $\hat{\mathcal{H}}$, where each $\hat{h} \in \hat{\mathcal{H}}$ defined by η-RERM optimizer on $\tilde{\mathcal{O}}\big(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H},\mathcal{O}(\epsilon/p))\big)$ points from S. The input cutoff parameters to the optimizer are $\eta(x,y)$, as computed in step 1.
   (ii) Each $(z,y) \in S_{\mathcal{U}}$ defines a function in the dual space, $f_{(z,y)} : \hat{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|^p$. Define $\bar{S}_{\mathcal{U}}$ to be the minimal cover of $S_{\mathcal{U}}$ under d_∞ norm at scale $\mathcal{O}(\epsilon)$.
4. Compute a modified Multiplicative Weights on $\bar{S}_{\mathcal{U}}$. Let $\left\{\hat{h}_1,\dots,\hat{h}_T\right\}$ be the returned set of classifiers.

Output: $\hat{h} = \frac{1}{T}\sum_{i=1}^T \hat{h}_i$.

We construct an adversarially robust sample compression scheme of a size determined by the fat-shattering dimension of the function class. Recall that uniform convergence does not necessarily hold. Instead, we derive generalization from sample compression bounds. Our proof crucially relies on Theorem 3.4.

Proof overview and algorithm outline. The complete proof is in Appendix B. We follow the steps in Algorithm 1.

(1) We start with computing a robust empirical risk minimizer (ERM) $h^\star$ on S for the ℓ_p robust loss. This defines the target loss we are aiming for at any point in S. In other words, the robust loss of $h^\star$ on (x, y) defines a *cutoff* $\eta(x, y)$ and our goal is to construct a predictor with a loss of $\eta(x, y) + \epsilon$ for any $(x, y) \in S$, which means that this predictor is an approximate robust ERM. In order to derive generalization, we cannot rely on uniform convergence. Instead, our predictor is based on a sample compression scheme from which we can generalize.

(2) Inflate the training set by including all possible perturbations. Whenever the same perturbation is mapped to more than one input, we assign the label of the input with the smallest index to prevent ambiguity. We denote this set by $S_{\mathcal{U}}$.

(3) Discretize the set $S_{\mathcal{U}}$ as follows: (i) Construct a set of functions $\hat{\mathcal{H}}$, such that each function is the output of η-RERM for $\mathcal{H}$ (defined in Equation (5)), performing on a subset $S' \subseteq S$ of size $d = \tilde{\mathcal{O}}\big(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H}, \mathcal{O}(\epsilon/p))\big)$. This means that for any $S' \subseteq S$ there exists $\hat{h} \in \hat{\mathcal{H}}$ that is an approximate robust ERM on S', that is, $\hat{h}$ is within $\eta(x, y) + \epsilon$ for any $(x, y) \in S'$. The size of $\hat{\mathcal{H}}$ is bounded $(m/d)^d$, where $|S| = m$.

(ii) Define a discretization $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ by taking a uniform cover of the dual space. In the dual space, each $(z, y) \in S_{\mathcal{U}}$ defines a function $f_{(z,y)} : \hat{\mathcal{H}} \to [0, 1]$ such that $f_{(z,y)}(h) = \big|h(z) - y\big|^p$. We take a minimal $\mathcal{O}(\epsilon)$-cover for $S_{\mathcal{U}}$ with the supremum norm, which is of size $\mathcal{N}(\mathcal{O}(\epsilon), S_{\mathcal{U}}, d_\infty)$. We use covering numbers arguments (Rudelson & Vershynin, 2006) to upper bound the size of $\bar{S}_{\mathcal{U}}$

(4) Compute a variant of Multiplicative Weights (MW) update (Algorithm 4) on $\bar{S}_{\mathcal{U}}$ for $T \approx \log\big|\bar{S}_{\mathcal{U}}\big|$ rounds as follows. From a corollary of Theorem 3.4 (see Corollary A.1), we know that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, upon receiving an i.i.d. sample S'' from $\mathcal{P}$ of size d, with probability $2/3$ over sampling S'' from $\mathcal{P}$, for any $h \in \mathcal{H}$ with $\forall (z, y) \in S'' : \ |h(z) - y|^p \le \eta(z, y)$, it holds that $\mathbb{P}_{(z,y)\sim\mathcal{P}}\{(z, y) : |h(z) - y|^p \le \eta(z, y) + \epsilon\} \ge 1 - \epsilon$, where $\eta(z, y)$ is the $\eta(x, y)$ for which $z \in \mathcal{U}(x)$ as defined in step 2. We can conclude that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, there exists such a set of points $S'' \subseteq \bar{S}_{\mathcal{U}}$. Then, we can find a set S' of d points in S that S'' originated from. Formally, $S'' \subseteq \bigcup_{(x,y)\in S'} \bigcup\{(z, y) : z \in \mathcal{U}(x)\}$. We execute the optimizer $\hat{h} \leftarrow \eta$-RERM on S' with the relevant cutoff parameters. $\hat{h}$ has error of $\eta(z, y) + \epsilon$ on a fraction of $(1 - \epsilon)$ points with respect to the distribution $\mathcal{P}$. We start with $\mathcal{P}_1$ as the uniform distribution over $\bar{S}_{\mathcal{U}}$ and find $\hat{h}_1$ respectively. We perform a multiplicative weights update on the distribution and find the next hypothesis w.r.t. the new distribution and so forth.

Following the analysis of MW (or α-Boost) from Schapire & Freund (2013, Section 6)), we know that for any point in $\bar{S}_{\mathcal{U}}$, roughly $(1 - \epsilon)$ base learners are within ϵ from the target cutoff. The rest ϵ fraction can contribute an error of at most ϵ since the loss is bounded by 1. We get that for any point in $\bar{S}_{\mathcal{U}}$, the average loss of hypotheses in the ensemble is within 2ϵ from the target cutoff. Crucially, we use strong base learners in the ensemble. By the covering argument, we get that for any point in $S_{\mathcal{U}}$, the average loss of the ensemble is within 4ϵ,

$$\forall (z, y) \in S_{\mathcal{U}} : \ \frac{1}{T}\sum_{i=1}^{T}\Big|\hat{h}_i(z) - y\Big|^p \le \eta(z, y) + 4\epsilon.$$

We are interested that the average prediction $\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i$ will be within the target cutoffs. For that reason, we use the convexity of the ℓ_p loss to show that

$$\left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(z) - y\right|^p \le \frac{1}{T}\sum_{i=1}^{T}\Big|\hat{h}_i(z) - y\Big|^p.$$

Therefore, we conclude that

$$\forall (z, y) \in S_{\mathcal{U}} : \ \left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(z) - y\right|^p \le \eta(z, y) + 4\epsilon,$$

which implies that we have an approximate robust ERM for S,

$$\forall (x, y) \in S : \ \sup_{z\in\mathcal{U}(x)}\left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(z) - y\right|^p \le \eta(x, y) + 4\epsilon.$$

The proof follows by applying a sample compression generalization bound in the agnostic case, bounding the compression size, and rescaling ϵ.

For convex classes, we have a proper learner. The output of the algorithm is a convex combination of functions from $\mathcal{H}$ which is also in the class.

4. Better Sample Complexity for the ℓ_1 Loss

In this section, we provide an algorithm with a substantial sample complexity improvement for the ℓ_1 robust loss. The key technical idea in this result is to note that, if we replace

base learners with weak learners in the improper ensemble predictor, we can still get an accurate prediction by taking the *median* aggregation of the ensemble. Thus, we incorporate a variant of median boosting for real-valued functions (Kégl, 2003; Hanneke et al., 2019) in our algorithm. Each base learner requires fewer samples and as a result, we improve the sample complexity. On the contrary, in Algorithm 1 we obtained accurate predictions for a $1 - \mathcal{O}(\epsilon)$ quantile of the predictors, and we output their average.

Theorem 4.1. *The sample complexity of Algorithm 2 for robust (ϵ, δ)-PAC learning a concept class $\mathcal{H}$ with the ℓ_1 robust loss is*

$$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H}, c\epsilon)\,\mathrm{fat}^*(\mathcal{H}, c\epsilon)}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

for some numerical constant $c \in (0, \infty)$.

Recall that $\mathrm{fat}^*(\mathcal{F}, \epsilon) \lesssim \frac{1}{\epsilon} 2^{\mathrm{fat}(\mathcal{F}, \epsilon/2)+1}$ by Equation (7).

We shall define the notion of weak learners in the context of real-valued learners.

Definition 4.2 (Weak real-valued learner)**.** Let $\xi \in (0, \frac{1}{2}]$, $\zeta \in [0, 1]$. We say that $f : \mathcal{X} \to [0, 1]$ is a (ζ, ξ)-weak learner for the ℓ_p loss, with respect to $\mathcal{D}$ and a target function $h^\star \in \mathcal{H}$ if

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\{(x, y) : |f(x) - y|^p > |h^\star(x) - y|^p + \zeta\} \leq \frac{1}{2} - \xi.$$

This notion of a weak learner must be formulated carefully. For example, taking a learner guaranteeing absolute loss at most $\frac{1}{2} - \xi$ is known to not be strong enough for boosting to work. On the other hand, by making the requirement too strong (for example, AdaBoost.R in Freund & Schapire (1997)), then the sample complexity of weak learning will be high that weak learners cannot be expected to exist for certain function classes. We can now present an overview of the proof and the algorithm.

Algorithm 2 Improper Robust Regressor

Input: $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$, $S = \{(x_i, y_i)\}_{i=1}^m$.
Parameters: ϵ.
Algorithms used: ℓ_1-RERM$_{\mathcal{H}}$ (Equation (6)), η-RERM$_{\mathcal{H}}$ (Equation (5)), a variant of median boosting: MedBoost (Algorithm 5), sparsification method (Algorithm 6).

1. Compute $h^\star \leftarrow \ell_1\text{-RERM}_{\mathcal{H}}(S)$.
   Denote $\eta(x, y) = \sup_{z\in\mathcal{U}(x)} |h^\star(z) - y|, \forall (x, y) \in S$.
2. Inflate S to $S_{\mathcal{U}}$ to include all perturbed points.
3. Discretize $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$: (i) Construct a function class $\hat{\mathcal{H}}$, where each $\hat{h} \in \hat{\mathcal{H}}$ defined by η-RERM optimizer on $\tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{H}, \mathcal{O}(\epsilon)))$ points from S. The input cutoff parameters to the optimizer are $\eta(x, y)$, as computed in step 1.
   (ii) Each $(z, y) \in S_{\mathcal{U}}$ defines a function in the dual space, $f_{(z,y)} : \hat{\mathcal{H}} \to [0, 1]$ such that $f_{(z,y)}(h) = \big|h(z) - y\big|$. Define $\bar{S}_{\mathcal{U}}$ to be the minimal cover of $S_{\mathcal{U}}$ under d_∞ norm at scale $\mathcal{O}(\epsilon)$.
4. Compute modified MedBoost on $\bar{S}_{\mathcal{U}}$, where $\hat{\mathcal{H}}$ consists of weak learners for any distribution over $\bar{S}_{\mathcal{U}}$. Let $\mathcal{F} = \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ be the returned set of classifiers.
5. Sparsify the set $\mathcal{F}$ to a smaller set $\left\{\hat{h}_1, \ldots, \hat{h}_k\right\}$.

Output: $\hat{h} = \mathrm{Median}\left(\hat{h}_1, \ldots, \hat{h}_k\right)$.

Proof overview and algorithm outline. The complete proof is in Appendix C.

We explain the main differences from Algorithm 1 and where the sample complexity improvement comes from. In the discretization step, we replace the base learners in $\hat{\mathcal{H}}$ with weak learners. We construct an improper ensemble predictor via a median boosting algorithm, where the weak learners are chosen from $\hat{\mathcal{H}}$. Specifically, each function in $\hat{\mathcal{H}}$ is the output of η-RERM for $\mathcal{H}$ (defined in Equation (5)), performing on a subset $S' \subseteq S$ of size $d = \tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{H}, \mathcal{O}(\epsilon)))$ This is in contrast to Algorithm 1, where we use Multiplicative Weights update that operates with strong base learners. We can make accurate predictions by taking the median aggregation due to the ℓ_1 loss.

Another improvement arises from sparsifying the ensemble (Hanneke et al., 2019) to be *independent* of the sample size while keeping the median accurate almost with the same resolution. The sparsification step uses sampling and uniform convergence in the dual space (with respect to the non-robust loss).

We elaborate on the steps in Algorithm 2. Steps (1),(2), and (3) are similar to Algorithm 1, besides the construction of $\hat{\mathcal{H}}$

as we explained above. In step (4), we compute a modified version of the real-valued boosting algorithm MedBoost (Kégl, 2003) on the discretized set $\bar{S}_{\mathcal{U}}$, see Algorithm 5. Hanneke et al. (2019) showed how to construct a sample compression scheme from MedBoost. From this step, we have that for any point in $\bar{S}_{\mathcal{U}}$, the median of the losses of each hypothesis in the ensemble is within ϵ of the target cutoff that was computed in step 1. By the covering argument, the median of the losses is within 3ϵ for any point in $(z, y) \in S_{\mathcal{U}}$,

$$\left|\mathrm{Med}\big(\hat{h}_1(z) - y, \ldots, \hat{h}_T(z) - y\big)\right| \leq \eta(z, y) + 3\epsilon.$$

The median is translation invariant, so we have

$$\left|\mathrm{Med}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z)\big) - y\right| \leq \eta(z, y) + 3\epsilon.$$

Finally, for any $(x, y) \in S$,

$$\sup_{z \in \mathcal{U}(x)} \left|\mathrm{Med}\big(\hat{h}_1(z) - y, \ldots, \hat{h}_T(z) - y\big)\right| \leq \eta(x, y) + 3\epsilon.$$

To further reduce the sample compression size, in step (5) we sparsify the ensemble to $k = \tilde{\mathcal{O}}(\mathrm{fat}^*(\mathcal{H}, c\epsilon))$ functions,

$$\sup_{z \in \mathcal{U}(x)} \left|\mathrm{Med}\big(\hat{h}_1(z) - y, \ldots, \hat{h}_k(z) - y\big)\right| \leq \eta(x, y) + 4\epsilon.$$

The proof follows by applying a sample compression generalization bound in the agnostic case, bounding the compression size, and rescaling ϵ.

5. Robust (η, β)-Regression

In this section, we study robust (η, β)-regression in realizable and agnostic settings. We provide an algorithm for the realizable setting and show how to reduce agnostic to realizable learning. We conclude by deriving sample complexity guarantees for both settings.

This model is different than regression which guarantees a small expected error (with high probability). In the robust (η, β)-regression, we aim for a small pointwise absolute error *almost everywhere* on the support of the distribution. Results for this model do not follow from the standard regression model. We first present our result for the realizable case. The proof is in Appendix D.

Theorem 5.1. *Assuming the distribution $\mathcal{D}$ is η-uniformly realizable (see Equation (4)) by a class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, the sample complexity of Algorithm 3 for robust $(\eta, \beta, \epsilon, \delta)$-PAC learning a concept class $\mathcal{H}$ is*

$$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H}, c\beta)\,\mathrm{fat}^*(\mathcal{H}, c\beta)}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right),$$

for some numerical constant $c \in (0, \infty)$.

Recall that $\mathrm{fat}^*(\mathcal{F}, \epsilon) \lesssim \frac{1}{\epsilon} 2^{\mathrm{fat}(\mathcal{F}, \epsilon/2)+1}$ by Equation (7).

Algorithm 3 Improper Robust (η, β)-Regressor

Input: $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, $S = \{(x_i, y_i)\}_{i=1}^m$.
Parameters: η, β.
Algorithms used: η-RERM$_{\mathcal{H}}$ (Equation (5)), a variant of median boosting: MedBoost (Algorithm 5), sparsification method (Algorithm 6).

1. Inflate S to $S_{\mathcal{U}}$ to include all perturbed points.
2. Discretize $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$: (i) Construct a function class $\hat{\mathcal{H}}$, where each $\hat{h} \in \hat{\mathcal{H}}$ defined by η-RERM optimizer on $\tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{H}, \mathcal{O}(\beta)))$ points from S. The input cutoff parameters to the optimizer are fixed η for all points. (ii) Each $(z, y) \in S_{\mathcal{U}}$ defines a function in the dual space, $f_{(z,y)} : \hat{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \big|h(z) - y\big|$. Define $\bar{S}_{\mathcal{U}}$ to be the minimal cover of $S_{\mathcal{U}}$ under d_∞ norm at scale $\mathcal{O}(\beta)$.
3. Compute modified MedBoost on $\bar{S}_{\mathcal{U}}$, where $\hat{\mathcal{H}}$ consists of weak learners for any distribution over $\bar{S}_{\mathcal{U}}$. Let $\mathcal{F} = \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ be the returned set of classifiers.
4. Sparsify the set $\mathcal{F}$ to a smaller set $\left\{\hat{h}_1, \ldots, \hat{h}_k\right\}$.

Output: $\hat{h} = \mathrm{Median}\left(\hat{h}_1, \ldots, \hat{h}_k\right)$.

We explain the main differences from Algorithm 2. This model is different from robust regression with the ℓ_1 loss. Our goal is to find a predictor with a prediction within $\eta + \beta$ of the true label *almost everywhere* the domain, assuming that the distribution is η-uniformly realizable by the function class (Equation (4)).

In this model, the cutoff parameter is given to us as a parameter and is *fixed* for all points. This is different from Algorithms 1 and 2, where we computed the changing cutoffs with a robust ERM oracle. Moreover, the weak learners in $\hat{\mathcal{H}}$ are defined as the output of η-RERM performing on a subset $S' \subseteq S$ of size $d = \tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{H}, \mathcal{O}(\beta)))$. Note that the scale of shattering depends on β and not ϵ. The resolution of discretization in the cover depends on β as well.

Agnostic setting We establish an upper bound on the sample complexity of the agnostic setting, by using a reduction to the realizable case. The main argument was originally suggested in (David et al., 2016) for the 0-1 loss and holds for the η-ball robust loss as well. The proof is in Appendix D.

Theorem 5.2. *The sample complexity for agnostic robust $(\eta, \beta, \epsilon, \delta)$-PAC learning a concept class $\mathcal{H}$ is*

$$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H}, c\beta)\,\mathrm{fat}^*(\mathcal{H}, c\beta)}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

for some numerical constant $c \in (0, \infty)$.

Recall that $\mathrm{fat}^*(\mathcal{F}, \epsilon) \lesssim \frac{1}{\epsilon} 2^{\mathrm{fat}(\mathcal{F}, \epsilon/2)+1}$ *by Equation* (7).

Remark. An agnostic learner for robust (η, β)-regression does not apply to the robust regression setting. The reason is that the optimal function in $\mathcal{H}$ may have different scales of robustness on different points, which motivates our approach of using changing cutoffs for different points. In Appendix D.3 we show that by using a fixed cutoff for all points we can obtain an error of only $\sqrt{\mathrm{OPT}_{\mathcal{H}}} + \epsilon$.

6. Discussion

In this paper, we studied the robustness of real-valued functions to test time attacks. We showed that finite fat-shattering is sufficient for learnability. we proved sample complexity for learning with the general ℓ_p losses and improved it for the ℓ_1 loss. We also studied a model of regression with a cutoff loss. We proved sample complexity in realizable and agnostic settings. We leave several interesting open questions for future research. (i) Improve the upper bound for learning with ℓ_p robust losses (if possible) and show a lower bound. There might be a gap between sample complexities of different values of p. More specifically, what is the sample complexity for learning with ℓ_2 robust loss? (ii) We showed that the fat-shattering dimension is a sufficient condition. What is a necessary condition? In the binary-valued case, we know that having a finite VC is not necessary. (iii) To what extent can we benefit from unlabeled samples for learning real-valued functions? This question was considered by Attias et al. (2022) for binary function classes, where they showed that the labeled sample complexity can be arbitrarily smaller compared to the fully-supervised setting. (iv) In this work we focused on the statistical aspect of robustly learning real-valued functions. It would be interesting to explore the computational aspect as well.

Acknowledgments

We thank Meni Sadigurschi for helpful discussions regarding sample compression schemes of real-valued functions. We are also grateful to Uri Sherman for discussions on the generalization of ERM in Stochastic Convex Optimization. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grants 993/17, 1602/19), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University. I.A. is supported by the Vatat Scholarship from the Israeli Council for Higher Education and by Kreitman School of Advanced Graduate Studies.

References

Agarwal, N., Bullins, B., Hazan, E., Kakade, S., and Singh, K. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pp. 111–119. PMLR, 2019.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Anava, O., Hazan, E., Mannor, S., and Shamir, O. Online learning for time series prediction. In *Conference on learning theory*, pp. 172–184. PMLR, 2013.

Angluin, D. Queries and concept learning. *Machine learning*, 2:319–342, 1988.

Anthony, M. and Bartlett, P. L. Function learning from interpolation. *Combinatorics, Probability and Computing*, 9(3):213–225, 2000.

Anthony, M., Bartlett, P. L., Bartlett, P. L., et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Ashtiani, H., Pathak, V., and Urner, R. Black-box certification and learning under adversarial perturbations. In *International Conference on Machine Learning*, pp. 388–398. PMLR, 2020.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pp. 162–183. PMLR, 2019.

Attias, I., Hanneke, S., and Mansour, Y. A characterization of semi-supervised adversarially-robust pac learnability. *arXiv preprint arXiv:2202.05420*, 2022.

Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pp. 431–441. PMLR, 2020.

Awasthi, P., Frank, N., Mao, A., Mohri, M., and Zhong, Y. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, 34, 2021a.

Awasthi, P., Frank, N., and Mohri, M. On the existence of the adversarial bayes classifier. *Advances in Neural Information Processing Systems*, 34, 2021b.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pp. 1117–1174. PMLR, 2022a.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Multi-class h-consistency bounds. *Advances in Neural Information Processing Systems*, 35:782–795, 2022b.

Awasthi, P., Mao, A., Mohri, M., and Zhong, Y. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pp. 10077–10094. PMLR, 2023.

Bartlett, P. L. and Long, P. M. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.

Bhattacharjee, R., Jha, S., and Chaudhuri, K. Sample complexity of robust linear classification on separated data. In *International Conference on Machine Learning*, pp. 884–893. PMLR, 2021.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pp. 831–840. PMLR, 2019.

Candes, E. and Recht, B. Exact matrix completion via convex optimization. *Communications of the ACM*, 55 (6):111–119, 2012.

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pp. 230–241, 2018.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355. PMLR, 2020.

Daniely, A. and Shalev-Shwartz, S. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pp. 287–316. PMLR, 2014.

Daniely, A., Sabato, S., Ben-David, S., and Shalev-Shwartz, S. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 207–232. JMLR Workshop and Conference Proceedings, 2011.

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

Diochnos, D., Mahloujifar, S., and Mahmoody, M. Adversarial risk and robustness: General definitions and implications for the uniform distribution. *Advances in Neural Information Processing Systems*, 31, 2018.

Dudley, R. M. A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, pp. 1–142. Springer, 1984.

Duffy, N. and Helmbold, D. Boosting methods for regression. *Machine Learning*, 47(2):153–200, 2002.

Feige, U., Mansour, Y., and Schapire, R. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pp. 637–657, 2015.

Floyd, S. and Warmuth, M. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gourdeau, P., Kanade, V., Kwiatkowska, M., and Worrell, J. On the hardness of robust classification. *The Journal of Machine Learning Research*, 22(1):12521–12549, 2021.

Graepel, T., Herbrich, R., and Shawe-Taylor, J. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

Hanneke, S. The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1):1319–1333, 2016.

Hanneke, S., Kontorovich, A., and Sadigurschi, M. Sample compression for real-valued learners. In *Algorithmic Learning Theory*, pp. 466–488. PMLR, 2019.

Hazan, E. and Ma, T. A non-generative framework and convex relaxations for unsupervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Hazan, E., Kale, S., and Shalev-Shwartz, S. Near-optimal algorithms for online matrix prediction. In *Conference on Learning Theory*, pp. 38–1. JMLR Workshop and Conference Proceedings, 2012.

Hazan, E., Livni, R., and Mansour, Y. Classification with low rank and missing data. In *International conference on machine learning*, pp. 257–266. PMLR, 2015.

Kearns, M. J. and Schapire, R. E. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

Kégl, B. Robust regression by boosting the median. In *Learning Theory and Kernel Machines*, pp. 258–272. Springer, 2003.

Khim, J. and Loh, P.-L. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Kleer, P. and Simon, H. Primal and dual combinatorial dimensions. *arXiv preprint arXiv:2108.10037*, 2021.

Kontorovich, A. and Attias, I. Fat-shattering dimension of k-fold maxima. *arXiv preprint arXiv:2110.04763*, 2021.

Littlestone, N. and Warmuth, M. Relating data compression and learnability. 1986.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mao, A., Mohri, M., and Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*, 2023.

Mendelson, S. An optimal unrestricted learning procedure. *Journal of the ACM*, 66(6):1—42, 2019.

Montasser, O., Hanneke, S., and Srebro, N. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

Montasser, O., Goel, S., Diakonikolas, I., and Srebro, N. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning*, pp. 7010–7021. PMLR, 2020a.

Montasser, O., Hanneke, S., and Srebro, N. Reducing adversarially robust learning to non-robust pac learning. *Advances in Neural Information Processing Systems*, 33: 14626–14637, 2020b.

Montasser, O., Hanneke, S., and Srebro, N. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory*, pp. 3452–3482. PMLR, 2021a.

Montasser, O., Hanneke, S., and Srebro, N. Transductive robust learning guarantees. *arXiv preprint arXiv:2110.10602*, 2021b.

Pollard, D. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

Rudelson, M. and Vershynin, R. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pp. 603–648, 2006.

Schapire, R. E. and Freund, Y. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Simon, H. U. Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, 26 (3):751–763, 1997.

Srebro, N., Rennie, J., and Jaakkola, T. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Xing, Y., Zhang, R., and Cheng, G. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 514–522. PMLR, 2021.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pp. 7085–7094. PMLR, 2019.

A. Auxiliary Results

Proof of Theorem 3.4. Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and let

$$\mathcal{H} = \{(x,y) \mapsto |f(x) - y| : f \in \mathcal{F}\}.$$

Define the function classes

$$\mathcal{F}_1 = \{(x,y) \mapsto |h(x) - y| - \eta(x,y) : h \in \mathcal{H}\},$$

and

$$\mathcal{F}_2 = \{(x,y) \mapsto \max\{f(x,y), 0\} : f \in \mathcal{F}_1\}.$$

We claim that $\mathrm{fat}(\mathcal{H}, \gamma) = \mathrm{fat}(\mathcal{F}_1, \gamma)$. Take a set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ that is γ-shattered by $\mathcal{H}$. There exists a witness $r = (r_1, \ldots, r_m) \in [0,1]^m$ such that for each $\sigma = (\sigma_1, \ldots, \sigma_m) \in \{-1, 1\}^m$ there is a function $h_\sigma \in \mathcal{H}$ such that

$$\forall i \in [m] \begin{cases} h_\sigma((x_i, y_i)) \geq r_i + \gamma, & \text{if } \sigma_i = 1 \\ h_\sigma((x_i, y_i)) \leq r_i - \gamma, & \text{if } \sigma_i = -1. \end{cases}$$

The set S is shattered by $\mathcal{F}_1$ by taking $\tilde{r} = (r_1 + \eta(x_1, y_1), \ldots, r_m + \eta(x_m, y_m))$. Similarly, any set that is shattered by $\mathcal{F}_1$ is also shattered by $\mathcal{H}$.

The class $\mathcal{F}_2$ consists of choosing a function from $\mathcal{F}_1$ and computing its pointwise maximum with the constant function 0. In general, for two function classes $\mathcal{G}_1, \mathcal{G}_2$, we can define the maximum aggregation class

$$\max(\mathcal{G}_1, \mathcal{G}_2) = \{x \mapsto \max\{g_1(x), g_2(x)\} : g_i \in \mathcal{G}_i\},$$

and Kontorovich & Attias (2021) showed that for any $\mathcal{G}_1, \mathcal{G}_2$

$$\mathrm{fat}(\max(\mathcal{G}_1, \mathcal{G}_2), \gamma) \lesssim (\mathrm{fat}(\mathcal{G}_1, \gamma) + \mathrm{fat}(\mathcal{G}_2, \gamma)) \log^2(\mathrm{fat}(\mathcal{G}_1, \gamma) + \mathrm{fat}(\mathcal{G}_2, \gamma)).$$

Taking $\mathcal{G}_1 = \mathcal{F}_1$ and $\mathcal{G}_2 \equiv 0$, we get

$$\mathrm{fat}(\mathcal{F}_2, \gamma) \lesssim \mathrm{fat}(\mathcal{F}_1, \gamma) \log^2(\mathrm{fat}(\mathcal{F}_1, \gamma)).$$

For the particular case $\mathcal{G}_2 \equiv 0$, we can show a better bound of

$$\mathrm{fat}(\mathcal{F}_2, \gamma) \lesssim \mathrm{fat}(\mathcal{F}_1, \gamma).$$

In words, it means that truncation cannot increase the shattering dimension. Indeed, take a set $S = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ that is γ-shattered by $\mathcal{F}_2 = \max(\mathcal{F}_1, 0)$, we show that this set is γ-shattered by $\mathcal{F}_1$. There exists a witness $r = (r_1, \ldots, r_k) \in [0,1]^m$ such that for each $\sigma = (\sigma_1, \ldots, \sigma_k) \in \{-1, 1\}^k$ there is a function $f_\sigma \in \mathcal{F}_1$ such that

$$\forall i \in [k] \begin{cases} \max\{f_\sigma((x_i, y_i)), 0\} \geq r_i + \gamma, & \text{if } \sigma_i = 1 \\ \max\{f_\sigma((x_i, y_i)), 0\} \leq r_i - \gamma, & \text{if } \sigma_i = -1. \end{cases}$$

For $\max\{f_\sigma((x_i, y_i)), 0\} \leq r_i - \gamma$, we simply have that $f_\sigma((x_i, y_i)) \leq r_i - \gamma$. Moreover, this implies that $r_i \geq \gamma$. As a result,

$$\begin{aligned} \max\{f_\sigma((x_i, y_i)), 0\} &\geq r_i + \gamma \\ &\geq 2\gamma \\ &> 0, \end{aligned}$$

which means that $f_\sigma((x_i, y_i)) \geq r_i + \gamma$. This shows that $\mathcal{F}_1$ γ-shatters S as well.

We can conclude the proof by applying Theorem A.2 to the class $\mathcal{F}_2$ and taking $\eta(x,y) = 0$. □

Corollary A.1. *Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ with a finite fat-shattering dimension (at any scale). For any $\beta, \epsilon, \delta \in (0,1)$, $1 \leq p < \infty$, any function $\eta : \mathcal{X} \times \mathcal{Y} \to [0,1]$, any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, for a random sample $S \sim \mathcal{D}^m$, if*

$$m(\eta, \beta, \epsilon, \delta) = \mathcal{O}\left(\frac{1}{\epsilon}\left(\mathrm{fat}(\mathcal{H}, \beta/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \beta/8)}{\beta\epsilon}\right) + \log\frac{1}{\delta}\right)\right),$$

then with probability at least $1-\delta$ over S, for any $h \in \mathcal{H}$ satisfying $|h(x)-y|^p \leq \eta(x,y)$, $\forall (x,y) \in S$, it holds that $\mathbb{P}_{(x,y)\sim\mathcal{D}}\{(x,y) : |h(x)-y|^p \leq \eta(x,y) + p\beta\} \geq 1-\epsilon$.

Proof. Following Theorem 3.4, we know that if $m = \mathcal{O}\left(\frac{1}{\epsilon}\left(\mathrm{fat}(\mathcal{H}, \beta/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \beta/8)}{\beta\epsilon}\right) + \log\frac{1}{\delta}\right)\right)$, with high probability over $S \sim \mathcal{D}^m$, for any $h \in \mathcal{H}$ satisfying $|h(x)-y| \leq (\eta(x,y))^{1/p}$ for all $(x,y) \in S$, it holds that $\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big\{(x,y) : |h(x)-y| \leq (\eta(x,y))^{1/p} + \beta\Big\} \geq 1-\epsilon$.

We show that for any (x,y) with $|h(x)-y| \leq (\eta(x,y))^{1/p} + \beta$, it holds that

$$\begin{aligned}|h(x)-y|^p &\overset{(i)}{\leq} \Big((\eta(x,y))^{1/p} + \beta\Big)^p \\ &\overset{(ii)}{\leq} \eta(x,y) + p\beta,\end{aligned}$$

and that will finish the proof. (i) Follows by just raising both sides to the power of p. (ii) Follows since the function $x \mapsto |x-y|^p$ is p-Lipschitz for $(x-y) \in [0,1]$, and so

$$\begin{aligned}\Big|\Big((\eta(x,y))^{1/p} + \beta\Big)^p - \eta(x,y)\Big| &= \Big|\Big((\eta(x,y))^{1/p} + \beta\Big)^p - \Big((\eta(x,y))^{1/p}\Big)^p\Big| \\ &\leq p\Big|(\eta(x,y))^{1/p} + \beta - (\eta(x,y))^{1/p}\Big| \\ &\leq p\beta.\end{aligned}$$

□

Theorem A.2 (**Generalization from approximate interpolation**). *(Anthony et al., 1999, Theorems 21.13 and 21.14) Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ with a finite fat-shattering dimension (at any scale). For any $\beta, \epsilon, \delta \in (0,1)$, $\eta \in [0,1]$, any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, for a random sample $S \sim \mathcal{D}^m$, if*

$$m(\eta, \beta, \epsilon, \delta) = \mathcal{O}\left(\frac{1}{\epsilon}\left(\mathrm{fat}(\mathcal{H}, \beta/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \beta/8)}{\beta\epsilon}\right) + \log\frac{1}{\delta}\right)\right),$$

then with probability at least $1-\delta$ over S, for any $h \in \mathcal{H}$ satisfying $|h(x)-y| \leq \eta$, $\forall (x,y) \in S$, it holds that $\mathbb{P}_{(x,y)\sim\mathcal{D}}\{(x,y) : |h(x)-y| \leq \eta + \beta\} \geq 1-\epsilon$.

The following is a bound on the covering numbers in d_∞.

Lemma A.3 (**Covering numbers for infinity metric**). *(Rudelson & Vershynin, 2006, Theorem 4.4) Let $\mathcal{F} \subseteq [0,1]^{\Omega}$ be a class of functions and $|\Omega| = n$. Then for any $0 < a \leq 1$ and $0 < t < 1/2$,*

$$\log \mathcal{N}(t, \mathcal{F}, d_\infty) \leq Cv\log(n/vt) \cdot \log^a(2n/v),$$

where $v = \mathrm{fat}_{cat}(\mathcal{F})$, and C, c are universal constants.

Lemma A.4 (**Fat-shattering of the loss class**). *Let $\mathcal{H} \subset \mathbb{R}^m$ be a real valued function class on m points. Denote the ℓ_p loss class of $\mathcal{H}$ by $L^p_{\mathcal{H}}$, for $1 \leq p < \infty$. Assume $L^p_{\mathcal{H}}$ is bounded by M. For any $\mathcal{H}$,*

$$\mathrm{fat}(L^p_{\mathcal{H}}, \gamma) \leq \mathcal{O}(\log^2(m)\mathrm{fat}(\mathcal{H}, \gamma/pM^{p-1}).$$

Proof. For any $\mathcal{X}$ and any function class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, define the *difference class* $\mathcal{H}^\Delta \subset \mathbb{R}^{\mathcal{X}\times\mathbb{R}}$ as

$$\mathcal{H}^\Delta = \{\mathcal{X} \times \mathbb{R} \ni (x,y) \mapsto \Delta_h(x,y) := h(x) - y; h \in \mathcal{H}\}.$$

In words, $\mathcal{H}^\Delta$ consists of all functions $\Delta_h(x,y) = h(x) - y$ indexed by $h \in \mathcal{H}$.

It is easy to see that for all $\gamma > 0$, we have $\mathrm{fat}_\gamma(\mathcal{H}^\Delta) \leq \mathrm{fat}_\gamma(\mathcal{H})$. Indeed, if $\mathcal{H}^\Delta$ γ-shatters some set $\{(x_1,y_1),\ldots,(x_k,y_k)\} \subset \mathcal{X}\times\mathbb{R}$ with shift $r \in \mathbb{R}^k$, then $\mathcal{H}$ γ-shatters the set $\{x_1,\ldots,x_k\} \subset \mathcal{X}$ with shift $r+(y_1,\ldots,y_k)$.

Next, we observe that taking the absolute value does not significantly increase the fat-shattering dimension. Indeed, for any real-valued function class $\mathcal{F}$, define $\mathrm{abs}(\mathcal{F}) := \{|f|;\ f \in \mathcal{F}\}$. Observe that $\mathrm{abs}(\mathcal{F}) \subseteq \max((F_j)_{j\in[2]})$, where $\mathcal{F}_1 = \mathcal{F}$ and $\mathcal{F}_2 = -\mathcal{F} =: \{-f; f \in \mathcal{F}\}$. It follows from Attias et al. (2019, Theorem 13) that

$$\mathrm{fat}_\gamma(\mathrm{abs}(\mathcal{F})) < \mathcal{O}(\log^2(m)(\mathrm{fat}_\gamma(\mathcal{F}) + \mathrm{fat}_\gamma(-\mathcal{F}))) < \mathcal{O}(\log^2(m)\mathrm{fat}_\gamma(\mathcal{F})). \tag{10}$$

Next, define $\mathcal{F}$ as the L_1 loss class of $\mathcal{H}$:

$$\mathcal{F} = \{\mathcal{X}\times\mathbb{R} \ni (x,y) \mapsto |h(x)-y)|; h \in \mathcal{H}\}.$$

Then

$$\begin{aligned}\mathrm{fat}_\gamma(\mathcal{F}) &= \mathrm{fat}_\gamma(\mathrm{abs}(\mathcal{H}^\Delta)) \\ &\leq \mathcal{O}(\log^2(m)\mathrm{fat}_\gamma(\mathcal{H}^\Delta)) \\ &\leq \mathcal{O}(\log^2(m)\mathrm{fat}_\gamma(\mathcal{H}));\end{aligned}$$

this proves the claim for L_1.

To analyze the L_2 case, consider $\mathcal{F} \subset [0,M]^{\mathcal{X}}$ and define $\mathcal{F}^{\circ p} := \{f^p; f \in \mathcal{F}\}$. We would like to bound $\mathrm{fat}_\gamma(\mathcal{F}^{\circ p})$ in terms of $\mathrm{fat}_\gamma(\mathcal{F})$. Suppose that $\mathcal{F}^{\circ p}$ γ-shatters some set $\{x_1,\ldots,x_k\}$ with shift $r^p = (r_1^p,\ldots,r_k^p) \in [0,M]^k$ (there is no loss of generality in assuming that the shift has the same range as the function class). Since for any y' the function $y \mapsto |y-y'|^p$ is pM^{p-1} Lipschitz for $(y-y') \in [0,M]$, we have

$$|a^p - b^p| \leq pM^{p-1}|a-b|, \qquad a,b \in [0,M],$$

we conclude that $\mathcal{F}$ is able to $\gamma/(pM^{p-1})$-shatter the same k points and thus $\mathrm{fat}_\gamma(\mathcal{F}^{\circ p}) \leq \mathrm{fat}_{\gamma/(pM^{p-1})}(\mathcal{F})$.

To extend this result to the case where $\mathcal{F} \subset [-M,M]^{\mathcal{X}}$, we use Equation (10).

In particular, define $\mathcal{F}$ as the L_p loss class of $\mathcal{H}$:

$$\mathcal{F} = \{\mathcal{X}\times\mathbb{R} \ni (x,y) \mapsto (h(x)-y)^p; h \in \mathcal{H}\}.$$

Then

$$\begin{aligned}\mathrm{fat}_\gamma(\mathcal{F}) &= \mathrm{fat}_\gamma((\mathcal{H}^\Delta)^{\circ p}) \\ &= \mathrm{fat}_\gamma((\mathrm{abs}(\mathcal{H}^\Delta))^{\circ p}) \\ &\leq \mathrm{fat}_{\gamma/(pM^{p-1})}(\mathrm{abs}(\mathcal{H}^\Delta)) \\ &\leq \mathcal{O}(\log^2(m)\mathrm{fat}_{\gamma/(pM^{p-1})}(\mathcal{H}^\Delta)) \\ &\leq \mathcal{O}(\log^2(m)\mathrm{fat}_{\gamma/(pM^{p-1})}(\mathcal{H})).\end{aligned}$$

□

The following generalization for sample compression in the realizable case was proven by Littlestone & Warmuth (1986); Floyd & Warmuth (1995). Their proof is for the 0-1 loss, but it applies similarly to bounded loss functions.

Lemma A.5 (**Sample compression generalization bound**). *Let a sample compression scheme (κ,ρ), and a loss function $\ell : \mathbb{R}\times\mathbb{R} \to [0,1]$. In the realizable case, for any $\kappa(S) \lesssim m$, any $\delta \in (0,1)$, and any distribution $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$, for $S \sim \mathcal{D}^m$, with probability $1-\delta$,*

$$\mathrm{Err}(\rho(\kappa(S)); \mathcal{D}) \leq \widehat{\mathrm{Err}}(\rho(\kappa(S)); S) + \mathcal{O}\left(\frac{|\kappa(S)|\log(m) + \log\frac{1}{\delta}}{m}\right).$$

The following generalization for sample compression in the agnostic case was proven by Graepel et al. (2005). Their proof is for the 0-1 loss, but it applies similarly to bounded loss functions. We use it with the η-ball robust loss.

Lemma A.6 (**Agnostic sample compression generalization bound**). *Let a sample compression scheme (κ, ρ), and a loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0,1]$. In the agnostic case, for any $\kappa(S) \lesssim m$, any $\delta \in (0,1)$, and any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, for $S \sim \mathcal{D}^m$, with probability $1-\delta$,*

$$\mathrm{Err}(\rho(\kappa(S)); \mathcal{D}) \leq \widehat{\mathrm{Err}}(\rho(\kappa(S)); S) + \mathcal{O}\left(\sqrt{\frac{\left(|\kappa(S)| \log(m) + \log \frac{1}{\delta}\right)}{m}}\right).$$

B. Proofs for Section 3: Robust Regression for ℓ_p Losses

Proof of Theorem 3.1. Fix $\epsilon, \delta \in (0,1)$. Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$. Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and let $S = \{(x_i, y_i)\}_{i=1}^m$ be an i.i.d. sample from $\mathcal{D}$. We elaborate on each one of the steps as described in Algorithm 1.

1. Compute $h^\star \leftarrow \ell_p\text{-RERM}_{\mathcal{H}}(S)$ in order to get the set of cutoffs $\eta(x,y) = \sup_{z \in \mathcal{U}(x)} |h^\star(z) - y|^p$ for $(x,y) \in S$. Let $\boldsymbol{\eta}_S = ((\eta(x_1,y_1)), \ldots, (\eta(x_m,y_m)))$. Our goal is to construct a predictor with an empirical robust loss of $\eta(x,y) + \epsilon$ for any $(x,y) \in S$, for the ℓ_p loss, which means that our predictor is an approximate Robust ERM.

2. Define the inflated training data set
$$S_{\mathcal{U}} = \bigcup_{i \in [n]} \left\{(z, y_{I(z)}) : z \in \mathcal{U}(x_i)\right\},$$
where $I(z) = \min\{i \in [m] : z \in \mathcal{U}(x_i)\}$. For $(z,y) \in S_{\mathcal{U}}$, let $\eta(z,y)$ be the $\eta(x,y)$ for which $z \in \mathcal{U}(x)$ and $y_{I(z)} = y$.

3. Discretize $S_{\mathcal{U}}$ to a finite set $\bar{S}_{\mathcal{U}}$ as following.

   (a) Define a set of functions, such that each function is defined by η-$\mathrm{RERM}_{\mathcal{H}}$ optimizer on $d = \mathcal{O}\left(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H}, \epsilon/8p) \log^2\left(\frac{p \cdot \mathrm{fat}(\mathcal{H}, \epsilon/8p)}{\epsilon^2}\right)\right)$ points from S.
   $$\hat{\mathcal{H}} = \{\eta\text{-}\mathrm{RERM}_{\mathcal{H}}(S', \boldsymbol{\eta}_{S'}, p) : S' \subseteq S, |S'| = d\}.$$
   Recall the definition of η-$\mathrm{RERM}_{\mathcal{H}}$, see Equation (5). The cardinality of this class is bounded as follows
   $$|\hat{\mathcal{H}}| \approx \binom{m}{d} \lesssim \left(\frac{m}{d}\right)^d. \tag{11}$$

   (b) A discretization $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ will be defined by covering of the dual class in d_∞ norm. Let $L^p_{\hat{\mathcal{H}}}$ be the L_p loss class of $\hat{\mathcal{H}}$, namely, $L^p_{\hat{\mathcal{H}}} = \left\{\mathcal{Z} \times \mathcal{Y} \ni (z,y) \mapsto |h(z) - y|^p : h \in \hat{\mathcal{H}}\right\}$. The *dual class* of $L^p_{\hat{\mathcal{H}}}$, ${L^p_{\hat{\mathcal{H}}}}^* \subseteq [0,1]^{\hat{\mathcal{H}}}$, is defined as the set of all functions $f_{(z,y)} : \hat{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|^p$, for any $(z,y) \in S_{\mathcal{U}}$. Formally, ${L^p_{\hat{\mathcal{H}}}}^* = \left\{f_{(z,y)} : (z,y) \in S_{\mathcal{U}}\right\}$, where $f_{(z,y)} = \left(f_{(z,y)}(h_1), \ldots, f_{(z,y)}(h_{|\hat{\mathcal{H}}|})\right)$. We take $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ to be a minimal ϵ-cover for $S_{\mathcal{U}}$ in d_∞,
   $$\sup_{(z,y) \in S_{\mathcal{U}}} \inf_{(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}}} \left\|f_{(z,y)} - f_{(\bar{z},\bar{y})}\right\|_\infty \leq \epsilon. \tag{12}$$
   Let $\mathrm{fat}^*(\mathcal{H}, \epsilon)$ be the dual ϵ-fat-shattering of $\mathcal{H}$. Using Lemma A.4, we can bound the dual fat-shattering of the $L^p_{\mathcal{H}}$ loss class by the fat-shattering of $\mathcal{H}$,
   $$\mathrm{fat}^*(L^p_{\mathcal{H}}, \epsilon) \leq \log^2(m)\, \mathrm{fat}^*(\mathcal{H}, \epsilon/p). \tag{13}$$

Applying a covering number argument from Lemma A.3 (taking $a = 1$) on the dual space, and upper bounding the dual fat-shattering of the L_p loss class as in Equation (13), we have the following bound

$$\begin{aligned}\left|\bar{S}_{\mathcal{U}}\right| &= \mathcal{N}(\epsilon, S_{\mathcal{U}}, d_\infty) \\ &\lesssim \exp\left(\mathrm{fat}^*(L^p_{\mathcal{H}}, c\epsilon)\log\left(\frac{|\hat{\mathcal{H}}|}{\epsilon\cdot\mathrm{fat}^*(L^p_{\mathcal{H}}, c\epsilon)}\right)\log\left(\frac{|\hat{\mathcal{H}}|}{\mathrm{fat}^*(L^p_{\mathcal{H}}, c\epsilon)}\right)\right) \\ &\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\log\left(\frac{|\hat{\mathcal{H}}|}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)\log\left(\frac{|\hat{\mathcal{H}}|}{\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)\log^2(m)\right) \\ &\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\log^2\left(\frac{|\hat{\mathcal{H}}|}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)\log^2(m)\right),\end{aligned} \tag{14}$$

where $c \in (0, \infty)$ is a numerical constant, derived from the covering argument in Lemma A.3.

4. Compute the following variant of Multiplicative Wights (MW) algorithm on the discretized set $\bar{S}_{\mathcal{U}}$ for $T \approx \log\left|\bar{S}_{\mathcal{U}}\right|$. Let $d = \mathcal{O}\left(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H}, \epsilon/8p)\log^2\left(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\right)\right)$, and let $\eta(z, y)$ be the $\eta(x, y)$ for which $z \in \mathcal{U}(x)$ as defined in step 2.

   From Corollary A.1, taking $\delta = 1/3$, $\beta = \epsilon/p$, we know that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, upon receiving an i.i.d. sample S'' from $\mathcal{P}$ of size d, with probability $2/3$ over sampling S'' from $\mathcal{P}$, for any $h \in \mathcal{H}$ with $\forall(z, y) \in S'' : |h(z) - y|^p \leq \eta(z, y)$, it holds that $\mathbb{P}_{(z,y)\sim\mathcal{P}}\{(z, y) : |h(z) - y|^p \leq \eta(z, y) + \epsilon\} \geq 1 - \epsilon$. We can conclude that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, there exists such a set of points $S'' \subseteq \bar{S}_{\mathcal{U}}$.

   Given that set, we can find the function with the aforementioned property in $\hat{\mathcal{H}}$. Let S' be the d points in S that the perturbed points S'' originated from. That is, $S'' \subseteq \bigcup_{(x,y)\in S'}\bigcup\{(z, y) : z \in \mathcal{U}(x)\}$. Take $\hat{\mathcal{H}} \ni \hat{h} = \eta\text{-RERM}_{\mathcal{H}}(S', \boldsymbol{\eta}_{S'}, p)$, it holds that $\forall(z, y) \in S'' : \left|\hat{h}(z) - y\right|^p \leq \eta(z, y)$, as a result we get $\mathbb{P}_{(z,y)\sim\mathcal{P}}\left\{(z, y) : \left|\hat{h}(z) - y\right|^p \leq \eta(z, y) + \epsilon\right\} \geq 1 - \epsilon$.

Algorithm 4 Modified Multiplicative Weights

Input: $\mathcal{H}, S, \bar{S}_{\mathcal{U}}$.
Parameters: $\epsilon, T, \boldsymbol{\eta}_S = (\eta(x_1, y_1), \ldots, \eta(x_m, y_m))$ for $(x_i, y_i) \in S$.
Algorithms used: Robust ERM for the η-ball robust loss: $\eta\text{-RERM}_{\mathcal{H}}$ (Equation (5)).
Initialize P_1 = Uniform($\bar{S}_{\mathcal{U}}$), $d = \mathcal{O}\left(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H}, \epsilon/8p)\log^2\left(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\right)\right)$, $\eta(z, y)$ is the $\eta(x, y)$ for which $z \in \mathcal{U}(x)$.
For $t = 1, \ldots, T$:

▷ Compute a strong base learner w.r.t. distribution $\mathcal{P}_t$ by finding n points in S and executing $\eta\text{-RERM}_{\mathcal{H}}$ on them.

(a) Find d points $S''_t \subseteq \bar{S}_{\mathcal{U}}$ such that any $h \in \mathcal{H}$ satisfying: $\forall(z, y) \in S''_t : |h(z) - y|^p \leq \eta(z, y)$, it holds that $\mathbb{E}_{(z,y)\sim\mathcal{P}_t}\left[\mathbb{I}\left\{|h(z) - y|^p \leq \eta(z, y) + \epsilon\right\}\right] \geq 1 - \epsilon$. (See the analysis for why this set exists).

(b) Let S'_t be the d points in S that S''_t originated from. Formally, $S''_t \subseteq \bigcup_{(x,y)\in S'_t}\bigcup\{(z, y) : z \in \mathcal{U}(x)\}$.

(c) Compute $\hat{h}_t = \eta\text{-RERM}_{\mathcal{H}}(S'_t, \boldsymbol{\eta}_{S'_t}, p)$.

▷ Make a multiplicative weights update on $\mathcal{P}_t$.

(d) For each $(z, y) \in \bar{S}_{\mathcal{U}}$:

$$P_{t+1}(z, y) \propto P_t(z, y)e^{-\xi\mathbb{I}\left\{\left|\hat{h}_t(z)-y\right|^p \leq \eta(z,y)\right\}}$$

Output: classifiers $\hat{h}_1, \ldots, \hat{h}_T$ and sets $S'_1, \ldots, S'_T$.

A uniformly 4ϵ-approximate adversarially robust sample compression scheme for S. The output of the algorithm is a sequence of functions $\hat{h}_1, \ldots, \hat{h}_T$, and the corresponding sets that encode them $S'_1, \ldots, S'_T$, where we predict with the

average of the returned hypotheses, $\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(\cdot)$. For $T \approx \log\left|\bar{S}_{\mathcal{U}}\right|$, we show that we have

$$\forall(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}} : \ \frac{1}{T}\sum_{i=1}^{T}\left|\hat{h}_i(\bar{z}) - \bar{y}\right|^p \leq \eta(\bar{z},\bar{y}) + 2\epsilon. \tag{15}$$

For any distribution $\mathcal{P}_t$ over $\bar{S}_{\mathcal{U}}$, we have a base learner $\hat{h}$, satisfying $\mathbb{E}_{(\bar{z},\bar{y})\sim\mathcal{P}_t}\left[\mathbb{I}\left\{\left|\hat{h}(\bar{z}) - \bar{y}\right|^p \leq \eta(\bar{z},\bar{y}) + \epsilon\right\}\right] \geq 1-\epsilon$, due to Theorem 3.4. Following standard analysis of MW / α-Boost (see Schapire & Freund (2013, Section 6)), for any $(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}}$, $1-\epsilon$ fraction of the base learners has an error within $\eta(\bar{z},\bar{y}) + \epsilon$. The loss is bounded by 1, so the other ϵ fraction can add an error of at most ϵ. The overall average loss of the base learners is upper bounded by $\eta(\bar{z},\bar{y}) + 2\epsilon$. Note that we can find these base learners in $\hat{\mathcal{H}}$, as defined in step $2(a)$ of the main algorithm. Crucially, we use strong base learners in order to ensure a low empirical loss of the average base learners.

From the covering argument (Equation (12)), we have

$$\forall(z,y) \in S_{\mathcal{U}} : \ \frac{1}{T}\sum_{i=1}^{T}\left|\hat{h}_i(z) - y\right|^p \leq \eta(z,y) + 4\epsilon. \tag{16}$$

Indeed, for any $(z,y) \in S_{\mathcal{U}}$ there exists $(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}}$, such that for any $h \in \mathcal{H}$,

$$\left|\left|h(z) - y\right|^p - \left|h(\bar{z}) - \bar{y}\right|^p\right| \leq \epsilon.$$

Specifically, it holds for $\left\{\hat{h}_1,\ldots,\hat{h}_T\right\} \subseteq \mathcal{H}$ and $h^\star \in \mathcal{H}$. Using the triangle inequality we have

$$\frac{1}{T}\sum_{i=1}^{T}\left|\hat{h}_i(z) - y\right|^p \leq \frac{1}{T}\sum_{i=1}^{T}\left|\hat{h}_i(\bar{z}) - \bar{y}\right|^p + \epsilon, \tag{17}$$

and

$$\eta(\bar{z},\bar{y}) = \left|h^\star(\bar{z}) - \bar{y}\right|^p \leq \left|h^\star(z) - y\right|^p + \epsilon = \eta(z,y). \tag{18}$$

Combining Equations (17) and (18) we get Equation (16). Finally, using the convexity of the ℓ_p loss, we have

$$\left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(z) - y\right|^p \leq \frac{1}{T}\sum_{i=1}^{T}\left|\hat{h}_i(z) - y\right|^p. \tag{19}$$

Finally, from Equations (16) and (19) we conclude a uniformly 4ϵ-approximate *adversarially robust* sample compression scheme for S,

$$\forall(z,y) \in S_{\mathcal{U}} : \ \left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(z) - y\right|^p \leq \eta(z,y) + 4\epsilon, \tag{20}$$

which implies that

$$\forall(x,y) \in S : \ \sup_{z\in\mathcal{U}(x)}\left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(x) - y\right|^p \leq \eta(x,y) + 4\epsilon.$$

We summarize the compression size. We have have $T = \mathcal{O}\big(\log\left|\bar{S}_{\mathcal{U}}\right|\big)$ predictors, where each one is representable by $d = \mathcal{O}\Big(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H},\epsilon/8p)\log^2\Big(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\Big)\Big)$ points. By counting the number of predictors using Equation (14), we get

$$
\begin{aligned}
\log\left(\left|\bar{S}_{\mathcal{U}}\right|\right) &\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon/p/p)\log^2\left(\frac{|\hat{\mathcal{H}}|}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p/p)}\right)\log^2(m)\right)\\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\log^2\left(\frac{|\hat{\mathcal{H}}|}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)\log^2(m)\\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\left(\frac{m}{d}\right)^d\right)\log^2(m)\\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\left(\log\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)+d\log\left(\frac{m}{d}\right)\right)^2\log^2(m)\\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\left(\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)+d\log\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)\log\left(\frac{m}{d}\right)+d^2\log^2\left(\frac{m}{d}\right)\right)\log^2(m)\\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\,d^2\log^2\left(\frac{m}{d}\right)\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)\log^2(m)\,.
\end{aligned}
\tag{21}
$$

We get a uniformly 4ϵ-approximate *adversarially robust* sample compression scheme for S of size

$$
\mathcal{O}\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)\,d^3\log^2\left(\frac{m}{d}\right)\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}\right)\log^2(m)\right).
$$

By plugging in $d=\mathcal{O}\left(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H},\epsilon/8p)\log^2\left(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\right)\right)$, we have

$$
\mathcal{O}\left(\frac{1}{\epsilon^3}\mathrm{fat}^3(\mathcal{H},\epsilon/8p)\,\mathrm{fat}^*(\mathcal{H},c\epsilon/p)\log^6\left(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\right)\log^2\left(\frac{m}{\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H},\epsilon/8p)\log^2\left(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\right)}\right)\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon/p)}\right)\log^2(m)\right).
$$

Let (κ,ρ) be the compression scheme and $|\kappa(S)|$ the compression size. Let $\widehat{\mathrm{Err}}_{\ell_p}(h;S)$ be the empirical loss of h on S with the ℓ_p robust loss. We can derive the error as follows,

$$
\begin{aligned}
\mathrm{Err}_{\ell_p}(\rho(\kappa(S));\mathcal{D}) &\overset{(i)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_p}(\rho(\kappa(S));S)+\sqrt{\frac{\left(|\kappa(S)|\log(m)+\log\frac{1}{\delta}\right)}{m}}\\
&\overset{(ii)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_p}(h^\star;S)+4\epsilon+\sqrt{\frac{\left(|\kappa(S)|\log(m)+\log\frac{1}{\delta}\right)}{m}}\\
&\overset{(iii)}{\lesssim} \mathrm{Err}_{\ell_p}(h^\star;\mathcal{D})+4\epsilon+\sqrt{\frac{\left(|\kappa(S)|\log(m)+\log\frac{1}{\delta}\right)}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{m}}\\
&\lesssim \mathrm{Err}_{\ell_p}(h^\star;\mathcal{D})+4\epsilon+\sqrt{\frac{\left(|\kappa(S)|\log(m)+\log\frac{1}{\delta}\right)}{m}},
\end{aligned}
$$

(i) follows from a generalization of sample compression scheme in the agnostic case, see Lemma A.6, (ii) follows Equation (20), (iii) follows from Hoeffding's inequality.

Take m sufficiently large such that

$$
\sqrt{\frac{|\kappa(S)|\log(m)+\log\frac{1}{\delta}}{m}}\lesssim\epsilon.
$$

Re-scale $\epsilon = \epsilon/5$ and plug in the compression size, we get sample complexity of size

$$\mathcal{M} = \mathcal{O}\left(\frac{1}{\epsilon^2}\left(|\kappa(S)|\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right),$$

where $|\kappa(S)|$ is upper bounded as follows

$$\mathcal{O}\left(\frac{1}{\epsilon^3}\mathrm{fat}^3(\mathcal{H},\epsilon/8p)\,\mathrm{fat}^*(\mathcal{H},c\epsilon/p)\log^6\left(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\right)\log^2\left(\frac{m}{\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H},\epsilon/8p)\log^2\left(\frac{p\cdot\mathrm{fat}(\mathcal{H},\epsilon/8p)}{\epsilon^2}\right)}\right)\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon/p)}\right)\log^2(m)\right).$$

We conclude the sample complexity

$$\mathcal{M} = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^5}\mathrm{fat}^3(\mathcal{H},c\epsilon/p)\,\mathrm{fat}^*(\mathcal{H},c\epsilon/p) + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right),$$

for some numerical constant $c \in (0,\infty)$. □

C. Proofs for Section 4: Better Sample Complexity for the ℓ_1 Loss

Proof of Theorem 4.1. Fix $\epsilon,\delta \in (0,1)$. Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$. Fix a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, and let $S = \{(x_i,y_i)\}_{i=1}^m$ be an i.i.d. sample from $\mathcal{D}$. We elaborate on each one of the steps as described in Algorithm 2.

1. Compute $h^\star \leftarrow \ell_1\text{-RERM}_{\mathcal{H}}(S)$ in order to get the set of cutoffs $\eta(x,y) = \sup_{z\in\mathcal{U}(x)}|h^\star(z)-y|$ for $(x,y)\in S$. Let $\boldsymbol{\eta}_S = (\eta(x_1,y_1),\ldots,\eta(x_m,y_m))$. Our goal is to construct a predictor with an empirical robust loss of $\eta(x,y)+\epsilon$ for any $(x,y)\in S$, which means that our predictor is an approximate Robust ERM.

2. Define the inflated training data set
$$S_{\mathcal{U}} = \bigcup_{i\in[n]}\left\{(z,y_{I(z)}) : z\in\mathcal{U}(x_i)\right\},$$
where $I(z) = \min\{i\in[m] : z\in\mathcal{U}(x_i)\}$. For $(z,y)\in S_{\mathcal{U}}$, let $\eta(z,y)$ be the $\eta(x,y)$ for which $z\in\mathcal{U}(x)$ and $y_{I(z)} = y$.

3. Discretize $S_{\mathcal{U}}$ to a finite set $\bar{S}_{\mathcal{U}}$ as following.

   (a) Define a set of functions, such that each function is defined by η-$\mathrm{RERM}_{\mathcal{H}}$ optimizer on $d = \mathcal{O}\left(\mathrm{fat}(\mathcal{H},\epsilon/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H},\epsilon/8)}{\epsilon^2}\right)\right)$ points from S.
   $$\hat{\mathcal{H}} = \{\eta\text{-}\mathrm{RERM}_{\mathcal{H}}(S',\boldsymbol{\eta}_{S'},1) : S'\subseteq S, |S'| = d\}.$$
   Recall the definition of η-$\mathrm{RERM}_{\mathcal{H}}$, see Equation (5).
   In order to understand what this definition of $\hat{\mathcal{H}}$ serves for, see step 4 below. The cardinality of this class is bounded as follows
   $$|\hat{\mathcal{H}}| \approx \binom{m}{d} \lesssim \left(\frac{m}{d}\right)^d. \tag{22}$$

   (b) A discretization $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ will be defined by covering of the dual class in d_∞ norm. Let $L^1_{\hat{\mathcal{H}}}$ be the L_1 loss class of $\hat{\mathcal{H}}$, namely, $L^1_{\hat{\mathcal{H}}} = \left\{\mathcal{Z}\times\mathcal{Y}\ni(z,y)\mapsto|h(z)-y| : h\in\hat{\mathcal{H}}\right\}$. The *dual class* of $L^1_{\hat{\mathcal{H}}}$, ${L^1_{\hat{\mathcal{H}}}}^* \subseteq [0,1]^{\hat{\mathcal{H}}}$, is defined as the set of all functions $f_{(z,y)} : \hat{\mathcal{H}}\to[0,1]$ such that $f_{(z,y)}(h) = \left|h(z)-y\right|$, for any $(z,y)\in S_{\mathcal{U}}$. Formally, ${L^1_{\hat{\mathcal{H}}}}^* = \left\{f_{(z,y)} : (z,y)\in S_{\mathcal{U}}\right\}$, where $f_{(z,y)} = \left(f_{(z,y)}(h_1),\ldots,f_{(z,y)}(h_{|\hat{\mathcal{H}}|})\right)$. We take $\bar{S}_{\mathcal{U}}\subseteq S_{\mathcal{U}}$ to be a minimal ϵ-cover for $S_{\mathcal{U}}$ in d_∞,
   $$\sup_{(z,y)\in S_{\mathcal{U}}}\inf_{(\bar{z},\bar{y})\in\bar{S}_{\mathcal{U}}}, \left\|f_{(z,y)} - f_{(\bar{z},\bar{y})}\right\|_\infty \le \epsilon. \tag{23}$$

Let $\mathrm{fat}^*(\mathcal{H}, \epsilon)$ be the dual ϵ-fat-shattering of $\mathcal{H}$. Using Lemma A.4, we can bound the dual fat-shattering of the $L^1_{\mathcal{H}}$ loss class by the fat-shattering of $\mathcal{H}$,

$$\mathrm{fat}^*\left(L^1_{\mathcal{H}}, \epsilon\right) \leq \log^2(m)\, \mathrm{fat}^*(\mathcal{H}, \epsilon)\,. \tag{24}$$

By applying a covering number argument from Lemma A.3 (taking $a = 1$) on the dual space, and upper bounding the dual fat-shattering of the L_1 loss class as in Equation (24), we have the following bound

$$\begin{aligned}
\left|\bar{S}_{\mathcal{U}}\right| &= \mathcal{N}(\epsilon, S_{\mathcal{U}}, d_\infty) \\
&\lesssim \exp\left(\mathrm{fat}^*\left(L^1_{\mathcal{H}}, c\epsilon\right) \log\left(\frac{|\hat{\mathcal{H}}|}{\epsilon \cdot \mathrm{fat}^*(L^1_{\mathcal{H}}, c\epsilon)}\right) \log\left(\frac{|\hat{\mathcal{H}}|}{\mathrm{fat}^*(L^1_{\mathcal{H}}, c\epsilon)}\right)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon) \log\left(\frac{|\hat{\mathcal{H}}|}{\epsilon \cdot \mathrm{fat}^*(\mathcal{H}, c\epsilon)}\right) \log\left(\frac{|\hat{\mathcal{H}}|}{\mathrm{fat}^*(\mathcal{H}, c\epsilon)}\right) \log^2(m)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon) \log^2\left(\frac{|\hat{\mathcal{H}}|}{\epsilon \cdot \mathrm{fat}^*(\mathcal{H}, c\epsilon)}\right) \log^2(m)\right),
\end{aligned} \tag{25}$$

where $c \in (0, \infty)$ is a numerical constant, derived from the covering argument in Lemma A.3.

4. Compute a modified version of the real-valued boosting algorithm $\mathrm{MedBoost}$ (Kégl, 2003; Hanneke et al., 2019) on the discretized set $\bar{S}_{\mathcal{U}}$. The output of the algorithm is a uniformly ϵ-approximate sample compression scheme for the set $\bar{S}_{\mathcal{U}}$, for $\approx \log\left(\left|\bar{S}_{\mathcal{U}}\right|\right)$ boosting rounds. Moreover, the weak learners are chosen from the set $\hat{\mathcal{H}}$. Once we have these weak learners, the guarantee of the algorithm follows from Hanneke et al. (2019, Corollary 6). We should explain why we have a weak learner for any distribution over $\bar{S}_{\mathcal{U}}$.

The existence of weak learners in $\hat{\mathcal{H}}$. Let $d = \mathcal{O}\left(\mathrm{fat}(\mathcal{H}, \epsilon/8) \log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \epsilon/8)}{\epsilon^2}\right)\right)$ and let $\eta(z, y)$ be the $\eta(x, y)$ for which $z \in \mathcal{U}(x)$ as defined in step 2. Taking $\delta = 1/3$, we know that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, upon receiving an i.i.d. sample S'' from $\mathcal{P}$ of size d, with probability $2/3$ over sampling S'' from $\mathcal{P}$, for any $h \in \mathcal{H}$ satisfying $\forall (z, y) \in S'' :\ |h(z) - y| \leq \eta(z, y)$, it holds that $\mathbb{P}_{(z,y)\sim\mathcal{P}}((z, y) : |h(z) - y| > \eta(z, y) + \epsilon) \leq 1/3$. That is, such a function is a $(\epsilon, 1/6)$-weak learner for $\mathcal{P}$ and $h^\star$ (computed in step 1). We can conclude that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, there exists a set of points $S'' \subseteq \bar{S}_{\mathcal{U}}$ of size d that defines a weak learner for $\mathcal{P}$ and $h^\star$.

Furthermore, we can find these weak learners in $\hat{\mathcal{H}}$ as follows. Let S' be the d points in S that the perturbed points S'' originated from. That is, $S'' \subseteq \bigcup_{(x,y)\in S'} \bigcup \{(z, y) : z \in \mathcal{U}(x)\}$. Therefore, we can conclude that $\hat{h} = \eta\text{-}\mathrm{RERM}_{\mathcal{H}}(S', \boldsymbol{\eta}_{S'})$ is a weak learner, and can be found in $\hat{\mathcal{H}}$. So, we can think of $\hat{\mathcal{H}}$ as a pool of weak learners for any possible distribution over the discretized set $\bar{S}_{\mathcal{U}}$.

Algorithm 5 Modified MedBoost

Input: $\mathcal{H}, S, \bar{S}_{\mathcal{U}}$.
Parameters: $\epsilon, T, \boldsymbol{\eta}_S = (\eta(x_1, y_1), \ldots, \eta(x_m, y_m))$ for $(x_i, y_i) \in S$.
Algorithms used: Robust ERM for the η-ball robust loss: η-$\mathrm{RERM}_{\mathcal{H}}$ (Equation (5)).
Initialize $\mathcal{P}_1$ = Uniform($\bar{S}_{\mathcal{U}}$), $d = \mathcal{O}\Big(\mathrm{fat}(\mathcal{H}, \epsilon/8) \log^2\Big(\frac{\mathrm{fat}(\mathcal{H}, \epsilon/8)}{\epsilon^2}\Big)\Big)$, $\eta(z, y)$ is the $\eta(x, y)$ for which $z \in \mathcal{U}(x)$ as defined in step 2 of the main algorithm.
For $t = 1, \ldots, T$:

▷ Compute a weak base learner w.r.t. distribution $\mathcal{P}_t$ by finding d points in S and executing η-$\mathrm{RERM}_{\mathcal{H}}$ on them.

(a) Find d points $S''_t \subseteq \bar{S}_{\mathcal{U}}$ such that any $h \in \mathcal{H}$ satisfying: $\forall (z, y) \in S''_t : \ |h(z) - y| \leq \eta(z, y)$, it holds that $\mathbb{E}_{(z,y)\sim\mathcal{P}_t}\Big[\mathbb{I}\Big\{|h(z) - y| \geq \eta(z, y) + \epsilon\Big\}\Big] \leq 1/3$. (See the analysis for why this set exists).

(b) Let S'_t be the d points in S that S''_t originated from. Formally, $S''_t \subseteq \bigcup_{(x,y)\in S'_t} \bigcup \{(z, y) : z \in \mathcal{U}(x)\}$.

(c) Compute $\hat{h}_t = \eta$-$\mathrm{RERM}_{\mathcal{H}}(S'_t, \boldsymbol{\eta}_{S'_t}, 1)$. From steps (a) and (b), it follows that $\hat{h}_t$ is a $(\epsilon, 1/6)$-weak learner with respect to the distribution $\mathcal{P}_t$ over $\bar{S}_{\mathcal{U}}$.

▷ Update the weight of the weak learner in the ensemble and make a multiplicative weights update on $\mathcal{P}_t$.

(d) For $i = 1, \ldots, n = \big|\bar{S}_{\mathcal{U}}\big|$:

i. Set
$$w_i^{(t)} = 1 - 2\mathbb{I}\Big[\Big|\hat{h}_t(z_i) - y_i\Big| > \eta(z_i, y_i) + \epsilon\Big].$$

ii. Set
$$\alpha_t = \frac{1}{2} \log\left(\frac{(1 - 1/6) \sum_{i=1}^n \mathcal{P}_t(z_i, y_i)\, \mathbb{I}\Big[w_i^{(t)} = 1\Big]}{(1 + 1/6) \sum_{i=1}^n \mathcal{P}_t(z_i, y_i)\, \mathbb{I}\Big[w_i^{(t)} = -1\Big]}\right).$$

iii.
- If $\alpha_t = \infty$: return T copies of h_t, $(\alpha_1 = 1, \ldots, \alpha_T = 1)$, and S'_t.
- Else:
$$P_{t+1}(z_i, y_i) = P_t(z_i, y_i) \frac{\exp(-\alpha_t w_i^t)}{\sum_{j=1}^n \mathcal{P}_t(z_j, y_j) \exp\big(-\alpha_t w_j^t\big)}.$$

Output: Hypotheses $\hat{h}_1, \ldots, \hat{h}_T$, coefficients $\alpha_1, \ldots, \alpha_T$ and sets $S'_1, \ldots, S'_T$.

A uniformly 3ϵ-approximate adversarially robust sample compression scheme for S. The output of MedBoost is a uniformly ϵ-approximate sample compression scheme for the set $\bar{S}_{\mathcal{U}}$. We show that this is a uniformly 2ϵ-approximate *adversarially robust* sample compression scheme for S, that is, a sample compression for S scheme with respect to the robust loss.

For $T \approx \log\big|\bar{S}_{\mathcal{U}}\big|$ boosting rounds, it follows from Hanneke et al. (2019, Corollary 6) that the output of the algorithm satisfy

$$\forall (\bar{z}, \bar{y}) \in \bar{S}_{\mathcal{U}} : \ \Big|\mathrm{Med}\big(\hat{h}_1(\bar{z}), \ldots, \hat{h}_T(\bar{z}); \alpha_1, \ldots, \alpha_T\big) - \bar{y}\Big| \leq \eta(\bar{z}, \bar{y}) + \epsilon, \tag{26}$$

$\mathrm{Med}\big(\hat{h}_1(\bar{z}), \ldots, \hat{h}_T(\bar{z}); \alpha_1, \ldots, \alpha_T\big)$ is the weighted median of $\hat{h}_1, \ldots, \hat{h}_T$ with weights $\alpha_1, \ldots, \alpha_T$. From the covering argument (Equation (23)), this implies that

$$\forall (z, y) \in S_{\mathcal{U}} : \ \Big|\mathrm{Med}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z); \alpha_1, \ldots, \alpha_T\big) - y\Big| \leq \eta(z, y) + 3\epsilon. \tag{27}$$

Indeed, for any $(z, y) \in S_{\mathcal{U}}$ there exists $(\bar{z}, \bar{y}) \in \bar{S}_{\mathcal{U}}$, such that for any $h \in \mathcal{H}$,

$$\Big|\big|h(z) - y\big| - \big|h(\bar{z}) - \bar{y}\big|\Big| \leq \epsilon.$$

Specifically, it holds for $\Big\{\hat{h}_1, \ldots, \hat{h}_T\Big\} \subseteq \mathcal{H}$ and $h^\star \in \mathcal{H}$.

So,

$$\begin{aligned}
\left|\mathrm{Med}\big(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\big)-y\right| &\overset{(a)}{=} \left|\mathrm{Med}\big(\hat{h}_1(z)-y,\ldots,\hat{h}_T(z)-y;\alpha_1,\ldots,\alpha_T\big)\right| \\
&\overset{(b)}{\leq} \left|\mathrm{Med}\big(\hat{h}_1(\bar{z})-\bar{y},\ldots,\hat{h}_T(\bar{z})-\bar{y};\alpha_1,\ldots,\alpha_T\big)\right|+\epsilon \\
&\overset{(c)}{=} \left|\mathrm{Med}\big(\hat{h}_1(\bar{z}),\ldots,\hat{h}_T(\bar{z});\alpha_1,\ldots,\alpha_T\big)-\bar{y}\right|+\epsilon \\
&\overset{(d)}{\leq} |h^\star(\bar{z})-\bar{y}|+2\epsilon \\
&\overset{(e)}{\leq} |h^\star(z)-y|+3\epsilon \\
&\overset{(f)}{=} \eta(z,y)+3\epsilon,
\end{aligned} \tag{28}$$

(a)+(c) follow since the median is translation invariant, (b)+(e) follow from the covering argument, (d) holds since the returned function by $\mathrm{MedBoost}$ is a uniformly ϵ-approximate sample compression for $\bar{S}_{\mathcal{U}}$, (f) follows from the definition in step 2 of the algorithm.

Finally, from Equation (28) we conclude a uniformly 3ϵ-approximate *adversarially robust* sample compression scheme for S,

$$\forall (x,y)\in S: \sup_{z\in\mathcal{U}(x)}\left|\mathrm{Med}\big(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\big)-y\right| \leq \eta(x,y)+3\epsilon. \tag{29}$$

We summarize the compression size. We have have $T=\mathcal{O}\big(\log\big|\bar{S}_{\mathcal{U}}\big|\big)$ predictors, where each one is representable by $d=\mathcal{O}\Big(\mathrm{fat}(\mathcal{H},\epsilon/8)\log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\epsilon/8)}{\epsilon}\Big)\Big)$ points. By counting the number of predictors using Equation (25), we get

$$\begin{aligned}
\log\big(\big|\bar{S}_{\mathcal{U}}\big|\big) &\lesssim \log\left(\exp\left(\mathrm{fat}^*(\mathcal{H},c\epsilon)\log^2\left(\frac{|\hat{\mathcal{H}}|}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon)}\right)\log^2(m)\right)\right) \\
&\lesssim \mathrm{fat}^*(\mathcal{H},c\epsilon)\log^2\left(\frac{|\hat{\mathcal{H}}|}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon)}\right)\log^2(m) \\
&\lesssim \mathrm{fat}^*(\mathcal{H},c\epsilon)\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon)}\left(\frac{m}{d}\right)^d\right)\log^2(m) \\
&\lesssim \mathrm{fat}^*(\mathcal{H},c\epsilon)\left(\log\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon)}\right)+d\log\left(\frac{m}{d}\right)\right)^2\log^2(m) \\
&\lesssim \mathrm{fat}^*(\mathcal{H},c\epsilon)\left(\log^2\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon)}\right)+d\log\left(\frac{1}{\epsilon\cdot\mathrm{fat}^*(\mathcal{H},c\epsilon)}\right)\log\left(\frac{m}{d}\right)+d^2\log^2\left(\frac{m}{d}\right)\right)\log^2(m)\,.
\end{aligned}$$

All together we have a compression of size $\mathcal{O}\big(d\log\big(\big|\bar{S}_{\mathcal{U}}\big|\big)\big)$, which is already sufficient for deriving generalization. We can reduce further the number of predictors to be *independent* of the sample size, thereby reducing the sample compression size and improving the sample complexity.

5. We follow the sparsification method suggested by Hanneke et al. (2019). The idea is that by sampling functions from the ensemble, we can guarantee via a uniform convergence for the dual space, that with high probability it is sufficient to have roughly $\approx \mathrm{fat}^*(\mathcal{H},\beta)$ predictors.

   For $\alpha_1,\ldots,\alpha_T\in[0,1]$ with $\sum_{t=1}^T\alpha_t=1$, we denote the categorical distribution by $\mathrm{Cat}(\alpha_1,\ldots,\alpha_T)$, which is a discrete distribution on the set $[T]$ with probability of α_t on $t\in[T]$. The inputs to the algorithm are $\tau(x,y)=\eta(x,y)+4\epsilon$ and $k=\mathcal{O}\big(\mathrm{fat}^*(\mathcal{H},c\epsilon)\log^2\big(\mathrm{fat}^*(\mathcal{H},c\epsilon)/\epsilon^2\big)\big)$, where $c\in(0,\infty)$ is a numerical constant.

Algorithm 6 Sparsify

Input: Hypotheses $\hat{h}_1, \ldots, \hat{h}_T$, coefficients $\alpha_1, \ldots, \alpha_T$, $S = \{(x_i, y_i\}_{i=1}^m$.
Parameter: $\epsilon, k, \boldsymbol{\tau} = (\tau(x_1, y_1), \ldots, \tau(x_m, y_m))$.

(a) Let $\alpha'_t = \alpha_t / \sum_{s=1}^T \alpha_s$.

(b) **Repeat**:

i. Sample $(J_1, \ldots, J_k) \sim \mathrm{Cat}(\alpha'_1, \ldots, \alpha'_{\mathrm{T}})^{\mathrm{k}}$.

ii. Let $\mathcal{F} = \{h_{J_1}, \ldots, h_{J_k}\}$.

iii. **Until** $\forall (x, y) \in S : \left| \left\{ f \in \mathcal{F} : \sup_{z \in \mathcal{U}(x)} |f(z) - y| > \tau(x, y) \right\} \right| < k/2$.

Output: Hypotheses $h_{J_1}, \ldots, h_{J_k}$.

The sparsification method returns with high probability function $\{f_1, \ldots, f_k\}$, such that

$$\forall (x, y) \in S : \sup_{z \in \mathcal{U}(x)} \left| \mathrm{Med}\big(f_1(x), \ldots, f_k(x)\big) - y \right| \leq \eta(x, y) + 4\epsilon. \tag{30}$$

We get a uniformly 4ϵ-approximate *adversarially robust* sample compression scheme for S, where we have $\mathcal{O}\big(\mathrm{fat}^*(\mathcal{H}, c\epsilon) \log^2\big(\mathrm{fat}^*(\mathcal{H}, c\epsilon)/\epsilon^2\big)\big)$ functions, and each function is representable by $\mathcal{O}\big(\mathrm{fat}(\mathcal{H}, \epsilon/8) \log^2\big(\mathrm{fat}(\mathcal{H}, \epsilon/8)/\epsilon^2\big)\big)$ points, therefore, the compression set size is

$$\mathrm{fat}(\mathcal{H}, \epsilon/8)\, \mathrm{fat}^*(\mathcal{H}, c\epsilon) \log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \epsilon/8)}{\epsilon^2}\right) \log^2\left(\frac{\mathrm{fat}^*(\mathcal{H}, c\epsilon)}{\epsilon^2}\right).$$

Let (κ, ρ) be the compression scheme and $|\kappa(S)|$ the compression size. Let $\widehat{\mathrm{Err}}_{\ell_1}(h; S)$ be the empirical loss of h on S with the ℓ_1 robust loss. We can derive the error as follows,

$$\begin{aligned}
\mathrm{Err}_{\ell_1}(\rho(\kappa(S)); \mathcal{D}) &\overset{(i)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_1}(\rho(\kappa(S)); S) + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log \frac{1}{\delta}\big)}{m}} \\
&\overset{(ii)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_1}(h^\star; S) + 4\epsilon + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log \frac{1}{\delta}\big)}{m}} \\
&\overset{(iii)}{\lesssim} \mathrm{Err}_{\ell_1}(h^\star; \mathcal{D}) + 4\epsilon + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log \frac{1}{\delta}\big)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{m}} \\
&\lesssim \mathrm{Err}_{\ell_1}(h^\star; \mathcal{D}) + 4\epsilon + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log \frac{1}{\delta}\big)}{m}},
\end{aligned}$$

(i) follows from a generalization of sample compression scheme in the agnostic case, see Lemma A.6, (ii) follows Equation (30), (iii) follows from Hoeffding's inequality.

Take m sufficiently large such that

$$\sqrt{\frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m}} \lesssim \epsilon.$$

Re-scale $\epsilon = \epsilon/5$ and plug in the compression size, we get sample complexity of size

$$\mathcal{M} = \mathcal{O}\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}(\mathcal{H}, c\epsilon)\, \mathrm{fat}^*(\mathcal{H}, c\epsilon) \log^2\left(\frac{\mathrm{fat}(\mathcal{H}, c\epsilon)}{\epsilon^2}\right) \log^2\left(\frac{\mathrm{fat}^*(\mathcal{H}, c\epsilon)}{\epsilon^2}\right) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right),$$

for some numerical constant $c \in (0, \infty)$. □

D. Proofs for Section 5: Robust (η, β)-Regression

D.1. Realizable

Proof of Theorem 5.1. Fix $\epsilon, \delta, \beta \in (0,1)$ and $\eta \in [0,1]$. Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$. Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and let $S = \{(x_i, y_i)\}_{i=1}^m$ be an i.i.d. sample from $\mathcal{D}$.

We elaborate on each one of the steps as described in Algorithm 3.

1. Define the inflated training data set
$$S_{\mathcal{U}} = \bigcup_{i \in [n]} \left\{ (z, y_{I(z)}) : z \in \mathcal{U}(x_i) \right\},$$
where $I(z) = \min\{i \in [m] : z \in \mathcal{U}(x_i)\}$.

2. Discretize $S_{\mathcal{U}}$ to a finite set $\bar{S}_{\mathcal{U}}$ as following.

   (a) Define a set of functions, such that each function is robustly accurate on $d = \mathcal{O}\left(\mathrm{fat}(\mathcal{H}, \beta/8) \log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \beta/8)}{\beta}\right)\right)$ points in S, with respect to the η-ball robust loss,
$$\hat{\mathcal{H}} = \{\eta\text{-}\mathrm{RERM}_{\mathcal{H}}(S') : S' \subseteq S, |S'| = d\}.$$
   Recall the definition of η-$\mathrm{RERM}_{\mathcal{H}}$, see Equation (5). In order to understand what this definition of $\hat{\mathcal{H}}$ serves for, see step 3 below. The cardinality of this class is bounded as follows
$$|\hat{\mathcal{H}}| \approx \binom{m}{d} \lesssim \left(\frac{m}{d}\right)^d. \tag{31}$$

   (b) A discretization $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ will be defined by covering of the dual class in d_∞ norm. Let $L^1_{\hat{\mathcal{H}}}$ be the L_1 loss class of $\hat{\mathcal{H}}$, namely, $L^1_{\hat{\mathcal{H}}} = \left\{\mathcal{Z} \times \mathcal{Y} \ni (z,y) \mapsto |h(z) - y| : h \in \hat{\mathcal{H}}\right\}$. The *dual class* of $L^1_{\hat{\mathcal{H}}}$, ${L^1_{\hat{\mathcal{H}}}}^* \subseteq [0,1]^{\hat{\mathcal{H}}}$, is defined as the set of all functions $f_{(z,y)} : \hat{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|$, for any $(z,y) \in S_{\mathcal{U}}$. Formally, ${L^1_{\hat{\mathcal{H}}}}^* = \left\{f_{(z,y)} : (z,y) \in S_{\mathcal{U}}\right\}$, where $f_{(z,y)} = \left(f_{(z,y)}(h_1), \ldots, f_{(z,y)}(h_{|\hat{\mathcal{H}}|})\right)$. We take $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ to be a minimal β-cover for $S_{\mathcal{U}}$ in d_∞,
$$\sup_{(z,y) \in S_{\mathcal{U}}} \inf_{(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}}}, \left\|f_{(z,y)} - f_{(\bar{z},\bar{y})}\right\|_\infty \leq \beta. \tag{32}$$
   Let $\mathrm{fat}^*(\mathcal{H}, \beta)$ be the dual β-fat-shattering of $\mathcal{H}$. Using Lemma A.4, we can bound the dual fat-shattering of the $L^1_{\mathcal{H}}$ loss class by the fat-shattering of $\mathcal{H}$,
$$\mathrm{fat}^*\left(L^1_{\mathcal{H}}, \beta\right) \leq \log^2(m)\, \mathrm{fat}^*(\mathcal{H}, \beta). \tag{33}$$
   By applying a covering number argument from Lemma A.3 (taking $a = 1$) on the dual space, and upper bounding the dual fat-shattering of the L_1 loss class as in Equation (33), we have the following bound
$$\begin{aligned}
\left|\bar{S}_{\mathcal{U}}\right| &= \mathcal{N}(\beta, S_{\mathcal{U}}, d_\infty) \\
&\lesssim \exp\left(\mathrm{fat}^*\left(L^1_{\mathcal{H}}, c\beta\right) \log\left(\frac{|\hat{\mathcal{H}}|}{\beta \cdot \mathrm{fat}^*(L^1_{\mathcal{H}}, c\beta)}\right) \log\left(\frac{|\hat{\mathcal{H}}|}{\mathrm{fat}^*(L^1_{\mathcal{H}}, c\beta)}\right)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\beta) \log\left(\frac{|\hat{\mathcal{H}}|}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)}\right) \log\left(\frac{|\hat{\mathcal{H}}|}{\mathrm{fat}^*(\mathcal{H}, c\beta)}\right) \log^2(m)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\left(\frac{|\hat{\mathcal{H}}|}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)}\right) \log^2(m)\right),
\end{aligned} \tag{34}$$

where $c \in (0, \infty)$ is a numerical constant, derived from the covering argument in Lemma A.3.

3. Compute a modified version of the real-valued boosting algorithm $\mathrm{MedBoost}$ (Algorithm 5) on the discretized set $\bar{S}_{\mathcal{U}}$. The inputs to the algorithm are as follows. Set $\epsilon = \beta$, $\boldsymbol{\eta}_S = (\eta, \ldots, \eta)$, and $T \approx \log\left(\left|\bar{S}_{\mathcal{U}}\right|\right)$ rounds of boosting.

   The output of the algorithm is a uniform ϵ-approximate sample compression scheme for the set $\bar{S}_{\mathcal{U}}$. Moreover, the weak learners are chosen from the set $\hat{\mathcal{H}}$. Once we have these weak learners, the guarantee of the algorithm follows from Hanneke et al. (2019, Corollary 6). We should explain why we have a weak learner for any distribution over $\bar{S}_{\mathcal{U}}$.

   The existence of weak learners in $\hat{\mathcal{H}}$. From Theorem A.2, taking $\epsilon = \delta = 1/3$, we know that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, upon receiving an i.i.d. sample S'' from $\mathcal{P}$ of size $\mathcal{O}\left(\mathrm{fat}(\mathcal{H}, \beta/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H},\beta/8)}{\beta}\right)\right)$, with probability $2/3$ over sampling S'' from $\mathcal{P}$, for any $h \in \mathcal{H}$ with $\forall (z,y) \in S'' : \ |h(z) - y| \leq \eta$, it holds that $\mathbb{P}_{(z,y)\sim\mathcal{P}}\{(z,y) : |h(z) - y| > \eta + \beta\} \leq 1/3$. That is, such a function is a $(\eta + \beta, 1/6)$-weak learner for $\mathcal{P}$ (see Definition 4.2). We can conclude that for any distribution $\mathcal{P}$ on $\bar{S}_{\mathcal{U}}$, there exists a set of points $S'' \subseteq \bar{S}_{\mathcal{U}}$ of size $\mathcal{O}\left(\mathrm{fat}(\mathcal{H}, \beta/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H},\beta/8)}{\beta}\right)\right)$ that defines a weak learner for $\mathcal{P}$.

   Moreover, we can find these weak learners in $\hat{\mathcal{H}}$ as follows. Let S' be the $\mathcal{O}\left(\mathrm{fat}(\mathcal{H}, \beta/8)\log^2\left(\frac{\mathrm{fat}(\mathcal{H},\beta/8)}{\beta}\right)\right)$ points in S that the perturbed points S'' originated from. That is, $S'' \subseteq \bigcup_{(x,y)\in S'} \bigcup \{(z,y) : z \in \mathcal{U}(x)\}$. Therefore, we can conclude that $\hat{h} = \eta\text{-}\mathrm{RERM}_{\mathcal{H}}(S')$ is a weak learner, and can be found in $\hat{\mathcal{H}}$. So, we can think of $\hat{\mathcal{H}}$ as a pool of weak learners for any possible distribution over the discretized set $\bar{S}_{\mathcal{U}}$.

   A uniformly 3β-approximate adversarially robust sample compression scheme for S. The output of $\mathrm{MedBoost}$ is a uniformly β-approximate sample compression scheme for the set $\bar{S}_{\mathcal{U}}$. We show that this is a uniformly 2β-approximate *adversarially robust* sample compression scheme for S, that is, a sample compression for S scheme with respect to the robust loss.

   For $T \approx \log\left|\bar{S}_{\mathcal{U}}\right|$ boosting rounds, it follows from Hanneke et al. (2019, Corollary 6) that the output of the algorithm satisfy

$$\forall (\bar{z}, \bar{y}) \in \bar{S}_{\mathcal{U}} : \ \left|\mathrm{Med}\big(\hat{h}_1(\bar{z}), \ldots, \hat{h}_T(\bar{z}); \alpha_1, \ldots, \alpha_T\big) - \bar{y}\right| \leq \eta(\bar{z}, \bar{y}) + \beta, \tag{35}$$

   $\mathrm{Med}\big(\hat{h}_1(\bar{z}), \ldots, \hat{h}_T(\bar{z}); \alpha_1, \ldots, \alpha_T\big)$ is the weighted median of $\hat{h}_1, \ldots, \hat{h}_T$ with weights $\alpha_1, \ldots, \alpha_T$. From the covering argument (Equation (23)), this implies that

$$\forall (z, y) \in S_{\mathcal{U}} : \ \left|\mathrm{Med}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z); \alpha_1, \ldots, \alpha_T\big) - y\right| \leq \eta(z, y) + 3\beta. \tag{36}$$

   Indeed, for any $(z, y) \in S_{\mathcal{U}}$ there exists $(\bar{z}, \bar{y}) \in \bar{S}_{\mathcal{U}}$, such that for any $h \in \mathcal{H}$,

$$\Big|\big|h(z) - y\big| - \big|h(\bar{z}) - \bar{y}\big|\Big| \leq \beta.$$

   Specifically, it holds for $\left\{\hat{h}_1, \ldots, \hat{h}_T\right\} \subseteq \mathcal{H}$ and $h^\star \in \mathcal{H}$.

   So,

$$\begin{aligned}
\left|\mathrm{Med}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z); \alpha_1, \ldots, \alpha_T\big) - y\right| &\overset{(a)}{=} \left|\mathrm{Med}\big(\hat{h}_1(z) - y, \ldots, \hat{h}_T(z) - y; \alpha_1, \ldots, \alpha_T\big)\right| \\
&\overset{(b)}{\leq} \left|\mathrm{Med}\big(\hat{h}_1(\bar{z}) - \bar{y}, \ldots, \hat{h}_T(\bar{z}) - \bar{y}; \alpha_1, \ldots, \alpha_T\big)\right| + \beta \\
&\overset{(c)}{=} \left|\mathrm{Med}\big(\hat{h}_1(\bar{z}), \ldots, \hat{h}_T(\bar{z}); \alpha_1, \ldots, \alpha_T\big) - \bar{y}\right| + \beta \\
&\overset{(d)}{\leq} |h^\star(\bar{z}) - \bar{y}| + 2\beta \\
&\overset{(e)}{\leq} |h^\star(z) - y| + 3\beta \\
&\overset{(f)}{=} \eta(z, y) + 3\beta,
\end{aligned}$$

   (a)+(c) follow since the median is translation invariant, (b)+(e) follow from the covering argument, (d) holds since the returned function by $\mathrm{MedBoost}$ is a uniformly β-approximate sample compression for $\bar{S}_{\mathcal{U}}$, (f) follows from the definition in step 2 of the algorithm.

Finally, from Equation (36) we conclude a uniformly 3β-approximate *adversarially robust* sample compression scheme for S,

$$\forall (x,y) \in S: \sup_{z \in \mathcal{U}(x)} \left| \mathrm{Med}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z); \alpha_1, \ldots, \alpha_T\big) - y \right| \leq \eta(x,y) + 3\beta. \tag{37}$$

We summarize the compression size. We have have $T = \mathcal{O}\big(\log\big|\bar{S}_{\mathcal{U}}\big|\big)$ predictors, where each one is representable by $d = \mathcal{O}\Big(\mathrm{fat}(\mathcal{H}, \beta/8) \log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\beta/8)}{\beta}\Big)\Big)$ points. By counting the number of predictors using Equation (25), we get

$$\begin{aligned}
\log\big(\big|\bar{S}_{\mathcal{U}}\big|\big) &\lesssim \log\left(\exp\left(\mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\left(\frac{|\hat{\mathcal{H}}|}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)}\right) \log^2(m)\right)\right) \\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\left(\frac{|\hat{\mathcal{H}}|}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)}\right) \log^2(m) \\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\left(\frac{1}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)} \left(\frac{m}{d}\right)^d\right) \log^2(m) \\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\beta) \left(\log\left(\frac{1}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)}\right) + d \log\left(\frac{m}{d}\right)\right)^2 \log^2(m) \\
&\lesssim \mathrm{fat}^*(\mathcal{H}, c\beta) \left(\log^2\left(\frac{1}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)}\right) + d \log\left(\frac{1}{\beta \cdot \mathrm{fat}^*(\mathcal{H}, c\beta)}\right) \log\left(\frac{m}{d}\right) + d^2 \log^2\left(\frac{m}{d}\right)\right) \log^2(m).
\end{aligned}$$

All together we have a compression of size $\mathcal{O}\big(d \log\big(\big|\bar{S}_{\mathcal{U}}\big|\big)\big)$, which is already sufficient for deriving generalization. We can reduce further the number of predictors to be *independent* of the sample size, thereby reducing the sample compression size and improving the sample complexity.

4. Compute the sparsification method (Algorithm 6). The idea is that by sampling functions from the ensemble, we can guarantee via a uniform convergence for the dual space, that with high probability it is sufficient to have roughly $\approx \mathrm{fat}^*(\mathcal{H}, \beta)$ predictors. Applying Hanneke et al. (2019, Theorem 10) with the parameters $\tau = \eta + 4\beta$ and $k = \mathcal{O}\big(\mathrm{fat}^*(\mathcal{H}, c\beta) \log^2(\mathrm{fat}^*(\mathcal{H}, c\beta)/\beta)\big)$, where $c \in (0, \infty)$ is a numerical constant, the sparsification method returns with high probability function $\{f_1, \ldots, f_k\}$, such that

$$\forall (x,y) \in S: \sup_{z \in \mathcal{U}(x)} \left| \mathrm{Med}\big(f_1(z), \ldots, f_k(z)\big) - y \right| \leq \eta + 4\beta.$$

We get a uniformly 4β-approximate *adversarially robust* sample compression scheme for S, where we have $\mathcal{O}\big(\mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\big(\mathrm{fat}^*(\mathcal{H}, c\beta)/\beta^2\big)\big)$ functions, and each function is representable by $\mathcal{O}\big(\mathrm{fat}(\mathcal{H}, \beta/8) \log^2\big(\mathrm{fat}(\mathcal{H}, \beta/8)/\beta^2\big)\big)$ points, therefore, the compression set size is

$$\mathrm{fat}(\mathcal{H}, \beta/8)\, \mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \beta/8)}{\beta^2}\right) \log^2\left(\frac{\mathrm{fat}^*(\mathcal{H}, c\beta)}{\beta^2}\right).$$

Let (κ, ρ) be the compression scheme and $|\kappa(S)|$ the compression size. Let $\widehat{\mathrm{Err}}_{\ell_1}(h; S)$ be the empirical loss of h on S with the ℓ_1 robust loss. We can derive the error as follows,

$$\begin{aligned}
\mathrm{Err}_{\ell_1}(\rho(\kappa(S)); \mathcal{D}) &\overset{(i)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_1}(\rho(\kappa(S)); S) + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log\frac{1}{\delta}\big)}{m}} \\
&\overset{(ii)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_1}(h^\star; S) + 4\beta + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log\frac{1}{\delta}\big)}{m}} \\
&\overset{(iii)}{\lesssim} \mathrm{Err}_{\ell_1}(h^\star; \mathcal{D}) + 4\beta + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log\frac{1}{\delta}\big)}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{m}} \\
&\lesssim \mathrm{Err}_{\ell_1}(h^\star; \mathcal{D}) + 4\beta + \sqrt{\frac{\big(|\kappa(S)| \log(m) + \log\frac{1}{\delta}\big)}{m}},
\end{aligned}$$

(i) follows from a generalization of sample compression scheme in the agnostic case, see Lemma A.6, (ii) follows Equation (30), (iii) follows from Hoeffding's inequality.

Take m sufficiently large such that

$$\sqrt{\frac{|\kappa(S)|\log(m)+\log\frac{1}{\delta}}{m}} \lesssim \beta.$$

Re-scale $\beta = \beta/5$ and plug in the compression size, we get sample complexity of size

$$\mathcal{M} = \mathcal{O}\left(\frac{1}{\beta^2}\left(\operatorname{fat}(\mathcal{H},c\beta)\operatorname{fat}^*(\mathcal{H},c\beta)\log^2\left(\frac{\operatorname{fat}(\mathcal{H},c\beta)}{\beta^2}\right)\log^2\left(\frac{\operatorname{fat}^*(\mathcal{H},c\beta)}{\beta^2}\right)\log\frac{1}{\beta}+\log\frac{1}{\delta}\right)\right),$$

for some numerical constant $c \in (0,\infty)$. □

D.2. Agnostic

Proof of Theorem 5.2. The construction follows a reduction to the realizable case similar to (David et al., 2016), which is for the non-robust zero-one loss. Moreover, we use a margin-based analysis of $\mathrm{MedBoost}$ algorithm (see Kégl (2003, Theorem 1)), and overcome some technical challenges.

Denote $\Lambda_{\mathrm{RE}} = \Lambda_{\mathrm{RE}}(\eta, 1/3, 1/3, \mathcal{H}, \mathcal{U}, \ell^{\eta}_{\mathcal{U}})$, the sample complexity of Robust (η,β)-regression for a class $\mathcal{H}$ with respect to a perturbation function $\mathcal{U}$, taking $\epsilon = \delta = 1/3$

Using a robust ERM, find the maximal subset $S' \subseteq S$ with zero empirical robust loss (for the η-ball loss), such that $\inf_{h\in\mathcal{H}} \widehat{\mathrm{Err}}_{\eta}(h,f;S') = 0$. Now, Λ_{RE} samples suffice for weak robust learning for any distribution $\mathcal{D}$ on S'.

Compute the $\mathrm{MedBoost}$ on S', with $T \approx \log(|S'|)$ boosting rounds, where each weak robust learner is trained on $\approx \Lambda_{\mathrm{RE}}$ samples. The returned weighted median $\hat{h} = \mathrm{Med}\big(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\big)$ satisfies $\widehat{\mathrm{Err}}_{\eta}\Big(\hat{h},f;S'\Big) = 0$, and each hypothesis $\hat{h}_t \in \Big\{\hat{h}_1,\ldots,\hat{h}_T\Big\}$ is representable as set of size $\mathcal{O}(\Lambda_{\mathrm{RE}})$. This defines a compression scheme of size $\Lambda_{\mathrm{RE}}T$, and $\hat{h_i}$ can be reconstructed from a compression set of points from S of size $\Lambda_{\mathrm{RE}}T$.

Recall that $S' \subseteq S$ is a maximal subset such that $\inf_{h\in\mathcal{H}} \widehat{\mathrm{Err}}_{\eta}(h,f;S') = 0$ which implies that $\inf_{\hat{h}\in\mathcal{H}} \widehat{\mathrm{Err}}_{\eta}(h,f;S') \leq \inf_{h\in\mathcal{H}} \widehat{\mathrm{Err}}_{\eta}(h,f;S')$. Plugging it into an agnostic sample compression bound Lemma A.6, we have a sample complexity of $\tilde{\mathcal{O}}\big(\frac{\Lambda_{\mathrm{RE}}}{\epsilon^2}\big)$, which translates into $\tilde{\mathcal{O}}\Big(\frac{\operatorname{fat}(\mathcal{H},c\eta)\operatorname{fat}^*(\mathcal{H},c\eta)}{\epsilon^2}\Big)$, for some numerical constant $c \in (0,\infty)$. □

D.3. Naive approach with a fixed cutoff

An agnostic learner for robust (η,β)-regression does not apply to the robust regression setting. The reason is that the optimal function in $\mathcal{H}$ has different scales of robustness on different points. we show that by using a fixed cutoff for all points we can obtain an error of $\sqrt{\mathrm{OPT}_{\mathcal{H}}} + \epsilon$.

Theorem D.1. *For any $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ with finite γ-fat-shattering for all $\gamma > 0$, any $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, and any $\epsilon,\delta \in (0,1)$, $\eta \in [0,1]$, for some numerical constant $c \in (0,\infty)$, with probability $1-\delta$, Algorithm 7 outputs a function with error at most $\sqrt{\mathrm{OPT}_{\mathcal{H}}} + \epsilon$, for the $\ell_{1,\mathcal{U}}(\cdot)$ robust loss, and using a sample of size*

$$\tilde{\mathcal{O}}\left(\frac{\operatorname{fat}(\mathcal{H},c\epsilon)\operatorname{fat}^*(\mathcal{H},c\epsilon)}{\epsilon^2}+\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right).$$

Recall that $\operatorname{fat}^*(\mathcal{F},\epsilon) \lesssim \frac{1}{\epsilon}2^{\operatorname{fat}(\mathcal{F},\epsilon/2)+1}$ *by Equation (7).*

Algorithm 7

Input: $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, $S = \{(x_i, y_i)\}_{i=1}^m$, $\tilde{S} = \{(x_i, y_i)\}_{i=1}^n$.
Algorithms used: Agnostic learner for Robust (η, β)-regression (see Theorem 5.2): Agnostic-η-Regressor.

1. Define a grid $\Theta = \left\{\frac{1}{m}, \frac{2}{m}, \frac{4}{m}, \frac{8}{m}, \ldots, 1\right\}$.
2. Define $\mathcal{H}_\Theta = \{h_\theta = \text{Agnostic-}\theta\text{-Regressor}(S) : \theta \in \Theta\}$.
3. Find an optimal function on the holdout set

$$\hat{h}_\theta = \operatorname*{argmin}_{h_\theta \in \mathcal{H}_\Theta} \frac{1}{|\tilde{S}|} \sum_{(x,y)\in\tilde{S}} \mathbb{I}\left[\sup_{z\in\mathcal{U}(x)} |h_\theta(z) - y| \geq \theta\right]$$

Output: $\hat{h}_\theta$.

Proof of Theorem D.1. Let

$$\mathrm{OPT}_{\mathcal{H}} = \inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\sup_{z\in\mathcal{U}(x)} |h(z) - y|\Big],$$

which is obtained by $h^\star \in \mathcal{H}$. By Markov Inequality we have

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\left(\sup_{z\in\mathcal{U}(x)} \left|h^\star(z) - y\right| > \eta\right) \leq \frac{\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{z\in\mathcal{U}(x)}\left|h^\star(z) - y\right|\right]}{\eta}.$$

Taking $\eta = \sqrt{\mathrm{OPT}_{\mathcal{H}}}$,

$$\begin{aligned}\mathbb{P}_{(x,y)\sim\mathcal{D}}\left(\sup_{z\in\mathcal{U}(x)} \left|h^\star(z) - y\right| > \sqrt{\mathrm{OPT}_{\mathcal{H}}}\right) &\leq \frac{\mathrm{OPT}_{\mathcal{H}}}{\sqrt{\mathrm{OPT}_{\mathcal{H}}}} \\ &= \sqrt{\mathrm{OPT}_{\mathcal{H}}}.\end{aligned}$$

This means that we can apply the algorithm for agnostic robust uniform η regression with $\eta = \sqrt{\mathrm{OPT}_{\mathcal{H}}}$, and obtain an error of $\sqrt{\mathrm{OPT}_{\mathcal{H}}} + \epsilon$. The problem is that $\mathrm{OPT}_{\mathcal{H}}$ is not known in advance. To overcome this issue, we can have a grid search on the scale of η, and then verify our choice using a holdout training set.

We define a grid,$\Theta = \left\{\frac{1}{m}, \frac{2}{m}, \frac{4}{m}, \frac{8}{m}, \ldots, 1\right\}$, such that one of its elements satisfies $\sqrt{\mathrm{OPT}_{\mathcal{H}}} < \hat{\theta} < 2\sqrt{\mathrm{OPT}_{\mathcal{H}}}$.

For each element in the grid, we compute the agnostic regressor for the η-robust loss. That is, we define $\mathcal{H}_\Theta = \{h_\theta = \text{Agnostic-}\theta\text{-Regressor}(S) : \theta \in \Theta\}$.

We choose the optimal function on a holdout labeled set $\tilde{S}$ of size $\approx \frac{1}{\epsilon^2}\log\frac{1}{\delta}$,

$$\hat{h}_\theta = \operatorname*{argmin}_{h_\theta \in \mathcal{H}_\Theta} \frac{1}{|\tilde{S}|} \sum_{(x,y)\in\tilde{S}} \mathbb{I}\left[\sup_{z\in\mathcal{U}(x)} |h_\theta(z) - y| \geq \theta\right].$$

With high probability, the algorithm outputs a function with error at most $\sqrt{\mathrm{OPT}_{\mathcal{H}}} + \epsilon$ for the ℓ_1 robust loss, using a sample of size

$$\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H}, c\epsilon)\,\mathrm{fat}^*(\mathcal{H}, c\epsilon)}{\epsilon^2}\right).$$

□