Agnostic Sample Compression Schemes for Regression

Idan Attias [1] **Steve Hanneke** [2] **Aryeh Kontorovich** [1] **Menachem Sadigurschi** [1]

Abstract

We obtain the first positive results for bounded sample compression in the agnostic regression setting with the ℓ_p loss, where $p \in [1, \infty]$. We construct a generic *approximate* sample compression scheme for real-valued function classes exhibiting exponential size in the fat-shattering dimension but independent of the sample size. Notably, for linear regression, an *approximate* compression of size linear in the dimension is constructed. Moreover, for ℓ_1 and ℓ_∞ losses, we can even exhibit an efficient *exact* sample compression scheme of size linear in the dimension. We further show that for every other ℓ_p loss, $p \in (1, \infty)$, there does not exist an exact agnostic compression scheme of bounded size. This refines and generalizes a negative result of David, Moran, and Yehudayoff (2016) for the ℓ_2 loss. We close by posing general open questions: for agnostic regression with ℓ_1 loss, does every function class admit an exact compression scheme of polynomial size in the pseudo-dimension? For the ℓ_2 loss, does every function class admit an approximate compression scheme of polynomial size in the fat-shattering dimension? These questions generalize Warmuth's classic sample compression conjecture for realizable-case classification (Warmuth, 2003).

1. Introduction

Sample compression is a central problem in learning theory, whereby one seeks to retain a "small" subset of the labeled sample that uniquely defines a "good" hypothesis. Quantifying *small* and *good* specifies the different variants of the problem. For instance, in the classification setting, taking *small* to mean "constant size" (i.e., depending only on the VC-dimension d of the concept class but not on the sample size m) and *good* to mean "consistent with the sample" specifies the classic realizable sample compression problem for VC classes. The feasibility of the latter was an open problem between its being posed by Littlestone and Warmuth (1986b) and its positive resolution by Moran & Yehudayoff (2016), with various intermediate steps in between (Floyd, 1989; Helmbold, Sloan, and Warmuth, 1992; Floyd and Warmuth, 1995b; Ben-David and Litman, 1998; Kuzmin and Warmuth, 2007; Rubinstein, Bartlett, and Rubinstein, 2009; Rubinstein and Rubinstein, 2012; Chernikov and Simon, 2013; Livni and Simon, 2013; Moran, Shpilka, Wigderson, and Yehudayoff, 2017). A stronger form of this problem, where *small* means $\mathcal{O}(d)$ (or even exactly d), remains open (Warmuth, 2003).

David, Moran, and Yehudayoff (2016) recently generalized the definition of *compression scheme* to the agnostic case, where it is required that the function reconstructed from the compression set obtains an average loss on the full data set nearly as small as the function in the class that minimizes this quantity. In Remark 2.2, we give a strong motivation for this criterion by arguing an equivalence to the generalization ability of the compression-based learning algorithm. Under this definition, David et al. (2016) extended the realizable-case result for VC classes to cover the agnostic case as well: a bounded-size compression scheme for the former implies such a scheme (in fact, of the same size) for the latter. They also generalized from binary to multiclass concept families, with the graph dimension in place of VC-dimension. Proceeding to real-valued function classes, David et al. (2016) came to a starkly negative conclusion: they established that there is *no* constant-size exact agnostic sample compression scheme for linear functions under the ℓ_2 loss. (*Realizable* linear regression in $\mathbb{R}^d$ trivially admits sample compression of size $d + 1$, under any loss, by selecting a minimal subset that spans the data.)

Main results. We are the first to construct bounded sample compression schemes for agnostic regression with ℓ_p loss, $p \in [1, \infty]$. Table 1 summarizes our contributions in the context of previous results. We refer to an α-approximate compression as one where the function reconstructed from the compression set achieves an average error at most α compared to the optimal function in the class. We consider the sample compression to be exact when we precisely recover this error. See Equations (3) and (4) for formal definitions.

[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>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Our approach begins with proposing a boosting method (Algorithm 1) to construct an α-*approximate* sample compression scheme for agnostic ℓ_p regression, within function classes characterized by a finite fat-shattering dimension. The scheme has a size of $\tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{F}, c\alpha/p)\,\mathrm{fat}^*(\mathcal{F}, c\alpha/p))$[1], for some numerical constant $c > 0$, as established by Theorem 3.1. Here, $\mathrm{fat}(\mathcal{F}, c\alpha/p)$ represents the fat-shattering dimension of function class $\mathcal{F}$ at scale $c\alpha/p$, and fat^* is the dimension of the dual-class, which is finite as long as the dimension of the primal class is finite and can be at most exponentially larger, see Equation (5). Notably, our compression size is independent of the sample size. A major open question is how to improve the exponential dependence in the dimension, even in the realizable binary classification setting (Warmuth, 2003). While such an approximate compression has been previously acknowledged in realizable regression (Hanneke et al., 2019), and exact compression in agnostic binary classification (David et al., 2016), in Section 3 we delve into the details of our techniques and elucidate why methods previously suggested fall short in addressing agnostic regression.

We proceed with exploring linear regression. The negative result of David et al. (2016) regarding the impossibility of achieving an *exact* compression for linear regression with the ℓ_2 (squared) loss raises a general doubt over whether exact sample compression is ever a viable approach to agnostic learning of real-valued functions. We address this concern by proving that, if we replace the ℓ_2 loss with the ℓ_1 or ℓ_∞ loss, then there *is* a simple exact agnostic compression scheme of size $d + 1$ for ℓ_1 linear regression and $d + 2$ for ℓ_∞ in $\mathbb{R}^d$, see Theorems 4.3 and 4.4. This is somewhat surprising, given the above negative result for the ℓ_2 loss. Computationally, our compression schemes for ℓ_1 and ℓ_∞ involve solving a linear program of linear size.

We then propose Algorithm 2 for an α-approximate sample compression for ℓ_p linear regression of size $\mathcal{O}(d \log(p/\alpha))$, where $p \in (1, \infty)$, see Theorem 4.2. Roughly speaking, we reduce the problem to realizable binary classification with linear functions. Our approach involves introducing a discretized dataset on which the optimal solution of Support Vector Machine (SVM) pointwise approximates an optimal regressor on the original dataset. We complement this result by showing that $p \in \{1, \infty\}$ are the *only two* ℓ_p losses for which a constant-size exact compression scheme exists (Theorem 4.6), generalizing the argument of David et al. (2016).

These appear to be the first positive results for bounded agnostic sample compression for real-valued function classes. We close by posing intriguing open questions generalizing our result to arbitrary function classes: under the ℓ_1 loss, does *every* function class admit an exact agnostic compression scheme of size equal to its pseudo-dimension? under the ℓ_2 loss, does *every* function class admit an approximate agnostic compression of size equal to its fat-shattering dimension? We argue that this represents a generalization of Warmuth's classic sample compression problem, which asks whether every space of classifiers admits a compression scheme of size VC-dimension in the realizable case.

Related work. Sample compression scheme is a classic technique for proving generalization bounds, introduced by Littlestone & Warmuth (1986a); Floyd & Warmuth (1995a). These bounds proved to be useful in numerous learning settings, particularly when the uniform convergence property does not hold or provides suboptimal rates, such as binary classification (Graepel et al., 2005b; Moran & Yehudayoff, 2016; Bousquet et al., 2020), multiclass classification (Daniely et al., 2015; Daniely & Shalev-Shwartz, 2014; David et al., 2016; Brukhim et al., 2022), regression (Hanneke et al., 2019; Attias et al., 2023), active learning (Wiener et al., 2015), density estimation (Ashtiani et al., 2020), adversarially robust learning (Montasser et al., 2019; 2020; 2021; 2022; Attias et al., 2022; Attias & Hanneke, 2023), learning with partial concepts (Alon et al., 2022), and showing Bayes-consistency for nearest-neighbor methods (Gottlieb et al., 2014; Kontorovich et al., 2017). As a matter of fact, compressibility and learnability are known to be equivalent for general learning problems (David et al., 2016). A remarkable result by Moran & Yehudayoff (2016) showed that VC classes enjoy a sample compression that is independent of the sample size.

David et al. (2016) introduced sample compression in the context of regression. They showed that an exact compression scheme for ℓ_2 agnostic linear regression requires a linear growth relative to the sample size. Additionally, they showed that it is feasible to have an α-approximate compression for zero-dimensional linear regression with a size of $\log(1/\alpha)/\alpha$. In a broader sense, they established the equivalence between learnability and the presence of an approximate compression in regression.

Hanneke et al. (2019) showed how to convert *consistent* real-valued learners into constant-size (i.e., independent of sample size) efficiently computable approximate compression schemes for the realizable (or nearly realizable) regression with the ℓ_∞ loss. This result was obtained via a weak-to-strong boosting procedure, coupled with a generic construction of weak learners out of abstract regressors. The *agnostic* variant of this problem remains open in its full generality.

Ashtiani et al. (2020) adapted the notion of a compression scheme to the distribution learning problem. They showed that if a class of distributions admits robust compressibility then it is agnostically learnable.

[1] $\tilde{\mathcal{O}}$ hides polylogarithmic factors in the specified expression.

Problem Setup	Compression Type	Compression Size	Reference
Realizable/Agnostic Binary Classification	Exact	$\mathcal{O}(\mathrm{VC} \cdot \mathrm{VC}^*)$	(Moran & Yehudayoff, 2016; David et al., 2016)
Realizable/Agnostic Multiclass Classification	Exact	$\mathcal{O}(\mathrm{d_G} \cdot \mathrm{d_G^*})$	(David et al., 2016)
		$\mathcal{O}\Big(\mathrm{DS}^{1.5} \cdot \mathrm{polylog}(m)\Big)$	(Brukhim et al., 2022)
		$\Omega\Big(\log(m)^{1-o(1)}\Big)$	(Pabbaraju, 2023)
Realizable ℓ_∞ Regression	α-Approximate	$\mathcal{O}\Big(\mathrm{fat}_{c\alpha} \cdot \mathrm{fat}^*_{c\alpha} \cdot \mathrm{polylog}\Big(\mathrm{fat}_{c\alpha}, \mathrm{fat}^*_{c\alpha}, \frac{1}{\alpha}\Big)\Big)$	(Hanneke et al., 2019)
Agnostic ℓ_p Regression: $p \in (1, \infty)$	α-Approximate	$\mathcal{O}\Big(\mathrm{fat}_{c\alpha} \cdot \mathrm{fat}^*_{c\alpha} \cdot \mathrm{polylog}\Big(\mathrm{fat}_{c\alpha}, \mathrm{fat}^*_{c\alpha}, p, \frac{1}{\alpha}\Big)\Big)$	This work
Agnostic ℓ_p Regression: $p \in \{1, \infty\}$		$\mathcal{O}\Big(\mathrm{fat}_{c\alpha} \cdot \mathrm{fat}^*_{c\alpha} \cdot \mathrm{polylog}\Big(\mathrm{fat}_{c\alpha}, \mathrm{fat}^*_{c\alpha}, \frac{1}{\alpha}\Big)\Big)$	
Agnostic ℓ_p Linear Regression: $p \in \{1, \infty\}$	Exact	$\mathcal{O}(d)$	This work
Agnostic ℓ_p Linear Regression: $p \in (1, \infty)$	α-Approximate	$\mathcal{O}\Big(d \cdot \log\Big(\frac{p}{\alpha}\Big)\Big)$	This work
Agnostic ℓ_2 Linear Regression	Exact	$\Omega(m)$	(David et al., 2016)
Agnostic ℓ_p Linear Regression: $p \in [1, \infty]$	Exact	$\Omega(\log(m))$	This work

Table 1. **Sample compression schemes for classification and regression.** We denote the sample size by m, $c > 0$ is a numerical constant. The $o(1)$ term vanishes as $m \to \infty$. **(i) Binary Classification:** VC is the Vapnik-Chervonenkis dimension that characterizes realizable and agnostic learnability. Any dimension with $(\cdot)^*$ denotes the dimension of the dual-class. **(ii) Multiclass Classification:** $\mathrm{d_G}$ is the Graph-dimension and DS is the Daniely-Shwartz dimension. For a finite set of labels, both dimensions characterize realizable and agnostic learnability. For an infinite set, only the finiteness of the DS dimension is equivalent to learnability. There exist learnable function classes with infinite graph dimension and finite DS dimension. **(iii) Regression:** $\mathrm{fat}_{c\alpha}$ is the fat-shattering dimension at scale $c\alpha$. A function class is agnostically learnable in this setting if and only if the fat-shattering dimension is finite for any scale. However, in the realizable case, there are learnable classes with infinite fat-shattering dimension. We comment that the results in (Hanneke et al., 2019) are stated for ℓ_∞, but still hold for any ℓ_p (with extra polylog factors in p) due to Lipschitzness of this loss. **(iv) Linear Regression:** d is the vector space dimension. We refer to Section 5 for open problems.

2. Preliminaries

We denote $[m] := \{1, \ldots, m\}$. Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. The ℓ_p loss incurred by a hypothesis $f \in \mathcal{F}$ on (x, y) is given by $(x, y) \mapsto |f(x) - y|^p$, where $p \in [1, \infty]$. For $p \in [1, \infty)$, the loss incurred by a hypothesis $f \in \mathcal{F}$ on a labeled sample $S = \{(x_i, y_i) : i \in [m]\}$ is given by

$$L_p(f, S) := \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - y_i|^p, \tag{1}$$

while for $p = \infty$,

$$L_\infty(f, S) := \max_{1 \le i \le m} |f(x_i) - y_i|. \tag{2}$$

Remark 2.1. The ℓ_p regression objective is typically written without taking the pth root so as to facilitate optimization algorithms. As we avoid taking the p-th root, the resulting p-norm formulation does not directly converge to ℓ_∞ as p approaches infinity. Consequently, our ℓ_p results explicitly depend on p, similar to results in the literature.

Now let us introduce a formal definition of sample compression, and a criterion we require of any valid *agnostic compression scheme*. Following the definition, we provide a strong motivation for this criterion in terms of an equivalence to the generalization ability of the learning algorithm under general conditions.

Approximate and exact sample compression schemes. Following David et al. (2016), we define a *selection scheme* (κ, ρ) for a hypothesis class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ is defined as follows. A k-*selection* function κ maps sequences $\{(x_1, y_1), \ldots, (x_m, y_m)\} \in \bigcup_{\ell \ge 1} \{\mathcal{X} \times \mathcal{Y}\}^\ell$ to elements in $\mathcal{K} = \bigcup_{\ell \le k'} \{\mathcal{X} \times \mathcal{Y}\}^\ell \times \bigcup_{\ell \le k''} \{0, 1\}^\ell$, where $k' + k'' \le k$. A *reconstruction* is a function $\rho : \mathcal{K} \to \mathcal{Y}^{\mathcal{X}}$. We say that (κ, ρ) is a k-size agnostic *exact* sample compression scheme for $\mathcal{F}$ if κ is a k-selection and for all $S = \{(x_i, y_i) : i \in [m]\}$, $f_S := \rho(\kappa(S))$ achieves $\mathcal{F}$-competitive empirical loss:

$$L_p(f_S, S) \le \inf_{f \in \mathcal{F}} L_p(f, S). \tag{3}$$

We also define a relaxed notion of agnostic α-*approximate* sample compression in which f_S should satisfy

$$L_p(f_S, S) \le \inf_{f \in \mathcal{F}} L_p(f, S) + \alpha. \tag{4}$$

In principle, the *size* k of an agnostic compression scheme may depend on the data set size m, in which case we may denote this dependence by $k(m)$. However, in this work we are primarily interested in the case when $k(m)$ is *bounded*: that is, $k(m) \le k$ for some m-independent value k. Note that the above definition is fully general, in that it defines a notion of agnostic compression scheme for *any* function class $\mathcal{F}$ and loss function L, though in the present work we focus on L_p loss for $1 \le p \le \infty$.

Remark 2.2. At first, it might seem unclear why this is an appropriate generalization of sample compression to the agnostic setting. To see that it is so, we note that one of the main interests in sample compression schemes is their ability to *generalize*: that is, to achieve *low excess risk* under a *distribution* P on $\mathcal{X} \times \mathcal{Y}$ when the data S are sampled iid according to P (Littlestone and Warmuth, 1986b; Floyd and Warmuth, 1995b; Graepel, Herbrich, and Shawe-Taylor, 2005a). Also, as mentioned, in this work we are primarily interested in sample compression schemes that have *bounded size*: $k(m) \le k$ for an m-independent value k. Furthermore, we are also focusing on the most-general case, where this size bound should be independent of everything else in the scenario, such as the data S or the underlying distribution P. Given these interests, we claim that the above definition is essentially the only reasonable choice. More specifically, for L_p loss with $1 \le p < \infty$, any compression scheme with $k(m)$ bounded such that its expected excess risk under any P converges to 0 as $m \to \infty$ necessarily satisfies the above condition (or is easily converted into one that does). To see this, note that for any data set S for which such a compression scheme fails to satisfy the above $\mathcal{F}$-competitive empirical loss criterion, we can define a distribution P that is simply uniform on S, and then the compression scheme's selection function would be choosing a bounded number of points from S and a bounded number of bits, while guaranteeing that excess risk under P approaches 0, or equivalently, excess empirical loss approaches 0. To make this argument fully formal, only a slight modification is needed, to handle having multiple copies of points from S in the compression set; given that the size is bounded, these repetitions can be encoded in a bounded number of extra bits, so that we can stick to strictly distinct points in the compression set.

In the converse direction, we also note that any bounded-size agnostic compression scheme (in the sense of the above definition) will be guaranteed to have excess risk under P converging to 0 as $m \to \infty$, in the case that S is sampled iid according to P, for losses L_p with $1 \le p < \infty$, as long as P guarantees that $(X, Y) \sim P$ has Y bounded (almost surely). This follows from classic arguments about the generalization ability of compression schemes, which includes results for the agnostic case (Graepel, Herbrich, and Shawe-Taylor, 2005a). For unbounded Y one cannot, in general, obtain distribution-free generalization bounds. However, one can still obtain generalization under certain broader restrictions (see, e.g., Mendelson, 2015 and references therein). The generalization problem becomes more subtle for the L_∞ loss: this cannot be expressed as a sum of pointwise losses and there are no standard techniques for bounding the deviation of the sample risk from the true risk. One recently-studied guarantee achieved by minimizing empirical L_∞ loss is a kind of "hybrid error" generalization, developed in Hanneke et al. (2019, Theorem 9). We refer the interested reader to that work for the details of those results, which can easily be extended to apply to our notion of an agnostic compression scheme.

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

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

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). We also define the dual dimension. Define the dual class $\mathcal{F}^* \subseteq [0,1]^{\mathcal{F}}$ 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 dimension* 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),

$$\mathrm{fat}^*(\mathcal{F}, \gamma) \lesssim \frac{1}{\gamma} 2^{\mathrm{fat}(\mathcal{F}, \gamma/2)+1}. \tag{5}$$

3. Approximate Agnostic Compression for Real-Valued Function Classes

In this section, we construct an approximate compression scheme for all real-valued function classes that are agnostically PAC learnable, that is, classes with finite fat-shattering dimension at any scale (Alon et al., 1997; Bartlett & Long, 1998). We prove the following main result.

Theorem 3.1 (Approximate compression for agnostic regression)**.** *Let* $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$*,* $S = \{(x_i, y_i) : i \in [m]\} \subseteq \mathcal{X} \times [0,1]$*, an approximation parameter* $\alpha \in [0,1]$*, a weak learner parameter* $\beta \in (0, 1/2]$*, and* ℓ_p *loss where* $p \in [1, \infty]$*. By setting Algorithm 1 with* $T \leftarrow \mathcal{O}\left(\frac{1}{\beta^2}\log(m)\right)$ *and*

$$\begin{cases} d \leftarrow \tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{F}, c\alpha/p)), n \leftarrow \tilde{\mathcal{O}}\left(\frac{\mathrm{fat}^*(\mathcal{F}, c\alpha/p)}{\beta^2}\right), p \in [1, \infty) \\ d \leftarrow \tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{F}, c\alpha)), n \leftarrow \tilde{\mathcal{O}}\left(\frac{\mathrm{fat}^*(\mathcal{F}, c\alpha)}{\beta^2}\right), p = \infty, \end{cases}$$

we get an α-approximate sample compression scheme of size

$$\begin{cases} \tilde{\mathcal{O}}\left(\frac{1}{\beta^2}\operatorname{fat}(\mathcal{F}, c\alpha/p)\operatorname{fat}^*(\mathcal{F}, c\alpha/p)\right), p \in [1,\infty) \\ \tilde{\mathcal{O}}\left(\frac{1}{\beta^2}\operatorname{fat}(\mathcal{F}, c\alpha)\operatorname{fat}^*(\mathcal{F}, c\alpha)\right), p = \infty, \end{cases}$$

for some universal constant $c > 0$. Recall that the dual fat-shattering is at most exponential in the primal dimension, see Equation (5). $\tilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors of $(\operatorname{fat}, \operatorname{fat}^*, p, 1/\alpha, 1/\beta)$.

Remark 3.2. Note that having an α-approximate compression of size k implies the following bound on the generalization error: $\alpha + \sqrt{\frac{k\log(m/k)}{m}}$ (David et al., 2016, Theorem 4.2).

Our algorithm incorporates a boosting approach for real-valued functions. Therefore, we need a definition of weak learners in this context.

Definition 3.3 (Approximate weak real-valued learners)**.** Let $\beta \in (0, \frac{1}{2}]$, $\alpha \in (0,1)$. We say that $g : \mathcal{X} \to [0,1]$ is an *approximate* (α, β)-weak learner, with respect to P and a target function $f^\star \in \mathcal{F}$ if

$$\mathbb{P}_{(x,y)\sim P}\{(x,y) : |g(x) - y| > |f^\star(x) - y| + \alpha\} \leq \frac{1}{2} - \beta.$$

This notion of a weak learner must be formulated carefully. For example, taking a learner guaranteeing absolute loss at most $\frac{1}{2} - \beta$ is known to not be strong enough for boosting to work, see the discussion in Hanneke et al. (2019, Section 4). On the other hand, by making the requirement too strong (for example, $\mathrm{AdaBoost.R}$ 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 the main algorithm.

The challenges beyond realizable regression and agnostic classification. There is a crucial difference from previous boosting algorithms for real-valued used by Kégl (2003); Hanneke et al. (2019) in the realizable case. In our approach, the cut-offs $\psi(x, y)$ are allowed to vary across different points, in contrast to a fixed cut-off applied uniformly across all points. This flexibility enables us to address the agnostic setting, wherein the loss of an optimal minimizer may differ across various points in the sample. To prove the existence of weak learners we are required to have a generalization theorem that is compatible with changing cut-offs, see Theorem A.1. A similar generalization result was used in the context of adversarially robust learning (Attias & Hanneke, 2023).

The compression approach for agnostic binary classification, as discussed in (David et al., 2016), encounters a similar

Algorithm 1 Approximate Agnostic Sample Compression for ℓ_p Regression, $p \in [1, \infty]$

```
Input: $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$, $S = \{(x_i, y_i) : i \in [m]\} \subseteq \mathcal{X} \times [0,1]$.
Parameters: Approximation parameter $\alpha \in (0,1)$, weak learner parameter $\beta \in (0,1/2]$, weak learner sample size $d \geq 1$, sparsification parameter $n \geq 1$, number of boosting rounds $T \geq 1$, loss parameter $p \in [1,\infty]$.
Initialize: $P_1 \leftarrow \mathrm{Uniform}(S)$.

▷ Find an optimal function in $f^\star \in \mathcal{F}$. Our goal is to construct a function that pointwise approximates $f^\star$ on $S$
1. Compute:
    (a) $f^\star \leftarrow \mathrm{argmin}_{f\in\mathcal{F}} L_p(f, S)$ (defined in Equations (1) and (2)).
    (b) $\psi(x, y) \leftarrow |f^\star(x) - y|$, $\forall (x, y) \in S$.

▷ Median boosting for real-valued functions
2. For $t = 1, \ldots, T$:
    (a) Get an $(2\alpha, \beta)$-approximate weak learner $\hat{f}_t$ with respect to distribution $P_t$:
        Find a multiset $S_t \subset S$ of $d$ points such that for any $f \in \mathcal{F}$ with $|f(x) - y| \leq \psi(x, y) + \alpha \ \forall (x, y) \in S_t$, it holds that $\mathbb{P}_{(x,y)\sim P_t}\{(x, y) : |f(x) - y| > \psi(x, y) + 2\alpha\} \leq 1/2 - \beta$. ($S_t$ exists from Theorem A.1).
    (b) For $i = 1, \ldots, m$:
        Set $w_i^{(t)} \leftarrow 1 - 2\mathbb{I}\left[\left|\hat{f}_t(x_i) - y_i\right| > \psi(x_i, y_i) + 2\alpha\right]$.
    (c) Set $\alpha_t \leftarrow \frac{1}{2}\log\left(\frac{(1-\beta)\sum_{i=1}^m P_t(x_i,y_i)\mathbb{I}\left[w_i^{(t)}=1\right]}{(1+\beta)\sum_{i=1}^m P_t(x_i,y_i)\mathbb{I}\left[w_i^{(t)}=-1\right]}\right)$.
    (d) If $\alpha_t = \infty$:
        return $T$ copies of $\hat{f}_t$, $(\alpha_1 = 1, \ldots, \alpha_T = 1)$, $S_t$.
        Else:
        $P_{t+1}(x_i, y_i) \leftarrow P_t(x_i, y_i)\frac{\exp\left(-\alpha_t w_i^t\right)}{\sum_{j=1}^m P_t(x_j,y_j)\exp\left(-\alpha_t w_j^t\right)}$.

▷ Sparsifying the weighted ensemble $\left\{\hat{f}_i\right\}_{i=1}^T$ returned from boosting via sampling
3. Repeat:
    (a) Sampling:
        $(J_1, \ldots, J_n) \sim \mathrm{Categorial}\left(\frac{\alpha_1}{\sum_{s=1}^T \alpha_s}, \ldots, \frac{\alpha_T}{\sum_{s=1}^T \alpha_s}\right)^n$.
    (b) Let $\tilde{\mathcal{F}} = \{f_{J_1}, \ldots, f_{J_n}\}$.
    (c) Until $\forall (x, y) \in S$ :
        $\left|\left\{f \in \tilde{\mathcal{F}} : |f(x) - y| > \psi(x, y) + 3\alpha\right\}\right| < n/2$.

Compression: Multisets $S_{J_1}, \ldots, S_{J_n}$ and cut-offs $\psi|_{S_{J_1}}, \ldots, \psi|_{S_{J_n}}$ corresponding to the weak learners in $\tilde{F}$.
Reconstruction: Reconstruct weak learners $f_{J_i}$ from $S_{J_i}$ and $\psi|_{S_{J_i}}$, $i \in [n]$, and output their median $\mathrm{Median}(f_{J_1}, \ldots, f_{J_n})$.
```

challenge. In this method, our initial emphasis is on identifying the points correctly classified by an optimal function in the class. Subsequently, we apply compression techniques for realizable classification. However, in regression, discarding points where the optimal function makes mistakes is not feasible, given that the loss is not strictly zero-one. Instead, we utilize the entire sample, targeting the error for each point and constructing a function with a similar approximated error on each point.

Proof overview. First, we show that the returned output of Algorithm 1 is a valid compression. Then we bound the size of this compression.

Approximate compression correctness. In step 1, we compute some $f^\star \in \mathcal{F}$ the minimizes the empirical ℓ_p error on the sample S,

$$f^\star \leftarrow \operatorname*{argmin}_{f\in\mathcal{F}} L_p(f,S),$$

as defined in Equations (1) and (2). Let $\psi : \mathcal{X}\times\mathcal{Y} \to [0,1]$ be the ℓ_1 loss of $f^\star$ on each point in S,

$$\psi(x,y) \leftarrow |f^\star(x) - y|,\ \forall(x,y)\in S.$$

In step 2, we implement a boosting algorithm, following Definition 3.3 of weak learners. By using Theorem A.1 with $\delta = 1/3$ and $\varepsilon = 1/2 - \beta$, for any distribution P_t on S, upon receiving an i.i.d. sample $S_t \subseteq S$ from P_t of size

$$d = \mathcal{O}\left(\operatorname{fat}(\mathcal{F},\alpha/8)\log^2\left(\frac{\operatorname{fat}(\mathcal{F},\alpha/8)}{\alpha(1/2-\beta)}\right)\right),$$

with probability $2/3$ over sampling S_t from P_t, for any $f\in\mathcal{F}$ satisfying $\forall(x,y)\in S_t :\ |f(x)-y| \le \psi(x,y)+\alpha$, it holds that

$$\mathbb{P}_{(x,y)\sim P_t}\{(x,y) : |f(x)-y| > \psi(x,y)+2\alpha\} \le \frac{1}{2} - \beta.$$

That is, such a function is an approximate $(2\alpha,\beta)$-weak learner for P_t and $f^\star$. Since this holds with probability $2/3$, there must be such $S_t \subseteq S$. In order to construct an approximate $(2\alpha,\beta)$-weak learner $\hat{f}_t$, we need to find $f\in\mathcal{F}$ such that $\forall(x,y)\in S_t :\ |f(x)-y| \le \psi(x,y)+\alpha$, and so the weak learner can be encoded by S_t of size d and the set of cut-offs $\psi(x,y)\in[0,1]$ for all $(x,y)\in S_t$. We encode only approximations of the cut-offs to keep the compression size bounded (see the paragraph about the compression size below). For $T = \mathcal{O}\left(\frac{1}{\beta^2}\log(m)\right)$ rounds of boosting, Lemma A.3 guarantees that for all $(x,y)\in S$ the output of the boosting algorithm satisfies

$$\left|\operatorname{Median}\left(\hat{f}_1,\dots,\hat{f}_T;\alpha_1,\dots,\alpha_T\right)(x) - y\right| \le \psi(x,y)+2\alpha.$$

Finally, we use sampling to reduce the number of hypotheses in the ensemble from $\mathcal{O}\left(\frac{1}{\beta^2}\log(m)\right)$ to size that is independent of m. Lemma A.4 implies that the sparsification method in Step 3 ensures that we can sample

$$n = \mathcal{O}\big(\operatorname{fat}^*(\mathcal{F},c\alpha)\log^2(\operatorname{fat}^*(\mathcal{F},c\alpha)/\alpha)\big)$$

such that for all $(x,y)\in S$

$$|\operatorname{Median}(f_{J_1}(x),\dots,f_{J_n}(x)) - y| \le \psi(x,y)+3\alpha,$$

where $c>0$ is an absolute constant. By rescaling 3α to α, this proves the ℓ_1 and ℓ_∞ losses. For $p\in(1,\infty)$, we use the Lipschitzness of the ℓ_p loss and rescale the approximate parameter accordingly. We constructed a function h with $|h(x)-y| \le \psi(x,y)+\alpha$ for any $(x,y)\in S$, which implies

$$|h(x)-y|^p \overset{(i)}{\le} ((\psi(x,y))+\alpha)^p \overset{(ii)}{\le} \psi(x,y)^p + p\alpha,$$

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

$$\begin{aligned}|(\psi(x,y)+\alpha)^p - \psi(x,y)^p| &\le p|\psi(x,y)+\alpha-\psi(x,y)| \\ &\le p\alpha.\end{aligned}$$

By rescaling $p\alpha$ to α, we get

$$|\operatorname{Median}(f_{J_1}(x),\dots,f_{J_n}(x)) - y|^p \le \psi(x,y)^p+\alpha,$$

where

$$n = \Theta\left(\frac{1}{\beta^2}\operatorname{fat}^*(\mathcal{F},c\alpha/p)\log^2\left(\frac{p\operatorname{fat}^*(\mathcal{F},c\alpha/p)}{\alpha}\right)\right),$$

and

$$d = \mathcal{O}\left(\operatorname{fat}(\mathcal{F},c\alpha/p)\log^2\left(\frac{p\operatorname{fat}(\mathcal{F},c\alpha/p)}{\alpha(1/2-\beta)}\right)\right).$$

We proved the correctness of an α-approximate compression

$$L_p(\operatorname{Median}(f_{J_1},\dots,f_{J_n}),S) \le \inf_{f\in\mathcal{F}} L_p(f,S)+\alpha.$$

Approximate compression size. Each weak learner is encoded by a multiset $S'\subseteq S$ of size d and is constructed by computing some $f'\in\mathcal{F}$ that solves the constrained optimization

$$|f'(x)-y| \le \psi(x,y)+\alpha,\ \forall(x,y)\in S'.$$

We encode each $\psi(x,y)$ by some approximation $\tilde{\psi}(x,y)$, such that $\left|\tilde{\psi}(x,y)-\psi(x,y)\right| \le \alpha$, by discretizing $[0,1]$ to $1/\alpha$ buckets of size α, and each $\psi(x,y)$ is rounded down to the closest value $\tilde{\psi}(x,y)$. Each approximation requires to encode $\log(1/\alpha)$ bits, and so each learner encodes

$d \log(1/\alpha)$ bits and d samples. We have n weak learners, and the compression size is

$$n(d + d\log(1/\alpha)) \leq 2nd\log(1/\alpha)\,.$$

By plugging in n and d, and rescaling α, we conclude

$$\begin{cases} \tilde{\mathcal{O}}\left(\frac{1}{\beta^2}\operatorname{fat}(\mathcal{F}, c\alpha/p)\operatorname{fat}^*(\mathcal{F}, c\alpha/p)\right), p \in [1,\infty) \\ \tilde{\mathcal{O}}\left(\frac{1}{\beta^2}\operatorname{fat}(\mathcal{F}, c\alpha)\operatorname{fat}^*(\mathcal{F}, c\alpha)\right), p = \infty. \end{cases}$$

4. Agnostic Compression for Linear Regression

In this section, our focus is on ℓ_p linear regression in $\mathbb{R}^d$. We begin by improving upon the construction of an approximate sample compression scheme for general classes, incorporating the structure of linear functions. Next, we demonstrate the feasibility of constructing an exact compression for $p \in \{1,\infty\}$ with a size linear in d. In sharp contrast, we exhibit that this holds only for $p \in \{1,\infty\}$. We prove an impossibility result of achieving a bounded-size exact compression scheme for $p \in (1,\infty)$.

We use the following notation. Vectors $\mathbf{v} \in \mathbb{R}^d$ are denoted by boldface, and their jth coordinate is indicated by $\mathbf{v}(j)$. (Thus, $\mathbf{v}_i(j)$ indicates the jth coordinate of the ith vector in a sequence.)

4.1. Approximate Compression for $p \in [1,\infty]$

In this subsection, our instance space is $\mathcal{X} = [0,1]^d$, label space is $\mathcal{Y} = [0,1]$, and hypothesis class is bounded homogeneous linear functions $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, consisting of all $f_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}$ given by $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$, indexed by $\mathbf{w} \in \mathbb{R}^d$, where $\|\mathbf{w}\|_2 \leq 1$.

In Section 3 we proved an approximate compression for general function classes with ℓ_p losses of size $\mathcal{O}\left(\operatorname{fat}_{c\alpha/p} \cdot \operatorname{fat}^*_{c\alpha/p} \cdot \operatorname{polylog}\left(\operatorname{fat}_{c\alpha/p}, \operatorname{fat}^*_{c\alpha/p}, p, 1/\alpha\right)\right)$. We have an immediate corollary for linear functions. Let $\operatorname{Pdim}(\mathcal{F})$ be the pseudo-dimension of a function class $\mathcal{F}$ (Pollard, 1990b; Haussler, 1992), that can be defined as $\operatorname{Pdim}(\mathcal{F}) = \lim_{\gamma \to 0} \operatorname{fat}_\gamma(\mathcal{F})$. The fat-shattering dimension (at any scale) is upper bounded by the pseudo-dimension. Moreover, the vector space dimension is of the same order as the pseudo-dimension (Anthony et al., 1999), and the dimension of the dual vector space is equal to the one of the primal space. This implies the following.

Corollary 4.1. *Algorithm 1 is a sample compression scheme of size $\mathcal{O}\left(d^2 \cdot \operatorname{polylog}\left(d, p, \frac{1}{\alpha}\right)\right)$ for bounded linear regression in dimension d with the ℓ_p loss, for $p \in [1,\infty]$.*

Another "baseline" solution involves encoding the coefficients of the linear regressor up to a certain approximation parameter. To achieve an α-approximate sample compression, each coefficient should be accurate up to an additive error of α/dp for $p \in [1,\infty)$, and α/d for $p = \infty$. Thus, in this solution, we will encode $d\log(dp/\alpha)$ bits without retaining any samples for $p \in [1,\infty)$, and $d\log(d/\alpha)$ for $p = \infty$.

In this section, Theorems 4.2 to 4.4 improve upon these bounds by using a dedicated algorithm for linear functions. We start with the following result:

Theorem 4.2 (Approximate compression for agnostic linear regression)**.** *Let $\mathcal{F} = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq 1\right\}$, $S = \{(\mathbf{x}_i, y_i) : \|\mathbf{x}_i\|_2 \leq 1, \forall i \in [m]\} \subseteq \mathcal{X} \times [0,1]$, and an approximation parameter $\alpha \in (0,1)$. Algorithm 2 is an α-approximate sample compression scheme for the ℓ_p loss of size*

$$\begin{cases} \mathcal{O}\left(d \cdot \log\left(\frac{p}{\alpha}\right)\right), & p \in [1,\infty) \\ \mathcal{O}\left(d \cdot \log\left(\frac{1}{\alpha}\right)\right), & p = \infty. \end{cases}$$

Algorithm 2 Approximate Agnostic Compression for ℓ_p Linear Regression, $p \in [1,\infty]$

Input: $\mathcal{F} = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq 1\right\}$,
$S = \{(\mathbf{x}_i, y_i) : \|\mathbf{x}_i\|_2 \leq 1, \forall i \in [m]\} \subseteq \mathcal{X} \times [0,1]$.
Parameters: Approximation parameter $\alpha \in [0,1]$.

```
▷ Find an optimal regressor for $S$
1. $f^\star \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} L_p(f, S)$
▷ Define a discretized dataset where the new labels are discretized to a resolution of $\alpha$
2. Define $S_\alpha = A \cup B$, where
   $A = \left\{(\mathbf{x}_i, j\alpha) : i \in [m], j \in \left\{-\frac{1}{\alpha}, \ldots, -1, 0, 1, \ldots, \frac{1}{\alpha}\right\}\right\}$
   $B = \left\{(\mathbf{x}_i, j(1+\alpha)) : i \in [m], j \in \{-1, +1\}\right\}$
▷ Label by ±1 the discretized dataset with $f^\star$
3. Define
   $S_\alpha(f^\star) = \{((\mathbf{x}_i, \tilde{y}), z) :$ for any $(\mathbf{x}_i, \tilde{y}) \in S_\alpha :$
   $z = +1$ if $f^\star(\mathbf{x}_i) - \tilde{y} \leq 0$, otherwise $z = -1\}$
```

Compression: Run `SVM` for realizable binary classification on $S_\alpha(f^\star)$ and return a set of *support vectors*.
Reconstruction: Run `SVM` on the compression set.

Proof. Let $\mathcal{F}$ be the set of homogeneous linear predictors bounded by 1, $\mathcal{F} = \left\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq 1\right\}$, and data a set $S = \{(\mathbf{x}_i, y_i) : \|\mathbf{x}_i\|_2 \leq 1, \forall i \in [m]\} \subseteq \mathcal{X} \times [0,1]$.

Approximate compression correctness. The algorithmic idea is as follows. We first compute in Step 1 an optimal linear regressor $f^\star \in \mathcal{F}$ for the ℓ_p loss. In step 2, we create a discretized dataset S_α of size $m(2/\alpha+3)$, where for each example $\mathbf{x}_i$ we create $(2/\alpha+3)$ real-valued labels $\{-1-\alpha, -1, \ldots, -2\alpha, -\alpha, 0, \alpha, 2\alpha, \ldots, 1, 1+\alpha\}$. Then in step 3, we use the regressor $f^\star$ for classifying the dataset S_α. That is, for any $(\mathbf{x}_i, \tilde{y}) \in S_\alpha$, we have $((\mathbf{x}_i, \tilde{y}), +1)$ whenever $f^\star(\mathbf{x}_i) - \tilde{y} \leq 0$, and $((\mathbf{x}_i, \tilde{y}), -1)$ otherwise. We denote this dataset by $S_\alpha(f^\star)$. Note that for each $\mathbf{x}_i$ we created a grid of binary labels of resolution α in the range $[-1-\alpha, 1+\alpha]$, and since $|f^\star(\mathbf{x}_i)| \leq 1$, for each vector $\mathbf{x}_i$ there exists $\tilde{y}_1, \tilde{y}_2$ such that $(\mathbf{x}_i, \tilde{y}_1)$, $(\mathbf{x}_i, \tilde{y}_2) \in S_\alpha(f^\star)$ have different labels. To obtain compression, we execute `Support Vector Machine(SVM)` for realizable classification on $S_\alpha(f^\star)$. Note that the classification problem is in $\mathbb{R}^{d+1}$ and the original regression problem is in $\mathbb{R}^d$. Applying Caratheodory's theorem allows us to express its output as a linear combination of $d+2$ support vectors (along with their labels). The set of returned support vectors constitutes the compression set. For reconstruction, we utilize `SVM` on these support vectors. The hyperplane returned by `SVM` can be re-interpreted as a function from $\mathbb{R}^d$ to $\mathbb{R}$ that pointwise approximates $f^\star$ on all $\mathbf{x}_i$ in S.

We proceed to prove the correctness. Denote the output of the compression scheme by $f_{\texttt{svm}} = \rho(\kappa(S)) = (\mathbf{w}_{\texttt{svm}}, b_{\texttt{svm}})$, which a affine linear function in $\mathbb{R}^{d+1}$. This function can be re-interpreted as an affine linear function $\hat{f}: \mathbb{R}^d \to \mathbb{R}$, for any $\mathbf{x} \in \mathbb{R}^d$ we compute $y \in \mathbb{R}$ by solving $\langle \mathbf{w}_{\texttt{svm}}, (\mathbf{x}, y)\rangle + b_{\texttt{svm}} = 0$,

$$\hat{f}(\mathbf{x}) = y = \frac{\langle \mathbf{w}^d_{\texttt{svm}}, \mathbf{x}\rangle + b_{\texttt{svm}}}{\mathbf{w}_{\texttt{svm}}(d+1)},$$

where $\mathbf{w}^d_{\texttt{svm}} = (\mathbf{w}_{\texttt{svm}}(1), \ldots, \mathbf{w}_{\texttt{svm}}(d))$. It holds that $\mathbf{w}_{\texttt{svm}}(d+1) \neq 0$, since for any $\mathbf{x}_i$ there exists $\tilde{y}_1, \tilde{y}_2$ such that $(\mathbf{x}_i, \tilde{y}_1)$, $(\mathbf{x}_i, \tilde{y}_2) \in S_\alpha(f^\star)$ have different labels. If $\mathbf{w}_{\texttt{svm}}(d+1) = 0$ it means that the `SVM` hyperplane cannot distinguish between these two points, and thus, it makes a mistake on a realizable dataset, which is a contradiction. Since the output of `SVM` is a valid compression scheme for realizable binary classification, $\hat{f}$ should classify correctly all points in $S_\alpha(f^\star)$. It follows that for any $\mathbf{x}_i$ in S,

$$\left| f^\star(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i) \right| \leq \alpha,$$

due to the two adjacent grid points with resolution α lying above and below both the hyperplane of $f^\star$ and the $\hat{f}$ hyperplane. Therefore, for any $(\mathbf{x}_i, y_i) \in S$

$$\begin{aligned}\left| |f^\star(\mathbf{x}_i) - y_i| - |\hat{f}(\mathbf{x}_i) - y_i| \right| &\overset{(i)}{\leq} \left| f^\star(\mathbf{x}_i) - y_i - \hat{f}(\mathbf{x}_i) + y_i \right| \\ &= \left| f^\star(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i) \right| \\ &\leq \alpha,\end{aligned}$$

where (i) follows from the triangle inequality, and so $\hat{f}$ is an α-approximate sample compression scheme for the ℓ_1 and ℓ_∞ losses. For $p \in (1, \infty)$, using Lipschitzness of the ℓ_p loss, we have

$$\begin{aligned}&\left| |f^\star(\mathbf{x}_i) - y_i|^p - |\hat{f}(\mathbf{x}_i) - y_i|^p \right| \\ &\leq \left| p\Big(|f^\star(\mathbf{x}_i) - y_i| - |\hat{f}(\mathbf{x}_i) - y_i|\Big) \right| \\ &= p\left| |f^\star(\mathbf{x}_i) - y_i| - |\hat{f}(\mathbf{x}_i) - y_i| \right| \\ &\leq p\alpha.\end{aligned}$$

By rescaling $p\alpha$ to α, we have an α-approximate compression scheme for the ℓ_p loss.

Approximate compression size. The `SVM` running on $S_\alpha(f^\star)$ returns a set of support vectors of size at most $d+2$, since the input is in dimension $d+1$. The $\mathbf{x}$ vectors are part of the original sample S. We need to keep the grid point labels of the support vectors as well, each one of them requires $\log(1/\alpha)$ bits, and each classification ±1 costs an extra bit. We get a compression of size $d+2+(d+2)\log(1/\alpha)+d+2 = \mathcal{O}(d\log(1/\alpha))$. □

4.2. Exact Compression for $p \in \{1, \infty\}$

In this section, we show that agnostic linear regression in $\mathbb{R}^d$ admits an *exact* compression scheme of size $d+1$ under ℓ_1 and $d+2$ under ℓ_∞. Our instance space is $\mathcal{X} = \mathbb{R}^d$, label space is $\mathcal{Y} = \mathbb{R}$, and hypothesis class is $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, consisting of all $f_{\mathbf{w},b}: \mathcal{X} \to \mathcal{Y}$ given by $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle + b$, indexed by $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$. Note that we allow unbounded norms for the linear functions and the data can be unbounded as well, as opposed the the results in Section 4.1.

Theorem 4.3. *There exists an efficiently computable (see the linear program in Equation (8)) exact compression scheme for agnostic ℓ_1 linear regression of size $d+1$.*

The optimization technique based on minimizing the sum of absolute deviations is known as Least Absolute Deviations (LAD) and was introduced by Boscovich in 1757 (see, for example, Dodge (2008)). We derive a compression scheme from this method. Similarly, we can obtain a compression scheme for ℓ_∞ loss via linear programming.

Theorem 4.4. *There exists an efficiently computable (see the linear program in Equation (9)) exact compression scheme for agnostic ℓ_∞ linear regression of size $d+2$.*

4.3. Exact Constant Size Compression Is Impossible for $p \in (1, \infty)$

We proceed to show that it is impossible to have an *exact* compression scheme of constant size (independent of the sample size) for $p \in (1, \infty)$, generalizing the result for the ℓ_2 loss by David et al. (2016, Theorem 4.1).

Theorem 4.5 (David et al. (2016))**.** *There is no exact agnostic sample compression scheme for zero-dimensional linear regression with size $k(m) \leq m/2$.*

Theorem 4.6. *There is no exact agnostic sample compression scheme for zero-dimensional linear regression under ℓ_p loss, $1 < p < \infty$, with size $k(m) < \log(m)$.*

5. Open Problems

The positive result for ℓ_1 loss may also lead us to wonder how general of a result might be possible. In particular, noting that the pseudo-dimension (Pollard, 1984; 1990a; Anthony et al., 1999) of linear functions in $\mathbb{R}^d$ is precisely $d+1$ (Anthony et al., 1999), there is an intriguing possibility for the following generalization. For any class $\mathcal{F}$ of real-valued functions, denote by $\mathrm{Pdim}(\mathcal{F})$ the pseudo-dimension of $\mathcal{F}$.

Open Problem: Compressing to pseudo-dimension Number of Points. Under the ℓ_1 loss, does every class $\mathcal{F}$ of real-valued functions admit an *exact* agnostic compression scheme of size $\mathrm{Pdim}(\mathcal{F})$?

It is also interesting, and perhaps more approachable as an initial aim, to ask whether there is an agnostic compression scheme of size at most *proportional to* $\mathrm{Pdim}(\mathcal{F})$. Even falling short of this, one can ask the more-basic question of whether classes with $\mathrm{Pdim}(\mathcal{F}) < \infty$ always have *bounded* agnostic compression schemes (i.e., independent of sample size m), and more specifically whether the bound is expressible purely as a function of $\mathrm{Pdim}(\mathcal{F})$ (Moran & Yehudayoff (2016) have shown this is always possible in the realizable classification setting).

These questions are directly related to (and inspired by) the well-known long-standing conjecture of Floyd & Warmuth (1995b); Warmuth (2003), which asks whether, for realizable-case binary classification, there is always a compression scheme of size at most linear in the VC dimension of the concept class. Indeed, it is clear that a positive solution of our open problem above would imply a positive solution to the original sample compression conjecture, since in the realizable case with a function class $\mathcal{F}$ of $\{0,1\}$-valued functions, the minimal empirical ℓ_1 loss on the data is zero, and any function obtaining zero empirical ℓ_1 loss on a data set labeled with $\{0,1\}$ values must be $\{0,1\}$-valued on that data set, and thus can be thought of as a sample-consistent classifier.[2] Noting that, for $\mathcal{F}$ containing $\{0,1\}$-valued functions, $\mathrm{Pdim}(\mathcal{F})$ is equal the VC dimension, the implication is clear.

The converse of this direct relation is not necessarily true. Specifically, for a set $\mathcal{F}$ of real-valued functions, consider the set $\mathcal{H}$ of subgraph sets: $h_f(x,y) = \mathbb{I}[y \leq f(x)]$, $f \in \mathcal{F}$. In particular, note that the VC dimension of $\mathcal{H}$ is precisely $\mathrm{Pdim}(\mathcal{F})$. It is *not* true that any realizable classification compression scheme for $\mathcal{H}$ is also an agnostic compression scheme for $\mathcal{F}$ under ℓ_1 loss. Nevertheless, this reduction-to-classification approach seems intuitively appealing, and it might possibly be the case that there is some way to *modify* certain types of compression schemes for $\mathcal{H}$ to convert them into agnostic compression schemes for $\mathcal{F}$. Following up on this line of investigation seems the natural next step toward resolving the above general open question.

Similarly, we ask the analogous question for the ℓ_2 loss and approximate sample compression schemes.

Open Problem: Compressing to fat-shattering Number of Points. Let $c > 0$ be an absolute constant. Under the ℓ_2 loss, does every class $\mathcal{F}$ of real-valued functions admit an α-*approximate* agnostic compression scheme of size $\mathrm{fat}(\mathcal{F}, c\alpha)$?

Impact Statement

This work builds upon the community's understanding of machine learning methods. This has a positive impact on the scientific advancement of the field, and may lead to further improvements in our understanding, methodologies and applications of machine learning and AI. While there are not obvious direct societal implications of the present work, the indirect and longer term impact on society may be positive, negative or both depending on how, where and for what machine learning method that will have benefited from our research are used in the future.

Acknowledgements

IA is supported by the Vatat Scholarship from the Israeli Council for Higher Education. AK was partially supported by the Israel Science Foundation (grant No. 1602/19), an Amazon Research Award, and the Ben-Gurion University Data Science Research Center. We thank Mark Kozlenko for helpful comments on an earlier manuscript.

References

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.

Alon, N., Hanneke, S., Holzman, R., and Moran, S. A theory of pac learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 658–671. IEEE, 2022.

Anthony, M. and Bartlett, P. L. Function learning from in-

[2] To make such a function actually binary-valued everywhere, it suffices to threshold at $1/2$.

terpolation. *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., Ben-David, S., Harvey, N. J., Liaw, C., Mehrabian, A., and Plan, Y. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6): 1–42, 2020.

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

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

Attias, I., Hanneke, S., Kalavasis, A., Karbasi, A., and Velegkas, G. Optimal learners for realizable regression: Pac learning and online learning. *Advances in Neural Information Processing Systems*, 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.

Ben-David, S. and Litman, A. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86 (1):3–25, 1998.

Bousquet, O., Hanneke, S., Moran, S., and Zhivotovskiy, N. Proper learning, helly number, and an optimal svm bound. In *Conference on Learning Theory*, pp. 582–609. PMLR, 2020.

Brukhim, N., Carmon, D., Dinur, I., Moran, S., and Yehudayoff, A. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 943–955. IEEE, 2022.

Chernikov, A. and Simon, P. Externally definable sets and dependent pairs. *Israel J. Math.*, 194(1):409–425, 2013.

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. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

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

Dodge, Y. Least absolute deviation regression. *The Concise Encyclopedia of Statistics*, pp. 299–302, 2008.

Floyd, S. Space-bounded learning and the vapnik-chervonenkis dimension. In *Proceedings of the second annual workshop on Computational learning theory*, pp. 349–364. Morgan Kaufmann Publishers Inc., 1989.

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

Floyd, S. and Warmuth, M. K. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995b.

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.

Gottlieb, L.-A., Kontorovich, A., and Nisnevitch, P. Near-optimal sample compression for nearest neighbors. *Advances in Neural Information Processing Systems*, 27, 2014.

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, 2005a.

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

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

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

Helmbold, D., Sloan, R., and Warmuth, M. K. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

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.

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.

Kontorovich, A., Sabato, S., and Weiss, R. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. *Advances in Neural Information Processing Systems*, 30, 2017.

Kuzmin, D. and Warmuth, M. K. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007.

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

Littlestone, N. and Warmuth, M. K. Relating data compression and learnability. Technical report, Department of Computer and Information Sciences, Santa Cruz, CA, Ju, 1986b.

Livni, R. and Simon, P. Honest compressions and their application to compression schemes. In *Conference on Learning Theory*, pp. 77–92, 2013.

Mendelson, S. Learning without concentration. *J. ACM*, 62 (3):21:1–21:25, 2015.

Montasser, O., Hanneke, S., and Srebro, N. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pp. 2512–2530. PMLR, 2019.

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, 2020.

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

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

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

Moran, S., Shpilka, A., Wigderson, A., and Yehudayoff, A. Teaching and compressing for low vc-dimension. In *A Journey Through Discrete Mathematics*, pp. 633–656. Springer, 2017.

Pabbaraju, C. Multiclass learnability does not imply sample compression. *arXiv preprint arXiv:2308.06424*, 2023.

Pollard, D. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

Pollard, D. *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, 1990a.

Pollard, D. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–86. JSTOR, 1990b.

Rubinstein, B. I. and Rubinstein, J. H. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13(4), 2012.

Rubinstein, B. I., Bartlett, P. L., and Rubinstein, J. H. Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1): 37–59, 2009.

Warmuth, M. K. Compressing to VC dimension many points. In *Proceedings of the* 16^{th} *Conference on Learning Theory*, 2003.

Wiener, Y., Hanneke, S., and El-Yaniv, R. A compression technique for analyzing disagreement-based active learning. *J. Mach. Learn. Res.*, 16:713–745, 2015.

A. Auxiliary Proofs for Section 3

Our proof relies on several auxiliary results.

Existence of approximate weak learners. We start with a result about generalization from interpolation. Anthony & Bartlett (2000) established such a result for interpolation models (Anthony et al. (1999, Section 21.4)), where the cut-off parameter $\eta > 0$ is fixed. The following results extend to cut-offs that may differ for different points. A similar result appeared in Attias & Hanneke (2023) in the context of adversarially robust learning.

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

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

then with probability at least $1-\delta$ over S, for any $f \in \mathcal{F}$ satisfying $|f(x) - y| \le \psi(x,y) + \alpha$, $\forall (x,y) \in S$, it holds that $\mathbb{P}_{(x,y)\sim P}\{(x,y) : |f(x) - y| \le \psi(x,y) + 2\alpha\} \ge 1 - \epsilon$.

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

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

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

Proof of Theorem A.1. 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

$$\mathcal{F}_1 = \{(x,y) \mapsto |f(x) - y| - \psi(x,y) : f \in \mathcal{F}\},$$

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

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

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

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

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 $r = (r_1,\ldots,r_k) \in [0,1]^k$ 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 $\mathcal{F}_2$ with $t(x) = 0$ and $\eta = \alpha$. □

The following boosting and sparsification claims were proven for the case of a fixed cut-off parameter. The proofs extend similarly to the case of a changing cut-off parameter $\psi : \mathcal{X}\times\mathcal{Y} \to [0,1]$.

Boosting. Following (Hanneke et al., 2019), we define the weighted median as

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

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

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

We define $Q_\beta^+(f_1,\ldots,f_T;\alpha_1,\ldots,\alpha_T)(x) = Q_\beta^+(f_1(x),\ldots,f_T(x);\alpha_1,\ldots,\alpha_T)$, and $Q_\beta^-(f_1,\ldots,f_T;\alpha_1,\ldots,\alpha_T)(x) = Q_\beta^-(f_1(x),\ldots,f_T(x);\alpha_1,\ldots,\alpha_T)$, and $\mathrm{Median}(f_1,\ldots,f_T;\alpha_1,\ldots,\alpha_T)(x) = \mathrm{Median}(f_1(x),\ldots,f_T(x);\alpha_1,\ldots,\alpha_T)$. We omit the weights α_i when they are equal to each other. The following guarantee holds for the boosting procedure.

Lemma A.3. *Let $S = \{(x_i,y_i)\}_{i=1}^m$, $T = O\Big(\frac{1}{\beta^2}\log(m)\Big)$. Let $\hat{f}_1,\ldots,\hat{f}_T$ and $\alpha_1,\ldots,\alpha_T$ be the functions and coefficients returned from the median boosting procedure with changing cut-offs (Step 2 in Algorithm 1). For any $i \in \{1,\ldots,m\}$ it holds that*

$$\max\left\{ \left| Q_{\beta/2}^+(\hat{f}_1,\ldots,\hat{f}_T;\alpha_1,\ldots,\alpha_T))(x_i) - y_i \right|, \left| Q_{\beta/2}^-\hat{f}_1,\ldots,\hat{f}_T;\alpha_1,\ldots,\alpha_T)(x_i) - y_i \right| \right\} \leq \psi(x,y) + 2\alpha.$$

Sparsification.

Lemma A.4. *Choosing*

$$n = \Theta\left(\frac{1}{\beta^2} \mathrm{fat}^*(\mathcal{F}, c\alpha) \log^2(\mathrm{fat}^*(\mathcal{F}, c\alpha)\,/\alpha) \right)$$

in Step 3 of Algorithm 1, we have for all $(x,y) \in S$ $|\mathrm{Median}(f_{J_1}(x),\ldots,f_{J_n}(x)) - y| \leq \psi(x,y) + 3\alpha$, where $c > 0$ is a universal constant.

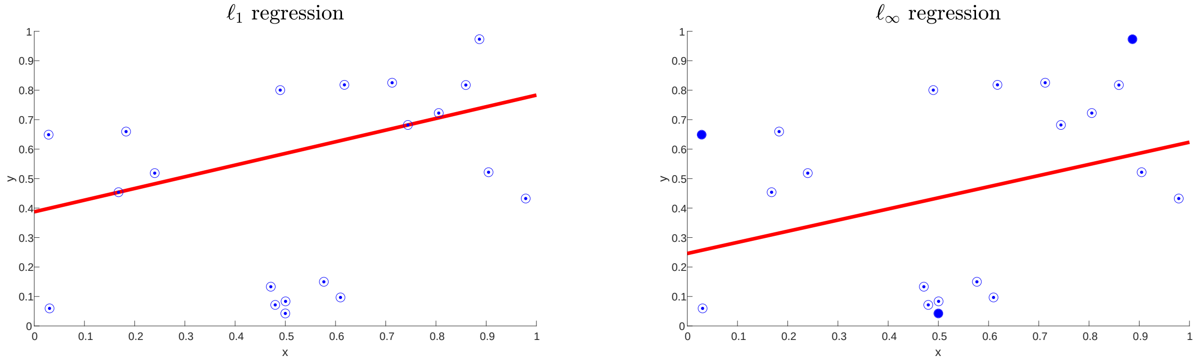

Figure 1. A sample S of $m = 20$ points (x_i, y_i) was drawn iid uniformly from $[0, 1]^2$. On this sample, ℓ_1 regression was performed by solving the LP in (7), shown on the left, and ℓ_∞ regression was performed by solving the LP in (9), on the right. In each case, the regressor provided by the LP solver is indicated by the thick (red) line. Notice that for ℓ_1, the line contains exactly 2 datapoints. For ℓ_∞, the regressor contains no datapoints; rather, the $d + 2 = 3$ "support vectors" are indicated by ●.

B. Missing Proofs for Section 4

Proof of Theorem 4.3. We start with $d = 0$. The sample then consists of $(y_1, \ldots, y_m)$ [formally: pairs (x_i, y_i), where $x_i \equiv 0$], and $\mathcal{F} = \mathbb{R}$ [formally, all functions $h : 0 \mapsto \mathbb{R}$]. We define f_S to be the median of $(y_1, \ldots, y_m)$, which for odd m is defined uniquely and for even m can be taken arbitrarily as the smaller of the two midpoints. It is well-known that such a choice minimizes the empirical ℓ_1 risk, and it clearly constitutes a compression scheme of size 1.

The case $d = 1$ will require more work. The sample consists of $(x_i, y_i)_{i\in[m]}$, where $x_i, y_i \in \mathbb{R}$, and $\mathcal{F} = \{\mathbb{R} \ni x \mapsto wx + b : a, b \in \mathbb{R}\}$. Let $(w^\star, b^\star)$ be a (possibly non-unique) minimizer of

$$L(w, b) := \sum_{i\in[m]} |(wx_i + b) - y_i|, \tag{6}$$

achieving the value $L^\star$. We claim that we can always find two indices $\hat{\imath}, \hat{\jmath} \in [m]$ such that the line determined by $(x_{\hat{\imath}}, y_{\hat{\imath}})$ and $(x_{\hat{\jmath}}, y_{\hat{\jmath}})$ also achieves the optimal empirical risk $L^\star$. More precisely, the line $(\hat{w}, \hat{b})$ induced by $((x_{\hat{\imath}}, y_{\hat{\imath}}), (x_{\hat{\jmath}}, y_{\hat{\jmath}}))$ via[3] $\hat{w} = (y_{\hat{\jmath}} - y_{\hat{\imath}})/(x_{\hat{\jmath}} - x_{\hat{\imath}})$ and $\hat{b} = y_{\hat{\imath}} - \hat{w}x_{\hat{\imath}}$, verifies $L(\hat{w}, \hat{b}) = L^\star$.

To prove this claim, we begin by recasting (6) as a linear program.

$$\begin{aligned} \min_{(\varepsilon_1,\ldots,\varepsilon_m,w,b)\in\mathbb{R}^{m+2}} \quad & \sum_{i=1}^{m} \varepsilon_i \quad \text{s.t.} \\ \forall i \in [m] \quad & \varepsilon_i \geq 0 \\ \forall i \in [m] \quad & wx_i + b - y_i \leq \varepsilon_i \\ \forall i \in [m] \quad & -wx_i - b + y_i \leq \varepsilon_i. \end{aligned} \tag{7}$$

We observe that the linear program in (7) is feasible with a finite solution (and actually, the constraints $\varepsilon_i \geq 0$ are redundant). Furthermore, any optimal value is achievable at one of the extreme points of the constraint-set polytope $\mathcal{P} \subset \mathbb{R}^{m+2}$. Next, we claim that the extreme points of the polytope $\mathcal{P}$ are all of the form $v \in \mathcal{P}$ with two (or more) of the ε_is equal to 0. This suffices to prove our main claim, since $\varepsilon_i = 0$ in $v \in \mathcal{P}$ iff the (w, b) induced by v verifies $wx_i + b = y_i$; in other words, the line induced by (w, b) contains the point (x_i, y_i). If a line contains two data points, it is uniquely determined by them: these constitute a compression set of size 2. (See illustration in Figure 1.)

Now we prove our claimed property of the extreme points. First, we claim that any extreme point of $\mathcal{P}$ must have at least one ε_i equal to 0. Indeed, let (w, b) define a line. Define

$$b^+ := \min\left\{\tilde{b} \in [b, \infty) : \exists i \in [m], wx_i + \tilde{b} = y_i)\right\}$$

[3] We ignore the degenerate possibility of vertical lines, which reduces to the 0-dimensional case.

and analogously,

$$b^- := \max\left\{\tilde{b} \in (-\infty, b] : \exists i \in [m], wx_i + \tilde{b} = y_i)\right\}.$$

In words, (w, b^+) is the line obtained by increasing b to a maximum value of b^+, where the line (w, b^+) touches a datapoint, and likewise, (w, b^-) is the line obtained by decreasing b to a minimum value of b^-, where the line (w, b^-) touches a datapoint.

Define by $S^+_{a,b} := \{i : |wx_i + b < y_i|\}$ the points above the line defined by (w, b) and $S^-_{a,b} := \{i : |wx_i + b > y_i|\}$ the points below the line defined by (w, b). For a line (w, b) which does not contain a data point we can rewrite the sample loss as

$$\begin{aligned}
L(w,b) &= \sum_{i \in S^+_{a,b}} (y_i - (wx_i + b)) + \sum_{i \in S^-_{a,b}} ((wx_i + b) - y_i) \\
&= \left(\sum_{i \in S^-_{a,b}} x_i - \sum_{i \in S^+_{a,b}} x_i \right) a + \left(|S^-_{a,b}| - |S^+_{a,b}| \right) b + \left(\sum_{i \in S^+_{a,b}} y_i - \sum_{i \in S^-_{a,b}} y_i \right) \\
&=: \lambda a + \mu b + \nu.
\end{aligned}$$

Since for fixed a and $b \in [b^-, b^+]$, the quantities $S^-_{a,b}, S^+_{a,b}$ are constant, it follows that the function $L(w, \cdot)$ is affine in b, and hence minimized at $b^\pm \in \{b^-, b^+\}$. Thus, there is no loss of generality in taking $b^\star = b^\pm$, which implies that the optimal solution's line $(w^\star, b^\star)$ contains a data point $(x_{\hat{\imath}}, y_{\hat{\imath}})$. If the line $(w^\star, b^\pm)$ contains other data points then we are done, so assume to the contrary that $\varepsilon_{\hat{\imath}}$ is the only ε_i that vanishes in the corresponding solution $v^\star \in \mathcal{P}$.

Let $\mathcal{P}_{\hat{\imath}} \subset \mathcal{P}$ consist of all v for which $\varepsilon_{\hat{\imath}} = 0$, corresponding to all feasible solutions whose line contains the data point $(x_{\hat{\imath}}, y_{\hat{\imath}})$. Let us say that two lines $(w_1, b_1), (w_2, b_2)$ are *equivalent* if they induce the same partition on the data points, in the sense of linear separation in the plane. The formal condition is $S^-_{w_1,b_1} = S^-_{w_1,b_1}$, which is equivalent to $S^+_{w_1,b_1} = S^+_{w_1,b_1}$.

Define $\mathcal{P}^\star_{\hat{\imath}} \subset \mathcal{P}_{\hat{\imath}}$ to consist of those feasible solutions whose line is equivalent to $(w^\star, b^\pm)$. Denote by $w^+ := \max\{a : (\varepsilon_1, .., \varepsilon_m, w, b) \in \mathcal{P}^\star_{\hat{\imath}}\}$ and define v^+ to be a feasible solution in $\mathcal{P}^\star_{\hat{\imath}}$ with slope w^+, and analogously, $w^- := \min\{w : (\varepsilon_1, .., \varepsilon_m, w, b) \in \mathcal{P}^\star_{\hat{\imath}}\}$ and $v^- \in \mathcal{P}^\star_{\hat{\imath}}$ with slope w^-. Geometrically this corresponds to rotating the line $(w^\star, b^\star)$ about the point $(x_{\hat{\imath}}, y_{\hat{\imath}})$ until it encounters a data point above and below.

Writing, as above, the sample loss in the form $L(w, b)$, we see that $L(\cdot, b^\pm)$ is affine in a over the range $w \in [w^-, w^+]$ and hence is minimized at one of the endpoints. This furnishes another datapoint $(x_{\hat{\jmath}}, y_{\hat{\jmath}})$ verifying $\hat{w}x_{\hat{\jmath}} + \hat{b} = y_{\hat{\jmath}}$ for $L(\hat{w}, \hat{b}) = L^\star$, and hence proves compressibility into two points for $d = 1$.

Generalizing to $d > 1$ is quite straightforward. We define

$$L(\mathbf{w}, b) = \sum_{i \in [m]} |(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i|$$

and express it as a linear program analogous to (7),

Linear programming for ℓ_1 regression:

$$\begin{aligned}
\min_{(\varepsilon_1, \ldots, \varepsilon_m, \mathbf{w}, b) \in \mathbb{R}^{m+d+1}} \quad & \sum_{i=1}^{m} \varepsilon_i \quad \text{s.t.} && (8) \\
\forall i \in [m] \quad & \varepsilon_i \geq 0 \\
\forall i \in [m] \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \leq \varepsilon_i \\
\forall i \in [m] \quad & - \langle \mathbf{w}, \mathbf{x}_i \rangle - b + y_i \leq \varepsilon_i.
\end{aligned}$$

Given an optimal solution $(\mathbf{w}^\star, b^\star)$, we argue exactly as above that $b^\star$ may be chosen so that the optimal regressor contains some datapoint — say, $(\mathbf{x}_1, y_1)$. Holding $b^\star$ and $\mathbf{w}(j)$, $j \neq 1$ fixed, we argue, as above, that $\mathbf{w}(1)$ may be chosen so that the optimal regressor contains another datapoint (say, $(\mathbf{x}_2, y_2)$). Proceeding in this fashion, we inductively argue that the optimal regressor may be chosen to contain some $d + 1$ datapoints, which provides the requisite compression scheme. □

Proof of Theorem 4.4. Given m labeled points in $\mathbb{R}^d \times \mathbb{R}$, $S = \{(\mathbf{x}_i, y_i) : i \in [m]\}$ and any $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ define the empirical risk

$$L(\mathbf{w}, b) \quad := \quad \max \left\{ |\langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i| : i \in [m] \right\}.$$

We cast the risk minimization problem as a linear program.

Linear programming for ℓ_∞ regression:

$$\begin{aligned} \min_{(\varepsilon, \mathbf{w}, b) \in \mathbb{R}^{d+2}} \ &: \ \varepsilon && (9) \\ s.t. \quad \forall i \ &: \ \varepsilon - \langle \mathbf{w}, \mathbf{x}_i \rangle - b + y_i \geq 0 \\ & \varepsilon + \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \geq 0. \end{aligned}$$

(As before, the constraint $\varepsilon \geq 0$ is implicit in the other constraints.) Introducing the Lagrange multipliers $\lambda_i, \mu_i \geq 0$, $i \in [m]$, we cast the optimization problem in the form of a Lagrangian:

$$\mathcal{L}(\varepsilon, \mathbf{w}, b, \mu_1 \ldots, \mu_m, \lambda_1 \ldots, \lambda_m) \quad = \quad \varepsilon - \sum_{i=1}^{m} \lambda_i \left(\varepsilon - \langle \mathbf{w}, \mathbf{x}_i \rangle - b + y_i \right) - \sum_{i=1}^{m} \mu_i \left(\varepsilon + \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \right).$$

The KKT conditions imply, in particular, that

$$\begin{aligned} \forall i : \quad & \lambda_i(\varepsilon - \langle \mathbf{w}, \mathbf{x}_i \rangle - b + y_i) = 0 \\ & \mu_i(\varepsilon + \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i) = 0. \end{aligned}$$

Geometrically, this means that either the constraints corresponding to the ith datapoint are inactive — in which case, omitting the datapoint does not affect the solution — or otherwise, the ith datapoint induces the active constraint

$$\langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i = \varepsilon. \tag{10}$$

On analogy with SVM, let us refer to the datapoints satisfying (10) as the *support vectors*; clearly, the remaining sample points may be discarded without affecting the solution. Solutions to (9) lie in $\mathbb{R}^{d+2}$ and hence $d + 2$ linearly independent datapoints suffice to uniquely pin down an optimal $(\varepsilon, \mathbf{w}, b)$ via the equations (10).

□

Proof of Theorem 4.6. Consider a sample $(y_1, \ldots, y_m) \in \{0, 1\}^m$. Partition the indices $i \in [m]$ into $S_0 := \{i \in [m] : y_i = 0\}$ and $S_1 := \{i \in [m] : y_i = 1\}$. The empirical risk minimizer is given by

$$\hat{r} := \operatorname*{argmin}_{s \in \mathbb{R}} \sum_{i=1}^{m} |y_i - s|^p.$$

To obtain an explicit expression for $\hat{r}$, define

$$F(s) = \sum_{i=1}^{m} |y_i - s|^p = |S_1|(1 - s)^p + |S_0| s^p =: N_1(1 - s)^p + N_0 s^p.$$

We then compute

$$F'(s) = p N_0 s^{p-1} - p N_1 (1 - s)^{p-1}$$

and find that $F'(s) = 0$ occurs at

$$\hat{s} = \frac{\mu^{1/(p-1)}}{1 + \mu^{1/(p-1)}},$$

where $\mu = N_1/N_0$. A straightforward analysis of the second derivative shows that $\hat{s} = \hat{r}$ is indeed the unique minimizer of F.

Thus, given a sample of size m, the unique minimizer $\hat{r}$ is uniquely determined by N_0 — which can take on any of integer $m+1$ values between 0 and m. On the other hand, every output of a k-selection function κ outputs a multiset $\hat{S} \subseteq S$ of size k' and a binary string of length $k'' = k - k'$. Thus, the total number of values representable by a k-selection scheme is at most

$$\sum_{k'=0}^{k} k' 2^{k-k'} < 2^{k+1} - k,$$

which, for $k < \log m$, is less than m. □

Remark B.1. A more refined analysis, along the lines of David et al. (2016, Theorem 4.1), should yield a lower bound of $k = \Omega(m)$. A technical complication is that unlike the $p = 2$ case, whose empirical risk minimizer has a simple explicit form, the general ℓ_p loss does not admit a closed-form solution and uniqueness must be argued from general convexity principles.