Advances in Robust Statistical Learning Theory

Thesis submitted in partial fulfillment
of the requirements for the degree of
"DOCTOR OF PHILOSOPHY"

by

Idan Attias

Submitted to the Senate of Ben-Gurion University of the Negev

Approved by the advisors
Prof. Aryeh Kontorovich and Prof. Yishay Mansour

Approved by the Dean of the Kreitman School of Advanced Graduate Studies
Prof. Orna Braun-Lewensohn

June 2024

Beer-Sheva

This work was carried out under the supervision of

Prof. Aryeh Kontorovich and **Prof. Yishay Mansour**

In the Department of Computer Science

Faculty of Natural Sciences

Acknowledgements

I would like to express my gratitude to everyone who has supported me throughout this journey. First and foremost, I am deeply grateful to my advisors, Prof. Aryeh Kontorovich and Prof. Yishay Mansour, for their unwavering guidance. Thank you for sharing your expertise with me and for always being available to discuss any topic, from the most academic discussions to personal matters. I couldn't have asked for better mentors.

I would like to express my gratitude to Steve Hanneke for being an invaluable mentor throughout this thesis. I truly enjoyed our conversations and learned something new with each meeting. Thank you for your patience and for sharing your knowledge with me.

I would like to thank my great collaborators, I enjoyed learning from each one of you – Idan Amir, Tomer Koren, Roi Livni, Matan Levi, Edith Cohen, Moshe Shechner, Uri Stemmer, Meni Sadigurschi, Eitan Mashiah, Angelos Assos, Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Alkis Kalavasis, Amin Karbasi, Grigoris Velegkas, Gintare Karolina Dziugaite, Mahdi Haghifam, Daniel M Roy, and Ziyi Liu.

I would like to acknowledge the financial support provided by the Israeli Council for Higher Education through their VATAT scholarship for outstanding PhD students in data science.

On a personal note, I am deeply thankful to my partner, Tannis, for her unconditional support throughout this journey. This PhD is as much yours as it is mine. And to Nina, who has brought light to every room she enters since she was born.

Abstract

Machine learning techniques were initially designed for stationary and benign environments, where the training and test data are assumed to be generated from the same statistical distribution. However, this assumption often doesn't hold in the real world. The existence of intelligent and adaptive adversaries can, to a significant extent, violate this statistical assumption, aiming to disrupt machine learning methods used in critical systems.

Our objective is to investigate the statistical and algorithmic aspects of learning models that are robust to test-time attacks on their inputs, commonly known as evasion attacks or adversarial examples. In our model, the learner receives training data with access to arbitrary perturbation sets, which represent an abstract set of possible perturbations that an adversary may utilize during test time. These perturbation sets effectively capture real-world modeling of adversarial examples, such as ℓ_p ball attacks. The goal of the learner is to make robust predictions even when faced with adversarial perturbations during testing.

In the first part of this thesis, we study the *sample complexity* of *adversarially robust PAC (Probably Approximately Correct) learning* in classification and regression for both supervised and semi-supervised settings. We investigate the following fundamental question:

How many training samples are required to learn a robust hypothesis in the presence of test-time adversarial attacks?

Our findings reveal a fundamentally different landscape compared to non-robust PAC learning. In the supervised setting (Attias, Kontorovich, and Mansour [1]), we applied a regret minimization algorithm to compute a near-optimal hypothesis for the learner and achieved improved sample complexity guarantees compared to prior works. We further investigated semi-supervised learning (Attias, Hanneke, and Mansour [2]), where the learner has access to both labeled and unlabeled examples. The learner aims to limit the amount of labeled data, a significantly more expensive resource than unlabeled data. Our study demonstrates that with sufficient unlabeled data, the labeled sample complexity can be arbitrarily smaller compared to previous works, sharply characterized by a different complexity measure. We provided nearly matching upper and lower bounds on this sample complexity, establishing a gap between supervised and semi-supervised

label complexities, which is known not to hold in standard non-robust PAC learning. Subsequently, we initiated the study of adversarially robust regression within the PAC framework (Attias and Hanneke [3]). We addressed the question of which hypothesis classes are PAC learnable in this setting. Our findings indicate that classes of finite fat-shattering dimension are learnable in both realizable and agnostic settings. Furthermore, convex hypothesis classes, are even properly learnable. In contrast, some non-convex hypothesis classes provably require improper learning algorithms.

The second part of this thesis consists of general techniques we developed that may interest the statistical learning theory community. First, we address a fundamental question in statistical learning theory regarding the complexity of ensemble methods for regression (Attias and Kontorovich [4]). Specifically, we provide estimates on the *fat-shattering dimension*—a property that characterizes learnability in regression—of aggregation rules for real-valued hypothesis classes. The latter consists of all ways of choosing multiple hypotheses from these classes, and computing a pointwise function of them, such as the median, mean, and maximum. Our bound is expressed in terms of the fat-shattering dimensions of the component classes. This contribution has numerous applications and fills an important gap in the literature, as this was previously known only for binary hypothesis classes.

In a different work (Attias, Hanneke, Kontorovich, and Sadigurschi [5]), we study *sample compression schemes* in the context of agnostic regression with the ℓ_p loss, $p \in [1, \infty]$. Sample compression schemes in machine learning are methods used to simplify a learned model by representing it with a smaller, essential subset of the training data and a method to reconstruct the full model from this subset. A notable example of such a scheme is the Support Vector Machine algorithm. We construct a generic sample compression scheme for real-valued hypothesis classes, exhibiting exponential size in the fat-shattering dimension but independent of the sample size. For linear regression, we construct a compression of size linear in the dimension and show that $p \in \{1, \infty\}$ provably provides better guarantees than other ℓ_p losses. Prior to our work, this question was investigated only for realizable regression and classification problems.

Attribution

The content of this thesis includes several papers developed in collaboration with Aryeh Kontorovich, Yishay Mansour, Steve Hanneke, and Meni Sadigurschi. The results presented here were made possible through the effort and insight of everyone involved.

- The content of Chapter 3 is based on the following publication:
  Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for adversarially robust learning. In *Journal of Machine Learning Research (JMLR), 2022.*
- The content of Chapter 4 is based on the following publication:
  Idan Attias, Steve Hanneke, and Yishay Mansour. A characterization of semi-supervised adversarially robust pac learnability. In *Advances in Neural Information Processing Systems (NeurIPS), 2022.*
- The content of Chapter 5 is based on the following publication:
  Idan Attias and Steve Hanneke. Adversarially robust pac learnability of real-valued functions. In *International Conference on Machine Learning (ICML), 2023.*
- The content of Chapter 6 is based on the following publication:
  Idan Attias and Aryeh Kontorovich. Fat-shattering dimension of k-fold aggregations. In *Journal of Machine Learning Research (JMLR), 2024*
- The content of Chapter 7 is based on the following publication:
  Idan Attias, Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Agnostic Sample Compression Schemes for Regression. *In International Conference on Machine Learning (ICML), 2024. Spotlight presentation.*

Contents

List of Algorithms

List of Tables

List of Figures

Chapter 1

Introduction

Machine learning focuses on developing algorithms and statistical models that enable computers to learn from and make decisions based on data. This capability has revolutionized fields such as image recognition, natural language processing, and speech recognition. However, traditional machine learning assumes that both the training and test data are drawn from the same distribution, an assumption that may not hold in adversarial environments where data can be intentionally manipulated to deceive the model.

In this thesis, we study theoretical questions regarding the development of robust machine learning models that can withstand test-time attacks on their inputs, commonly known as adversarial examples. These examples are often imperceptible to humans but can cause significant degradation in the performance of the learned model. Our learning framework extends the traditional PAC learning model to address the challenges posed by adversarial attacks.

1.1 PAC Learning with Test Time Adversarial Attacks

Probably Approximately Correct (PAC) learning has long been a cornerstone for the theoretical understanding of the feasibility and efficiency of learning algorithms (Vapnik and Chervonenkis [6], Valiant [7]). This framework provides a rigorous foundation for understanding the learning process and formalizes the concept of learning a hypothesis that generalizes well from a finite sample of training data. An algorithm is said to successfully learn if, with high probability (over a random sample), it outputs a hypothesis with a low error rate on unseen data at test time. The algorithm is evaluated based on two key parameters: sample complexity, which is the number of training examples needed to achieve a desired level of accuracy, and computational complexity, which refers to the time and resources required to find the hypothesis.

The PAC model assumes that both the training and test data are drawn i.i.d. from the same unknown

probability distribution. In this thesis, we study a variant of this model, called *adversarially robust PAC learning* (Feige, Mansour, and Schapire [8], Montasser, Hanneke, and Srebro [9]), which addresses the challenges posed by adversarial attacks on the inputs observed by the learner at test time. The model is defined as follows.

Let $\mathcal{X}$ be the input space, $\mathcal{Y}$ be the output (label) space, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class that we aim to learn (e.g., hyperplanes, neural networks). Suppose that there is an unknown probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. An adversarial attack is formalized by a perturbation (corruption) function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, where $\mathcal{U}(x)$ is the set of possible perturbations (attacks) on input x. We have no restriction on $\mathcal{U}$, besides $x \in \mathcal{U}(x)$. In practice, we usually consider $\mathcal{U}(x)$ to be the ℓ_p ball centered at x. The goal is to find hypothesis h that has a low *robust generalization error* with respect to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, distribution $\mathcal{D}$, and adversary $\mathcal{U}$, defined as

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

For classification we consider the zero-one loss $\ell(y, y') = \mathbb{I}[y \neq y']$ and for regression we consider the ℓ_p loss $\ell(y, y') = |y - y'|^p$.

A hypothesis class $\mathcal{H}$ is said to be *adversarially robust PAC learnable*, if there exists a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ such that for any $\epsilon, \delta > 0$, any distribution $\mathcal{D}$, and any perturbation function $\mathcal{U}$, the algorithm $\mathcal{A}$ will output hypothesis $\hat{h} = \mathcal{A}(S)$, where S is an i.i.d. sample from $\mathcal{D}$ of size $m = m(\epsilon, \delta, \mathcal{H}, \mathcal{U}) \in \mathbb{N}$, such that

$$\mathrm{Err}_{\ell}(\hat{h}; \mathcal{D}, \mathcal{U}) \leq \inf_{h\in\mathcal{H}} \mathrm{Err}_{\ell}(h; \mathcal{D}, \mathcal{U}) + \epsilon, \quad \text{with probability } 1 - \delta. \tag{1.1}$$

The *sample complexity* is defined as the minimal i.i.d. sample size for which Eq. (1.1) holds. If there exists $h \in \mathcal{H}$ with $\mathrm{Err}_{\ell}(h; \mathcal{D}, \mathcal{U}) = 0$, we call it the *realizable* setting, and otherwise, the *agnostic* setting. This setting recovers the non-robust PAC setting, by just taking the perturbation function to be identity $\mathcal{U}(x) = x$.

1.2 Our Contributions: The Sample Complexity of Adversarially Robust PAC Learning

In the standard PAC setting, hypothesis classes are considered learnable if their VC dimension [see Definition 2.5] in binary classification, or the fat-shattering dimension [see Definition 2.9] in regression, is finite. In the following lines of work, we investigate learnability in the adversarially robust PAC learning model in several settings. While we provide some algorithmic principles in these works, the main focus is on the sample

complexity.

Improved Generalization Bounds for Adversarially Robust Learning [1]

We study a robust learning framework for classification and regression, introduced by Feige, Mansour, and Schapire [8]. In this game-theoretic framework, the learner and the adversary play a half-infinite zero-sum game, where the learner chooses a mixture of hypotheses from a hypothesis class (possibly of infinite size), and the adversary chooses a mixture of perturbations, which is finite for each possible input. We improve the sample complexity bounds for binary classification and also provide bounds for regression.

More formally, we employ a regret minimization algorithm that uses an ERM oracle as a black box. The algorithm provides near-optimal policies for the players on a given training sample. For binary classification, we provide an improved sample complexity guarantee for agnostic (ϵ, δ)-PAC learning of size $\tilde{\mathcal{O}}\left(\frac{k\,\mathrm{VC}(\mathcal{H})+\log(1/\delta)}{\epsilon^2}\right)$[1] where the adversary is limited to k possible corruptions (sometimes referred to as perturbations) and $\mathrm{VC}(\mathcal{H})$ is the Vapnik-Chervonenkis dimension of hypothesis class $\mathcal{H}$. This result extends to multiclass classification where the VC dimension is replaced by the graph dimension or the Natarajan dimension. We extend the algorithm to the regression setting and prove the following upper bound on the sample complexity $\tilde{\mathcal{O}}\left(\frac{k\left(\int_0^1 \mathrm{fat}(\mathcal{H},\gamma)d\gamma\right)+\log(1/\delta)}{\epsilon^2}\right)$[2], where $\mathrm{fat}(\mathcal{H},\gamma)$ is the fat-shattering dimension of hypothesis class $\mathcal{H}$ with a scale of γ. In the course of our work, we introduced the notion of partial concept classes, which had also been implicitly studied previously by Long [10]. This concept has since become highly influential in the learning theory community, largely due to the contributions of Alon, Hanneke, Holzman, and Moran [11].

A Characterization of Semi-Supervised Adversarially Robust PAC Learnability [2]

Adversarial robustness has been empirically shown to significantly benefit from semi-supervised learning, where the learner has access to both labeled and unlabeled examples. We address the question of how many labeled and unlabeled examples are required to ensure learnability in the PAC model. We show that with enough unlabeled data (the size of a labeled sample that a fully-supervised method would require), the labeled sample complexity can be arbitrarily smaller compared to previous works, and is sharply characterized by a different complexity measure. We prove nearly matching upper and lower bounds on this sample complexity. This establishes a gap between supervised and semi-supervised label complexities, which does not hold in standard non-robust PAC learning.

More formally, denote by Λ^s the sample complexity for supervised robust learning. We show that in the

[1] $\tilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors in the specified expression.
[2] If the integral diverges at 0, we can truncate the integral and the bound still holds.

realizable setting, having Λ^s *unlabeled* samples, and $\Lambda^{ss} = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})+\log(1/\delta)}{\epsilon}\right)$ labeled samples suffice for (ϵ, δ) robust PAC learnability, where $\mathrm{VC}_{\mathcal{U}}$ is an extension to the VC dimension, $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$, and for some hypothesis classes, there can be an infinite gap between the dimensions (see Definition 4.1). Moreover, we show that for some hypothesis classes, $\Lambda^{ss} \ll \Lambda^s$ and the gap can be arbitrarily large. We also show a similar behavior in the agnostic setting.

Adversarially Robust PAC Learnability of Real-Valued Functions [3]

We initiated the study of adversarially robust regression within the PAC framework. This model is different from the one in Attias et al. [1] for the following reasons: the adversary is allowed to choose perturbations from an arbitrary set, possibly of infinite size, and a richer set of loss functions. We addressed the question of which hypothesis classes are PAC learnable in this setting. Our findings indicate that classes of finite fat-shattering dimension are learnable in both realizable and agnostic settings. Furthermore, convex hypothesis classes are even properly (i.e., the algorithm is constrained to return a hypothesis within the hypothesis class) learnable. In contrast, some non-convex function classes provably require improper learning algorithms [9]. On the other hand, non-robust regression with the mean squared error is properly learnable [12].

More formally, our main result is an upper bound on the sample complexity of agnostic robust regression with the ℓ_p loss: $\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}(\mathcal{H},\epsilon/p)\mathrm{fat}^*(\mathcal{H},\epsilon/p)}{\epsilon^2}\right)$, where $\mathrm{fat}^*(\mathcal{H}, \epsilon/p)$ is the fat-shattering dimension of the dual-class, which is finite as long as the primal dimension is finite (see Definition 2.17). We also show that convex classes are properly learnable with a larger sample complexity: $\tilde{\mathcal{O}}\left(\frac{\mathrm{fat}^3(\mathcal{H},\epsilon/p)\mathrm{fat}^*(\mathcal{H},\epsilon/p)}{\epsilon^5}\right)$.

1.3 Contributions to Statistical Learning Theory

Fat-Shattering Dimension of k-fold Aggregations [4]

In many scenarios in machine learning, the output of the model consists of an ensemble of models. This idea can be formalized by aggregation rules of hypothesis classes, which consists of all ways of choosing k functions, one from each of the k classes, and computing pointwise an "aggregate" function of these, such as the median, mean, and maximum. This finds application in adversarially robust learning, clustering methods, learning polyhedra with a margin, and ensemble methods such as bootstrap aggregation and boosting.

We provide near-optimal estimates on the fat-shattering dimension of such aggregation rules of real-valued hypothesis classes, where our bounds are stated in terms of the fat-shattering dimensions of the component classes. For linear and affine hypothesis classes, we provide a considerably sharper upper bound and a matching lower bound, achieving, in particular, an optimal dependence on k. Prior to our work, this question has been investigated mainly for binary hypothesis classes.

More formally, we prove the following upper bounds for a natural class of aggregation rules. Assuming $\mathrm{fat}(\mathcal{H}_i, \gamma) \leq d$, for $1 \leq i \leq k, \gamma > 0$, then the fat-shattering of the aggregated class over $\mathcal{H}_1, \ldots, \mathcal{H}_k$ is upper bounded by $\mathcal{O}\left(dk \log^2(dk)\right)$. Moreover, if the hypothesis classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$ are bounded in $[-R, R]$ and $\mathrm{fat}(\mathcal{H}_i, \epsilon\gamma) \leq d$, we show an upper bound of $O\left(dk \log^{1+\epsilon} \frac{Rk}{\gamma}\right)$, where $\epsilon, \gamma > 0$.

For affine hypothesis classes, we can provide tighter bounds. If the affine hypothesis classes $\mathcal{H}_1, \ldots, \mathcal{H}_k$ are bounded in $[-R, R]$, we prove a dimension-free upper bound on the fat-shattering dimension at scale γ of the aggregated classes: $\mathcal{O}\left(\frac{R^2 k \log(k)}{\gamma^2}\right)$. Additionally, for the maximum aggregation rule and affine hypothesis classes (not necessarily bounded), we show a tight dimension-dependent bound $\Theta(dk \log(k))$, where d is the Euclidean dimension.

Agnostic Sample Compression Schemes for Regression [5]

Sample compression is a central problem in learning theory, introduced by Littlestone and Warmuth [13], Floyd and Warmuth [14], whereby one seeks to retain a "small" subset of the labeled sample (called compression set) from which a "good" hypothesis can be reconstructed. 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) and good to mean "consistent with the sample" specifies the classic realizable sample compression problem for VC classes.

The benefits of sample compression schemes are the following. *Generalization*: Models that can be effectively compressed often generalize better to unseen data. Specifically in the PAC model, sample compression generalization bounds still hold when uniform convergence bounds become vacuous. These bounds have proven useful in a wide range of learning settings, such as multiclass classification, regression, adversarially robust learning, active learning, and density estimation. *Computation and space efficiency*: By reducing the size of the dataset, computational resources required for training and making predictions are significantly reduced, and smaller models require less storage space, which is beneficial for applications with limited memory or storage capacity. A classic example of a sample compression scheme is the Support Vector Machine (SVM), where the model can be represented by a small subset of the training examples that determine the decision boundary, known as support vectors.

We studied sample compression schemes in the agnostic regression setting with the ℓ_p loss, $p \in [1, \infty]$. Prior to our work, this question was investigated only for classification and realizable regression. First, we construct a generic α-*approximate* sample compression scheme for real-valued function class $\mathcal{H}$ of size $\tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{H}, \alpha/p)\, \mathrm{fat}^*(\mathcal{H}, \alpha/p))$, where α-approximation means that the function reconstructed from the compression set achieves an average error at most α compared to the optimal hypothesis in the class. Notably,

for linear regression, we construct an approximate compression of size linear in the Euclidean dimension. Finally, we show a separation between ℓ_1 and ℓ_∞ linear regression to other ℓ_p losses. While for ℓ_1 and ℓ_∞ we can construct an *exact* sample compression scheme of smaller size, we show that such compression cannot exist for $p \in (1, \infty)$.

1.4 Excluded Work

During my PhD studies, I have also contributed to other projects that extend beyond the scope of this dissertation. These additional works include:

- Idan Amir, Idan Attias, Tomer Koren, Yishay Mansour, and Roi Livni. Prediction with corrupted expert advice. In *Advances in Neural Information Processing Systems (NeurIPS), 2020. Spotlight presentation. [15]*
- Matan Levi, Idan Attias, and Aryeh Kontorovich. Domain invariant adversarial learning. In *Transactions on Machine Learning Research (TMLR), 2022.* [16]
- Idan Attias, Edith Cohen, Moshe Shechner, and Uri Stemmer. A framework for adversarial streaming via differential privacy and difference estimators. In *Innovations in Theoretical Computer Science Conference (ITCS), 2023*, and *Algorithmica, 2024.* [17]
- Eitan-Hai Mashiah, Idan Attias, and Yishay Mansour. Learning revenue maximization using posted prices for stochastic strategic patient buyers. In *Conference on Artificial Intelligence (AAAI), 2023.* [18]
- Angelos Assos, Idan Attias, Yuval Dagan, Constantinos Daskalakis, and Maxwell Fishelson. Online learning and solving infinite games with an erm oracle. In *Conference on Learning Theory (COLT), 2023.* [19]
- Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: Pac learning and online learning. In *Advances in Neural Information Processing Systems (NeurIPS), 2023. Oral presentation.* [20]
- Idan Attias, Gintare Karolina Dziugaite, Mahdi Haghifam, Roi Livni, and Daniel M Roy. Information complexity of stochastic convex optimization: Applications to generalization, memorization, and tracing. In *International Conference on Machine Learning (ICML), 2024. Best paper award.* [21]
- Ziyi Liu, Idan Attias, and Daniel M Roy. Causal bandits: The Pareto optimal frontier of adaptivity, a reduction to linear bandits, and limitations around unknown marginals. In *International Conference on Machine Learning (ICML), 2024.* [22]

- Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Universal rates for regression: Separations between cut-off and absolute loss. In *Conference on Learning Theory (COLT), 2024.* [23]

- Ziyi Liu, Idan Attias, and Daniel M Roy. Sequential probability assignment with contexts: Minimax regret, contextual Shtarkov sums, and contextual normalized maximum likelihood. In *Advances in Neural Information Processing Systems (NeurIPS), 2024.* [24]

Chapter 2

Background and Preliminaries

2.1 The PAC Model

Let $\mathcal{X}$ be the input space, $\mathcal{Y}$ be the output (label) space, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class that we aim to learn (e.g., hyperplanes, neural networks). Suppose there is an unknown probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The goal is to find hypothesis that has a low *generalization error* with respect to a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ and distribution $\mathcal{D}$:

Definition 2.1 (Generalization error) *The generalization error of function $f : \mathcal{X} \to \mathcal{Y}$ with respect to distribution $\mathcal{D}$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is defined by*

$$\mathrm{Err}_\ell(f; \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]\,.$$

For classification we consider the zero-one loss $\ell(y, y') = \mathbb{I}[y \neq y']$ and for regression we consider the ℓ_p loss $\ell(y, y') = |y - y'|^p$ for $p \geq 1$.

We define the average error of a function with respect to a set of labeled examples as the empirical error:

Definition 2.2 (Empirical error) *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a loss function, and let $S = \{(x_i, y_i)\}_{i=1}^m$. The empirical error of a function f on S is defined by*

$$\widehat{\mathrm{Err}}_\ell(f; S) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i).$$

Definition 2.3 (PAC learnability [7]) *For any $\epsilon, \delta \in (0, 1)$, the sample complexity of realizable (ϵ, δ)-PAC learning for a class $\mathcal{H}$, denoted by $\mathcal{M}_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H})$, is the smallest integer m for which there exists a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ realizable by $\mathcal{H}$, namely*

$\inf_{h\in\mathcal{H}} \mathrm{Err}_\ell\left(h;\mathcal{D}\right)=0$, *for a random sample* $S\sim\mathcal{D}^m$, *it holds that*

$$\mathbb{P}\left(\mathrm{Err}_\ell\left(\mathcal{A}(S);D\right)\leq\epsilon\right)>1-\delta.$$

If no such m *exists, define* $\mathcal{M}_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H})=\infty$, *and* $\mathcal{H}$ *is not* (ϵ,δ)-PAC *learnable in the realizable case.*

The agnostic setting extends PAC learning for general distributions, where there is not necessarily a hypothesis in $\mathcal{H}$ with zero generalization error on $\mathcal{D}$, namely, $\inf_{h\in\mathcal{H}} \mathrm{Err}_\ell\left(h;\mathcal{D}\right)>0$.

Definition 2.4 (Agnostic PAC learnability [25]) *For any* $\epsilon,\delta\in(0,1)$, *the sample complexity of agnostic* (ϵ,δ)-PAC *learning for a class* $\mathcal{H}$, *denoted by* $\mathcal{M}_{\mathrm{AG}}(\epsilon,\delta,\mathcal{H})$, *is the smallest integer* m, *for which there exists a learning algorithm* $\mathcal{A}:(\mathcal{X}\times\mathcal{Y})^*\to\mathcal{Y}^{\mathcal{X}}$, *such that for every distribution* $\mathcal{D}$ *over* $\mathcal{X}\times\mathcal{Y}$, *for a random sample* $S\sim\mathcal{D}^m$, *it holds that*

$$\mathbb{P}\left(\mathrm{Err}_\ell\left(\mathcal{A}(S);\mathcal{D}\right)\leq\inf_{h\in\mathcal{H}}\mathrm{Err}_\ell\left(h;\mathcal{D}\right)+\epsilon\right)>1-\delta.$$

If no such m *exists, define* $\Lambda_{\mathrm{AG}}(\epsilon,\delta,\mathcal{H})=\infty$, *and* $\mathcal{H}$ *is not* (ϵ,δ)-PAC *learnable in the agnostic setting.*

2.2 The Sample Complexity of PAC Learning in Binary Classification

In this setting, the label space is $\mathcal{Y}=\{0,1\}$ and the loss function is the zero-one: $(y,y')\mapsto\mathbb{I}[y\neq y']$. The VC dimension characterizes the PAC learnability of binary-valued hypothesis classes.

Definition 2.5 (VC dimension [26]) *Denote the projection of a hypothesis class* $\mathcal{H}$ *on set* $S=\{x_1,\ldots,x_k\}$ *by* $\mathcal{H}|_S=\{(h(x_1),\ldots,h(x_k)):h\in\mathcal{H}\}$. *We say that a set* $S\subseteq\mathcal{X}$ *is shattered by* $\mathcal{H}$ *if* $\{0,1\}^S=\mathcal{H}|_S$. *The* VC*-dimension of* $\mathcal{H}$ *is defined as the maximal size of a shattered set* S *(possibly* ∞*).*

Theorem 2.6 (Sample complexity of binary classification [27],[28],[29]) *The sample complexity of* (ϵ,δ)-PAC *learning a hypothesis class* $\mathcal{H}\subseteq\{0,1\}^{\mathcal{X}}$*:*

$$\Theta\left(\frac{\mathrm{VC}(\mathcal{H})+\log\frac{1}{\delta}}{\epsilon}\right),$$

and the sample complexity of agnostic (ϵ,δ)-PAC *learning is*

$$\Theta\left(\frac{\mathrm{VC}(\mathcal{H})+\log\frac{1}{\delta}}{\epsilon^2}\right),$$

Every hypothesis class with a finite VC dimension is learnable by empirical risk minimization with near-optimal sample complexity.

Definition 2.7 (Empirical risk minimizer (ERM)) *The empirical risk minimizer learning algorithm* $\mathrm{ERM} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ *for a class* $\mathcal{H}$ *with respect to a sequence* S *is defined by* $\mathrm{ERM}_{\mathcal{H}}(S) \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathrm{Err}}_\ell(h; S)$.

Theorem 2.8 (VC classes are learnable by empirical risk minimization [27]) *Any hypothesis class* $\mathcal{H}$ *is* (ϵ, δ)-PAC *learnable by an ERM with sample of size* $\Omega\left(\frac{\mathrm{VC}(\mathcal{H}) \log(1/\epsilon) + \log(1/\delta))}{\epsilon}\right)$, *and agnostic* (ϵ, δ)-PAC *learnable with sample of size* $\Omega\left(\frac{\mathrm{VC}(\mathcal{H}) + \log(1/\delta))}{\epsilon}\right)$.

2.3 The Sample Complexity of PAC Learning in Regression

In this setting, the label space is $\mathcal{Y} = [0, 1]$ and the loss function is the ℓ_p loss: $(y, y') \mapsto |y - y'|^p$, $p \geq 1$. The fat-shattering dimension characterizes the PAC learnability of real-valued hypothesis classes.

Definition 2.9 (Fat-shattering dimension [30–32]) *For* $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ *and* $\gamma > 0$, *we say that* $\mathcal{F}$ γ-*shatters a set* $S = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ *if there exists an* $r = (r_1, \ldots, r_m) \in \mathbb{R}^m$ *such that for each* $b \in \{-1, 1\}^m$ *there is a function* $f_b \in \mathcal{F}$ *such that*

$$\forall i \in [m] : \begin{cases} f_b(x_i) \geq r_i + \gamma & \text{if } b_i = 1 \\ f_b(x_i) \leq r_i - \gamma & \text{if } b_i = -1 \end{cases}$$

We refer to r *as the* shift. *The* γ*-fat-shattering dimension, denoted by* $\mathrm{fat}(\mathcal{F}, \gamma)$, *is the size of the largest* γ*-shattered set (possibly* ∞*). Sometimes we use the notation* $\mathrm{fat}_\gamma(\mathcal{F})$ *instead of* $\mathrm{fat}(\mathcal{F}, \gamma)$.

Also, sometimes we use the following compact way to describe a shattered set. A set $S = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ *is said to be* γ*-shattered by* $\mathcal{F}$ *if*

$$\sup_{r \in \mathbb{R}^m} \min_{y \in \{-1,1\}^m} \sup_{f \in F} \min_{i \in [m]} b_i (f(x_i) - r_i) \geq \gamma. \tag{2.1}$$

Theorem 2.10 (Sample complexity for (agnostic) regression [32],[33]) *Let* $\mathcal{L}^p_{\mathcal{H}}$ *be the* ℓ_p *loss class of hypothesis class* $\mathcal{H}$,

$$L^p_{\mathcal{H}} := \{\mathcal{X} \times [0, 1] \ni (x, y) \mapsto |h(x) - y|^p : h \in \mathcal{H}\}.$$

The sample complexity of (ϵ, δ)-PAC *learning a hypothesis class* $L^p_{\mathcal{H}}$ *is*

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

where $c > 0$ *is a universal constant.*

2.4 Other Complexity Measures and Generalization Bounds

It is a classic fact [28, Theorem 3.3] that the Rademacher complexity controls generalization bounds in a wide range of supervised learning settings.

Definition 2.11 (Rademacher complexity) *Let $\mathcal{F}$ be a real-valued function class on the domain space $\mathcal{W}$. Define the empirical Rademacher complexity of $\mathcal{F}$ on a given sequence $(w_1, \ldots, w_n) \in \mathcal{W}^n$ as*

$$\mathcal{R}_n(\mathcal{F}|w_1, \ldots, w_n) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(w_i),$$

where $\sigma = (\sigma_1, \ldots \sigma_n)$ are independent random variables uniformly chosen from $\{-1, 1\}$. The Rademacher complexity of $\mathcal{F}$ with respect to a distribution $\mathcal{D}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{w_1, \ldots, w_n \sim \mathcal{D}} \mathcal{R}_n(\mathcal{F}|w_1, \ldots, w_n).$$

Theorem 2.12 (Rademacher complexity generalization bound [28]) *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{W}$ to $[0, 1]$. Then, for any $\delta > 0$ with probability at least $1 - \delta$ over the draw of an i.i.d. sample $\mathbf{w} = (w_1, \cdots, w_n)$ from distribution D, for all $g \in \mathcal{G}$:*

$$E_{w \sim D}[g(w)] - \frac{1}{n} \sum_{i=1}^{n} g(w_i) \leq 2R_n(\mathcal{G}|\mathbf{w}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Theorem 2.13 ([34, 35]) *For any $\mathcal{F} \subseteq [-1, 1]^X$, any $\gamma \in (0, 1)$ and $\mathbf{w} = (\mathbf{w_1}, \ldots, \mathbf{w_n}) \in \mathcal{W}^n$,*

$$\mathcal{R}_n(\mathcal{F}|\mathbf{w}) \leq \sqrt{\frac{C}{n}} \int_0^1 \sqrt{\mathrm{fat}_{c\gamma}(\mathcal{F}) \log\left(\frac{2}{\gamma}\right)} d\gamma,$$

where c and C are universal constants.

When the integral above diverges, the bound can be refined by

$$\mathcal{R}_n(\mathcal{F}|\mathbf{w}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \sqrt{\frac{C}{n}} \int_\alpha^1 \sqrt{\mathrm{fat}_{c\gamma}(\mathcal{F}) \log\left(\frac{2}{\gamma}\right)} d\gamma \right\}.$$

Theorem 2.14 ([36]) *For any $\mathcal{F} \subseteq [-1, 1]^X$ and $\mathbf{w} = (\mathbf{w_1}, \ldots, \mathbf{w_n}) \in \mathcal{W}^n$,*

$$\mathcal{R}_n(\mathcal{F}|\mathbf{w}) \leq C\sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}},$$

where C is a universal constant.

Definition 2.15 (Uniform covering numbers) *We say that $\mathcal{G} \subseteq [0, 1]^\Omega$ is ϵ-cover for $\mathcal{F} \subseteq [0, 1]^\Omega$ in $\|\cdot\|_\infty$*

norm, if for any $f \in \mathcal{F}$ the exists $g \in \mathcal{G}$ such that for any $x \in \Omega$, $|f(x) - g(x)| \leq \epsilon$. The ϵ-covering number of $\mathcal{F}$ is the minimal cardinality of any ϵ-cover, and denoted by $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$.

Lemma 2.16 (Uniform covering numbers upper bound [37] (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 $0 < t < 1/2$,*

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

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

Definition 2.17 (Dual dimensions) *Let $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$ be a hypothesis class. The dual hypothesis class $\mathcal{F}^* \subseteq \mathcal{Y}^\mathcal{F}$ is defined as the set of all functions $g_x : \mathcal{F} \to \mathcal{Y}$ where $g_x(f) = f(x)$. 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.*

For $\mathcal{Y} = \{0,1\}$, we denote the VC*-dimension of the dual class by $\mathrm{VC}^*(\mathcal{F})$. It is known that $\mathrm{VC}^*(\mathcal{F}) < 2^{\mathrm{VC}(\mathcal{F})+1}$ (Assouad [38]). For $\mathcal{Y} = [0,1]$ we denote the fat-shattering dimension at scale γ of the dual-class by $\mathrm{fat}^*(\mathcal{F}, \gamma)$. We have the following bound due to Kleer and Simon [39, Corollary 3.8 and inequality 3.1],*

$$\mathrm{fat}^*(\mathcal{F}, \gamma) \lesssim \frac{1}{\gamma} 2^{\mathrm{fat}(\mathcal{F}, \gamma/2)+1}. \tag{2.2}$$

2.5 Sample Compression Schemes

Definition 2.18 (Approximate and exact sample compression schemes) *Following David, Moran, and Yehudayoff [40], 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 $\{(x_1, y_1), \ldots, (x_m, y_m)\} \in \bigcup_{\ell \geq 1} \{\mathcal{X} \times \mathcal{Y}\}^\ell$ to elements in $\mathcal{K} = \bigcup_{\ell \leq k'} \{\mathcal{X} \times \mathcal{Y}\}^\ell \times \bigcup_{\ell \leq k''} \{0,1\}^\ell$, where $k' + k'' \leq k$. 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 $\mathcal{F}$ and loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ 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:*

$$\widehat{\mathrm{Err}}_\ell(f_S; S) \leq \inf_{f \in \mathcal{F}} \widehat{\mathrm{Err}}_\ell(f; S).$$

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

$$\widehat{\mathrm{Err}}_\ell(f_S; S) \leq \inf_{f \in \mathcal{F}} \widehat{\mathrm{Err}}_\ell(f; S) + \alpha.$$

The classic sample compression definition for realizable classification by Littlestone and Warmuth [13], Floyd and Warmuth [14] is for exact compression scheme where ℓ is the 0-1 *loss and* $\inf_{f\in\mathcal{F}} \widehat{\mathrm{Err}}_\ell(f;S) = 0$

One of the advantages of the sample compression schemes is the ability to generalize beyond uniform convergence. The following generalization bound is a variation of the classic bound by Graepel et al. [41]. It follows the same arguments while using the empirical Bernstein bound instead of Hoeffding's inequality. A variation of this bound, with respect to the 0-1 loss, appears in [11, Lemma 42], and [42, Section 5]. For full proof see the lecture notes on sample compression in the context of adaptive data analysis [43].

Lemma 2.19 (Data-Dependent Sample Compression Generalization Bound) *For any sample compression scheme (κ,ρ), for any $m\in\mathbb{N}$, $\delta\in(0,1)$, loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,1]$, for any distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, for $S\sim\mathcal{D}^m$, with probability $1-\delta$,*

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

Note that this bound applies for any bounded loss, in particular to the robust loss of ℓ: $(x,y)\mapsto\sup_{z\in\mathcal{U}(x)}\ell(f(z),y)$.

2.6 Adversarially Robust PAC Learning

Recall the extension of PAC learning to handle test-time adversarial attacks, as discussed in Section 1.1. The main difference is that at test time, the learner will only observe the adversarially perturbed example and not the original one, while the training time does not change. As a result, we are now interested in minimizing the robust generalization error:

$$\mathrm{Err}_\ell(f;\mathcal{D},\mathcal{U}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{z\in\mathcal{U}(x)}\ell(f(z),y)\right].$$

In Theorem 3.1 (Chapter 3) we show that in the case of finite perturbation sets, $|\mathcal{U}(x)|\leq k$, and when ℓ is the zero-one loss, the VC of the robust loss class, $L^{\mathcal{U}}_{\mathcal{H}} = \left\{(x,y)\mapsto\max_{z\in\mathcal{U}(x)}\mathbb{I}\left[h(z)\neq y\right] : h\in\mathcal{H}\right\}$, is upper bounded by $\mathcal{O}(\mathrm{VC}(\mathcal{H})\log(k))$. This means that we can minimize the empirical robust error and the sample complexity grows with $\mathcal{O}(\mathrm{VC}(\mathcal{H})\log(k))$.

On the other hand, Montasser, Hanneke, and Srebro [9] showed that when $\mathcal{U}(x)$ is infinite, this is no longer true, and as a result, robust empirical risk (error) minimization might fail. At the same time, generalization bounds based on sample compression schemes still hold.

Chapter 3

Improved Generalization Bounds for Adversarially Robust Learning

We consider a model of robust learning in an adversarial environment. The learner gets uncorrupted training data with access to possible corruptions that may be affected by the adversary during testing. The learner's goal is to build a robust classifier, which will be tested on future adversarial examples. The adversary is limited to k possible corruptions for each input. We model the learner-adversary interaction as a zero-sum game. This model is closely related to the adversarial examples model of Schmidt et al. [44], Madry et al. [45].

Our main results consist of generalization bounds for the binary and multiclass classification, as well as the real-valued case (regression). For the binary classification setting, we both tighten the generalization bound of Feige et al. [8], and are also able to handle infinite hypothesis classes. The sample complexity is improved from $\mathcal{O}(\frac{1}{\epsilon^4}\log(\frac{|\mathcal{H}|}{\delta}))$ to $\mathcal{O}\big(\frac{1}{\epsilon^2}(k\,\mathrm{VC}(\mathcal{H})\log^{\frac{3}{2}+\alpha}(k\,\mathrm{VC}(\mathcal{H}))+\log(\frac{1}{\delta})\big)$ for any $\alpha > 0$. Additionally, we extend the algorithm and generalization bound from the binary to the multiclass and real-valued cases. Along the way, we obtain results on fat-shattering dimension and Rademacher complexity of k-fold maxima over function classes; these may be of independent interest.

For binary classification, the algorithm of Feige et al. [8] uses a regret minimization algorithm and an ERM oracle as a black box; we adapt it for the multiclass and regression settings. The algorithm provides us with near-optimal policies for the players on a given training sample.

3.1 Introduction

We study the classification and regression problems in a setting of adversarial examples. This setting is different from standard supervised learning in that examples, at testing time, may be corrupted in an adversarial manner to disrupt the learner's performance. As standard supervised learning methods have demonstrated vulnerabilities, the challenge to design reliable robust models has gained significant attention and has been termed *adversarial examples*. We study the adversarially robust learning paradigm from a generalization point of view.

We consider the following robust learning framework for multiclass and real-valued functions of Feige et al. [8]. There is an unknown distribution over the uncorrupted inputs domain. The learner receives a labeled uncorrupted sample (the labels can be categorical or real-valued) and has knowledge during the training phase of all possible corruptions that the adversary might effect. The learner selects a hypothesis from a fixed hypothesis class (in our case, a mixture of hypotheses from base class $\mathcal{H}$) that gives a prediction (a distribution over predictions) for a corrupted input. The learner's accuracy is measured by predicting the true label of the uncorrupted input while they observe only the corrupted input during test time. Thus, their goal is to find a policy that is robust against those corruptions. The adversary is capable of corrupting each future input, but there are only k possible corruptions for each point in the instance space. This suggests the game-theoretic framework of a zero-sum game between the learner and the adversary. The model is closely related to the one proposed by Schmidt et al. [44], Madry et al. [45] and other common robust optimization approaches [46], which deal with bounded worst-case perturbations (under ℓ_∞ norm) on the samples. In this work, we do not assume any metric for the corruptions: the adversary can map an input from the instance space to any other space, but is limited with finitely many possible corruptions for each input.

Our main results are generalization bounds for classification and regression. For the binary classification setting, we improve the generalization bound given in Feige et al. [8]. In particular, we allow for the use of infinite base hypothesis classes $\mathcal{H}$. The sample complexity has been improved from $\mathcal{O}(\frac{1}{\epsilon^4}\log(\frac{|\mathcal{H}|}{\delta}))$ to $\mathcal{O}\big(\frac{1}{\epsilon^2}(k\,\mathrm{VC}(\mathcal{H})\log^{\frac{3}{2}+\alpha}(k\,\mathrm{VC}(\mathcal{H})) + \log(\frac{1}{\delta})\big)$, for any $\alpha > 0$. Roughly speaking, the core of all proofs is a bound on the Rademacher complexity of the k-fold maximum of the convex hull of the loss class of $\mathcal{H}$. The k-fold maximum captures the k possible corruptions for each input. In the regression setting, we provide three different generalization bounds. One of the main contributions in this setting is an upper bound on the empirical fat-shattering dimension of k-fold maximum class.

Our algorithm is an adaptation of the regret minimization algorithm proposed for binary classification by Feige et al. [8] for computing near-optimal policies for the players on the training data to the multiclass classification settings. It is a variant of the algorithm found in Cesa-Bianchi et al. [47] and based on the ideas of Freund and Schapire [48]. An ERM (empirical risk minimization) oracle is repeatedly used to return a

hypothesis from a fixed hypothesis class $\mathcal{H}$ that minimizes the error rate on a given sample, while weighting samples differently every time. The learner uses a randomized classifier chosen uniformly from the mixture of hypotheses returned by the algorithm.

Thus, we extend the ERM paradigm by using *adversarial training* techniques instead of merely find a hypothesis that minimizes the empirical risk. In contradistinction to "standard" learning, ERM often does not yield models that are robust to adversarially corrupted examples [49–54].

Subsequent Work: Montasser, Hanneke, and Srebro [9, 55].

Following the conference version [56] of this work, Montasser, Hanneke, and Srebro [9] have proved that VC classes are robustly PAC-learnable only improperly (that is, the hypothesis is selected from a broader class than that of the true concept), with respect to any arbitrary perturbation set, possibly of infinite size. The sample complexity[1] is independent of the number of allowed perturbations, $\tilde{\mathcal{O}}\Big(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\Big)$ in the realizable setting and $\tilde{\mathcal{O}}\Big(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\Big)$ in the agnostic setting, where $\mathrm{VC}^*(\mathcal{H})$ denotes the dual VC-dimension. Their approach relies on sample compression arguments whereas uniform convergence does not hold. As a by-product, for the case of $k < \infty$ possible corruptions for each input, they obtained a sample complexity of size $\mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})\log k}{\epsilon^2}\mathrm{polylog}(\frac{\mathrm{VC}(\mathcal{H})\log k}{\epsilon}) + \frac{1}{\epsilon^2}\log(\frac{1}{\delta})\right)$ for the zero-one robust loss (which is defined below).

The main difference between the two works is the definition of the loss function. Specifically, for functions $h_1, \ldots, h_T$, in the binary classification setting, we define the loss $\ell : \Delta(\mathcal{H}) \times \mathcal{X} \times \mathcal{Y} \to [0,1]$ by

$$\ell_1(h_1, \ldots, h_T, x, y) = \max_{z \in \mathcal{U}(x)} \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}\left[h_i(z) \neq y\right] = \max_{z \in \mathcal{U}(x)} \left|\frac{1}{T}\sum_{i=1}^{T} h_i(z) - y\right|, \tag{3.1}$$

which we refer to as the $[0,1]$-*robust loss*. Montasser et al. [9, 55] defined a loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ as follows

$$\ell_2(h, x, y) = \max_{z \in \mathcal{U}(x)} \mathbb{I}\left[h(z) \neq y\right], \tag{3.2}$$

which we refer to as the *zero-one robust loss*. More specifically, they consider for functions $h_1, \ldots, h_T$ the loss

$$\ell_3(h_1, \ldots, h_T, x, y) = \max_{z \in \mathcal{U}(x)} \mathbb{I}\left[\mathrm{Majority}(h_1(z), \ldots, h_T(z)) \neq y\right], \tag{3.3}$$

where Majority takes the majority of its Boolean inputs (assuming that T is odd). Clearly, if

[1] $\tilde{\mathcal{O}}(\cdot)$ hides poly-logarithmic factors of VC, VC^*, $1/\epsilon$, $1/\delta$.

$\ell_1(h_1,\ldots,h_T,x,y) < 1/2$ then $\ell_3(h_1,\ldots,h_T,x,y) = 0$. However, if $\ell_3(h_1,\ldots,h_T,x,y) = 0$ it only guarantees that $\ell_1(h_1,\ldots,h_T,x,y) < 1/2$ but can be very far from zero. This is why an upper bound on the sample complexity of ℓ_1 implies an upper bound on the sample complexity of ℓ_3, but not vice versa. We summarize the main results for both definitions in Section 3.1.

The work of Montasser et al. [9], that considers the zero-one robust loss, improper learning is necessary due to the lack of uniform convergence, which may arise in the case of an infinite set of corruptions. The learner competes with the single optimal hypothesis in the class and outputs a mixture of hypotheses to do so. In this work, considering the $[0,1]$-robust loss, we would like to guarantee an ϵ-optimal value for the learner in a zero-sum game, via a mixed strategy, and so we find an ϵ-optimal mixture of hypotheses. That is, we compete with the optimal mixture of hypotheses from the function class. In that sense, we have a proper learning algorithm, with respect to the convex hull of the hypothesis class.

In another closely related work from the computational perspective, Montasser et al. [55] reduced the problem of robust learning to non-robust learning. Namely, their algorithm uses access to only a black-box PAC learner, similar to the algorithm of Feige et al. [8] that we employ in this paper. They provided an algorithm that achieves small robust risk in the realizable setting with sample complexity (that is independent of k) of $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})(\mathrm{VC}^*(\mathcal{H}))^2}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$, and uses $\mathcal{O}\left(\log^2(nk) + \log\frac{1}{\delta}\right)$ black-box oracle calls to any PAC-learner, where n is the sample size. Their result relies on sample compression and not uniform convergence.

Uniform Convergence of the Zero-One Robust Loss Class

For the case of a finite set of corruptions, and learning with respect to the zero-one robust loss, we show that the VC dimension of the robust loss class remains finite (as opposed to the case of infinite corruptions). As a result, we have uniform convergence, and robust ERM suffices to ensure learning. (The proof is in Section 3.7).

Lemma 3.1 *For any class $\mathcal{H}$ of* VC *dimension d, and any adversary $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ such that $|\mathcal{U}(x)| \leq k$, the* VC*-dimension of the zero-one robust loss class $L^{\mathcal{U}}_{\mathcal{H}} = \left\{(x,y) \mapsto \max_{z\in\mathcal{U}(x)} \mathbb{I}\left[h(z) \neq y\right] : h \in \mathcal{H}\right\}$ is at most $\mathcal{O}(d\log k)$.*

Via a standard uniform convergence argument, we have the following result.

Theorem 3.2 *For any class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ of* VC *dimension d, and any adversary $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ such that $|\mathcal{U}(x)| \leq k$. For the robust zero-one loss function $\ell(h,x,y) = \max_{z\in\mathcal{U}(x)} \mathbb{I}[h(z) \neq y]$, the sample complexity for the realizable setting is $\mathcal{M}_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U}) = \mathcal{O}\left(\frac{d\log k}{\epsilon}\log\frac{1}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$, and the sample complexity for the agnostic setting is $\mathcal{M}_{\mathrm{AG}}(\epsilon,\delta,\mathcal{H},\mathcal{U}) = \mathcal{O}\left(\frac{d\log k}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$.*

Main Results

We provide a summary of the results for the $[0,1]$-robust loss and the zero-one robust loss (see Eqs. (3.1) and (3.2) for the definitions) for robust (ϵ, δ)-PAC learning with a finite set of possible corruptions.

Sample complexity for agnostic learning with $[0,1]$-robust loss		
GENERALIZATION	BINARY CLASSIFICATION	REFERENCE
Uniform Convergence	$\mathcal{O}\left(\frac{1}{\epsilon^4}\log\frac{\lvert\mathcal{H}\rvert}{\delta}\right)$	Feige et al. [8]
Sample Compression	$\tilde{\mathcal{O}}\left(\frac{dd^*}{\epsilon^4}+\frac{1}{\epsilon^4}\log\frac{1}{\delta}\right)$ $\tilde{\mathcal{O}}\left(\frac{d\log k}{\epsilon^4}+\frac{1}{\epsilon^4}\log\frac{1}{\delta}\right)$	Montasser et al. [9]
Uniform Convergence	$\tilde{\mathcal{O}}\left(\frac{kd}{\epsilon^2}+\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$	This work
	REGRESSION	
Uniform Convergence	$\tilde{\mathcal{O}}\left(\inf_{\beta\geq 0}\left\{\beta+\sqrt{\frac{k}{n}}\int_\beta^1\sqrt{\mathrm{fat}_\gamma(\mathcal{H})}d\gamma\right\}+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right)$	This work

Table 3.1: Sample complexity for agnostic learning with $[0,1]$-robust loss for a finite set of corruptions. Notation: d denotes the VC dimension, d^* denote the dual dual-VC dimension, $\mathrm{fat}_\gamma(\cdot)$ is the $\gamma-$fat shattering dimension, and k is the size of possible corruptions for each input. $\tilde{\mathcal{O}}(\cdot)$ stands for for omitting poly-logarithmic factors of $d, d^*, 1/\epsilon, 1/\delta$.

Sample complexity for binary classification with zero-one robust loss			
GENERALIZATION	REALIZABLE	AGNOSTIC	REFERENCE
Sample Compression	$\tilde{\mathcal{O}}\left(\frac{dd^*}{\epsilon}+\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ $\tilde{\mathcal{O}}\left(\frac{d\log k}{\epsilon}+\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$	$\tilde{\mathcal{O}}\left(\frac{dd^*}{\epsilon^2}+\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ $\tilde{\mathcal{O}}\left(\frac{d\log k}{\epsilon^2}+\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$	Montasser et al. [9]
Uniform Convergence	$\mathcal{O}\left(\frac{d\log k}{\epsilon}\log\frac{1}{\epsilon}+\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$	$\mathcal{O}\left(\frac{d\log k}{\epsilon^2}+\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$	This work

Table 3.2: Sample complexity for binary classification with zero-one robust loss for a finite set of corruptions. The notations are as in Table 2.1.

Whether we can achieve a sample complexity of $\approx \frac{d\log k}{\epsilon^2}$ or $\frac{dd^*}{\epsilon^2}$ for agnostic learning with the $[0,1]$-robust loss remains an open question. The method of Montasser et al. [9] can be modified to accommodate learning with respect to the $[0,1]$ robust loss. Specifically, taking the majority of weak learners is not sufficient for obtaining an ϵ-optimal mixed strategy. Rather, we take a majority of strong learners (each with ϵ error), each of which takes $\approx \frac{d}{\epsilon^2}$ samples (and not $\approx d$). This implies a sample complexity (via sample compression scheme) of $\frac{dd^*}{\epsilon^4}$ or $\frac{d\log(k)}{\epsilon^4}$.

Other Related Work

The most closely related works studying robust learning with adversarial examples are Schmidt et al. [44], Madry et al. [45]. Their model deals with bounded worst-case perturbations (under ℓ_∞ norm) on the samples. This is slightly different from our model as we mentioned above. Other related works that analyze the theoretical aspects of adversarial robust generalization are Attias et al. [2], Attias and Hanneke [3], Montasser et al. [9], Levi et al. [16], Yin et al. [57], Awasthi et al. [58], Cullina et al. [59], Khim and Loh [60], Raghunathan et al. [61], Diochnos et al. [62], Balda et al. [63], Pydi and Jog [64], Tu et al. [65], Chen et al. [66], Carmon et al. [67], Alayrac et al. [68], Zhai et al. [69], Najafi et al. [70]. A different notion of robustness by Xu and Mannor [71] is shown to be sufficient and necessary for standard generalization. Learning with adversarial examples is extensively studied from the computational point of view as well [72–83].

All of our results are based on a robust learning model for binary classification suggested by Feige et al. [8]. The works of Feige et al. [8], Mansour et al. [84], Feige et al. [85] consider *robust inference* for the binary and multiclass case. The robust inference model assumes that the learner knows both the distribution and the target function. The main task is, for a given corrupted input, to derive a classification in a computationally efficient way that will minimize the error. In this work, we consider only the learning setting, where the learner has only access to an uncorrupted sample, and need to approximate the target function on possibly corrupted inputs, using a restricted hypothesis class $\mathcal{H}$.

The work of Globerson and Roweis [86] and its extensions Teo et al. [87], Dekel et al. [88] discuss a robust learning model where an uncorrupted sample is drawn from an unknown distribution, and the goal is to learn a linear classifier resilient against missing attributes in future test examples. They discuss both the static model (where the set of missing attributes is selected independently from the uncorrupted input) and the dynamic model (where the set of missing attributes may depend on the uncorrupted input). The model we use [8] extends the robust learning model to handle corrupted inputs (and not only missing attributes) and an arbitrary hypothesis class (rather than only linear classifiers).

There is a vast literature on statistics, operation research and machine learning regarding various noise models. Typically, most noise models assume a random process that generates the noise. In computational learning theory, popular noise models include random classification noise [89] and malicious noise [90, 91]. In the malicious noise model, the adversary gets to arbitrarily corrupt some small fraction of the examples; in contrast, in our model the adversary can always corrupt every example, but only in a limited way.

3.2 Model

There is an unknown distribution D over some domain $\mathcal{X}$ of uncorrupted examples and a finite domain of corrupted examples $\mathcal{Z}$, possibly the same as $\mathcal{X}$. Our setting is the *agnostic PAC-learning* framework in a deterministic scenario. The label of each input is uniquely determined by an arbitrary unknown target function $c : \mathcal{X} \to \mathcal{Y}$. The function c maps each uncorrupted input $x \in \mathcal{X}$ to a label $c(x) = y$, where the set of labels $\mathcal{Y}$ can be $\{1, \ldots, l\}$ or $\mathbb{R}$.

The adversary can corrupt an input by mapping an uncorrupted input $x \in \mathcal{X}$ to a corrupted one $z \in \mathcal{Z}$. There is a mapping $\mathcal{U}$ which for every $x \in \mathcal{X}$ defines a set $\mathcal{U}(x) \subseteq \mathcal{Z}$, such that $|\mathcal{U}(x)| \leq k$. The adversary can map an uncorrupted input x to any corrupted input $z \in \mathcal{U}(x)$. We assume that the learner has access to $\mathcal{U}(\cdot)$ during the training phase.

There is a fixed hypothesis class $\mathcal{H}$ of hypothesis $h : \mathcal{Z} \mapsto \mathcal{Y}$ over corrupted inputs. The learner observes an uncorrupted sample $S_u = \{\langle x_1, c(x_1)\rangle, \ldots, \langle x_m, c(x_m)\rangle\}$, where x_i is drawn i.i.d. from D, and selects a mixture of hypotheses from $\mathcal{H}$, $\tilde{h} \in \Delta(\mathcal{H})$. In the classification setting, $\tilde{h} : \mathcal{Z} \to \Delta(\mathcal{Y})$ is a mixture $\{h_i | \mathcal{H} \ni h_i : \mathcal{Z} \to \mathcal{Y}\}_{i=1}^{T}$ such that label $y \in \mathcal{Y} = \{1, \ldots, l\}$ gets a mass of $\sum_{i=1}^{T} \alpha_i \mathbb{I}[h_i(z) = y]$ where $\sum_{i=1}^{T} \alpha_i$=1. For each hypothesis $h \in \mathcal{H}$ in the mixture we use the zero-one loss to measure the quality of the classification, i.e., $\ell(h(z), y) = \mathbb{I}[h(z) \neq y]$. The loss of $\tilde{h} \in \Delta(\mathcal{H})$ is defined by $\ell(\tilde{h}(z), y) = \sum_{i=1}^{T} \alpha_i \ell(h_i(z), y)$. In the regression setting, $\tilde{h} : \mathcal{Z} \to \mathbb{R}$ is a mixture $\{h_i | \mathcal{H} \ni h_i : \mathcal{Z} \to \mathbb{R}\}_{i=1}^{T}$ and is defined by $\tilde{h}(z) = \sum_{i=1}^{T} \alpha_i h_i(z)$. For each hypothesis $h \in \mathcal{H}$ in the mixture we use L_1 and L_2 loss functions, i.e., $\ell(h(z), y) = |h(z) - y|^p$, for $p = 1, 2$. We assume the L_1 loss is bounded by 1. Again, the loss of $\tilde{h} \in \Delta(\mathcal{H})$ is defined by $\ell(\tilde{h}(z), y) = \sum_{i=1}^{T} \alpha_i \ell(h_i(z), y)$.

The test phase proceeds as follows. First, an uncorrupted input $x \in \mathcal{X}$ is drawn from D. Then, the adversary selects $z \in \mathcal{U}(x)$, given $x \in \mathcal{X}$. The learner observes a corrupted input $\mathcal{Z}$ and outputs a prediction, as dictated by $\tilde{h} \in \Delta(\mathcal{H})$. Finally, the learner incurs a loss as described above. The main difference from the classical learning models is that the learner will be tested on adversarially corrupted inputs $z \in \mathcal{U}(x)$. When selecting a strategy this needs to be taken into consideration.

The goal of the learner is to minimize the expected loss, while the adversary would like to maximize it. This defines a zero-sum game which has a value v which is the learner's error rate. We say that the learner's hypothesis is ϵ-optimal if it guarantees a loss which is at most $v + \epsilon$, and the adversary policy is ϵ-optimal if it guarantees a loss which is at least $v - \epsilon$. We refer to a 0-optimal policy as an optimal policy.

Formally, the error (risk) of the learner when selecting a hypothesis $\tilde{h} \in \Delta(\mathcal{H})$ is

$$\mathrm{Err}(\tilde{h}; \mathcal{D}, \mathcal{U}) = \mathbb{E}_{x \sim D}[\max_{z \in \mathcal{U}(x)} \ell(\tilde{h}(z), c(x))],$$

and their goal is to choose $\tilde{h} \in \Delta(\mathcal{H})$ with an error close to

$$\min_{\tilde{h}\in\Delta(\mathcal{H})} \mathrm{Err}(\tilde{h}; \mathcal{D}, \mathcal{U}) = \min_{\tilde{h}\in\Delta(\mathcal{H})} \mathbb{E}_{x\sim D}[\max_{z\in\mathcal{U}(x)} \ell(\tilde{h}(z), c(x))] = v.$$

3.3 Definitions and Notation

For a function class $\mathcal{H}$ with domain $\mathcal{Z}$ and range $\mathcal{Y} = \{1, \ldots, l\}$, denote the zero-one loss class

$$L_{\mathcal{H}} := \{Z \times \{1, \ldots, l\} \ni (z, y) \mapsto \mathbb{I}[h(z) \neq y] : h \in \mathcal{H}\}.$$

For $\mathcal{H}$ with domain $\mathcal{Z}$ and range $\mathbb{R}$, denote the L_p loss class

$$L^p_{\mathcal{H}} := \{Z \times \mathbb{R} \ni (z, y) \mapsto |h(z) - y|^p : h \in \mathcal{H}\}.$$

Throughout the article, we assume a bounded loss $\ell(h(z), y) \leq M$. Without the loss of generality, we use $M = 1$, since otherwise, M can be re-scaled.

We define the operator conv as the convex hull of a real-valued function class,

$$\mathrm{conv}(\mathcal{F}) := \left\{ W \ni w \mapsto \sum_{t=1}^{T} \alpha_t f_t(w) : T \in \mathbb{N}, \alpha_t \in [0,1], \sum_{t=1}^{T} \alpha_t = 1, f_t \in \mathcal{F} \right\}.$$

We also define the convex hull of loss class L, where the data is corrupted by $\mathcal{U}(\cdot)$,

$$\overset{\mathcal{U}}{\mathrm{conv}}(L) := \left\{ \mathcal{X} \times \mathcal{Y} \ni (x, y) \mapsto \max_{z\in\mathcal{U}(x)} \sum_{t=1}^{T} \alpha_t f_t(z, y) : T \in \mathbb{N}, \alpha_t \in [0,1], \sum_{t=1}^{T} \alpha_t = 1, f_t \in L \right\}.$$

For $1 \leq j \leq k$ define,

$$\mathcal{F}^{(j)}_{\mathcal{H}} := \{\mathcal{X} \times \mathcal{Y} \ni (x, y) \mapsto \mathbb{I}[h(z_j) \neq y] : h \in \mathcal{H},\ \mathcal{U}(x) = \{z_1, \ldots, z_k\}\}, \tag{3.4}$$

where we treat the set-valued output of $\mathcal{U}(x)$ as an ordered list, and $\mathcal{F}^{(j)}_{\mathcal{H}}$ is constructed by taking the jth element in this list, for each input x.

For a set W and k function classes $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(k)} \subseteq \mathbb{R}^W$, define the $\max$ operator

$$\max\left((\mathcal{A}^{(j)})_{j\in[k]}\right) := \left\{ W \ni w \mapsto \max_{j\in[k]} f^{(j)}(w) : f^{(j)} \in \mathcal{A}^{(j)} \right\}.$$

The composition of $\max$ and conv operators $\max\left((\mathrm{conv}(\mathcal{A}^{(j)}))_{j\in[k]}\right)$ is well-defined, note that

$$\overset{\mathcal{U}}{\mathrm{conv}}(L_{\mathcal{H}}) \subseteq \max\left((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]}\right). \tag{3.5}$$

Denote the error (risk) of hypothesis $h : \mathcal{Z} \mapsto \mathcal{Y}$ under corruption of $\mathcal{U}(\cdot)$ by

$$\mathrm{Err}(h;\mathcal{D},\mathcal{U}) = \mathbb{E}_{x\sim D}[\max_{z\in\mathcal{U}(x)} \ell(h(z),c(x))],$$

and the empirical error on sample S under corruption of $\mathcal{U}(\cdot)$ by

$$\widehat{\mathrm{Err}}(h;S,\mathcal{U}) = \frac{1}{|S|}\sum_{(x,y)\in S} \max_{z\in\mathcal{U}(x)} \ell(h(z),c(x)).$$

3.4 Algorithm

We have a base hypothesis class $\mathcal{H}$ with domain $\mathcal{Z}$ and range $\mathcal{Y}$ that can be $\{1,\ldots,l\}$ or $\mathbb{R}$. The learner receives a labeled uncorrupted sample and has access during the training to possible corruption by the adversary. We employ the regret minimization algorithm proposed by Feige et al. [8] for binary classification, and extend it to the regression and multiclass classification settings.

A brief description of the algorithm is as follows. Given $x \in \mathcal{X}$, we define a $|\mathcal{U}(x)| \times \mathcal{H}$ loss matrix M_x such that $M_x(z,h) = \mathbb{I}[h(z) \neq y]$, where $y = c(x)$. The learner's strategy is a distribution Q over $\mathcal{H}$. The adversary's strategy $P_x \in \Delta(\mathcal{U}(x))$, for a given $x \in \mathcal{X}$, is a distribution over the corrupted inputs $\mathcal{U}(x)$. We can treat P as a vector of distributions P_x over all $x \in \mathcal{X}$. Via the minimax principle, the value of the game is

$$v = \min_Q \max_P \mathbb{E}_{x\sim D}[P_x^T M_x Q] = \max_P \min_Q \mathbb{E}_{x\sim D}[P_x^T M_x Q]$$

For a given P, a learner's minimizing Q is simply a hypothesis that minimizes the error when the distribution over pairs $(z,y) \in \mathcal{Z}\times\mathcal{Y}$ is D^P, where

$$D^P(z,y) = \sum_{x:\ c(x)=y\wedge z\in\mathcal{U}(x)} P_x(z)D(x).$$

Hence, the learner selects

$$h^P = \arg\min_{h\in\mathcal{H}} \mathbb{E}_{(z,y)\sim D^P}[\ell(h(z),y)].$$

A hypothesis h^P can be found using the ERM oracle when D^P is the empirical distribution over a training sample.

Repeating this process multiple times yields a mixture of hypotheses $\tilde{h} \in \Delta(\mathcal{H})$ (mixed strategy- a distribution Q over $\mathcal{H}$) for the learner. The learner uses a randomized classifier chosen uniformly from this mixture. This also yields a mixed strategy for the adversary, defined by an average of vectors P. Therefore, for a given $x \in \mathcal{X}$, the adversary uses a distribution $P_x \in \Delta(\mathcal{U}(x))$ over corrupted inputs.

Algorithm 1 Approximate Minimax Strategy for a Half-Infinite Zero-Sum Game

Input: $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{Z}}$, $S = \{(x_i, y_i)\}_{i=1}^n, \mathcal{U} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$.
parameter: $\eta > 0$.
Algorithm used: ERM oracle.

▷ Initialize weights for each $(x, y) \in S$ and $z \in \mathcal{U}(x)$ and a distribution vector over $\mathcal{U}(x)$ for each $(x, y) \in S$.

1. For $(x, y) \in S, z \in \mathcal{U}(x)$:
   (a) $w_1(z, (x, y)) \leftarrow 1, \forall (x, y) \in S, \forall z \in \mathcal{U}(x)$.
   (b) $P^1(z, (x, y)) \leftarrow \frac{w_1(z,(x,y))}{\sum_{z' \in \mathcal{U}(x)} w_1(z',(x,y))}$.

▷ Compute approximate minimax strategy for the learner and maximin strategy for the adversary using regret minimization. The learner determines their best response using an ERM oracle, while the adversary employs multiplicative updates.

2. For $t = 1, \ldots, T$:
   (a) $h_t \leftarrow \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(z,y)\sim D^{P^t}}[\ell(h(z), y]$.
   (b) For $(x, y) \in S, z \in \mathcal{U}(x)$:
      i. $w_{t+1}(z, (x, y)) \leftarrow (1 + \eta \cdot [\ell(h_t(z), y)]) \cdot w_t(z, (x, y))$.
      ii. $P^{t+1}(z, (x, y)) \leftarrow \frac{w_{t+1}(z,(x,y))}{\sum_{z' \in \mathcal{U}(x)} w_{t+1}(z',(x,y))}$.

Output: $h_1, \ldots, h_T$ for the learner, $\frac{1}{T}\sum_{t=1}^T P^t$ for the adversary.

Similar to Feige et al. [8, Theorem 1], for the binary classification case and zero-one loss we have:

Theorem 3.3 *[8, Theorem 1] Fix a sample S of size n, and let $T \geq \frac{4n \log k}{\epsilon^2}$, where k is the number of possible corruptions for each input. For an uncorrupted sample S, we have that the strategies $P = \frac{1}{T}\sum_{t=1}^T P^t$ for the adversary and $h_1, \ldots, h_T$ (each one of them chosen uniformly) for the learner are ϵ-optimal strategies on S.*

Assuming a bounded loss, i.e., $\ell(h(z), y) \leq 1$, $\forall x \in \mathcal{X}, \forall z \in \mathcal{Z}, \forall h \in \mathcal{H}$, the result remains the same for the other settings.

3.5 Generalization Bound for Classification

We would like to show that if the sample S is large enough, then the policy achieved by the algorithm above will generalize well. We both improve a generalization bound, previously found in Feige et al. [8], which handles any mixture of hypotheses from $\mathcal{H}$, and also are able to handle an infinite hypothesis class $\mathcal{H}$. The sample complexity is improved from $\mathcal{O}(\frac{1}{\epsilon^4}\log(\frac{|\mathcal{H}|}{\delta}))$ to $\mathcal{O}\big(\frac{1}{\epsilon^2}(k\,\mathrm{VC}(\mathcal{H})\log^{\frac{3}{2}+\alpha}(k\,\mathrm{VC}(\mathcal{H}))+\log(\frac{1}{\delta})\big)$ for any $\alpha > 0$.

Theorem 3.4 (Generalization bound for binary classification) *Let $\mathcal{H} : \mathcal{Z} \mapsto \{0,1\}$ be a hypothesis class with finite* VC*-dimension. For any $\alpha > 0$ there exists a constant C_α and there is a sample complexity $n_0 = \frac{C_\alpha}{\epsilon^2}\left(k\,\mathrm{VC}(\mathcal{H})\log^{\frac{3}{2}+\alpha}(k\,\mathrm{VC}(\mathcal{H}))+\log(\frac{1}{\delta})\right)$, such that for $|S| \geq n_0$, for every $\tilde{h} \in \Delta(\mathcal{H})$*

$$|\mathrm{Err}(\tilde{h};\mathcal{D},\mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h};S,\mathcal{U})| \leq \epsilon$$

with probability at least $1-\delta$.

Theorem 3.4 provides an improvement to Theorem 7 in [92], where they considered learning with the intersection of hyperplanes for an imbalanced binary classification problem.

Theorem 3.5 *[93] Let $\mathcal{F}$ be a $\mathbb{R}^k$-valued function class, such that the coordinate projection class is denoted by $\mathcal{F}_j = \{w \mapsto f(w)_j \,|\, f \in \mathcal{F}\}$, for $1 \leq j \leq k$. Let $(\varphi_t)_{t\leq n}$ be a sequence of functions such that each φ_t is L-Lipschitz with respect to ℓ_∞ norm. For any $\alpha > 0$, there exists a constant $C_\alpha > 0$ such that if $|\varphi_t(f(w))| \vee ||f(w)||_\infty \leq B$, then it holds for any sequence $\mathbf{w} = (w_1, \cdots, w_n)$,*

$$\begin{aligned}\mathcal{R}_n(\varphi\circ\mathcal{F}|\mathbf{w}) &:= E_\sigma \sup_{f\in F}\frac{1}{n}\sum_{t=1}^{n}\sigma_t\varphi_t(f_t(w_t)) \\ &\leq C_\alpha L\sqrt{k}\cdot\max_{i\in[k]}\sup_{\mathbf{a}=(a_1,\ldots,a_n)}\mathcal{R}_n(\mathcal{F}_i|\mathbf{a})\cdot\log^{\frac{3}{2}+\alpha}\left(\frac{Bn}{\max_{i\in[k]}\sup_{\mathbf{a}=(a_1,\ldots,a_n)}\mathcal{R}_n(\mathcal{F}_i|\mathbf{a})}\right).\end{aligned}$$

Proof (of Theorem 3.4). Our strategy is to bound the empirical Rademacher complexity (over the sample points) of the loss class of $\tilde{h} \in \Delta(\mathcal{H})$. As we mentioned in Eq. (3.5), $\mathrm{conv}^{\mathcal{U}}(L_{\mathcal{H}}) \subseteq \max(\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})$. Recall that functions contained in $\mathcal{F}_{\mathcal{H}}^{(j)}$ are loss functions of the learner when the adversary corrupts input x to $z_j \in \mathcal{U}(x)$. We are left to bound the Rademacher complexity of the function class $\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})$. Formally,

$$\begin{aligned}|\mathrm{Err}(\tilde{h};\mathcal{D},\mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h};S,\mathcal{U})| &= |E_{(x,y)\sim D}\max_{j\in[k]}\sum_{t=1}^{T}\alpha_t f_t^{(j)}(x,y) - \frac{1}{n}\sum_{(x,y)\in S}\max_{j\in[k]}\sum_{t=1}^{T}\alpha_t f_t^{(j)}(x,y)| \\ &\leq 2\mathcal{R}_n\left(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x}\times\mathbf{y}\right) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}},\end{aligned}$$

where the inequality stems from applying Theorem 2.12 on the function class $\mathrm{conv}^{\mathcal{U}}(L_{\mathcal{H}})$ and Eq. (3.5). By taking $\varphi(z_1,\cdots,z_k)=\max_{j\in[k]} z_j$, which is a 1-Lipschitz with respect to ℓ_∞, and $\mathcal{F}=\{(x,y)\mapsto(f_1(x,y),\cdots,f_k(x,y))\,|\,f_j\in\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}),1\le j\le k\}$ we can apply Theorem 3.5, for any $\alpha>0$, there exists a constant $C_\alpha>0$ such that

$$\begin{aligned}
&\mathcal{R}_n\left(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x}\times\mathbf{y}\right)\\
&\le C_\alpha\sqrt{k}\cdot\max_{j\in[k]}\max_{\mathbf{w}=w_{1:n}}\mathcal{R}_n(\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)})|\mathbf{w})\cdot\log^{\frac{3}{2}+\alpha}\left(\frac{n}{\max_{j\in[k]}\max_{\mathbf{w}=w_{1:n}}\mathcal{R}_n(\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)})|\mathbf{w})}\right)\\
&=C_\alpha\sqrt{k}\cdot\max_{j\in[k]}\max_{\mathbf{w}=w_{1:n}}\mathcal{R}_n(\mathcal{F}_{\mathcal{H}}^{(j)}|\mathbf{w})\cdot\log^{\frac{3}{2}+\alpha}\left(\frac{n}{\max_{j\in[k]}\max_{\mathbf{w}=w_{1:n}}\mathcal{R}_n(\mathcal{F}_{\mathcal{H}}^{(j)}|\mathbf{w})}\right),
\end{aligned}$$

where the last equality follows from the well-known identity $\mathcal{R}_n(\mathcal{F}|\mathbf{w})=\mathcal{R}_n(\mathrm{conv}(\mathcal{F})|\mathbf{w})$, (see, e.g., Boucheron et al. [94, Theorem 3.3]).

The function $x\mapsto x\log^{3/2+\alpha}(n/x)$ has a maximum point at $x=n/e^{3/2+\alpha}$, and for $x\in(0,n/e^{3/2+\alpha}]$ is monotonic increasing. We bound the empirical Rademacher complexity (on any given sequence) via the VC-dimension [36]: $\mathcal{R}_n(\mathcal{F}|\mathbf{w})\le C\sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}}$,
and for $\left(C\sqrt{\mathrm{VC}(\mathcal{F}_{\mathcal{H}})}e^{3/2+\alpha}\right)^{2/3}\le n$, by the monotonicity of the function $x\log^{3/2+\alpha}(n/x)$ we get an upper bound of

$$\begin{aligned}
C_\alpha C\sqrt{\frac{k\max_{j\in[k]}\mathrm{VC}(\mathcal{F}_{\mathcal{H}}^{(j)})}{n}}\cdot\log^{\frac{3}{2}+\alpha}\left(\frac{n^{\frac{3}{2}}}{C\sqrt{\max_{j\in[k]}\mathrm{VC}(\mathcal{F}_{\mathcal{H}}^{(j)})}}\right)&\\
=C_\alpha C\sqrt{\frac{k\,\mathrm{VC}(\mathcal{H})}{n}}\cdot\log^{\frac{3}{2}+\alpha}\left(\frac{n^{\frac{3}{2}}}{C\sqrt{\mathrm{VC}(\mathcal{H})}}\right)&\\
=\mathcal{O}\left(C_\alpha\sqrt{\frac{k\,\mathrm{VC}(\mathcal{H})}{n}}\cdot\log^{\frac{3}{2}+\alpha}(n)\right),&
\end{aligned}$$

where the inequality follows from Lemma 3.7. We require that

$$C_\alpha\sqrt{\frac{k\,\mathrm{VC}(\mathcal{H})}{n}}\cdot\log^{\frac{3}{2}+\alpha}(n)+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\le\epsilon,$$

and a standard inversion of this inequality yields sample complexity $n_0=\mathcal{O}\big(\frac{C_\alpha}{\epsilon^2}(k\,\mathrm{VC}(\mathcal{H})\log^{\frac{3}{2}+\alpha}(k\,\mathrm{VC}(\mathcal{H}))+\log(\frac{1}{\delta})\big)$. □

We find it instructive to provide an alternative (albeit worse) bound of

$$\mathcal{R}_n\left(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x}\times\mathbf{y}\right) \leq \mathcal{O}\left(\sqrt{\frac{\mathrm{VC}(\mathcal{H})\log^2(\mathrm{VC}(\mathcal{H}))k\log k\log^9(n)}{n}}\right) \tag{3.6}$$

on the Rademacher complexity, via a different technique (In Appendix 3.7).

Multiclass Classification

Let $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{Z}}$ be a function class such that $\mathcal{Y}=[l]=\{1\ldots,l\}$. We follow similar arguments to the binary case, where the graph dimension replaces the VC dimension.

Graph Dimension. Let $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$ be a categorical function class such that $\mathcal{Y}=[l]=\{1\ldots,l\}$. Let $S\subseteq\mathcal{X}$. We say that $\mathcal{H}$ G-shatters S if there exists an $f:S\mapsto\mathcal{Y}$ such that for every $T\subseteq S$ there is a $g\in\mathcal{H}$ such that

$$\forall x\in T,\ g(x)=f(x) \text{ and } \forall x\in S\setminus T,\ g(x)\neq f(x).$$

The graph dimension of $\mathcal{H}$, denoted $d_G(\mathcal{H})$, is the maximal cardinality of a set that is G-shattered by $\mathcal{H}$.

Theorem 3.6 (Generalization bound for multiclass classification) *Let $\mathcal{H}$ be a function class with domain $\mathcal{Z}$ and range $\mathcal{Y}=[l]$ with finite Graph-dimension $d_G(\mathcal{H})$. For any $\alpha>0$ there exists a constant C_α and there is a sample complexity $n_0=\frac{C_\alpha}{\epsilon^2}\left(k\mathrm{d}_G(\mathcal{H})\log^{\frac{3}{2}+\alpha}(k\mathrm{d}_G(\mathcal{H}))+\log(\frac{1}{\delta})\right)$, such that for $|S|\geq n_0$, for every $\tilde{h}\in\Delta(\mathcal{H})$,*

$$|\mathrm{Err}(\tilde{h};\mathcal{D},\mathcal{U})-\widehat{\mathrm{Err}}(\tilde{h};S,\mathcal{U})|\leq\epsilon$$

with probability at least $1-\delta$.

The following Lemma is standard and holds for the function classes $\mathcal{F}_{\mathcal{H}}^{(j)}$ (defined in Eq. (3.4)).

Lemma 3.7 *Let $\mathcal{H}$ be a function class with domain $\mathcal{Z}$ and range $\mathcal{Y}=[l]$. Denote the Graph-dimension of $\mathcal{H}$ by $d_G(\mathcal{H})$. Then for all $j\in[k]$*

$$\mathrm{VC}(\mathcal{F}_{\mathcal{H}}^{(j)})\leq d_G(\mathcal{H}).$$

In particular, for binary-valued classes, $\mathrm{VC}(\mathcal{F}_{\mathcal{H}}^{(j)})\leq\mathrm{VC}(\mathcal{H})$ — since for these, the VC- and Graph-dimensions coincide.

Proof. Suppose that the binary function class $\mathcal{F}_{\mathcal{H}}^{(j)}$ shatters the points $\{(x_1, y_1), \ldots, (x_d, y_d)\} \subset \mathcal{X} \times \mathcal{Y}$. That means that for each $b \in \{0,1\}^d$, there is an $h_b \in \mathcal{H}$ such that $\mathbb{I}\left[h_b(z_j(x_i)) \neq y_i\right] = b_i$ for all $i \in [d]$, where $z_j(x)$ is the jth element in the (ordered) set-valued output of $\mathcal{U}$ on input x. We claim that $\mathcal{H}$ is able to G-shatter $S = \{z_j(x_1), \ldots, z_j(x_d)\} \subset \mathcal{Z}$. Indeed, for each $T \subseteq S$, let $b = b(T) \in \{0,1\}^S$ be its characteristic function. Taking $f : S \to \mathcal{Y}$ to be $f(x_i) = y_i$, we see that the definition of G-shattering holds. □

For the proof of Theorem 3.6, we follow the same proof of Theorem 3.4 and use the Graph-dimension property of Lemma 3.7.

A similar bound to that of Theorem 3.4 can be achieved by using the Natarajan dimension and the fact that

$$d_G(\mathcal{H}) \leq 4.67 \log_2(|\mathcal{Y}|) d_N(\mathcal{H})$$

as previously shown Ben-David et al. [95], where the Natarajan dimension is defined as follows

Natarajan Dimension. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a categorical function class such that $\mathcal{Y} = [l] = \{1 \ldots, l\}$. Let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$ N-shatters S if there exist $f_1, f_2 : S \mapsto \mathcal{Y}$ such that for every $y \in S$ $f_1(y) \neq f_2(y)$, and for every $T \subseteq S$ there is a $g \in \mathcal{H}$ such that

$$\forall x \in T,\ g(x) = f_1(x),\ \text{and}\ \forall x \in S \setminus T,\ g(x) = f_2(x).$$

The Natarajan dimension of $\mathcal{H}$, denoted $d_N(\mathcal{H})$, is the maximal cardinality of a set that is N-shattered by $\mathcal{H}$.

3.6 Generalization Bounds For Regression

Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ be a hypothesis class of real functions. In the following, we provide three different generalization bounds, which, as far as we can tell, are mutually incomparable uniformly over the parameter regimes.

Theorem 3.8 (Generalization bound for Regression) *Let $\mathcal{H}$ be a function class with domain $\mathcal{Z}$ and range $[0,1]$. Assume $\mathcal{H}$ has a finite γ-fat-shattering dimension for all $\gamma > 0$. Denote the sample size $|S| = n$ and*

$$m_n(\mathcal{H}) = \inf_{\beta \geq 0} \left\{ 4\beta + \mathcal{O}\left(\sqrt{\frac{\log^4(n)}{n}} \int_{\beta}^{1} \sqrt{\mathrm{fat}_{c\gamma}(\mathcal{H}) \log\left(\frac{1}{\gamma}\right)} d\gamma \right) \right\},$$

where c is a universal constant. For the L_1 loss function and for every $\tilde{h} \in \Delta(\mathcal{H})$, for any $\alpha > 0$ there exist

a constant C_α such that,

$$|\mathrm{Err}(\tilde{h};\mathcal{D},\mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h};S,\mathcal{U})| \leq \mathcal{O}\left(C_\alpha\sqrt{k}\cdot m_n(\mathcal{H})\cdot\log^{\frac{3}{2}+\alpha}\left(\frac{n}{m_n(\mathcal{H})}\right)+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right),$$

with probability at least $1-\delta$.

Moreover, in the case of L_2 loss function, the same result holds with $\mathrm{fat}_{\frac{c\gamma}{2}}(\mathcal{H})$ plugged into $m_n(\mathcal{H})$.

In the following corollary (proof is in Appendix 3.7) we derive a simplified bound for hyperplanes.

Corollary 3.9 *Let $\mathcal{H}$ be a function class of homogeneous hyperplanes with domain $\mathbb{R}^m$. Using the same assumptions as in Theorem 3.8, we have*

$$|\mathrm{Err}(\tilde{h};\mathcal{D},\mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h};S,\mathcal{U})| \leq \mathcal{O}\left(C_\alpha\sqrt{\frac{k}{n}}\log^{\frac{7}{2}}(n)\log^{\frac{3}{2}+\alpha}\left(\frac{n}{\log^{\frac{7}{2}}(n)}\right)+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right),$$

with probability at least $1-\delta$.

The class of hyperplanes can be learned with SGD, as the maximum of finite convex functions remains convex. However, our bound works for an arbitrary hypotheses class.

Theorem 3.10 (Generalization bound for Regression) *Let $\mathcal{H}$ be a function class with domain $\mathcal{Z}$ and range $[0,1]$. Assume $\mathcal{H}$ has a finite γ-fat-shattering dimension for all $\gamma>0$. Denote the sample size $|S|=n$ and*

$$m_n(\mathcal{H}) = \inf_{\alpha\geq 0}\left\{4\alpha+\mathcal{O}\left(\sqrt{\frac{k\log(k)\log^4(n)}{n}}\int_\alpha^1\sqrt{\log\left(\frac{1}{\gamma}\right)\left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{H})}{\gamma^2}\log^2\left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{H})}{\gamma}\right)\right)}d\gamma\right)\right\}.$$

For the L_1 loss function and for every $\tilde{h}\in\Delta(\mathcal{H})$,

$$|\mathrm{Err}(\tilde{h};\mathcal{D},\mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h};S,\mathcal{U})| \leq \mathcal{O}\left(m_n(\mathcal{H})+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right),$$

with probability at least $1-\delta$.

Moreover, in the case of L_2 loss function, the same result holds with $\mathrm{fat}_{\frac{c\gamma}{8}}(\mathcal{H})$ plugged into $m_n(\mathcal{H})$.

Theorem 3.11 (Generalization bound for Regression) *Let $\mathcal{H}$ be a function class with domain $\mathcal{Z}$ and range $[0,1]$. Assume $\mathcal{H}$ has a finite γ-fat-shattering dimension for all $\gamma>0$. Denote the sample size $|S|=n$ and $d=\mathrm{fat}_{\frac{\epsilon}{4}}(\mathcal{H})$. For the L_1 loss function, there is a sample complexity*

$$n_0 = \mathcal{O}\left(\frac{1}{\epsilon^2}\left(k\log(k)\frac{d}{\epsilon^2}\log^2\frac{d}{\epsilon}\log^2\frac{1}{\epsilon}\log^2\left(k\log(k)\frac{d}{\epsilon^4}\log^2\frac{d}{\epsilon}\log^2\frac{1}{\epsilon}\right)+\log\frac{1}{\delta}\right)\right),$$

such that for $|S| \geq n_0$*, for every* $\tilde{h} \in \Delta(\mathcal{H})$

$$|\mathrm{Err}(\tilde{h}; \mathcal{D}, \mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h}; S, \mathcal{U})| \leq \epsilon$$

with probability at least $1 - \delta$*.*

We would like to compare the bounds in Theorems 3.8, 3.10 and 3.11. In terms of dependence in the fat-shattering dimension and k, Theorem 3.8 would give a better bound than Theorem 3.10. However, the latter has a better dependence on $\log(n)$ factors. Regarding Theorem 3.11, on the one hand, the dependence in n (sample size) is $1/n^{1/4}$. On the other hand, we have the fat-shattering dimension with a specific scale (the error parameter, ϵ). In some cases, we can obtain an improved learning rate. For example, by taking $\mathrm{fat}_\gamma(\mathcal{H}) = 1/\gamma^6$, Theorem 3.8 guarantees learning rate of $1/n^{1/6}$ and so Theorem 3.11 provides a sharper bound.

Shattering Dimension of the Class $\max\left((\mathcal{A}^{(j)})_{j\in[k]}\right)$

The main result of this section is bounding the fat-shattering dimension of $\max\left((\mathcal{A}^{(j)})_{j\in[k]}\right)$ class.

Theorem 3.12 (Fat-shattering of *k*-fold maxima) *Let* $S = \{x_1, \ldots, x_m\}$*. For any* k *real-valued functions classes* $\mathcal{F}_1, \ldots, \mathcal{F}_k \subseteq \mathbb{R}^S$*,*

$$\mathrm{fat}_\gamma\left(\max\left((\mathcal{F}_j)_{j\in[k]}\right)\right) \leq \mathcal{O}\left(\log(k)\log^2(m)\sum_{j=1}^{k}\mathrm{fat}_\gamma(\mathcal{F}_j)\right).$$

Before presenting the proof, we introduce some auxiliary notions. We say that $\mathcal{F}$ "γ-shatters a set S at zero" if the shift r is constrained to be 0 in the the usual γ-shattering definition (has appeared previously in Gottlieb et al. [96]). The analogous dimension will be denoted by $\mathrm{fat}^0_\gamma(\mathcal{F})$.

Lemma 3.13 *For all* $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ *and* $\gamma > 0$*, we have*

$$\mathrm{fat}_\gamma(\mathcal{F}) = \max_{r\in\mathbb{R}^{\mathcal{X}}} \mathrm{fat}^0_\gamma(\mathcal{F} - r), \tag{3.7}$$

where $\mathcal{F} - r = \{f - r : f \in \mathcal{F}\}$ *is the* r*-shifted class; in particular, the maximum is always achieved.*

Proof. Fix $\mathcal{F}$ and γ. For any choice of $r \in \mathbb{R}^{\mathcal{X}}$, if $\mathcal{F} - r$ γ-shatters some set $S \subseteq \mathcal{X}$ at zero, then then $\mathcal{F}$ γ-shatters S in the usual sense with shift $r_S \in \mathbb{R}^S$ (i.e., the restriction of r to S). This proves that the left-hand side of Eq. (3.7) is at least as large as the right-hand side. Conversely, suppose that $\mathcal{F}$ γ-shatters some $S \subseteq \mathcal{X}$ in the usual sense, with some shift $r \in \mathbb{R}^S$. Choosing $r' \in \mathbb{R}^{\mathcal{X}}$ by $r'_S = r$ and $r'_{\mathcal{X}\setminus S} = 0$, we see that $\mathcal{F} - r'$ γ-shatters S at zero. This proves the other direction and hence the claim. □

Consider an *ambiguous* function class $F^\star \subseteq \{0,1,\star\}^X$. We say that $F^\star$ *shatters* a set $S \subseteq X$ if $F^\star(S) \supseteq \{0,1\}^S$. We say that $\bar{f} \in \{0,1\}^X$ is a *disambiguation* of $f^\star \in F^\star$ if the two functions agree on $x \in X$ whenever $f^\star(x) \neq \star$. We say that $\bar{F} \subseteq \{0,1\}^X$ is a disambiguation of $F^\star$ if each $\bar{f} \in \bar{F}$ is a disambiguation of *some* $f^\star \in F^\star$ and *every* $f^\star \in F^\star$ has a disambiguated representative $\bar{f} \in \bar{F}$. We define $\mathrm{VC}(F^\star)$ as the maximum size of a shattered set (possibly, ∞).

It will be convenient to visually represent such function classes as (possibly infinite) matrices, where the rows correspond to $f \in F$ and the columns correspond to $x \in X$.

Example 3.14 *It might be the case that* $\mathrm{VC}(F^\star) = 1$ *while any disambiguation* $\bar{F}$ *verifies* $\mathrm{VC}(\bar{F}) = 2$*:*

$$\begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ \star & 0 & 0 \\ 0 & \star & 0 \end{pmatrix} \end{array}.$$

It was mistakenly claimed in the conference version [56, Lemma 14] that one can always find a disambiguation $\bar{F}$ *such that* $\mathrm{VC}(\bar{F}) \leq \mathrm{VC}(F^\star)$*. We thank Yann Guermeur for pointing out this error.*

The following result provides a generic disambiguation rule that upper bounds the size of any disambiguated function classes. We reproduce it in Section 3.7 for completeness.

Lemma 3.15 *[97, Theorem 13] For* $X = \mathbb{N} = \{1,2,\ldots\}$ *and any* $F^\star \subseteq \{0,1,\star\}^X$ *with* $\mathrm{VC}(F^\star) \leq d$*, there is a disambiguation* $\bar{F} \subseteq \{0,1\}^X$ *with the following property: For each prefix* $X_m := [m] = \{1,2,\ldots,m\}$*, we have*

$$|\bar{F}(X_m)| \leq m^{\mathcal{O}(d \log m)}.$$

Example 3.16 *[97] Consider the following ambiguous class* $F^\star$ *consisting of* 5 *functions acting on the* 3

points $X = \{x_1, x_2, x_3\}$*:*

$$\begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ \star & 1 & 0 \\ 0 & \star & 1 \\ 1 & 0 & \star \end{pmatrix} \end{array}.$$

It is straightforward to verify that $\mathrm{VC}(F^*) = 1$ *and further that any disambiguation* $\bar{F}$ *verifies* $|\bar{F}(X)| = 5$*. Contrast this with the Sauer-Shelah lemma, which upper-bounds the number of behaviors that a class of VC-dimension* 1 *can achieve on* 3 *points by* 4*.*

There exists an ambiguous function class $F^\star$, such that for any disambiguation $\bar{F}$ it holds that $\mathrm{VC}(\bar{F}) = \infty$. See [97, Theorem 1].

Lemma 3.17 *Let* $G : \{-1,1\}^k \to \{-1,1\}$ *and let* $\mathcal{F}_1, \ldots, \mathcal{F}_k \subseteq \{-1,1\}^{\mathcal{X}}$ *be hypothesis classes with* $\mathrm{VC}(\mathcal{F}_j) = d_j$*. Denote* $\bar{d} := \frac{1}{k}\sum_{i=1}^k d_j$*. Define the function class*

$$G(\mathcal{F}_1, \ldots, \mathcal{F}_k) =: \{\mathcal{X} \ni x \mapsto G(f_1(x), \ldots, f_k(x)) : f_i \in \mathcal{F}_i\},$$

Then,

$$\mathrm{VC}(G(\mathcal{F}_1, \ldots, \mathcal{F}_k)) \leq 2k\log(3k)\bar{d}$$

Proof. We adapt the argument of Blumer et al. [98, Lemma 3.2.3], which is stated therein for k-fold unions and intersections. The $k = 1$ case is trivial, so assume $k \geq 2$. For any $S \subseteq \mathcal{X}$, define $G(\mathcal{F}_1, \ldots, \mathcal{F}_k)(S) \subseteq \{-1,1\}^S$ to be the restriction of $G(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ to S. The key observation is that

$$\begin{aligned} |G(\mathcal{F}_1, \ldots, \mathcal{F}_k)(S)| &\leq \prod_{j=1}^k |\mathcal{F}_j(S)| \\ &\leq \prod_{j=1}^k (e|S|/d_j)^{d_j} \\ &\leq (e|S|/\bar{d})^{\bar{d}k}. \end{aligned}$$

The last inequality requires proof. After taking logarithms and dividing both sides by k, it is equivalent to the claim that

$$\bar{d}\log\bar{d} \leq \frac{1}{k}\sum_{j=1}^k d_j \log d_j \ ,$$

an immediate consequence of Jensen's inequality applied to the convex function $f(x) = x \log x$.

The rest of the argument is identical to that of Blumer et al. [98]: one readily verifies that for $m = |S| = 2\bar{d}k \log(3k)$, we have $(em/\bar{d})^{\bar{d}k} < 2^m$. □

Proof (of Theorem 3.12). To prove the Theorem, it suffices to show that for all $\mathcal{F}_j \subseteq \mathbb{R}^S$

$$\mathrm{fat}^0_\gamma(\max((\mathcal{F}_j)_{j\in[k]})) \leq \mathcal{O}(\log(k)\log^2(m)\sum_{j=1}^{k}\mathrm{fat}^0_\gamma(\mathcal{F}_j)). \tag{3.8}$$

Indeed, we observe that r-shift commutes with the max operator:

$$\max((\mathcal{F}_j - r)_{j\in[k]}) = \max((\mathcal{F}_j)_{j\in[k]}) - r. \tag{3.9}$$

By applying Lemma 3.13 to the function class $\max((\mathcal{F}_j)_{j\in[k]})$ and using Eq. (3.9), we have

$$\mathrm{fat}_\gamma(\max((\mathcal{F}_j)_{j\in[k]})) = \max_r \mathrm{fat}^0_\gamma(\max((\mathcal{F}_j)_{j\in[k]}) - r) = \max_r \mathrm{fat}^0_\gamma(\max((\mathcal{F}_j - r)_{j\in[k]})).$$

Applying Eq. (3.8) to classes $F_j - r$ obtains

$$\max_r \mathrm{fat}^0_\gamma(\max((\mathcal{F}_j - r)_{j\in[k]}) \leq \max_r \mathcal{O}(\log(k)\log^2(m)\sum_{j=1}^{k}\mathrm{fat}^0_\gamma(\mathcal{F}_j - r)),$$

Then,

$$\begin{aligned}\max_r \mathcal{O}(\log(k)\log^2(m)\sum_{j=1}^{k}\mathrm{fat}^0_\gamma(\mathcal{F}_j - r)) &\leq \mathcal{O}(\log(k)\log^2(m)\sum_{j=1}^{k}\max_{r_j}\mathrm{fat}^0_\gamma(\mathcal{F}_j - r_j)) \\ &= \mathcal{O}(\log(k)\log^2(m)\sum_{j=1}^{k}\mathrm{fat}_\gamma(\mathcal{F}_j)),\end{aligned}$$

where the last identity follows from Lemma 3.13.

Now we proceed to prove Eq. (3.8). First, convert $\mathcal{F}_j \subseteq \mathbb{R}^S$ to a finite class $\mathcal{F}_j^\star \subseteq \{-\gamma, \gamma, \star\}^S$ for $S = \{x_1, \ldots, x_m\}$, as follows. For every vector in $v \in \mathcal{F}_j$, define $v^\star \in \mathcal{F}_j^\star$ by: $v_i^\star = \mathrm{sgn}(v_i)\gamma$ if $|v_i| \geq \gamma$ and $v_i^\star = \star$ else. The notion of shattering (at zero) remains the same: a set $T \subseteq S$ is shattered if $\{-\gamma, \gamma\}^T \subseteq \mathcal{F}_j^\star(T)$. Note that $\mathcal{F}_j^\star$ and $\mathcal{F}_j$ have the same γ-shattering dimension at zero.

Lemma 6.11 furnishes a mapping $\varphi : \mathcal{F}_j^\star \to \{-\gamma, \gamma\}^S$ such that (i) for all $v \in \mathcal{F}_j^\star$ and all $i \in [m]$, we have $v_i \neq \star \implies (\varphi(v))_i = v_i$ and (ii) $\varphi(\mathcal{F}_j^\star)$ does not shatter more points than $\mathcal{F}_j^\star$ times $\log^2(m)$. Together, properties (i) and (ii) imply that for all $j \in [k]$,

$$\mathrm{fat}^0_\gamma(\varphi(\mathcal{F}_j^\star)) \leq \mathcal{O}(\mathrm{fat}^0_\gamma(\mathcal{F}_j) \cdot \log^2(m)).$$

Finally, observe that any set of points in S γ-shattered by $\max((\mathcal{F}_j)_{j\in[k]})$ are also shattered by $\max((\varphi(\mathcal{F}_j^\star))_{j\in[k]})$. Applying Lemma 3.17 with $G(f_1,\ldots,f_k)(x) = \max_{j\in[k]} f_j(x)$ shows that $\max((\varphi(\mathcal{F}_j^\star))_{j\in[k]})$ cannot shatter $2\log(3k)\sum_{j=1}^k d_j$ points, where

$$d_j = \mathrm{fat}_\gamma^0(\varphi(\mathcal{F}_j^\star)) \le \mathcal{O}(\mathrm{fat}_\gamma^0(\mathcal{F}_j)\cdot\log^2(m)).$$

We have shown that,

$$\mathrm{fat}_\gamma^0(\max((\mathcal{F}_j)_{j\in[k]})) \le \mathcal{O}(\log(k)\log^2(m)\sum_{j=1}^k \mathrm{fat}_\gamma^0(\mathcal{F}_j)),$$

this concludes the proof of Eq. (3.8). □

Shattering Dimension of L_1 and L_2 Loss Classes

Lemma 3.18 *Let $\mathcal{H}\subset\mathbb{R}^m$ be a real valued function class on m points. denote $L^1_\mathcal{H}$ and $L^2_\mathcal{H}$ the L_1 and L_2 loss classes of $\mathcal{H}$ respectively. Assume $L^2_\mathcal{H}$ is bounded by M. For any $\mathcal{H}$,*

$$\mathrm{fat}_\gamma(L^1_\mathcal{H}) \;\le\; \mathcal{O}(\log^2(m)\mathrm{fat}_\gamma(\mathcal{H})), \quad \textit{and} \quad \mathrm{fat}_\gamma(L^2_\mathcal{H}) \le \mathcal{O}(\log^2(m)\mathrm{fat}_{\gamma/2M}(\mathcal{H})).$$

Lemma 3.19 *Let $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ be an arbitrary loss function. For $j\in[k]$ define*

$$\mathcal{F}_\mathcal{H}^{(j),\ell} := \{\mathcal{X}\times\mathcal{Y}\ni(x,y)\mapsto\ell(h(z_j),y) : h\in\mathcal{H},\, \mathcal{U}(x)=\{z_1,\ldots,z_k\}\},$$

and

$$L^\ell_\mathcal{H} := \{Z\times\mathcal{Y}\ni(z,y)\mapsto\ell(h(z),y) : h\in\mathcal{H}\}.$$

Then, for all $\gamma>0$,

$$\mathrm{fat}_\gamma(\mathcal{F}_\mathcal{H}^{(j),\ell}) \le \mathrm{fat}_\gamma(L^\ell_\mathcal{H}).$$

Proof. The claim stems from the inclusion $\mathcal{F}_\mathcal{H}^{(j),\ell}\subseteq L^\ell_\mathcal{H}$. □

Proof (of Lemma 3.18). 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 Theorem 3.12 that

$$\mathrm{fat}_\gamma(\mathrm{abs}(\mathcal{F})) < \mathcal{O}(\log^2(m)(\mathrm{fat}_\gamma(\mathcal{F}) + \mathrm{fat}_\gamma(-\mathcal{F}))) < \mathcal{O}(\log^2(m)\mathrm{fat}_\gamma(\mathcal{F})). \tag{3.10}$$

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 $\mathcal{F}^{\circ 2} := \left\{f^2; f \in \mathcal{F}\right\}$. We would like to bound $\mathrm{fat}_\gamma(\mathcal{F}^{\circ 2})$ in terms of $\mathrm{fat}_\gamma(\mathcal{F})$. Suppose that $\mathcal{F}^{\circ 2}$ γ-shatters some set $\{x_1,\ldots,x_k\}$ with shift $r^2 = (r_1^2,\ldots,r_k^2) \in [0,M]^k$ (there is no loss of generality in assuming that the shift has the same range as the function class). Using the elementary inequality

$$|a^2 - b^2| \leq 2M|a-b|, \qquad a,b \in [0,M],$$

we conclude that $\mathcal{F}$ is able to $\gamma/(2M)$-shatter the same k points and thus $\mathrm{fat}_\gamma(\mathcal{F}^{\circ 2}) \leq \mathrm{fat}_{\gamma/(2M)}(\mathcal{F})$.

To extend this result to the case where $\mathcal{F} \subset [-M,M]^{\mathcal{X}}$, we use Eq. (3.10). In particular, define $\mathcal{F}$ as the L_2 loss class of $\mathcal{H}$:

$$\mathcal{F} = \left\{\mathcal{X} \times \mathbb{R} \ni (x,y) \mapsto (h(x) - y)^2; h \in \mathcal{H}\right\}.$$

Then

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

□

Auxiliary Results

Finally, before providing formal proofs, we use the following result on the fat-shattering of convex hulls. We then conclude a bound on the fat-shattering dimension of k-fold maximum of convex hulls using Theorem 3.12.

Theorem 3.20 *[99, Theorem 1.5] There is an absolute constant C, such that for every function class F bounded by $[0,1]$ and every $\gamma > 0$,*

$$\mathrm{fat}_\gamma(\mathrm{conv}(F)) \leq C\frac{\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F})}{\gamma^2}\log^2\left(\frac{2\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F})}{\gamma}\right)$$

Corollary 3.21 *Let $S = \{x_1, \ldots, x_m\}$. For any k real-valued functions classes $\mathcal{F}_1, \ldots, \mathcal{F}_k \subseteq [0,1]^S$,*

$$\mathrm{fat}_\gamma(\max((\mathrm{conv}(\mathcal{F}_j))_{j\in[k]})) \leq \mathcal{O}\left(k\log(k)\log^2(m)\max_{j\in[k]}\left(\frac{\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F}_j)}{\gamma^2}\log^2\left(\frac{\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F}_j)}{\gamma}\right)\right)\right).$$

Proof.

$$\begin{aligned}
\mathrm{fat}_\gamma(\max((\mathrm{conv}(\mathcal{F}_j))_{j\in[k]})(S)) &\overset{(i)}{\leq} \mathcal{O}\left(\log(k)\log^2(m)\sum_{j=1}^k \mathrm{fat}_\gamma(\mathrm{conv}(\mathcal{F}_j))\right) \\
&\overset{(ii)}{\leq} \mathcal{O}\left(\log(k)\log^2(m)\sum_{j=1}^k \frac{\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F}_j)}{\gamma^2}\log^2\left(\frac{\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F}_j)}{\gamma}\right)\right) \\
&\leq \mathcal{O}\left(k\log(k)\log^2(m)\max_{j\in[k]}\left(\frac{\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F}_j)}{\gamma^2}\log^2\left(\frac{\mathrm{fat}_{\frac{\gamma}{4}}(\mathcal{F}_j)}{\gamma}\right)\right)\right),
\end{aligned}$$

where (i) stems from Theorem 3.12 and (ii) stems from Theorem 3.20. □

Proofs

We now formally prove our main results for this section, generalization bounds in the case of real-valued functions.

Proof (of Theorem 3.8). We follow the same steps as in the proof of Theorem 3.4 with two changes. The first one is bounding the empirical Rademacher complexity via the fat-shattering dimension (instead of the VC-dimension in the binary case), using Theorem 2.13,

$$R_n(\mathcal{F}|\mathbf{w}) \leq \inf_{\beta \geq 0} \left\{ 4\beta + \sqrt{\frac{C}{n}} \int_{\beta}^{1} \sqrt{\mathrm{fat}_{c\gamma}(\mathcal{F}) \log\left(\frac{2}{\gamma}\right)} d\gamma \right\} := g_n(\mathcal{F}),$$

this bound holds for every sequence of points. The second difference is that we now need to bound the maximum fat-shattering dimension (instead of the VC-dimension) over the classes $\mathcal{F}_{\mathcal{H}}^{(j)}$, for that purpose we use Lemma 3.18 and Lemma3.19,

$$\max_{j \in [k]} \mathrm{fat}_{\gamma}(\mathcal{F}_{\mathcal{H}}^{(j)}) \leq \mathcal{O}(\log^2(n) \mathrm{fat}_{\gamma}(\mathcal{H})).$$

Denote

$$m_n(\mathcal{H}) = \inf_{\beta \geq 0} \left\{ 4\beta + \mathcal{O}\left(\sqrt{\frac{\log^4(n)}{n}} \int_{\beta}^{1} \sqrt{\mathrm{fat}_{c\gamma}(\mathcal{H}) \log\left(\frac{1}{\gamma}\right)} d\gamma \right) \right\}.$$

Similar to Theorem 3.4, the function $x \log^{3/2+\alpha}(n/x)$ is monotonic increasing for $x \in (0, n/e^{3/2+\alpha}]$. For sufficiently large n $\left(g_n(\mathcal{F}) \leq n/e^{3/2+\alpha}\right)$ and considering the aforementioned changes we have that for any $\alpha > 0$ there exists a constant $C_\alpha > 0$ such that

$$\begin{aligned}
&\mathcal{R}_n(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j \in [k]})|\mathbf{x} \times \mathbf{y}) \\
&\leq C_\alpha \sqrt{k} \cdot \max_{j \in [k]} \max_{\mathbf{w} = w_{1:n}} \mathcal{R}_n(\mathcal{F}_{\mathcal{H}}^{(j)}|\mathbf{w}) \cdot \log^{\frac{3}{2}+\alpha}\left(\frac{n}{\max_{j \in [k]} \max_{\mathbf{w}=w_{1:n}} \mathcal{R}_n(\mathcal{F}_{\mathcal{H}}^{(j)}|\mathbf{w})} \right) \\
&\leq \mathcal{O}\left(C_\alpha \sqrt{k} \max_{j \in [k]} g_n(\mathcal{F}_{\mathcal{H}}^{(j)}) \cdot \log^{\frac{3}{2}+\alpha}\left(\frac{n}{\max_{j \in [k]} g_n(\mathcal{F}_{\mathcal{H}}^{(j)})} \right) \right) \\
&= \mathcal{O}\left(C_\alpha \sqrt{k} \cdot m_n(\mathcal{H}) \cdot \log^{\frac{3}{2}+\alpha}\left(\frac{n}{m_n(\mathcal{H})} \right) \right).
\end{aligned}$$

We conclude that

$$|\mathrm{Err}(\tilde{h}; \mathcal{D}, \mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h}; S, \mathcal{U})| \leq \mathcal{O}\left(C_\alpha \sqrt{k} \cdot m_n(\mathcal{H}) \cdot \log^{\frac{3}{2}+\alpha}\left(\frac{n}{m_n(\mathcal{H})} \right) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}} \right). \qquad \square$$

Proof (of Theorem 3.10). Similar to the proof for binary case, we bound the empirical Rademacher complexity of the loss class of $\tilde{h} \in \Delta(\mathcal{H})$.

$$\begin{aligned}|\mathrm{Err}(\tilde{h};\mathcal{D},\mathcal{U}) - \widehat{\mathrm{Err}}(\tilde{h};S,\mathcal{U})| &= |E_{(x,y)\sim D}\max_{j\in[k]}\sum_{t=1}^{T}\alpha_t f_t^{(j)}(x,y) - \frac{1}{n}\sum_{(x,y)\in S}\max_{j\in[k]}\sum_{t=1}^{T}\alpha_t f_t^{(j)}(x,y)| \\ &\leq 2\mathcal{R}_n(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x}\times\mathbf{y}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}},\end{aligned}$$

where the inequality stems from applying Theorem 2.12 on the function class $\mathrm{conv}^{\mathcal{U}}(L_{\mathcal{H}})$ and Eq. (3.5). From Theorem 2.13 we have

$$\begin{aligned}&\mathcal{R}_n(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x}\times\mathbf{y}) \\ &\leq \inf_{\alpha\geq 0}\left\{4\alpha + \sqrt{\frac{C_1}{n}}\int_{\alpha}^{1}\sqrt{\mathrm{fat}_{c\gamma}(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]}))\log\left(\frac{2}{\gamma}\right)}d\gamma\right\}.\end{aligned}$$

Using Corollary 3.21 we upper bound the inner term by

$$\mathcal{O}\left(\sqrt{\frac{k\log(k)\log^2(n)}{n}}\int_{\alpha}^{1}\sqrt{\log\left(\frac{1}{\gamma}\right)\max_{j\in[k]}\left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{F}_{\mathcal{H}}^{(j)}(S))}{\gamma^2}\log^2\left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{F}_{\mathcal{H}}^{(j)}(S))}{\gamma}\right)\right)}d\gamma\right).$$

Lemmas 3.18 and 3.19 concludes the proof with

$$\mathcal{O}\left(\sqrt{\frac{k\log(k)\log^4(n)}{n}}\int_{\alpha}^{1}\sqrt{\log\left(\frac{1}{\gamma}\right)\left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{H})}{\gamma^2}\log^2\left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{H})}{\gamma}\right)\right)}d\gamma\right). \qquad \square$$

Proof (of Theorem 3.11). Denote the sample size by $|S| = n$. We start off with a known generalization bound by Bartlett and Long [32], showing that for any function class $\mathcal{H} : \mathcal{Z} \to [0,1]$, the sample size is at least

$$n \leq \mathcal{O}\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_{\frac{\epsilon}{5}}(\mathcal{H})\log^2\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right).$$

In our case, the function class we are interested in is $\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})$. by Corollary 3.21 we have that

$$\mathrm{fat}_{\epsilon}(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})) \leq \mathcal{O}\left(k\log(k)\log^2(n)\left(\frac{\mathrm{fat}_{\frac{\epsilon}{4}}(\mathcal{H})}{\epsilon^2}\log^2\left(\frac{\mathrm{fat}_{\frac{\epsilon}{4}}(\mathcal{H})}{\epsilon}\right)\right)\right).$$

Thus, it suffices to solve the following

$$n \leq \mathcal{O}\left(\left(\frac{1}{\epsilon^2}\left(k\log(k)\log^2(n)\left(\frac{\mathrm{fat}_{\frac{\epsilon}{4}}(\mathcal{H})}{\epsilon^2}\log^2\left(\frac{\mathrm{fat}_{\frac{\epsilon}{4}}(\mathcal{H})}{\epsilon}\right)\right)\log^2\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)\right).$$

Denote $d = \mathrm{fat}_{\frac{\epsilon}{4}}(\mathcal{H})$, $A = \frac{1}{\epsilon^2}\log\frac{1}{\delta}$, and $B = k\log(k)\frac{d}{\epsilon^4}\log^2\frac{d}{\epsilon}\log^2\frac{1}{\epsilon}$. It suffices to take $n_0 = \mathcal{O}\left(B\log^2 B + A\right)$, therefore,

$$n \leq \mathcal{O}\left(\frac{1}{\epsilon^2}\left(k\log(k)\frac{d}{\epsilon^2}\log^2\frac{d}{\epsilon}\log^2\frac{1}{\epsilon}\log^2\left(k\log(k)\frac{d}{\epsilon^4}\log^2\frac{d}{\epsilon}\log^2\frac{1}{\epsilon}\right)+\log\frac{1}{\delta}\right)\right). \qquad \square$$

3.7 Deferred Proofs

Proof (of Lemma 3.1). Take an arbitrary sample $S = \{(x_1, y_1), ..., (x_n, y_n)\}$. Construct the set that contains all possible corrupted examples on inputs from S, $S_{\mathcal{U}} = \bigcup_{i\in[n]}\{z : z \in \mathcal{U}(x_i)\}$, the size of $S_{\mathcal{U}}$ is at most nk. Denote by $L^{\mathcal{U}}_{\mathcal{H}}(S)$ the set of all possible behaviors on S using functions in $L^{\mathcal{U}}_{\mathcal{H}}$, and by $\mathcal{H}(S_{\mathcal{U}})$, the set of all possible behaviors on $S_{\mathcal{U}}$ using functions in $\mathcal{H}$. Namely, $L^{\mathcal{U}}_{\mathcal{H}}(S) = \left\{(\ell(x_1, y_1), \ldots, \ell(x_n, y_n)) : \ell \in L^{\mathcal{U}}_{\mathcal{H}}\right\}$ and $\mathcal{H}(S_{\mathcal{U}}) = \{(h(z_1), \ldots, h(z_m)) : h \in \mathcal{H}\}$. Observe that each pattern in the set $L^{\mathcal{U}}_{\mathcal{H}}(S)$ will map to at least one pattern in $\mathcal{H}(S_{\mathcal{U}})$, implying that the size of $L^{\mathcal{U}}_{\mathcal{H}}(S)$ is at most the size of $\mathcal{H}(S_{\mathcal{U}})$. Using Sauer's lemma, the size of $\mathcal{H}(S_{\mathcal{U}})$ is at most $(nk)^d$, solving for n such that $(nk)^d < 2^n$ yields the stated bound. $\square$

Proof (of Corollary 3.9). We seek an upper bounds on the following term in the case of homogeneous hyperplanes with norm bounded by 1.

$$\begin{aligned} m_n(\mathcal{H}) &= \inf_{\beta\geq 0}\left\{4\beta + \mathcal{O}\left(\sqrt{\frac{\log^4(n)}{n}}\int_\beta^1\sqrt{\mathrm{fat}_{c\gamma}(\mathcal{H})\log\left(\frac{1}{\gamma}\right)}d\gamma\right)\right\}, \\ &\leq \inf_{\beta\geq 0}\left\{4\beta + \mathcal{O}\left(\sqrt{\frac{\log^4(n)}{n}}\int_\beta^1\frac{1}{\gamma}\sqrt{\log\left(\frac{2}{\gamma}\right)}d\gamma\right)\right\}, \end{aligned}$$

where the inequality stems from the bound $\mathrm{fat}_\delta(H) \leq \frac{1}{\delta^2}$ [100].

Compute

$$\int_\beta^1\frac{1}{t}\sqrt{\log\frac{2}{t}}dt = \frac{2}{3}\left((\log 2/\beta)^{3/2} - (\log 2)^{3/2}\right),$$

choosing $\beta = 1/\sqrt{n}$ yields

$$m_n(\mathcal{H}) \quad \leq \quad \mathcal{O}\left(\sqrt{\frac{1}{n}}\log^{\frac{7}{2}}(n)\right).$$

The function $x \log^{3/2+\alpha}(n/x)$ is monotonic increasing for $x \in (0, n/e^{3/2+\alpha}]$. Then, for sufficiently large n, $\left(\log^{7/2}(n) e^{3/2+\alpha}\right)^{2/3} \leq n$ we have

$$m_n(\mathcal{H}) \cdot \log^{\frac{3}{2}+\alpha}\left(\frac{n}{m_n(\mathcal{H})}\right) \leq \mathcal{O}\left(\sqrt{\frac{1}{n}} \log^{\frac{7}{2}}(n) \log^{\frac{3}{2}+\alpha}\left(\frac{n}{\log^{\frac{7}{2}}(n)}\right)\right). \qquad \square$$

Proof (of Lemma 3.15). For any finite sequence $(x_1, y_1), \ldots, (x_k, y_k)$ with $x_i \in X$, $y_i \in \{0, 1\}$, and $x_1 < \ldots < x_k$, denote by $F^\star|_{(x_1,y_1),\ldots,(x_k,y_k)}$ the subfamily of those members of $F^\star$ that label the point x_i with y_i, for all i. For such a constrained subfamily, we define its *weight*:

$$w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k)}) \quad = \quad \sum_S \frac{1}{n(S)^{d+1}},$$

where the summation is over all nonempty subsets S of $\mathbb{N} \setminus \{1, \ldots, x_k\}$ that are shattered by this subfamily, and $n(S)$ denotes the largest element of S. The definition applies verbatim to the special case where $k = 0$, i.e., $F^\star|_\emptyset = F^\star$. Clearly, if c is a prefix of c', then $w(F^\star|_c) \geq w(F^\star|_{c'})$, and hence the maximum weight is achieved by $F^\star|_\emptyset = F^\star$. The latter is upper-bounded by

$$w(F^\star) \leq \sum_{n \in \mathbb{N}} \frac{n^{d-1}}{n^{d+1}} = \sum_{n \in \mathbb{N}} \frac{1}{n^2} = \frac{\pi^2}{6}, \tag{3.11}$$

where the numerator n^{d-1} accounts for the number of of subsets of $[n]$ of size at most d which have n as their largest element.

Any constrained subfamily $F^\star|_{(x_1,y_1),\ldots,(x_k,y_k)}$ induces the "majority" classifier $M[F^\star|_{(x_1,y_1),\ldots,(x_k,y_k)}] : X \to \{0, 1\}$ as follows:

$$M[F^\star|_{(x_1,y_1),\ldots,(x_k,y_k)}](x) \quad = \quad \mathbb{1}[w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k),(x,1)}) > w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k),(x,0)})] \tag{3.12}$$

(ties may be broken arbitrarily, and the rule above favors 0 in such cases). We observe that

$$w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k)}) \quad \geq \quad w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k),(x,1)}) + w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k),(x,0)}),$$

with equality occurring if and only if no $f^\star \in F^\star$ verifies $f^\star(x) = \star$.

We now describe the disambiguation procedure. We proceed one "row" $f^\star \in F^\star$ at a time. For a given $f^\star \in F^\star$, initialize the "constraint" sequence c to be empty (i.e., to be of length $k = 0$). Predict the label at $x = 1$ via $y = M[F^\star|_c](x)$. The prediction is said to be a *mistake* if $f^\star(x) \neq \star$ and $y \neq f(x)$. In case of a mistake, append $(x, f^\star(x))$ to the end of the constraint sequence c and leave c unchanged otherwise. Repeat the procedure for $x = 2$: predict $y = M[F^\star|_c](x)$ and append $(x, f^\star(x))$ to c in case of a mistake.

Repeating these steps for $x = 1, 2, \ldots, m$ produces a disambiguation $\bar{f}$ of $f^\star$. To disambiguate the next "row" of $F^\star$, re-initialize $c := \emptyset$ and repeat the procedure above for $x = 1, 2, \ldots, m$.

Having described the construction of $\bar{F}$, it remains to analyze the number of behaviors that it can possibly attain on a prefix of length m — that is, to bound $|\bar{F}(X_m)|$. The first key observation is that if c is the constraint before a mistake and c' immediately after, then (3.12) implies that $w(F^\star|_c) \geq \frac{1}{2} w(F^\star|_{c'})$ (i.e., the weight of the constrained family is reduced by a half or more). This is because a mistake is caused by the *majority* being wrong, and the updated constraint effectively removes those members of $F^\star$ that contributed to the mistake. The second key observation is that if some $x \leq m$ witnesses the last[2] mistake when disambiguating a given $f^\star$, the weight prior to updating the constraint on this mistake is at least $1/m^{d+1}$ — because in this case, $\{x\}$ must be a shattered set.

Together with (3.11), these two estimates on the weight immediately prior to the last update imply that the number of updates u satisfies

$$\frac{1}{m^{d+1}} 2^{u-1} \leq w(F^\star) \leq \frac{\pi^2}{6},$$

which implies that $u = \mathcal{O}(d \log m)$. To translate this into an estimate on $|\bar{F}(X_m)|$, observe that any $\bar{f} \in \bar{F}$ is uniquely defined by the indices on which a mistake was made during its disambiguation procedure. It follows that $|\bar{F}(X_m)| \leq \mathcal{O}(\binom{m}{u}) \leq m^{\mathcal{O}(d \log m)}$. □

Additional Generalization Bound for Binary Classification

We derive the result in Eq. (3.6). Denote the sample size $|S| = n$ and $\mathrm{VC}(\mathcal{H}) = d$. Using Theorem 3.10 for binary valued function classes we upper bound the empirical Rademacher complexity on the sample $R_n(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x} \times \mathbf{y})$ by

$$\inf_{\alpha \geq 0} \left\{ 4\alpha + \mathcal{O}\left(\sqrt{\frac{k \log(k) \log^4(n)}{n}} \int_\alpha^1 \sqrt{\log\left(\frac{2}{\gamma}\right) \left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{H})}{\gamma^2} \log^2\left(\frac{\mathrm{fat}_{\frac{c\gamma}{4}}(\mathcal{H})}{\gamma} \right) \right)} d\gamma \right) \right\}.$$

For a binary-valued class, this is upper-bounded by

$$\begin{aligned} &\inf_{\alpha \geq 0} \left\{ 4\alpha + \mathcal{O}\left(\sqrt{\frac{dk \log(k) \log^4(n)}{n}} \int_\alpha^1 \sqrt{\log\left(\frac{2}{\gamma}\right) \left(\frac{1}{\gamma^2} \log^2\left(\frac{d}{\gamma} \right) \right)} d\gamma \right) \right\} \\ &= \inf_{\alpha \geq 0} \left\{ 4\alpha + \mathcal{O}\left(\sqrt{\frac{dk \log(k) \log^4(n)}{n}} \int_\alpha^1 \frac{1}{\gamma} \log\left(\frac{d}{\gamma} \right) \sqrt{\log\left(\frac{2}{\gamma}\right)} d\gamma \right) \right\}. \end{aligned}$$

[2]The case where no mistakes are made is trivial.

Computing

$$\begin{aligned}
\int_\alpha^1 \frac{1}{\gamma}\log\left(\frac{d}{\gamma}\right)\sqrt{\log\left(\frac{2}{\gamma}\right)}d\gamma &= \log(d)\int_\alpha^1 \frac{1}{\gamma}\sqrt{\log\left(\frac{2}{\gamma}\right)}d\gamma + \int_\alpha^1 \frac{1}{\gamma}\log\left(\frac{1}{\gamma}\right)\sqrt{\log\left(\frac{2}{\gamma}\right)}d\gamma \\
&\leq \log(d)\int_\alpha^1 \frac{1}{\gamma}\sqrt{\log\left(\frac{2}{\gamma}\right)}d\gamma + \int_\alpha^1 \frac{1}{\gamma}\log^{\frac{3}{2}}\left(\frac{2}{\gamma}\right)d\gamma \\
&= \frac{2}{3}\log(d)\left(\log^{\frac{3}{2}}\left(\frac{2}{\alpha}\right) - \log^{\frac{3}{2}}(2)\right) - \frac{2}{5}\left(\log^{\frac{5}{2}}(2) - \log^{\frac{5}{2}}\left(\frac{2}{\alpha}\right)\right) \\
&\leq \log(d)\log^{\frac{3}{2}}\left(\frac{2}{\alpha}\right) + \log^{\frac{5}{2}}\left(\frac{2}{\alpha}\right)
\end{aligned}$$

and choosing $\alpha = \frac{1}{\sqrt{n}}$ yields

$$\log(d)\log^{\frac{3}{2}}(2\sqrt{n}) + \log^{\frac{5}{2}}\left(2\sqrt{n}\right) \leq \mathcal{O}\left(\log(d)\log^{\frac{5}{2}}(n)\right)$$

and

$$\mathcal{R}_n(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x}\times\mathbf{y}) \leq \mathcal{O}\left(\sqrt{\frac{d\log^2(d)k\log(k)\log^9(n)}{n}}\right).$$

Chapter 4

A Characterization of Semi-Supervised Adversarially Robust PAC Learnability

We study the problem of learning an adversarially robust predictor to test time attacks in the *semi-supervised* PAC model. We address the question of how many *labeled* and *unlabeled* examples are required to ensure learning. We show that having enough unlabeled data (the size of a labeled sample that a fully-supervised method would require), the labeled sample complexity can be arbitrarily smaller compared to previous works, and is sharply characterized by a *different* complexity measure. We prove nearly matching upper and lower bounds on this sample complexity. This shows that there is a significant benefit in semi-supervised robust learning even in the worst-case distribution-free model, and establishes a gap between supervised and semi-supervised label complexities which is known not to hold in standard non-robust PAC learning.

4.1 Introduction

The problem of learning predictors that are immune to adversarial corruptions at inference time is central in modern machine learning. The phenomenon of fooling learning models by adding imperceptible perturbations to their input illustrates a basic vulnerability of learning-based models and is named *adversarial examples*. We study the model of adversarially-robust PAC learning, in a *semi-supervised* setting.

Adversarial robustness has been shown to significantly benefit from semi-supervised learning, mostly empirically, but also theoretically in some specific cases of distributions [e.g., 16, 67–70, 101, 102]. In this paper, we ask the following natural question. To what extent can we benefit from *unlabeled* data in the learning process of robust models in the general case? More specifically, what is the sample complexity in a distribution-free model?

Our semi-supervised model is formalized as follows. Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. We formalize the adversarial attack 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_p ball centered at x. In this paper, we have no restriction on $\mathcal{U}$, besides $x \in \mathcal{U}(x)$. The robust error of hypothesis h on a pair (x, y) is $\sup_{z \in \mathcal{U}(x)} \mathbb{I}[h(z) \neq y]$. The learner has access to both *labeled* and *unlabeled* examples drawn i.i.d. from unknown distribution $\mathcal{D}$, and the goal is to find $h \in \mathcal{H}$ with low robust error on a random point from $\mathcal{D}$. The sample complexity in semi-supervised learning has two parameters, the number of labeled examples and the number of unlabeled examples which suffice to ensure learning. The learner would like to restrict the amount of labeled data, which is significantly more expensive to obtain than unlabeled data.

In this paper, we show a gap between supervised and semi-supervised label complexities of adversarially robust learning in a distribution-free model. The label complexity in semi-supervised may be arbitrarily smaller compared to the supervised case and is characterized by a different complexity measure. Importantly, we are not using more data, just less labeled data. The unlabeled sample size is the same as how much labeled data a fully-supervised method would require, so this is a strict improvement. This kind of gap is known not to hold in standard (non-robust) PAC learning, this is a unique property of robust learning.

Background

The following complexity measure $\mathrm{VC}_{\mathcal{U}}$ was introduced by Montasser et al. [9] (and denoted there by $\dim_{\mathcal{U}\times}$) as a candidate for determining the sample complexity of supervised robust learning. It was shown that indeed its finiteness is necessary, but not sufficient. This parameter is our primary object in this work, as we will show that it characterizes the labeled sample complexity of *semi-supervised* robust PAC-learning.

Definition 4.1 ($\mathrm{VC}_{\mathcal{U}}$-dimension) *A sequence of points $\{x_1, \ldots, x_k\}$ is $\mathcal{U}$-shattered by $\mathcal{H}$ if $\forall y_1, \ldots, y_k \in \{0,1\}$, $\exists h \in \mathcal{H}$ such that $\forall i \in [k], \forall z \in \mathcal{U}(x_i)$, $h(z) = y_i$. The $\mathrm{VC}_{\mathcal{U}}(\mathcal{H})$ is largest integer k for which there exists a sequence $\{x_1, \ldots, x_k\}$ $\mathcal{U}$-shattered by $\mathcal{H}$.*

Intuitively, this dimension relates to a shattering of the entire perturbation sets, instead of one point in the standard VC-dimension. When $\mathcal{U}(x) = \{x\}$, this parameter coincides with the standard VC. Moreover, for any hypothesis class $\mathcal{H}$, it holds that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$, and the gap can be arbitrarily large. That is, there exist $\mathcal{H}_0$ such that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{VC}(\mathcal{H}_0) = \infty$ (see Proposition 4.8).

For an improved lower bound on the sample complexity, Montasser et al. [9, Theorem 10] introduced the Robust Shattering dimension, denoted by $\mathrm{RS}_{\mathcal{U}}$ (and denoted there by $\dim_{\mathcal{U}}$).

Definition 4.2 ($\mathrm{RS}_{\mathcal{U}}$-dimension) *A sequence $x_1, \ldots, x_k$ is said to be $\mathcal{U}$-robustly shattered by $\mathcal{F}$ if $\exists z_1^+, z_1^-, \ldots, z_k^+, z_k^-$ such that $x_i \in \mathcal{U}\left(z_i^+\right) \cap \mathcal{U}\left(z_i^-\right) \forall i \in [k]$ and $\forall y_1, \ldots, y_k \in \{+,-\}, \exists f \in \mathcal{F}$ with*

$f(\zeta) = y_i$, $\forall \zeta \in \mathcal{U}\left(z_i^{y_i}\right), \forall i \in [k]$. *The $\mathcal{U}$-robust shattering dimension* $\mathrm{RS}_{\mathcal{U}}(\mathcal{H})$ *is defined as the maximum size of a set that is $\mathcal{U}$-robustly shattered by $\mathcal{H}$.*

Specifically, the lower bound on the sample complexity is $\Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ for realizable robust learning, and $\Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ for agnostic robust learning. They also showed upper bounds of $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}\cdot\mathrm{VC}^*}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$[1] in the realizable case and $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}\cdot\mathrm{VC}^*}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ in the agnostic case, where VC^* is the dual VC dimension (definitions are in Section 4.7). Montasser et al. [9] showed that for any $\mathcal{H}$, $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \le \mathrm{RS}_{\mathcal{U}}(\mathcal{H}) \le \mathrm{VC}(\mathcal{H})$, and there can be an arbitrary gap between them. Specifically, there exists $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = \infty$, and there exists $\mathcal{H}_1$ with $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_1) = 0$ and $\mathrm{VC}(\mathcal{H}_1) = \infty$.

Main Contributions

- In Section 4.3, we first analyze the simple case where the support of the marginal distribution on the inputs is fully known to the learner. In this case, we show a tight bound of $\Theta\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$ on the labeled complexity for learning $\mathcal{H}$.

- In Section 4.4, we present a generic algorithm that can be applied both for the realizable and agnostic settings. We prove an upper bound and nearly matching lower bounds on the sample complexity in the realizable case. For semi-supervised robust learning, we prove a labeled sample complexity bound Λ^{ss} and compare it to the sample complexity of supervised robust learning Λ^{s}. Our algorithm uses $\Lambda^{\mathrm{ss}} = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ *labeled* examples and $\mathcal{O}(\Lambda^{\mathrm{s}})$ *unlabeled* examples. Recall that $\Lambda^{\mathrm{s}} = \Omega(\mathrm{RS}_{\mathcal{U}})$, and since $\mathrm{RS}_{\mathcal{U}}$ can be arbitrarily larger than $\mathrm{VC}_{\mathcal{U}}$, this means our labeled sample complexity represents a significant improvement over the sample complexity of supervised learning.

- In Section 4.5, we prove upper and lower bounds on the sample complexity in the agnostic setting. We reveal an interesting structure, which is inherently different than the realizable case. Let η be the minimal agnostic error. If we allow an error of $3\eta+\epsilon$, it is sufficient for our algorithm to have $\Lambda^{\mathrm{ss}} = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ *labeled* examples and $\mathcal{O}(\Lambda^{\mathrm{s}})$ *unlabeled* examples (as in the realizable case). If we insist on having error $\eta + \epsilon$, then there is a lower bound of $\Lambda^{\mathrm{ss}} = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ labeled examples. Furthermore, an error of $(\frac{3}{2} - \gamma)\eta + \epsilon$ is unavoidable if the learner is restricted to $\mathcal{O}(\mathrm{VC}_{\mathcal{U}})$ labeled examples, for any $\gamma > 0$. We also show that *improper* learning is necessary, similar to the supervised case. We summarize the results in Section 4.1 showing for which labeled and unlabeled samples we have a robust learner.

- The above results show that there is a significant benefit in semi-supervised robust learning. For example, take $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = n$. The labeled sample size for learning $\mathcal{H}_0$ in supervised

[1] $\tilde{\mathcal{O}}(\cdot)$ stands for omitting poly-logarithmic factors of $\mathrm{VC}, \mathrm{VC}^*, \mathrm{VC}_{\mathcal{U}}, \mathrm{RS}_{\mathcal{U}}, 1/\epsilon, 1/\delta$.

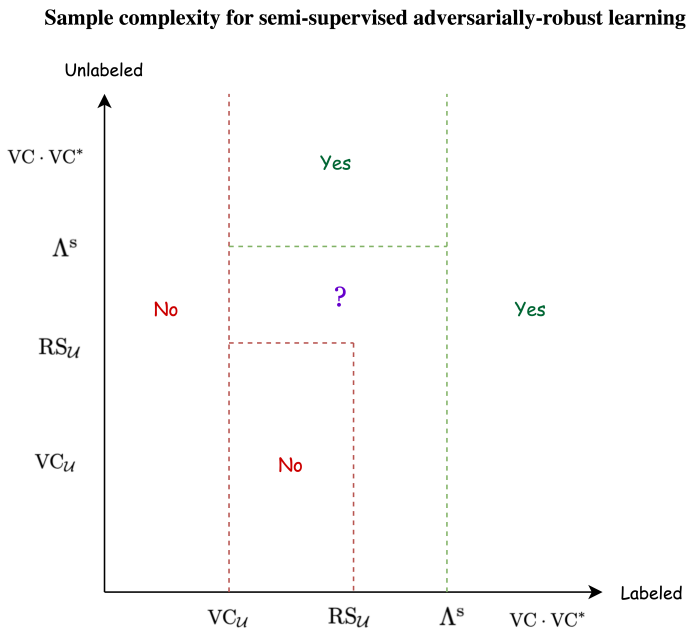

Figure 4.1: Sample complexity regimes for semi-supervised robust learning, for the realizable model and the agnostic model with error $3\eta + \epsilon$, where η is the minimal agnostic error in the hypothesis class. Obtaining an error of $\eta + \epsilon$ requires at least $\mathrm{RS}_{\mathcal{U}}$ labeled examples, as in the supervised case. Λ^{s} denotes the sample complexity of supervised robust learning. It is an open question whether Λ^{s} equals $\mathrm{RS}_{\mathcal{U}}$.

learning is $\Omega(n)$. In contrast, in semi-supervised learning our algorithms require only $\mathcal{O}(1)$ *labeled* examples and $\mathcal{O}(n)$ *unlabeled* examples. We are not using more data, just less labeled data. Note that n can be arbitrarily large.

- A byproduct of our result is that if we assume that the distribution is robustly realizable by a hypothesis class (i.e., there exists a hypothesis with zero robust error) then, with respect to the non-robust loss (i.e., the standard 0-1 loss) we can learn with only $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$ labeled examples, even if the VC is infinite. Recall that there exists $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$, $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = \infty$ and $\mathrm{VC}(\mathcal{H}_0) = \infty$. Learning linear functions with a margin is a special case of this data-dependent assumption. Moreover, we show that this is obtained only by *improper* learning. (See Section 4.6.)

Related Work

Adversarially robust learning. The work of Montasser et al. [9] studied the setting of fully-supervised robust PAC learning. In this paper, we propose a semi-supervised method with a significant improvement in the labeled sample size. We show that the labeled and unlabeled sample complexities are controlled by different complexity measures. Adversarially robust learning has been extensively studied in several supervised learning models [e.g., 1, 3, 4, 8, 44, 55, 57, 59, 60, 81, 83, 103–111]. For semi-supervised robust learning, Ashtiani et al. [83] showed that under some assumptions, robust PAC learning is possible

with $\mathcal{O}(\mathrm{VC}(\mathcal{H}))$ labeled examples and additional unlabeled samples. Carmon et al. [67] studied a robust semi-supervised setting where the distribution is a mixture of Gaussians and the hypothesis class is linear separators.

Semi-supervised (non-robust) learning. There is substantial interest in semi-supervised (non-robust) learning, and many contemporary practical problems significantly benefit from it [e.g., 112–114]. This was formalized in theoretical frameworks. Urner et al. [115] suggested a semi-supervised learning (non-robust) framework, with an algorithmic idea that is similar to our method. Their framework consists of two steps; using labeled data to learn a classifier with small error (not necessarily a member of the target class $\mathcal{H}$), and then labeling an unlabeled input sample in order to use a fully-supervised proper learner. They investigate scenarios where the saving of labeled examples occurs. In our paper, we are interested in the robust loss function. We use labeled data in order to learn a classifier (with the 0-1 loss function) from a class with a potentially smaller complexity measure, then we label an unlabeled input sample and use a fully-supervised method using the robust loss function. The sample complexity of learning the robust loss class is controlled by a larger complexity measure. Fortunately, this affects our unlabeled sample size and not the labeled sample size as in the fully-supervised setting. Göpfert et al. [116] studied circumstances where the learning rate can be improved given unlabeled data. Darnstädt et al. [117] showed that the label complexity gap between the semi-supervised and the fully supervised setting can become arbitrarily large for concept classes of infinite VC-dimension and that this gap is bounded when a function class contains the constant zero and the constant one functions. Balcan and Blum [118, 119] introduced an augmented version of the PAC model designed for semi-supervised learning and analyzed when unlabeled data can help. The main idea is to augment the notion of learning a concept class, with a notion of compatibility between a function and the data distribution that we hope the target function will satisfy.

4.2 Preliminaries

Let $\mathcal{X}$ be the instance space, $\mathcal{Y}$ a label space, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. A perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ maps an input to a set $\mathcal{U}(x) \subseteq \mathcal{X}$. Denote the 0-1 loss of hypothesis h on (x, y) by $\ell_{0\text{-}1}(h; x, y) = \mathbb{I}[h(x) \neq y]$, and the robust loss with respect to $\mathcal{U}$ by $\ell_{\mathcal{U}}(h; x, y) = \sup_{z \in \mathcal{U}(x)} \mathbb{I}[h(z) \neq y]$. Denote the support of a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ by $\mathrm{supp}(\mathcal{D}) = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : \mathcal{D}(x, y) > 0\}$. Denote the marginal distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$ and its support by $\mathrm{supp}(\mathcal{D}_{\mathcal{X}}) = \{x \in \mathcal{X} : \mathcal{D}(x, y) > 0\}$. Define the *robust risk (error)* of a hypothesis $h \in \mathcal{H}$ with respect to distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

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

The approximation error of $\mathcal{H}$ on $\mathcal{D}$, namely, the optimal robust error achievable by a hypothesis in $\mathcal{H}$ on $\mathcal{D}$ is denoted by,

$$\mathrm{Err}(\mathcal{H};\mathcal{D},\mathcal{U}) = \inf_{h\in\mathcal{H}} \mathrm{Err}\left(h;\mathcal{D},\mathcal{U}\right).$$

We say that a distribution $\mathcal{D}$ is *robustly realizable* by a class $\mathcal{H}$ if $\mathrm{Err}(\mathcal{H};\mathcal{D},\mathcal{U}) = 0$.

Define the *empirical robust error* of a hypothesis $h \in \mathcal{H}$ with respect to a sequence $S \in (\mathcal{X}\times\mathcal{Y})^*$,

$$\widehat{\mathrm{Err}}\left(h;S,\mathcal{U}\right) = \frac{1}{|S|}\sum_{(x,y)\in S} \ell_{\mathcal{U}}(h;x,y) = \frac{1}{|S|}\sum_{(x,y)\in S}\left[\sup_{z\in\mathcal{U}(x)} \mathbb{I}\left[h(z)\neq y\right]\right].$$

The *robust empirical risk minimizer* learning algorithm $\mathrm{RERM} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{H}$ for a class $\mathcal{H}$ on a sequence S is defined by

$$\mathrm{RERM}_{\mathcal{H}}(S) \in \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{\mathrm{Err}}\left(h;S,\mathcal{U}\right).$$

When the perturbation function is the identity, $\mathcal{U}(x) = \{x\}$, we recover the standard notions. The *risk* of a hypothesis $h \in \mathcal{H}$ with respect to distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ is defined by $\mathrm{Err}\left(h;\mathcal{D}\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_{0\text{-}1}(h;x,y)\right] = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\left[h(x)\neq y\right]\right]$, and the *empirical risk* of a hypothesis $h \in \mathcal{H}$ with respect to a sequence $S \in (\mathcal{X}\times\mathcal{Y})^*$ is defined by $\widehat{\mathrm{Err}}\left(h;S\right) = \frac{1}{|S|}\sum_{(x,y)\in S}\ell_{0\text{-}1}(h;x,y) = \frac{1}{|S|}\sum_{(x,y)\in S}\left[\mathbb{I}\left[h(x)\neq y\right]\right]$. The *empirical risk minimizer* learning algorithm $\mathrm{ERM} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{H}$ for a class $\mathcal{H}$ on a sequence S is defined by $\mathrm{ERM}_{\mathcal{H}}(S) \in \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\mathrm{Err}}\left(h;S\right)$.

A learning algorithm $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ for a class $\mathcal{H}$ is called *proper* if it always outputs a hypothesis in $\mathcal{H}$, otherwise it is called *improper*.

Realizable robust PAC **learning.** We define the supervised and semi-supervised settings.

Definition 4.3 (Realizable robust PAC learnability) *For any $\epsilon,\delta \in (0,1)$, the sample complexity of realizable robust (ϵ,δ)-PAC learning for a class $\mathcal{H}$, with respect to perturbation function $\mathcal{U}$, denoted by $\Lambda_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U})$, is the smallest integer m for which there exists a learning algorithm $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ robustly realizable by $\mathcal{H}$, namely $\mathrm{Err}\left(\mathcal{H};D,\mathcal{U}\right) = 0$, for a random sample $S \sim \mathcal{D}^m$, it holds that*

$$\mathbb{P}\left(\mathrm{Err}\left(\mathcal{A}(S);D,\mathcal{U}\right) \leq \epsilon\right) > 1-\delta.$$

If no such m exists, define $\Lambda_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U}) = \infty$, and $\mathcal{H}$ is not robustly (ϵ,δ)-PAC learnable with respect to $\mathcal{U}$.

For the standard (non-robust) learning with the 0-1 loss function, we omit the dependence on $\mathcal{U}$ and denote the sample complexity of class $\mathcal{H}$ by $\Lambda_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H})$.

Definition 4.4 (Realizable semi-supervised robust PAC learnability) *A hypothesis class $\mathcal{H}$ is semi-supervised realizable robust (ϵ,δ)-PAC learnable, with respect to perturbation function $\mathcal{U}$, if for any $\epsilon,\delta\in(0,1)$, there exists $m_u, m_l \in \mathbb{N}\cup\{0\}$, and a learning algorithm $\mathcal{A}:(\mathcal{X}\times\mathcal{Y})^*\cup(\mathcal{X})^*\to\mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ robustly realizable by $\mathcal{H}$, namely $\operatorname{Err}(\mathcal{H};D,\mathcal{U})=0$, for random samples $S^l\sim\mathcal{D}^{m_l}$ and $S^u_{\mathcal{X}}\sim\mathcal{D}^{m_u}_{\mathcal{X}}$, it holds that*

$$\mathbb{P}\left(\operatorname{Err}\left(\mathcal{A}(S^l,S^u_{\mathcal{X}});D,\mathcal{U}\right)\leq\epsilon\right)>1-\delta.$$

The sample complexity $\mathcal{M}_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U})$ includes all such pairs (m_u,m_l). If no such (m_u,m_l) exist, then $\mathcal{M}_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U})=\emptyset$.

Agnostic robust PAC **learning.** In this case we have $\operatorname{Err}(\mathcal{H};\mathcal{D},\mathcal{U})>0$, and we would like to compete with the optimal $h\in\mathcal{H}$. We add a parameter to the sample complexity, denoted by η, which is the optimal robust error of a hypothesis in $\mathcal{H}$, namely $\eta=\operatorname{Err}(\mathcal{H};\mathcal{D},\mathcal{U})$. We say that a function f is (α,ϵ)-optimal if $\operatorname{Err}(f;D,\mathcal{U})\leq\alpha\eta+\epsilon$.

Definition 4.5 (Agnostic robust PAC learnability) *For any $\epsilon,\delta\in(0,1)$, the sample complexity of agnostic robust (α,ϵ,δ)-PAC learning for a class $\mathcal{H}$, with respect to perturbation function $\mathcal{U}$, denoted by $\Lambda_{\mathrm{AG}}(\alpha,\epsilon,\delta,\mathcal{H},\mathcal{U},\eta)$, is the smallest integer m, for which there exists a learning algorithm $\mathcal{A}:(\mathcal{X}\times\mathcal{Y})^*\to\mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, for a random sample $S\sim\mathcal{D}^m$, it holds that*

$$\mathbb{P}\left(\operatorname{Err}\left(\mathcal{A}(S);D,\mathcal{U}\right)\leq\alpha\inf_{h\in\mathcal{H}}\operatorname{Err}(h;\mathcal{D},\mathcal{U})+\epsilon\right)>1-\delta.$$

If no such m exists, define $\Lambda_{\mathrm{AG}}(\alpha,\epsilon,\delta,\mathcal{H},\mathcal{U},\eta)=\infty$, and $\mathcal{H}$ is not robustly (α,ϵ,δ)-PAC learnable in the agnostic setting with respect to $\mathcal{U}$. Note that for $\alpha=1$ we recover the standard agnostic definition, our notation allows for a more relaxed approximation.

Analogously, we define the semi-supervised case.

Definition 4.6 (Agnostic semi-supervised robust PAC learnability) *A hypothesis class $\mathcal{H}$ is semi-supervised agnostically robust (α,ϵ,δ)-PAC learnable, with respect to perturbation function $\mathcal{U}$, if for any $\epsilon,\delta\in(0,1)$, there exists $m_u, m_l \in \mathbb{N}\cup\{0\}$, and a learning algorithm $\mathcal{A}:(\mathcal{X}\times\mathcal{Y})^*\cup(\mathcal{X})^*\to\mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, for random samples $S^l\sim\mathcal{D}^{m_l}$ and $S^u_{\mathcal{X}}\sim\mathcal{D}^{m_u}_{\mathcal{X}}$, it holds that*

$$\mathbb{P}\left(\operatorname{Err}\left(\mathcal{A}(S^l,S^u_{\mathcal{X}});\mathcal{D},\mathcal{U}\right)\leq\alpha\inf_{h\in\mathcal{H}}\operatorname{Err}(h;\mathcal{D},\mathcal{U})+\epsilon\right)>1-\delta.$$

The sample complexity $\mathcal{M}_{\mathrm{AG}}(\alpha,\epsilon,\delta,\mathcal{H},\mathcal{U},\eta)$ includes all such pairs (m_u,m_l). If no such (m_u,m_l)

exist, then $\mathcal{M}_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta) = \emptyset$.

Partial concept classes [11]. Let a partial concept class $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$. For $h \in \mathcal{H}$ and input x such that $h(x) = \star$, we say that h is undefined on x. The support of a partial hypothesis $h : \mathcal{X} \to \{0, 1, \star\}$ is the preimage of $\{0, 1\}$, formally, $h^{-1}(\{0, 1\}) = \{x \in \mathcal{X} : h(x) \neq \star\}$. The main motivation for introducing partial concept classes is that data-dependent assumptions can be modeled in a natural way that extends the classic theory of total concepts. The VC dimension of a partial class $\mathcal{H}$ is defined as the maximum size of a shattered set $S \subseteq \mathcal{X}$, where S is shattered by $\mathcal{H}$ if the projection of $\mathcal{H}$ on S contains all possible binary patterns, $\{0, 1\}^S \subseteq \mathcal{H}|_S$. The VC-dimension also characterizes verbatim the PAC learnability of partial concept classes, even though uniform convergence does not hold in this setting.

We use the notation $\tilde{\mathcal{O}}(\cdot)$ for omitting poly-logarithmic factors of $\mathrm{VC}, \mathrm{VC}^*, \mathrm{VC}_{\mathcal{U}}, \mathrm{RS}_{\mathcal{U}}, 1/\epsilon, 1/\delta$. See Section 4.7 for additional preliminaries on complexity measures, sample compression schemes, and partial concept classes.

4.3 Warm-Up: Knowing the Support of the Marginal Distribution

In this section, we provide a tight bound on the labeled sample complexity when the support of marginal distribution is fully known to the learner, under the robust realizable assumption. Studying this setting gives an intuition for the general semi-supervised model. The main idea is that as long as we know the support of the marginal distribution, $\mathrm{supp}(\mathcal{D}_{\mathcal{X}}) = \{x \in \mathcal{X} : \exists y \in \mathcal{Y}, \text{ s.t. } \mathcal{D}(x, y) > 0\}$, we can restrict our search to a subspace of functions that are robustly self-consistent, $\mathcal{H}_{\mathcal{U}\text{-cons}} \subseteq \mathcal{H}$, where

$$\mathcal{H}_{\mathcal{U}\text{-cons}} = \{h \in \mathcal{H} : \forall x \in \mathrm{supp}(\mathcal{D}_{\mathcal{X}}), \forall z, z' \in \mathcal{U}(x), h(z) = h(z')\}\,.$$

As long as the distribution is robustly realizable, i.e., $\mathrm{Err}(\mathcal{H}; \mathcal{D}, \mathcal{U}) = 0$, we are guaranteed that the target hypothesis belongs to $\mathcal{H}_{\mathcal{U}\text{-cons}}$. As a result, it suffices to learn the class $\mathcal{H}_{\mathcal{U}\text{-cons}}$ with the 0-1 loss function, in order to robustly learn the original class $\mathcal{H}$. We observe that,

$$\mathrm{VC}(\mathcal{H}_{\mathcal{U}\text{-cons}}) = \mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H}).$$

Moreover, there exits $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{VC}(\mathcal{H}_0) = \infty$ (see Proposition 4.8). Fortunately, moving from $\mathrm{VC}(\mathcal{H})$ to $\mathrm{VC}_{\mathcal{U}}(\mathcal{H})$ implies a significant sample complexity improvement. Since $\mathrm{supp}(\mathcal{D}_{\mathcal{X}})$ is known, we can now employ any algorithm for learning the hypothesis class $\mathcal{H}_{\mathcal{U}\text{-cons}}$. [2] This leads eventually to

[2] See Mohri et al. [28, Chapter 3] for standard upper and lower bounds. In order to remove the superfluous $\log \frac{1}{\epsilon}$ factor of the standard uniform convergence based upper bound, $\mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} \log \frac{1}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon}\right)$, we can use the learning algorithm and its analysis

robustly learning $\mathcal{H}$ with labeled sample complexity that scales linearly with $\mathrm{VC}_{\mathcal{U}}$ (instead of the VC). Formally,

Theorem 4.7 *For hypothesis class $\mathcal{H}$ and adversary $\mathcal{U}$, when the support of the marginal distribution $\mathcal{D}_{\mathcal{X}}$ is known, the labeled sample complexity is $\Theta\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$.*

The following Proposition demonstrates that semi-supervised robust learning requires much fewer labeled samples compared to the supervised counterpart. Recall the lower bound on the sample complexity of supervised robust learning, $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}, \mathcal{U}) = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ given by Montasser et al. [9, Theorem 10]. For completeness, we prove the following in Section 4.7.

Proposition 4.8 ([9], Proposition 9) *There exists a hypothesis class $\mathcal{H}_0$ such that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$, $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = \infty$, and $\mathrm{VC}(\mathcal{H}_0) = \infty$.*

We can now conclude the following separation result on supervised and semi-supervised label complexities.

Corollary 4.9 *The hypothesis class in Proposition 4.8 is not learnable in supervised robust learning (i.e., we need to see the entire data distribution). However, when $\mathrm{supp}(\mathcal{D}_{\mathcal{X}})$ is known, this class can be learned with $\mathcal{O}(\frac{1}{\epsilon}\log\frac{1}{\delta})$ labeled examples.*

In the next section, we prove a stronger separation in the general semi-supervised setting. The size of the labeled data required in the supervised case is lower bounded by $\mathrm{RS}_{\mathcal{U}}$, whereas in the semi-supervised case, the *labeled* sample complexity depends only on $\mathrm{VC}_{\mathcal{U}}$ and the *unlabeled* data is lower bounded by $\mathrm{RS}_{\mathcal{U}}$. Moreover, note that in Theorem 4.7, when $\mathrm{supp}(\mathcal{D}_{\mathcal{X}})$ is known, we can use any proper learner. In Section 4.4 we show that in the general semi-supervised model this is not the case, and sometimes improper learning is necessary, similar to supervised robust learning.

4.4 Near-Optimal Semi-Supervised Sample Complexity

In this section, we present our algorithm and its guarantees for the realizable setting. We also prove nearly matching lower bounds on the sample complexity. Finally, we show that improper learning is necessary in semi-supervised robust learning, similar to the supervised case.

We present a generic semi-supervised robust learner, that can be applied in both realizable and agnostic settings. The algorithm uses the following two subroutines. The first one is any algorithm for learning partial concept classes, which controls our *labeled* sample size. (In Section 4.7 we discuss in detail the algorithm suggested by Alon et al. [11].) The second subroutine is any algorithm for the agnostic adversarially robust

from Hanneke [29] that applies for any $\mathcal{H}$ and $\mathcal{D}$, or some other algorithms that are doing so while restricting the hypothesis class or the data distribution [e.g., 120–127].

supervised learning, which controls our *unlabeled* sample size. (In Section 4.7 we discuss in detail the algorithm suggested by Montasser et al. [9].) Any progress on one of these problems directly improves the guarantees of our algorithm. We use the following definition that explains how to convert a total concept class into a partial one, in a way that preserves the idea of the robust loss function.

Definition 4.10 *Let a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ and a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$. For any $h \in \mathcal{H}$, we define a corresponding partial concept $h^\star : \mathcal{X} \to \{0,1,\star\}$, and denote this mapping by $\varphi(h) = h^\star$. For $x \in \mathcal{X}$, whenever h is not consistent on the entire set $\mathcal{U}(x)$, i.e., $\exists z, z' \in \mathcal{U}(x), h(z) \neq h(z')$, define $h^\star(x) = \star$. Otherwise, h is robustly self-consistent on x, i.e., $\forall z, z' \in \mathcal{U}(x), h(z) = h(z')$ and h remains unchanged, $h^\star(x) = h(x)$. The corresponding partial concept class is defined by $\mathcal{H}^\star_{\mathcal{U}} = \{h^\star : \varphi(h) = h^\star,\ \forall h \in \mathcal{H}\}$.*

The main motivation for the above definition is the following. Fix a hypothesis h. For any point x, as defined above, the adversary can force a mistake on h, regardless of the prediction of h. We would like to mark such points as *mistake*. We do this by defining a partial concept $h^\star$ and setting $h^\star(x) = \star$, which, for partial concepts, implies a mistake. The benefit of this preprocessing is that we reduce the complexity of the hypothesis class from VC to $\mathrm{VC}_{\mathcal{U}}$, which potentially can reduce the labeled sample complexity.

We are now ready to describe the algorithm.

Algorithm 2 Generic Adversarially-Robust Semi-Supervised (GRASS) learner

Input: Labeled data set $S^l \sim \mathcal{D}^{m_l}$, unlabeled data set $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$, hypothesis class $\mathcal{H}$, perturbation function $\mathcal{U}$, parameters ϵ, δ.
Algorithms used: PAC learner $\mathcal{A}$ for partial concept classes, agnostic adversarially robust supervised PAC learner $\mathcal{B}$.

1. Given the class $\mathcal{H}$, construct the hypothesis class $\mathcal{H}^\star_{\mathcal{U}}$ using Definition 4.10.
2. Execute the learning algorithm for partial concepts $\mathcal{A}$ on $\mathcal{H}^\star_{\mathcal{U}}$ and sample S^l, with the 0-1 loss and parameters $\frac{\epsilon}{3}, \frac{\delta}{2}$. Denote the resulting hypothesis h_1.
3. Label the unlabeled data set $S^u_{\mathcal{X}}$ with h_1, denote the labeled sample by S^u. (On points where h_1 predicts $\star$, we can arbitrarily choose a label of 0 or 1.)
4. Execute the agnostic adversarially robust supervised PAC learner $\mathcal{B}$ on S^u with parameters $\frac{\epsilon}{3}, \frac{\delta}{2}$. Denote the resulting hypothesis h_2.

Output: h_2.

Algorithm motivation. The main idea behind the algorithm is the following. Given the class $\mathcal{H}^\star_{\mathcal{U}}$, we would like to find a hypothesis $h_1 \in \mathcal{H}^\star_{\mathcal{U}}$ which has a small error, whose existence follows from our realizability assumption. The required sample size scales with $\mathrm{VC}_{\mathcal{U}}$, which is the complexity of $\mathcal{H}^\star_{\mathcal{U}}$, rather than VC. This is where we make a significant gain in the labeled sample complexity. Note that h_1 does not guarantee a small robust error, although it does guarantee a small non-robust error. We utilize an additional unlabeled sample for this task, which we label using h_1. If we would simply minimize the non-robust error

on this sample we would simply get back h_1. The main insight is that we would like to minimize the robust error over this sample, which will result in hypothesis h_2. We now need to bound the robust error of h_2. The optimal function h_{opt} has only a slightly increased robust error on this sample, namely, at most on the sample points where it disagrees with h_1. Note that h_1 might have a large robust error due to the perturbation $\mathcal{U}$. However, a robust supervised PAC learner would return a hypothesis h_2 which has robust error similar to h_{opt}, which is at most ϵ.

Algorithm outline and guarantees. In the first step, we convert $\mathcal{H}$ to $\mathcal{H}^\star_{\mathcal{U}}$. Then we employ a learning algorithm $\mathcal{A}$ for partial concepts on $\mathcal{H}^\star_{\mathcal{U}}$ with a labeled sample $S^l \sim \mathcal{D}^{m_l}$. The output of the algorithm is a function h_1 with $\epsilon/3$ on the 0-1 error. Crucially, we needed for this step $|S^l| = \tilde{\mathcal{O}}(\text{VC}_{\mathcal{U}}(\mathcal{H})/\epsilon)$ labeled examples for learning the partial concept $\mathcal{H}^\star_{\mathcal{U}}$, since $\text{VC}(\mathcal{H}^\star_{\mathcal{U}}) = \text{VC}_{\mathcal{U}}(\mathcal{H})$. So our labeled sample size is controlled by the sample complexity for learning partial concepts with the 0-1 loss. In step 3, we label an independent unlabeled sample $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$ with h_1, denote his labeled sample by S^u. Define a distribution $\tilde{\mathcal{D}}$ over $\mathcal{X} \times \mathcal{Y}$ by $\tilde{\mathcal{D}}(x, h_1(x)) = \mathcal{D}_{\mathcal{X}}(x)$, and so S^u is an i.i.d. sample from $\tilde{\mathcal{D}}$. We argue that the robust error of $\mathcal{H}$ with respect to $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$, i.e., $\text{Err}(\mathcal{H}; \tilde{\mathcal{D}}, \mathcal{U}) = \frac{\epsilon}{3}$. Indeed, the function with zero robust error on $\mathcal{D}$, $h_{\text{opt}} \in \text{argmin}_{h \in \mathcal{H}} \text{Err}(h; \mathcal{D}, \mathcal{U})$ has a robust error of at most $\frac{\epsilon}{3}$ on $\tilde{\mathcal{D}}$. Finally, we employ an agnostic adversarially robust supervised PAC learner $\mathcal{B}$ for the class $\mathcal{H}$ on $S^u \sim \tilde{\mathcal{D}}^{m_u}$, that should be of size of the sample complexity of agnostically robust learn $\mathcal{H}$ with respect to $\mathcal{U}$, when the optimal robust error of hypothesis from $\mathcal{H}$ on $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$. Moreover, the total variation distance between $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$. We are guaranteed that the resulting hypothesis h_2 has a robust error of at most $\frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$ on $\mathcal{D}$. We conclude that a size of $|S^u_{\mathcal{X}}| = m_u = \Lambda_{\text{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right)$ unlabeled samples suffices, this completes the proof for Theorem 4.11. For a specific instantiation of such algorithm ([9]), we deduce the sample complexity in Theorem 4.13. A simple analysis of the latter yields a dependence of ϵ^2 for the unlabeled sample size. However, by applying a suitable data-dependent generalization bound, we reduce this dependence to ϵ. (Full proofs appear in Section 4.7).

We now formally present the sample complexity of the generic semi-supervised learner for the robust realizable setting. First, in the case of using a generic agnostic robust supervised learner as a subroutine (step 4 in the algorithm). Then we deduce the sample complexity of a specific instantiation of such an algorithm.

Theorem 4.11 *For any hypothesis class $\mathcal{H}$ and adversary $\mathcal{U}$, algorithm* GRASS *(ϵ, δ)-*PAC *learns $\mathcal{H}$ with respect to the robust loss function, in the realizable robust case, with samples of size*

$$m_l = \mathcal{O}\left(\frac{\text{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} \log^2 \frac{\text{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon}\right) \ , \ m_u = \Lambda_{\text{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right),$$

where $\Lambda_{\mathrm{AG}}(\alpha,\epsilon,\delta,\mathcal{H},\mathcal{U},\eta)$ *is the sample complexity of adversarially-robust agnostic supervised* (α,ϵ,δ)-PAC *learning, such that* η *is the error of the optimal hypothesis in* $\mathcal{H}$*, i.e.,* $\eta = \mathrm{Err}(\mathcal{H};\mathcal{D},\mathcal{U})$.

Remark 4.12 *Note that if we simply invoke a* PAC *learner (for total concept classes) on* $\mathcal{H}$*, with the 0-1 loss, instead of steps 1 and 2 in the algorithm, we would get a labeled sample complexity of roughly* $\mathcal{O}(\mathrm{VC}(\mathcal{H}))$*. This is already an exponential improvement upon previous results that require roughly* $\mathcal{O}\left(2^{\mathrm{VC}(\mathcal{H})}\right)$ *labeled samples. The purpose of using partial concept classes is to further reduce the labeled sample complexity to* $\mathcal{O}(\mathrm{VC}_{\mathcal{U}}(\mathcal{H}))$.

The following result follows by using the agnostic supervised robust learner suggested by Montasser et al. [9]. A simple analysis of the latter yields a dependence of ϵ^2 for the unlabeled sample size. However, by applying a suitable data-dependent generalization bound, we reduce this dependence to ϵ.

Theorem 4.13 *For any hypothesis class* $\mathcal{H}$ *and adversary* $\mathcal{U}$*, Algorithm* GRASS (ϵ,δ)-PAC *learns* $\mathcal{H}$ *with respect to the robust loss function, in the realizable robust case, with samples of size*

$$m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right),\ m_u = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right).$$

We present nearly matching lower bounds for the realizable setting. The following Corollary stems from Theorem 4.7 and Montasser et al. [9, Theorem 10].

Corollary 4.14 *For any* $\epsilon,\delta\in(0,1)$*, the sample complexity of realizable robust* (ϵ,δ)-PAC *learning for a class* $\mathcal{H}$*, with respect to perturbation function* $\mathcal{U}$ *is*

$$m_l = \Omega\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right),\ m_u = \infty, \quad \textit{or} \quad m_l + m_u = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right).$$

Proper vs. improper. In Section 4.3, we have seen that when the support of the marginal distribution $\mathcal{D}_{\mathcal{X}}$ is known, the labeled sample complexity is $\Theta\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$. This was obtained by a proper learner: keep the robustly self-consistent hypotheses, $\mathcal{H}_{\mathcal{U}\text{-cons}} \subseteq \mathcal{H}$, and then use ERM on this class. The case when $\mathcal{D}_{\mathcal{X}}$ is unknown is different. We know that there exists a perturbation function $\mathcal{U}$ and a hypothesis class $\mathcal{H}$ with finite VC-dimension that cannot be robust PAC learned with any proper learning rule [9, Lemma 3]. The same proof holds in the semi-supervised case. Note that both algorithms $\mathcal{A}$ and $\mathcal{B}$ used in Algorithm 2 are improper. (The proof appears in Section 4.7.)

Theorem 4.15 *There exists* $\mathcal{H}$ *with* $\mathrm{VC}(\mathcal{H}) = 0$ *such that for any proper learning rule* $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \cup (\mathcal{X})^* \to \mathcal{H}$*, there exists a distribution* $\mathcal{D}$ *over* $\mathcal{X}\times\mathcal{Y}$ *that is robustly realizable by* $\mathcal{H}$*, i.e.,* $\mathrm{Err}(\mathcal{H};\mathcal{D},\mathcal{U}) = 0$*. It holds that* $\mathrm{Err}\left(\mathcal{A}(S^l, S^u_{\mathcal{X}});D,\mathcal{U}\right) > \frac{1}{8}$ *with probability at least* $\frac{1}{7}$ *over* $S^l \sim \mathcal{D}^{m_l}$ *and* $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}$*, where*

$m_l, m_u \in \mathbb{N} \cup \{0\}$ *is the size of the labeled and unlabeled samples respectively. Moreover, when the marginal distribution* $\mathcal{D}_{\mathcal{X}}$ *is known, there exists a proper learning rule for any* $\mathcal{H}$.

4.5 Agnostic Robust Learning

In this section, we prove the guarantees of Algorithm 2 in the more challenging agnostic robust setting. We then prove lower bounds on the sample complexity which exhibits that it is inherently different from the realizable case.

We follow the same steps as in the proof of the realizable case, with the following important difference. In the first two steps of the algorithm, we learn a partial concept class with respect to the 0-1 loss and obtain a hypothesis with an error of $\eta + \epsilon/3$ (η is the optimal robust error of a hypothesis in $\mathcal{H}$ and not 0). This leads eventually to the error of $3\eta + \epsilon$ for learning with respect to the robust loss.

We then present two negative results. In Theorem 4.17 we show that for obtaining error $\eta + \epsilon$ there is a lower bound of $\Omega(\mathrm{RS}_{\mathcal{U}})$ labeled examples, this result coincides with the lower bound of supervised robust learning. In Theorem 4.18, we show that for any $\gamma > 0$, there exists a hypothesis class, such that having access only to $\mathcal{O}(\mathrm{VC}_{\mathcal{U}})$ labeled examples, leads to an error $(\frac{3}{2} - \gamma)\eta + \epsilon$. (All proofs for this section are in Section 4.7.)

We start with the upper bounds. First, we analyze the case of using a generic agnostic robust learner, then we deduce the sample complexity of a specific instantiation of such algorithm.

Theorem 4.16 *For any hypothesis class* $\mathcal{H}$ *and adversary* $\mathcal{U}$*, Algorithm* GRASS $(3, \epsilon, \delta)$-PAC *learns* $\mathcal{H}$ *with respect to the robust loss function, in the agnostic robust case, with samples of size*

$$m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right), \quad m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, 2\eta + \frac{\epsilon}{3}\right),$$

where $\Lambda_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta)$ *is the sample complexity of adversarially-robust agnostic supervised learning, such that* η *is error of the optimal hypothesis in* $\mathcal{H}$*, namely* $\eta = \mathrm{Err}(\mathcal{H}; \mathcal{D}, \mathcal{U})$.

By using the agnostic supervised robust learner suggested by Montasser et al. [9], we have the following upper bound on the unlabeled sample size, $m_u = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$.

We now present two negative results.

Theorem 4.17 *For any* $\epsilon, \delta \in (0, 1)$*, the sample complexity of agnostic robust* $(1, \epsilon, \delta)$-PAC *learning for a class* $\mathcal{H}$*, with respect to perturbation function* $\mathcal{U}$ *is (even if* $\mathcal{D}_{\mathcal{X}}$ *is known),*

$$m_l = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right), \; m_u = \infty.$$

Theorem 4.18 *For any $\gamma > 0$, there exists a hypothesis class $\mathcal{H}$ and adversary $\mathcal{U}$, such that the sample complexity for $(\frac{3}{2} - \gamma, \epsilon, \delta)$-*PAC *learn $\mathcal{H}$ is*

$$m_l = \Omega\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right) \ , \ m_u = \infty.$$

Open question. What is the optimal error rate in the agnostic setting when using only $\mathcal{O}(\mathrm{VC}_{\mathcal{U}})$ labeled examples?

4.6 Learning with the 0-1 Loss Assuming Robust Realizability

In this section, we learn with respect to the 0-1 loss, under robust realizability assumption. A Distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is robustly realizable by $\mathcal{H}$ given a perturbation function $\mathcal{U}$ if there is $h \in \mathcal{H}$ such that not only h classifies all points in $\mathcal{D}$ correctly, it also does so with respect to the robust loss function, that is, $\mathrm{Err}(\mathcal{H}; \mathcal{D}, \mathcal{U}) = 0$. Note that our guarantees, only in this section, are with respect to the non-robust error. The formal definition is in Section 4.7. A simple example of this model is the following. Let $\mathcal{H}$ be linear separators on $\mathcal{X}$ the unit ball in $\mathbb{R}^d$, and $\mathcal{U}$ as ℓ_2 balls of radius γ, the robustly realizable distributions are separable with margin γ, where $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = \frac{1}{\gamma^2}$ but $\mathrm{VC}(\mathcal{H}) = d + 1$ can be arbitrarily larger. Moreover, we have the following example. (All proofs are in appendix Section 4.7.)

Proposition 4.19 *For any $m \in \mathbb{N}$, there exist a hypothesis class $\mathcal{H}_m$ and distribution $\mathcal{D}$, such that $\mathcal{D}$ is robustly realizable by $\mathcal{H}_m$, $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_m) = 1$, and $\mathrm{VC}(\mathcal{H}_m) = 2m$.*

Standard VC theory does not ensure learning in this case. In this section, we explain how we can learn in such a scenario with a small sample complexity (scales linearly in $\mathrm{VC}_{\mathcal{U}}$). Moreover, we show that it cannot be achieved via proper learners.

Theorem 4.20 *The sample complexity for learning a hypothesis class $\mathcal{H}$ with respect to the 0-1 loss, for any distribution $\mathcal{D}$ that is robustly realizable by $\mathcal{H}$, namely $\mathrm{Err}\left(\mathcal{H}; D, \mathcal{U}\right) = 0$,*

$$\mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right), \Omega\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right).$$

This Theorem was an intermediate step in the proof of Theorem 4.11, and the sample complexity is the same as Theorem 4.22, $\mathcal{O}\left(\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H})\right)$. We show that there exists a robust ERM that fails in this setting (Proposition 4.25 in Section 4.7). Then, we claim that every proper learner fails.

Theorem 4.21 *There exists $\mathcal{H}$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 1$, such that for any proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is robustly realizable by $\mathcal{H}$, i.e., $\mathrm{Err}\left(\mathcal{H}; \mathcal{D}, \mathcal{U}\right) = 0$, and it*

holds that $\mathrm{Err}\left(\mathcal{A}(S); D\right) > \frac{1}{8}$ *with probability at least* $\frac{1}{7}$ *over* $S \sim \mathcal{D}^m$.

4.7 Deferred Preliminaries and Proofs

Additional Preliminaries for Section 6.2

Complexity measures. The capacity measures, $\mathrm{VC}_{\mathcal{U}}$, $\mathrm{RS}_{\mathcal{U}}$ and VC, play an important role in our results. See Definitions 4.1 and 4.2 for the $\mathrm{VC}_{\mathcal{U}}$ and $\mathrm{RS}_{\mathcal{U}}$ dimensions. It holds that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{RS}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$, in Proposition 4.8 we demonstrate an arbitrary gap between $\mathrm{VC}_{\mathcal{U}}$ and $\mathrm{RS}_{\mathcal{U}}$, the key parameters controlling the sample complexity of robust learnability.

Partial concept classes - [11]. Let a partial concept class $\mathcal{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$. For $h \in \mathcal{H}$ and input x such that $h(x) = \star$, we say that h is undefined on x. The support of a partial hypothesis $h : \mathcal{X} \to \{0,1,\star\}$ is the preimage of $\{0,1\}$, formally, $h^{-1}(\{0,1\}) = \{x \in \mathcal{X} : h(x) \neq \star\}$. The main motivation for introducing partial concept classes is that data-dependent assumptions can be modeled in a natural way that extends the classic theory of total concepts.

The VC-dimension of a partial class $\mathcal{H}$ is defined as the maximum size of a shattered set $S \subseteq \mathcal{X}$, where S is shattered by $\mathcal{H}$ if the projection of $\mathcal{H}$ on S contains all possible binary patterns, $\{0,1\}^S \subseteq \mathcal{H}|_S$. The VC dimension also characterizes verbatim the PAC learnability of partial concept classes. However, the uniform convergence argument does not hold, and the ERM principle does not ensure learning. The proof hinges on a combination of sample compression scheme and a variant of the *one-Inclusion-Graph* algorithm [128]. In Section 4.4 we elaborate on the sample complexity of partial concept classes, and in Section 4.7 we elaborate on the learning algorithms. The definitions of realizability and agnostic learning in the partial concepts sense generalize the classic definitions for total concept classes. See [11, Section 2 and Appendix C] for more details.

Proofs for Section 4.3

Proof (of Proposition 4.8). We overview the construction by Montasser et al. [9], which exemplifies an arbitrarily large gap between $\mathrm{VC}_{\mathcal{U}}$ and $\mathrm{RS}_{\mathcal{U}}$. In this example $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 0$, $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}) = \infty$, and $\mathrm{VC}(\mathcal{H}) = \infty$.

Define the Euclidean ball of radius r perturbation function $\mathcal{U}(x) = B_r(x)$. Consider infinite sequences $(x_n)_{n\in\mathbb{N}}$ and $(z_n)_{n\in\mathbb{N}}$ of points such that $\forall i \neq j,\ \mathcal{U}(x_i) \cap \mathcal{U}(x_j) = U(x_i) \cap \mathcal{U}(z_j) = \mathcal{U}(x_j) \cap \mathcal{U}(z_i) = \emptyset$, and $\forall i,\ \left|\mathcal{U}(x_i) \cap \mathcal{U}(z_i)\right| = 1$.

For a bit string $b \in \{0,1\}^{\mathbb{N}}$, define a hypothesis $h_b : \{\mathcal{U}(x_i) \cup \mathcal{U}(z_i)\}_{i\in\mathbb{N}} \to \{0,1\}$ as follows.

$$h_b = \begin{cases} h_b\Big(\mathcal{U}(x_i)\Big) = 1 \ \wedge \ h_b\Big(\mathcal{U}(z_i) \setminus \mathcal{U}(x_i)\Big) = -1, \quad b_i = 0 \\ h_b\Big(\mathcal{U}(z_i)\Big) = 1 \ \wedge \ h_b\Big(\mathcal{U}(x_i) \setminus \mathcal{U}(z_i)\Big) = -1, \quad b_i = 1. \end{cases}$$

Define the hypothesis class $\mathcal{H} = \left\{h_b : b \in \{0,1\}^{\mathbb{N}}\right\}$. It holds that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 0$ and $\mathrm{RS}_{\mathcal{U}} = \infty$. □

Proofs for Section 4.4

Before proceeding to the proof, we present the following result on learning partial concept classes. Recall the definition of VC is in the context of partial concepts (see Section 4.7).

Theorem 4.22 ([11], Theorem 34) *Any partial concept class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) < \infty$ is* PAC *learnable in the realizable setting with sample complexity,*

- $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\min\left\{\frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\log\frac{1}{\delta}, \frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\log^2\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\right) + \frac{1}{\epsilon}\log\frac{1}{\delta}\right\}\right)$
- $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \Omega\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$.

Proof (of Theorem 4.11). At first, we convert the hypothesis class $\mathcal{H}$ to $\mathcal{H}^\star_{\mathcal{U}}$ as described in Definition 4.10. Then, we employ the learning algorithm $\mathcal{A}$ for partial concepts on the partial concept class $\mathcal{H}^\star_{\mathcal{U}}$ and S^l, denote the resulting hypothesis by h_1. Note that we reduced the complexity of the class, since $\mathrm{VC}(\mathcal{H}^\star_{\mathcal{U}}) = \mathrm{VC}_{\mathcal{U}}(\mathcal{H})$. Theorem 4.22 implies that whenever $m_l = |S^l| \geq \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$, the hypothesis h_1 has a non-robust error at most $\frac{\epsilon}{3}$ with probability $1 - \frac{\delta}{2}$, with respect to the 0-1 loss. Note that there exists $h \in \mathcal{H}$ that classifies correctly any point in $\mathcal{D}$ with respect to the robust loss function. So when we convert $\mathcal{H}$ to $\mathcal{H}^\star_{\mathcal{U}}$, the "partial version" of h still classifies correctly any point in S^l, and does not return any $\star$, which always counts as a mistake. Algorithm $\mathcal{A}$ guarantees to return a hypothesis that is ϵ-optimal with respect to the 0-1 loss, with high probability. Observe that after these two steps, we obtain the following intermediate result. Whenever a distribution $\mathcal{D}$ is robustly realizable by a hypothesis class $\mathcal{H}$, i.e., $\mathrm{Err}(\mathcal{H}; \mathcal{D}, \mathcal{U}) = 0$, we have an algorithm that learns this class with respect to the 0-1 loss, with sample complexity of

$$\Upsilon(\epsilon, \delta, \mathcal{H}, \mathcal{U}) = \mathcal{O}\left(\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H})\right) = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right). \tag{4.1}$$

The sample complexity of this model is defined formally in Definition 4.24. In Section 4.6 we present more results for this model.

In the third step, we label an independent unlabeled sample $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$ with h_1, denote this labeled

sample by S^u. Define a distribution $\tilde{\mathcal{D}}$ over $\mathcal{X} \times \mathcal{Y}$ by

$$\tilde{\mathcal{D}}(x, h_1(x)) = \mathcal{D}_{\mathcal{X}}(x),$$

and so S^u is an i.i.d. sample from $\tilde{\mathcal{D}}$. We argue that the robust error of $\mathcal{H}$ with respect to $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$, i.e., $\mathrm{Err}(\mathcal{H}; \tilde{\mathcal{D}}, \mathcal{U}) \leq \frac{\epsilon}{3}$. Indeed, we show that $h_{\text{opt}} \in \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{Err}(h; \mathcal{D}, \mathcal{U})$ has a robust error of at most $\frac{\epsilon}{3}$ on $\tilde{\mathcal{D}}$. Note that,

$$\mathrm{Err}(\mathcal{H}; \tilde{\mathcal{D}}, \mathcal{U}) \leq \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))\right] = \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}} \left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)\right]. \tag{4.2}$$

Observe that the following holds for any (x, y),

$$\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x)) \leq \ell_{\mathcal{U}}(h_{\text{opt}}; x, y) + \ell_{0\text{-}1}(h_1; x, y). \tag{4.3}$$

Indeed, the right-hand side is 0, whenever h_1 classifies (x, y) correctly, and h_{opt} robustly classifies (x, y) correctly, which implies that the left-hand side is 0 as well.

By taking the expectation on Eq. (4.3) we have,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{0\text{-}1}(h_1; x, y)]. \tag{4.4}$$

By combining it together, we obtain

$$\begin{aligned}
\mathrm{Err}(\mathcal{H}; \tilde{\mathcal{D}}, \mathcal{U}) &\leq \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}} \left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{0\text{-}1}(h_1; x, y)] \\
&\leq \frac{\epsilon}{3}
\end{aligned}$$

where (i) follows from Eq. (4.2) and (ii) follows from Eq. (4.4).

Finally, we employ an agnostic adversarially robust supervised PAC learner $\mathcal{B}$ for the class $\mathcal{H}$ on $S^u \sim \tilde{\mathcal{D}}^{m_u}$, that should be of size of the sample complexity of agnostically robust learn $\mathcal{H}$ with respect to $\mathcal{U}$, when the optimal robust error of hypothesis from $\mathcal{H}$ on $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$. We are guaranteed that the resulting hypothesis h_2 has a robust error of at most $\frac{\epsilon}{3} + \frac{\epsilon}{3} = \frac{2\epsilon}{3}$ on $\tilde{\mathcal{D}}$, with probability $1 - \frac{\delta}{2}$. We observe that the total variation distance between $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$, and as a result, h_2 has a robust error of at most $\frac{2\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$ on $\mathcal{D}$, with probability $1 - \delta$.

We conclude that a size of $|S^u_{\mathcal{X}}| = m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right)$ unlabeled samples suffices, in addition to $m_l = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ labeled samples which are required in the first 2 steps. □

We now prove Theorem 4.13 using a data-dependent compression-based generalization bound Lemma 2.19:

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

This bound includes the empirical error factor, and as soon as we call the compression-based learner on a sample that is "nearly" realizable (Step 4 in the algorithm), we can improve the sample complexity of the agnostic robust supervised learner, such that the dependence on ϵ^2 is reduced to ϵ, for the unlabeled sample size.

Proof (of Theorem 4.13). Montasser et al. [9, Theorem 6] introduced an agnostic robust supervised learner that requires the following labeled sample size,

$$\Lambda_{\mathrm{AG}}\left(1, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta\right) = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right).$$

Their argument for generalization is based on classic compression generalization bound by Graepel et al. [41], adapted to the robust loss. See Montasser et al. [9, Lemma 11].

We show that in our use case, we can deduce a stronger bound. We employ the agnostic learner on a distribution that is "close" to realizable, the error of the optimal $h \in \mathcal{H}$ is at most $\eta = \frac{\epsilon}{3}$, and so we need $\Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right)$ unlabeled examples. As a result, we obtain an improved bound by using a data-dependant generalization bound described in Lemma 2.19.

This improves the unlabeled sample size (denoted by m_u) and reduces its dependence on ϵ^2 to ϵ. Overall we obtain a sample complexity of

$$m_u = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right), \quad m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right). \qquad \square$$

Proof (of Theorem 4.15). This proof is identical to [9, Lemma 3], We overview the idea of the proof. If the proof is true for a labeled sample, it remains true when some of the labels are missing.

Define the following hypothesis class $\mathcal{H}_m \subseteq [0,1]^{\mathcal{X}}$. Define the instance space $\mathcal{X} = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}$ and a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, such that the perturbation sets of the instances do not intersect, that is, $\forall i, j \in [m] : \mathcal{U}(x_i) \cap \mathcal{U}(x_j)$. We can simply take the perturbations sets to be ℓ_2 unit balls, $\mathcal{U}(x) = \{z \in \mathbb{R} : \|z - x\|_2 \leq 1\}$ such that $\forall i, j \in [m] : \ \|x_i - x_j\|_2 > 2$. Now, each $h_b \in \mathcal{H}_m$ is represented by a bit string $b = \{0,1\}^m$, such that if $b_i = 1$, then there exists an adversarial example in $\mathcal{U}(x_i)$ that is unique for each h_b, and otherwise, the function is consistent on $\mathcal{U}(x_i)$.

Formally, for each $i \in [m]$ define a bijection $\psi_i : x_i \times \mathcal{H}_m \to \mathcal{U}(x_i) \setminus \{x_i\}$. Define $\mathcal{H}_m = \{h_b : b \in \{0,1\}^m\}$,

such that for any $x_i \in \mathcal{X}$, h_b is defined by

$$h_b(x_i) = \begin{cases} h_b\Big(\mathcal{U}(x_i) \setminus \psi_i(x_i, h_b)\Big) = 0 \ \wedge \ h_b\Big(\psi_i(x_i, h_b)\Big) = 1, & b_i = 1, \\ h_b\Big(\mathcal{U}(x_i)\Big) = 0, & b_i = 0. \end{cases}$$

Note that since ψ_i is a bijection and different functions with $b_i = 1$ have a different perturbation for x_i that causes a misclassification.

For a function class $\mathcal{H}$, define the robust loss class $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}} = \Big\{(x, y) \mapsto \sup_{z \in \mathcal{U}(x)} \mathbb{I}\{h(z) \neq y\} : h \in \mathcal{H}\Big\}$. It holds that $\mathrm{VC}(\mathcal{H}_m) \leq 1$ and $\mathrm{VC}\left(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}_m}\right) = m$ (see [9, Lemma 2]).

We define a function class $\tilde{\mathcal{H}}_{3m} = \big\{h_b \in \mathcal{H}_{3m} : \sum_{i=1}^{3m} b_i = m\big\}$. In words, we are keeping only functions in $\mathcal{H}_{3m}$ that are robustly correct on exactly $2m$ points. Note that the function $h_{\mathbf{0}}$ (bit string of all zeros) which is robustly correct on all $3m$ points, is not the class.

The idea is that we can construct a family of $\binom{3m}{2m}$ distributions, such that each distribution is supported on $2m$ points from $\mathcal{X} = \{x_1, \ldots, x_{3m}\}$. Now, if we have a proper learning rule, observing only m points, the algorithm has no information which are the remaining m points in the support (out of $2m$ possible points in $\mathcal{X}$). For each such a distribution there exists $h \in \tilde{\mathcal{H}}_{3m}$, with zero robust error. We can follow a standard proof of the no-free-lunch theorem [e.g., 27, Section 5], showing via the probabilistic method, that there exists a distribution on which the algorithm has a constant error, although there is an optimal function in $\tilde{\mathcal{H}}_{3m}$. See [9, Lemma 3] for the full proof. □

Proofs for Section 4.5

Before proceeding to the proof, we present the following result on agnostic learning partial concept classes. Recall the definition of VC is in the context of partial concepts (see Section 4.7).

Theorem 4.23 ([11], Theorem 41) *Any partial concept class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) < \infty$ is agnostically* PAC *learnable with sample complexity,*

- $\Lambda_{\mathrm{AG}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon^2} \log^2\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon^2}\right) + \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$.
- $\Lambda_{\mathrm{AG}}(\epsilon, \delta, \mathcal{H}) = \Omega\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$.

Proof (of Theorem 4.16). We follow the same steps as in the proof of the realizable case, with the following difference. In the first two steps of the algorithm, we learn with respect to the 0-1 loss, with an error of η (the optimal robust error of a hypothesis in $\mathcal{H}$) and not 0, which leads eventually to an approximation of 3η for learning with the robust loss.

At first, we convert the class $\mathcal{H}$ into $\mathcal{H}^{\star}_{\mathcal{U}}$, on which we employ the learning algorithm $\mathcal{A}$ for partial concepts with the sample S^l. Theorem 4.23 implies that whenever $m_l = |S^l| \geq \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$, the resulting hypothesis h_1 returned by algorithm $\mathcal{A}$ has a non-robust error at most $\eta + \frac{\epsilon}{3}$ with probability $1 - \frac{\delta}{2}$, with respect to the 0-1 loss, where $\eta = \mathrm{Err}(\mathcal{H}; \mathcal{D}, \mathcal{U})$. Note that there exists $h \in \mathcal{H}$ with robust error of η on $\mathcal{D}$. The "partial version" of h has an error of η on $\mathcal{D}$ with respect to the 0-1 loss. As a result, algorithm $\mathcal{A}$ guarantees to return a hypothesis that is ϵ-optimal with respect to the 0-1 loss, with high probability.

We label an independent unlabeled sample $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$ with h_1, denote this labeled sample by S^u. Similarly to the realizable case, define a distribution $\tilde{\mathcal{D}}$ over $\mathcal{X} \times \mathcal{Y}$ by

$$\tilde{\mathcal{D}}(x, h_1(x)) = \mathcal{D}_{\mathcal{X}}(x),$$

and so S^u is an i.i.d. sample from $\tilde{\mathcal{D}}$. We argue that the robust error of $\mathcal{H}$ with respect to $\tilde{\mathcal{D}}$ is at most $2\eta + \frac{\epsilon}{3}$, i.e., $\mathrm{Err}(\mathcal{H}; \tilde{\mathcal{D}}, \mathcal{U}) = 2\eta + \frac{\epsilon}{3}$, by showing that $h_{\text{opt}} = \mathrm{argmin}_{h \in \mathcal{H}} \mathrm{Err}(h; \mathcal{D}, \mathcal{U})$ has a robust error of at most $2\eta + \frac{\epsilon}{3}$ on $\tilde{\mathcal{D}}$.

Eqs. (4.2) to (4.4) still hold as in the realizable case proof. Combining it together, we have

$$\begin{aligned}
\mathrm{Err}(\mathcal{H}; \tilde{\mathcal{D}}, \mathcal{U}) &\leq \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}} \left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{0\text{-}1}(h_1; x, y)] \\
&\leq \eta + \eta + \frac{\epsilon}{3} \\
&= 2\eta + \frac{\epsilon}{3},
\end{aligned}$$

where (i) follows from Eq. (4.2) and (ii) follows from Eq. (4.4).

Finally, we employ an agnostic adversarially robust supervised PAC learner $\mathcal{B}$ for the class $\mathcal{H}$ on $S^u \sim \tilde{\mathcal{D}}^{m_u}$, that should be of size of the sample complexity of agnostically robust learn $\mathcal{H}$ with respect to $\mathcal{U}$, when the optimal robust error of hypothesis from $\mathcal{H}$ on $\tilde{\mathcal{D}}$ is at most $2\eta + \frac{\epsilon}{3}$. We are guaranteed that the resulting hypothesis h_2 has a robust error of at most $2\eta + \frac{\epsilon}{3} + \frac{\epsilon}{3} = 2\eta + \frac{2\epsilon}{3}$ on $\tilde{\mathcal{D}}$, with probability $1 - \frac{\delta}{2}$. We observe that the total variation distance between $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is at most $\eta + \frac{\epsilon}{3}$, and as a result, h_2 has a robust error of at most $2\eta + \frac{2\epsilon}{3} + \eta + \frac{\epsilon}{3} = 3\eta + \epsilon$ on $\mathcal{D}$, with probability $1 - \delta$.

We conclude that a size of $|S^u_{\mathcal{X}}| = m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, 2\eta + \frac{\epsilon}{3}\right)$ unlabeled sample suffices, in addition to the $m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} \log^2 \frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$ labeled samples which are required in the first 2 steps. We remark that the best known value of $\Lambda_{\mathrm{AG}}\left(1, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta\right)$ is $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$. □

Proof (of Theorem 4.17). We give a proof sketch, this is similar to [9, Theorem 10], knowing the marginal distribution $\mathcal{D}_{\mathcal{X}}$ does not give more power to the learner. The argument is based on the standard lower bound for VC classes (for example [28, Section 3]). Let $S = \{x_1, \ldots, x_k\}$ be a maximal set that is $\mathcal{U}$-robustly shattered by $\mathcal{H}$.

Let $z^+_1, z^-_1, \ldots, z^+_k, z^-_k$ be as in Definition 4.2, and note that for $i \neq j$, $z^+_i \neq z^+_j$ and $z^-_i \neq z^-_j$. Define a distribution $\mathcal{D}_{\boldsymbol{\sigma}}$ for any possible labeling $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k) \in \{0,1\}^k$ of S.

$$\forall j \in [k]: \begin{cases} \mathcal{D}_{\boldsymbol{\sigma}}(z^+_j, 1) = \frac{1-\alpha}{2k} \ \wedge \ \mathcal{D}_{\boldsymbol{\sigma}}(z^-_j, 0) = \frac{1+\alpha}{2k} & \sigma_j = 0, \\ \mathcal{D}_{\boldsymbol{\sigma}}(z^+_j, 1) = \frac{1+\alpha}{2k} \ \wedge \ \mathcal{D}_{\boldsymbol{\sigma}}(z^-_j, 0) = \frac{1-\alpha}{2k} & \sigma_j = 1. \end{cases}$$

We can now choose α as a function of ϵ, δ in order to get a lower bound on the sample complexity $|S| \gtrsim \frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon^2}$. □

Proof (of Theorem 4.18). We take the construction in Proposition 4.8, where there is an arbitrary gap between $\mathrm{VC}_{\mathcal{U}}$ and

$\mathrm{RS}_{\mathcal{U}}$.

Recall that on every pair (x, z) in Proposition 4.8 the optimal error is $\eta = 1/2$. On such unlabeled pairs, the learner can only randomly choose a prediction, and the error is $3/4$. We have $\mathrm{VC}_{\mathcal{U}} = 0$, and the labeled sample size is $\frac{1}{\epsilon^2}\log\frac{1}{\delta}$. As $(\mathrm{RS}_{\mathcal{U}} - \frac{1}{\epsilon^2}\log\frac{1}{\delta})$ grows, the gap between the learner and the optimal classifier is approaching $3/2$, which means that for any $\gamma > 0$ we can pick $\mathrm{RS}_{\mathcal{U}}$ such that error of $(\frac{2}{3} - \gamma)\eta$ is not possible.

In order to prove the case of any $0 < \eta \leq 1/2$, we can just add points such that their perturbation set does not intersect with any other perturbation set, and follow the same argument. □

Auxiliary Definitions and Proofs for Section 4.6

Definition of the model.

Definition 4.24 ((non-robust) PAC learnability for robustly realizable distributions) *For any $\epsilon, \delta \in (0,1)$, the sample complexity of (ϵ,δ)-PAC learning for a class $\mathcal{H}$, denoted by $\Upsilon(\epsilon,\delta,\mathcal{H},\mathcal{U})$, is the smallest integer m for which there exists a learning algorithm $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ robustly realizable by $\mathcal{H}$ with respect to a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, namely $\mathrm{Err}(\mathcal{H}; D, \mathcal{U}) = 0$, for a random sample $S \sim \mathcal{D}^m$, it holds that*

$$\mathbb{P}\left(\mathrm{Err}\left(\mathcal{A}(S); D\right) \leq \epsilon\right) > 1 - \delta.$$

If no such m exists, define $\Upsilon(\epsilon,\delta,\mathcal{H},\mathcal{U}) = \infty$, and $\mathcal{H}$ is not (ϵ,δ)-PAC for distributions that are robustly realizable by $\mathcal{H}$ with respect to $\mathcal{U}$.

Proof (of Proposition 4.19). Define the uniform distribution $\mathcal{D}$ over the support $\{(x_1,1),\ldots,(x_{2m},1)\}$, such that $\bigcap_{i=1}^{2m}\mathcal{U}(x_i) \neq \emptyset$. Define $\mathcal{H} : \mathcal{X} \to 2^{\mathcal{X}}$ to be all binary functions over $\mathcal{X}$. Note that the $\mathcal{D}$ is robustly realizable by $\mathcal{H}$, the constant function that always returns 1 has no error. Moreover we have $\mathrm{VC}_{\mathcal{U}} = 1$, and $\mathrm{VC} = 2m$, for any $m \in \mathbb{N}$. □

Proof (of Theorem 4.20). We follow only the first two steps of the generic Algorithm 2. Namely, take a labeled sample S and a hypothesis class $\mathcal{H}$ and create the partial hypothesis class $\mathcal{H}^\star_{\mathcal{U}}$. Assuming that the distribution is robustly realizable by $\mathcal{H}$, we end up in a realizable setting of learning a partial concept class $\mathcal{H}^\star_{\mathcal{U}}$.

In the second step of the algorithm, we call a learning algorithm for partial concept classes (Section 4.7) in order to do so. The sample complexity is the same as Theorem 4.22, $\Upsilon(\epsilon,\delta,\mathcal{H},\mathcal{U}) = \mathcal{O}\left(\Lambda_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H})\right)$. Has we have shown in the proof of Theorem 4.11, Eq. (4.1), this implies the Theorem. □

Proposition 4.25 *Consider the distribution $\mathcal{D}$ and the hypothesis class $\mathcal{H}$ in Theorem 4.19. There exists a robust ERM algorithm returning a hypothesis $h_{ERM} \in \mathcal{H}$, such that $\mathrm{Err}(h_{ERM}; \mathcal{D}) \geq \frac{1}{4}$ with probability 1 over $S \sim \mathcal{D}^m$.*

Proof. Consider the following robust ERM. For any sample of size m, return 1 on the sample points and randomly choose a label for out-of-sample points. The error rate of such a robust ERM is at least $1/4$ with probability 1. □

Proof (of Theorem 4.21). This follows from a similar no-free-lunch argument for VC classes [e.g., 27, Section 5]. We briefly explain the proof idea.

Take the distribution $\mathcal{D}$, and the class $\mathcal{H}$ from Proposition 4.25 with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 1$ and $\mathrm{VC}(\mathcal{H}) = 3m$. Keep functions that are robustly self-consistent only on $2m$ points. Construct all of the distributions on $2m$ points from the support of $\mathcal{D}$. We have $\binom{3m}{2m}$ such distributions, and each one of them is robustly realizable by different $h \in \mathcal{H}$. The idea is that a proper leaner observing only m points should guess which are the remaining m points in the support of the distribution. There rest of the proof follows from the no-free-lunch proof. It can be shown formally via the probabilistic method, that for every proper rule, there exists a distribution on which the error is constant with fixed probability. □

Learning Algorithms for Partial Concept Classes

Here we overview the algorithmic techniques from Alon et al. [11, Theorem 34 and 41], for learning partial concepts in realizable and agnostic settings. We use these algorithms in step 2 of our Algorithm 2.

One-inclusion graph algorithm for partial concept classes. We briefly discuss the algorithm, for the full picture, see [128, 129]. The one-inclusion algorithm for a class $\mathcal{F} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ gets an input of unlabeled examples $S = (x_1, \ldots, x_m)$ and labels $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)$ that are consistent with some $f \in \mathcal{F}$, that is, $f(x_k) = y_k$ for all $k \neq i$. It guarantees an (ϵ, δ)- PAC learner in the realizable setting, with sample complexity of $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon} \log \frac{1}{\delta}\right)$ as mentioned in Theorem 4.22.

Here is a description of the algorithm. First, construct the one-inclusion graph. For any $j \in [m]$ and $f \in \mathcal{F}|_S$ define $E_{j,f} = \{f' \in \mathcal{F}|_S : f'(x_k) = f(x_k), \forall k \neq j\}$, that is, all functions in $\mathcal{F}|_S$ that are consistent with f on S, except the point x_j. Define the set of edges $E = \{E_{j,f} : j \in [m], f \in \mathcal{F}|_S\}$, and the set vertices $V = \mathcal{F}|_S$ of the one-inclusion graph $G = (V, E)$. An orientation function $\psi : E \to V$ for an undirected graph G is an assignment of a direction to each edge, turning G into a directed graph. Find an orientation ψ that minimizes the out-degree of G. For prediction of x_i, pick $f \in V$ such that $f(x_k) = y_k$ for all $k \neq i$, and output $\psi(E_{i,f})(x_i)$.

Note that this algorithm is transductive, in the sense that in order to predict the label of a test point, it uses the entire training sample to compute its prediction.

Boosting and compression schemes. Recall the well-known boosting algorithm, α-Boost [130, pages 162-163], which is a simplified version of AdaBoost, where the returned function is a simple majority over weak learners, instead of a weighted majority. For a hypothesis class $\mathcal{H}$ and a sample of size m, the algorithm yields a compression scheme of size $\mathcal{O}(\mathrm{VC}(\mathcal{H}) \log(m))$. Recall the generalization bound based on a sample compression scheme in Lemma 2.19.

The learning algorithm for the realizable setting is α-Boost, where the weak learners are taken from the one-inclusion graph algorithm. As mentioned in Theorem 4.22, this obtains an upper bound of $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon} \log^2\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\right) + \frac{1}{\epsilon} \log \frac{1}{\delta}\right)$.

For the agnostic setting, follow a reduction to the realizable case suggested by David et al. [40]. The reduction requires the construction of a compression scheme based on a Boosting algorithm. Roughly speaking, the reduction works as follows. Denote $\Lambda_{\mathrm{RE}} = \Lambda_{\mathrm{RE}}(1/3, 1/3, \mathcal{H})$, the sample complexity of $(1/3, 1/3)$-PAC learn $\mathcal{H}$, in the realizable case. Now, Λ_{RE} samples suffice for weak learning for any distribution $\mathcal{D}$ on a given sample S.

Find the maximal subset $S' \subseteq S$ such that $\inf_{h \in \mathcal{H}} \widehat{\mathrm{Err}}(h; S') = 0$. Now, Λ_{RE} samples suffice for weak robust

learning for any distribution $\mathcal{D}$ on S'. Execute the α-boost algorithm on S', with parameters $\alpha = \frac{1}{3}$ and number of boosting rounds $T = \mathcal{O}(\log(|S'|))$, where each weak learner is trained on Λ_{RE} samples. The returned hypothesis $\bar{h} = \mathrm{Majority}\left(\hat{h}_1, \ldots, \hat{h}_T\right)$ satisfies that $\widehat{\mathrm{Err}}\left(\bar{h}; S'\right) = 0$, and each hypothesis $\hat{h}_t \in \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ is representable as set of size $\mathcal{O}(\Lambda_{\mathrm{RE}})$. This defines a compression scheme of size $\Lambda_{\mathrm{RE}} T$, and $\bar{h}$ 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 $\inf_{h \in \mathcal{H}} \widehat{\mathrm{Err}}(h; S') = 0$ which implies that $\widehat{\mathrm{Err}}\left(\bar{h}; S\right) \leq \inf_{h \in \mathcal{H}} \widehat{\mathrm{Err}}(h; S)$. Plugging it into a data-dependent compression generalization bound (Lemma 2.19), we obtain a sample complexity of
$\Lambda_{\mathrm{AG}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon^2} \log^2\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon^2}\right) + \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$, as mentioned in Theorem 4.23.

Supervised Robust Learning Algorithms

We overview the algorithms of Montasser et al. [9, proofs of Theorems 4 and 8]. Their construction is based on sample compression methods explored in [131, 132].

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, fix a distribution $\mathcal{D}$ over the input space $\mathcal{X} \times \mathcal{Y}$. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be an i.i.d. training sample from a robustly realizable distribution $\mathcal{D}$ by $\mathcal{H}$, namely $\inf_{h \in \mathcal{H}} \mathrm{Err}(h; \mathcal{D}, \mathcal{U}) = 0$. Denote $d = \mathrm{VC}(\mathcal{H})$, $d^* = \mathrm{VC}^*(\mathcal{H})$ is the *dual VC-dimension*. Fix $\epsilon, \delta \in (0,1)$.

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 $I(z) = \min\{i \in [n] : z \in \mathcal{U}(x_i)\}$. The goal is to construct a compression scheme that is consistent with $S_{\mathcal{U}}$.

2. Discretize $S_{\mathcal{U}}$ to a finite set $\bar{S}_{\mathcal{U}}$. Define the class of hypotheses with zero robust error on every d points in S,

$$\hat{\mathcal{H}} = \left\{\mathrm{RERM}_{\mathcal{H}}(S') : S' \subseteq S, |S'| = d\right\},$$

where $\mathrm{RERM}_{\mathcal{H}}$ maps any labeled set to a hypothesis in $\mathcal{H}$ with zero robust loss on this set. The cardinality of this class is bounded as follows

$$|\hat{\mathcal{H}}| = \binom{n}{d} \leq \left(\frac{en}{d}\right)^d.$$

Discretize $S_{\mathcal{U}}$ to a finite set using the finite class $\hat{\mathcal{H}}$. Define the *dual class* $\mathcal{H}^* \subseteq \{0,1\}^{\mathcal{H}}$ of $\mathcal{H}$ as the set of all functions $f_{(x,y)} : \mathcal{H} \to \{0,1\}$ defined by $f_{(x,y)}(h) = \mathbb{I}[h(x) \neq y]$, for any $h \in \mathcal{H}$ and $(x,y) \in S_{\mathcal{U}}$. If we think of a binary matrix where the rows consist of the distinct hypotheses and the columns are points, then the dual class corresponds to the transposed matrix where the distinct rows are points and the columns are hypotheses. A discretization $\bar{S}_{\mathcal{U}}$ will be defined by the dual class of $\hat{\mathcal{H}}$. Formally, $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ consists of exactly one $(x,y) \in S_{\mathcal{U}}$ for each distinct classification $\left\{f_{(x,y)}(h)\right\}_{h \in \hat{\mathcal{H}}}$. In other words, $\hat{\mathcal{H}}$ induces a finite partition of $S_{\mathcal{U}}$ into regions where every $\hat{h} \in \hat{\mathcal{H}}$ suffers a constant loss $\mathbb{I}\left[\hat{h}(x) \neq y\right]$ in each region, and the discretization $\bar{S}_{\mathcal{U}}$ takes one point per

region. By Sauer's lemma [26, 133], for $n > 2d$,

$$|\bar{S}_{\mathcal{U}}| \leq \left(\frac{e|\hat{\mathcal{H}}|}{d^*}\right)^{d^*} \leq \left(\frac{e^2 n}{dd^*}\right)^{dd^*},$$

3. Execute the following modified version of the algorithm α-boost [130, pages 162-163] on the discretized set $\bar{S}_{\mathcal{U}}$, with parameters $\alpha = \frac{1}{3}$ and number of boosting rounds $T = \mathcal{O}\left(\log\left(|\bar{S}_{\mathcal{U}}|\right)\right) = \mathcal{O}\left(dd^* \log(n)\right)$.

Modified α-boost

Input: $\mathcal{H}, S, \bar{S}_{\mathcal{U}}, d, \mathrm{RERM}_{\mathcal{H}}$.
Parameters: α, T.
Initialize P_1 = Uniform($\bar{S}_{\mathcal{U}}$).
For $t = 1, \ldots, T$:

(a) Find $\mathcal{O}(d)$ points $S_t \subseteq \bar{S}_{\mathcal{U}}$ such that every $h \in \mathcal{H}$ with $\widehat{\mathrm{Err}}(h; S_t) = 0$ has $\mathrm{Err}(h; P_t) \leq 1/3$.

(b) Let S'_t be the original $\mathcal{O}(d)$ points in S with $S_t \subseteq \bigcup_{(x,y)\in S'_t} \bigcup \{(z, y) : z \in \mathcal{U}(x)\}$.

(c) Let $\hat{h}_t = \mathrm{RERM}_{\mathcal{H}}(S'_t)$.

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

$$P_{t+1}(x, y) \propto P_t(x, y) e^{-\alpha \mathbb{I}\left\{\hat{h}_t(x) = y\right\}}$$

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

---

4. Output the majority vote $\bar{h} = \mathrm{Majority}\left(\hat{h}_1, \ldots, \hat{h}_T\right)$.

We are guaranteed that $\widehat{\mathrm{Err}}\left(\bar{h}; S, \mathcal{U}\right) = 0$, and each hypothesis $\hat{h}_t \in \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ is representable as set S'_t of size $\mathcal{O}(d)$. This defines a compression function $\kappa(S) = \bigcup_{t\in[T]} S'_t$. Thus, $\bar{h}$ can be reconstructed from a compression set of size

$$dT = \mathcal{O}\left(d^2 d^* \log(n)\right).$$

This compression size can be further reduced to $\mathcal{O}\left(dd^*\right)$, using a sparsification technique introduced by Hanneke et al. [131], Moran and Yehudayoff [132], by randomly choosing $\mathcal{O}(d^*)$ hypotheses from $\left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$. The proof follows via a standard uniform convergence argument. Plugging it into a compression generalization bound, we have a sample complexity of $\tilde{\mathcal{O}}\left(\frac{dd^*}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$, in the realizable robust case.

Agnostic case. The construction follows a reduction to the realizable case suggested by David et al. [40]. Denote $\Lambda_{\mathrm{RE}} = \Lambda_{\mathrm{RE}}(1/3, 1/3, \mathcal{H}, \mathcal{U})$, the sample complexity of $(1/3, 1/3)$-PAC learn $\mathcal{H}$ with respect to a perturbation function $\mathcal{U}$, in the realizable robust case.

Using a robust ERM, find the maximal subset $S' \subseteq S$ such that $\inf_{h\in\mathcal{H}} \widehat{\mathrm{Err}}\left(h; S', \mathcal{U}\right) = 0$. Now, Λ_{RE} samples suffice for weak robust learning for any distribution $\mathcal{D}$ on S'.

Execute the α-boost algorithm [130, pages 162-163] on S' for the robust loss function, with parameters $\alpha = \frac{1}{3}$ and number of boosting rounds $T = \mathcal{O}\left(\log\left(|S'|\right)\right)$, where each weak learner is trained on Λ_{RE} samples. The returned hypothesis $\bar{h} = \mathrm{Majority}\left(\hat{h}_1, \ldots, \hat{h}_T\right)$ satisfies that $\widehat{\mathrm{Err}}\left(\bar{h}; S', \mathcal{U}\right) = 0$, and each hypothesis $\hat{h}_t \in \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ is representable as set of size $\mathcal{O}(\Lambda_{\mathrm{RE}})$. This defines a compression scheme of size $\Lambda_{\mathrm{RE}} T$, and $\bar{h}$ 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 $\inf_{h\in\mathcal{H}} \widehat{\mathrm{Err}}\left(h; S', \mathcal{U}\right) = 0$ which implies that $\widehat{\mathrm{Err}}\left(\bar{h}; S, \mathcal{U}\right) \leq \inf_{h\in\mathcal{H}} \widehat{\mathrm{Err}}\left(h; S, \mathcal{U}\right)$. Plugging it into a compression generalization bound, we have a sample complexity of $\tilde{\mathcal{O}}\left(\frac{\Lambda_{\mathrm{RE}}}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$, which translates into $\tilde{\mathcal{O}}\left(\frac{dd^*}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$, in the agnostic robust case.

Chapter 5

Adversarially Robust PAC Learnability of Real-Valued Functions

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.

5.1 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., Madry et al. [45], Szegedy et al. [49], Biggio et al. [50], Goodfellow et al. [51]). From the theoretical standpoint, there has been a lot of effort to provide provable guarantees of such methods (e.g., [1, 2, 8, 44, 55, 57, 59, 60, 81, 83, 103–110, 134–137]), 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. [9]. 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. [1] 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. [9]. On the other hand, Mendelson [12] 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

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

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

$$\mathrm{Err}_{\ell_p}(\hat{h}; \mathcal{D}, \mathcal{U}) \le \inf_{h\in\mathcal{H}} \mathrm{Err}_{\ell_p}(h; \mathcal{D}, \mathcal{U}) + \epsilon. \tag{5.1}$$

The sample complexity for learning $\mathcal{H}$ is the size of a minimal i.i.d. sample from $\mathcal{D}$ such that there exists a learning algorithm with output as in Eq. (5.1). We refer to this model as *Robust Regression* with ℓ_p robust loss. This is a robust formulation of the classic nonparametric regression setting.

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

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

We refer to the loss function in this model as *cutoff loss*, where $\eta > 0$ 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,

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

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 and Bartlett [138], Simon [139]. See also Anthony et al. [33, section 21.4] and references therein.

Main Results

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 and Simon [39], and Eq. (2.2)).

- In Section 5.3 we provide a learning algorithm for robust regression with ℓ_p losses, with sample complexity [1]

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

  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 [9].

- In Section 5.4 we provide a learning algorithm with a substantial sample complexity improvement,

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

- In Section 5.5, 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. [1]. Subsequently, improved bounds appeared in Attias and Kontorovich [4]. Adversarially robust PAC learnability of binary-valued function classes with arbitrary perturbation sets was studied by Montasser et al. [9]. 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. [9] 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 robust *boosting algorithm for real-valued functions*. In the non-robust setting, Hanneke et al. [131] showed how to convert a boosting algorithm (originally introduced by Kégl [140]), into a sample compression scheme. In order to find weak learners (and prove their existence), we rely on *generalization from approximate interpolation*

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

(see Anthony and Bartlett [138] and Anthony and Bartlett [138, 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. [131]: "Duffy and Helmbold [141, 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. [9] 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 handling *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 loss that constructs an ensemble and predicts 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 [9]. Moreover, we show how to reduce the sample complexity substantially 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 [131]. We believe that this technique may be of independent interest.

5.2 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 [142, section 10.3.1] and Pollard [143, 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 $1 \leq p \leq \infty$, define the ℓ_p *robust loss* function of h on (x, y) with

respect to a perturbation function $\mathcal{U}$,

$$\ell_{p,\mathcal{U}}(h;(x,y)) = \sup_{z\in\mathcal{U}(x)} |h(z)-y|^p \,. \tag{5.2}$$

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

$$\ell^{\eta}_{\mathcal{U}}(h;(x,y)) = \mathbb{I}\left\{ \sup_{z\in\mathcal{U}(x)} |h(z)-y| \geq \eta \right\}. \tag{5.3}$$

The non-robust version of this loss function is also known as η-ball or η-tube loss (see for example Anthony et al. [33, Section 21.4]).

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

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

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

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

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 Diochnos et al. [62], Gourdeau et al. [144], Bubeck et al. [145].

Learning models. We precisely define the models for robustly learning real-valued functions. Our first model is learning with the ℓ_p robust loss (see Eq. (5.2)), we refer to this model as *Robust Regression*.

Definition 5.1 (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 the ℓ_p robust loss, 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*

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

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 Eq. (5.3)) 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

$$\mathrm{Err}_\eta(h^\star; \mathcal{D}, \mathcal{U}) = 0. \tag{5.4}$$

Definition 5.2 (Robust (η, β)-regression) *For any $\eta, \beta, \epsilon, \delta \in (0,1)$, 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 Eq. (5.4)), for an i.i.d. random sample $S \sim \mathcal{D}^m$, with probability at least $1-\delta$ over S, it holds that*

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

If no such m exists, define $\mathcal{M}_{\mathrm{RE}}(\eta, \beta, \epsilon, \delta, \mathcal{H}, \mathcal{U}) = \infty$, 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 $1-\delta$,

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

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 re-scaling 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 [9, 29, 146–155].

Oracles. We rely on the following robust empirical risk minimizers. Let a set $S = \{(x_i, y_i)\}_{i=1}^m$. Define an ϵ-approximate ψ-robust empirical risk minimizer $\psi\text{-}\,\mathrm{RERM}_{\mathcal{H}} : (\mathcal{X} \times \mathcal{Y})^m \times [0,1]^m \times (0,1) \to \mathcal{H}$,

$$\psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S, \psi|_S, \epsilon) := \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{m} \sum_{(x,y) \in S} \mathbb{I}\left[\sup_{z \in \mathcal{U}(x)} |h(z) - y| \geq \psi(x,y) + \epsilon \right], \tag{5.5}$$

where $\psi|_S = (\psi(x_1, y_1), \ldots, \psi(x_m, y_m))$. We refer to $\psi(x, y)$ as *cutoff* parameters. Note that ψ is a function of (x, y) and not necessarily a constant.

For $p \in [1, \infty)$, define the empirical robust loss of a function h on a labeled sample S by

$$L_p(h, S, \mathcal{U}) = \frac{1}{m} \sum_{(x,y)\in S} \sup_{z\in\mathcal{U}(x)} |h(z) - y|^p ,$$

and for $p = \infty$

$$L_\infty(h, S, \mathcal{U}) = \max_{(x,y)\in S} \sup_{z\in\mathcal{U}(x)} |h(z) - y| .$$

For the empirical non-robust loss, we omit $\mathcal{U}$ from the notation and use $L_p(h, S)$.

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

$$\ell_p\text{-}\operatorname{RERM}_{\mathcal{H}}(S) := \operatorname*{argmin}_{h\in\mathcal{H}} L_p(h, S, \mathcal{U}). \tag{5.6}$$

Approximate sample compression schemes. Recall Definition 2.18 of a sample compression scheme, which we use here for the ℓ_p loss. We say that a compression scheme (κ, ρ) is a k-size agnostic α-*approximate sample compression scheme* for $\mathcal{H}$ if κ is a k-selection and for all $S = \{(x_i, y_i) : i \in [m]\}$, and $f_S := \rho(\kappa(S))$ achieves $\mathcal{H}$-competitive empirical loss:

$$L_p(f_S, S) \le \inf_{h\in\mathcal{H}} L_p(h, S) + \alpha. \tag{5.7}$$

Similarly, we define an agnostic α-*approximate adversarially robust sample compression scheme* if $f_S := \rho(\kappa(S))$ achieves $\mathcal{H}$-competitive empirical robust loss:

$$L_p(f_S, S, \mathcal{U}) \le \inf_{h\in\mathcal{H}} L_p(h, S, \mathcal{U}) + \alpha. \tag{5.8}$$

When $p = \infty$ we refer to the sample compression as *uniformly* α-*approximate*.

Notation. We use the notation $\tilde{\mathcal{O}}(\cdot)$ for omitting poly-logarithmic factors of $(\operatorname{fat}(\mathcal{H}, \gamma), \operatorname{fat}^*(\mathcal{H}, \gamma), 1/\epsilon, 1/\delta, 1/\eta, 1/\beta)$. We denote $[n] = \{1, \ldots, n\}$, 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.

5.3 Robust Regression

In this section, we provide an algorithm and prove its sample complexity for robust regression with the ℓ_p loss. 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. [9]; there exist, non-convex function classes, where proper learning is impossible.

Theorem 5.3 *Algorithm 3 implies that the sample complexity for robust (ϵ, δ)-PAC learning a concept class $\mathcal{H}$ with the ℓ_p robust loss is*

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

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 Eq. (2.2).

Remark 5.4 *The output of Algorithm 3 is a convex combination of the functions from the concept class, which is a proper predictor, assuming convexity of the function class.*

Remark 5.5 *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 and Bartlett [138], where the cutoff parameter is fixed for all points (Theorem 5.12). The proof is in Section 5.7.

Theorem 5.6 (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 $\beta, \epsilon, \delta \in (0,1)$, any function $\psi : \mathcal{X} \times \mathcal{Y} \to [0,1]$, any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, for a random sample $S \sim \mathcal{D}^m$, if*

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

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

Algorithm 3 Improper Robust Regressor with High-Vote

Input: $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, $S = \{(x_i, y_i)\}_{i=1}^m$.
Parameters: Approximation parameter ϵ, base learner sample size $d \ge 1$, number of Multiplicative Weights rounds $T \ge 1$, loss parameter $p \in [1, \infty]$.
Algorithms used: ℓ_p-$\mathrm{RERM}_{\mathcal{H}}$ (Eq. (5.6)), ψ-$\mathrm{RERM}_{\mathcal{H}}$ (Eq. (5.5)), a robust variant of Multiplicative Weights (Algorithm 6).

1. Compute $h^\star \leftarrow \ell_p$-$\mathrm{RERM}_{\mathcal{H}}(S)$.
   Denote $\psi(x, y) = \sup_{z \in \mathcal{U}(x)} |h^\star(z) - y|, \forall (x, y) \in S$.
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 $\hat{h} \in \hat{\mathcal{H}}$ is obtained by ψ-RERM optimizer operating on d points from S. The input cutoff parameters to the optimizer are $\psi(x, y)$, as computed in step 1.
   (ii) Let $\tilde{\mathcal{H}} = \hat{\mathcal{H}} \cup \{h^\star\}$. Each $(z, y) \in S_{\mathcal{U}}$ defines a function in the dual space, $f_{(z,y)} : \tilde{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|^p$. Define $\bar{S}_{\mathcal{U}}$ to be the minimal cover of $S_{\mathcal{U}}$ at scale $\mathcal{O}(\epsilon/p)$ under the supremum norm.
4. Compute a robust Multiplicative Weights algorithm on $\bar{S}_{\mathcal{U}}$. Let $\left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ 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.

Proof overview and algorithm outline. The complete proof is in Section 5.7. We follow the steps in Algorithm 3.

1. We start with computing a robust empirical risk minimizer (ERM) $h^\star$ on S for the ℓ_p robust loss, $p \in [1, \infty]$. 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* $\psi(x, y)$ and our goal is to construct a predictor with a loss of $\psi(x, y)^p + \epsilon$ 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 ψ- RERM for $\mathcal{H}$ (defined in Eq. (5.5)), performing on a subset $S' \subseteq S$ of size

$$\begin{cases} d \leftarrow \tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H}, c\epsilon/p)\right), p \in [1, \infty) \\ d \leftarrow \tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H}, c\epsilon)\right), p = \infty. \end{cases}$$

   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 $\psi(x, y)^p + \epsilon$ for any $(x, y) \in S'$. The size of $\hat{\mathcal{H}}$ is bounded $(m/d)^d$, where $|S| = m$.

   (ii) Let $\tilde{\mathcal{H}} = \hat{\mathcal{H}} \cup \{h^\star\}$. Define a discretization $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ by taking a uniform cover of the dual space defined on $\tilde{\mathcal{H}}$. In the dual space, each $(z, y) \in S_{\mathcal{U}}$ defines a function $f_{(z,y)} : \tilde{\mathcal{H}} \to [0, 1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|^p$. We take a minimal $\mathcal{O}(\epsilon/p)$-cover for $S_{\mathcal{U}}$ with the supremum norm, which is of size $\mathcal{N}\left(\mathcal{O}(\epsilon/p), S_{\mathcal{U}}, \|\cdot\|_\infty\right)$. We use covering numbers arguments [37] to upper bound the size of $\bar{S}_{\mathcal{U}}$

4. Compute a variant of Multiplicative Weights (MW) update (Algorithm 6) on $\bar{S}_{\mathcal{U}}$ for $T \approx \log\left|\bar{S}_{\mathcal{U}}\right|$ rounds as follows. From Theorem 5.6 and using the Lipschitzness of the ℓ_p loss, 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 $\forall (z, y) \in S'' : \ |h(z) - y|^p \leq \psi(z, y)^p + \epsilon$, it holds that
   $\mathbb{P}_{(z,y)\sim\mathcal{P}}\{(z, y) : |h(z) - y|^p \leq \psi(z, y)^p + 2\epsilon\} \geq 1 - \epsilon$, where $\psi(z, y)$ is the $\psi(x, y)$ for which $z \in \mathcal{U}(x)$. 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 $\hat{h} \leftarrow \psi$- RERM on S' with the relevant cutoff parameters. $\hat{h}$ has error of $\psi(z, y)^p + \epsilon$ 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 and Freund [130, 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 3ϵ 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 5ϵ,

$$\forall (z,y) \in S_{\mathcal{U}} : \frac{1}{T}\sum_{i=1}^{T}\left|\hat{h}_i(z) - y\right|^p \leq \psi(z,y)^p + 5\epsilon.$$

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

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

Therefore, we conclude that

$$\forall (z,y) \in S_{\mathcal{U}} : \left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(z) - y\right|^p \leq \psi(z,y)^p + 5\epsilon,$$

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

$$\forall (x,y) \in S : \sup_{z\in\mathcal{U}(x)}\left|\frac{1}{T}\sum_{i=1}^{T}\hat{h}_i(z) - y\right|^p \leq \psi(x,y)^p + 5\epsilon.$$

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.

5.4 Improved Sample Complexity via Median Boosting and Sparsification

In this section, we provide an algorithm with a substantial sample complexity improvement. 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 [131, 140] in our algorithm. Each base learner requires fewer samples and as a result, we improve the sample complexity. On the contrary, in Algorithm 3 we obtained accurate predictions for a $1-\mathcal{O}(\epsilon)$ quantile of the predictors, and we output their average.

Theorem 5.7 *Algorithm 4 implies that the sample complexity for robust (ϵ,δ)-PAC learning a concept class $\mathcal{H}$ with the*

ℓ_p robust loss is

$$\begin{cases} \tilde{\mathcal{O}}\left(\dfrac{\mathrm{fat}(\mathcal{H}, c\epsilon/p)\,\mathrm{fat}^*(\mathcal{H}, c\epsilon/p)}{\epsilon^2} + \dfrac{1}{\epsilon^2}\log\dfrac{1}{\delta} \right), p \in [1,\infty) \\ \tilde{\mathcal{O}}\left(\dfrac{\mathrm{fat}(\mathcal{H}, c\epsilon)\,\mathrm{fat}^*(\mathcal{H}, c\epsilon)}{\epsilon^2} + \dfrac{1}{\epsilon^2}\log\dfrac{1}{\delta} \right), p = \infty, \end{cases}$$

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 Eq. (2.2).

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

Definition 5.8 (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 with respect to $\mathcal{D}$ and a target function $h^\star \in \mathcal{H}$ if*

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

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 and Schapire [156]), 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 4 Improper Robust Median Regressor

Input: $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, $S = \{(x_i, y_i)\}_{i=1}^m$.
Parameters: Approximation parameter $\epsilon \in (0,1)$, weak learner sample size $d \ge 1$, sparsification parameter $k \ge 1$, number of boosting rounds $T \ge 1$, loss parameter $p \in [1,\infty]$.
Algorithms used: ℓ_p-$\mathrm{RERM}_{\mathcal{H}}$ (Eq. (5.6)), ψ-$\mathrm{RERM}_{\mathcal{H}}$ (Eq. (5.5)), a variant of median boosting: MedBoost (Algorithm 7), sparsification method (Algorithm 8).

1. Compute $h^\star \leftarrow \ell_p$-$\mathrm{RERM}_{\mathcal{H}}(S)$.
   Denote $\psi(x,y) = \sup_{z\in\mathcal{U}(x)} |h^\star(z) - y|, \forall (x,y) \in S$.
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 $\hat{h} \in \hat{\mathcal{H}}$ is obtained by ψ-RERM optimizer operating on d points from S. The input cutoff parameters to the optimizer are $\psi(x,y)$, as computed in step 1.
   (ii) Let $\tilde{\mathcal{H}} = \hat{\mathcal{H}} \cup \{h^\star\}$. Each $(z,y) \in S_{\mathcal{U}}$ defines a function in the dual space, $f_{(z,y)} : \tilde{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|^p$. Define $\bar{S}_{\mathcal{U}}$ to be the minimal cover of $S_{\mathcal{U}}$ at scale $\mathcal{O}(\epsilon/p)$ under the supremum norm.
4. Compute robust 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 Section 5.7.

We explain the main differences from Algorithm 3 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 ψ- RERM for $\mathcal{H}$ (defined in Eq. (5.5)), performing on a subset $S' \subseteq S$ of size

$$\begin{cases} d \leftarrow \tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{H}, c\epsilon/p)) \,, p \in [1, \infty) \\ d \leftarrow \tilde{\mathcal{O}}(\mathrm{fat}(\mathcal{H}, c\epsilon)) \,, p = \infty. \end{cases}$$

This is in contrast to Algorithm 3, where we use $\mathrm{Multiplicative\ Weights}$ update that operates with stronger base learners. We can make accurate predictions by aggregating the outputs of the weak learners by taking their median instead of the average. Another improvement arises from sparsifying the ensemble [131] 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 4. Steps (1),(2), and (3) are similar to Algorithm 3, besides the construction of $\hat{\mathcal{H}}$ as we explained above. In step (4), we compute a robust version of the real-valued boosting algorithm $\mathrm{MedBoost}$ [140] on the discretized set $\bar{S}_{\mathcal{U}}$, see Algorithm 7. Hanneke et al. [131] showed how to construct a sample compression scheme from $\mathrm{MedBoost}$. 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 2ϵ of the target cutoff that was computed in step 1. By the covering argument, the median of the losses is within 4ϵ for any point in $(z, y) \in S_{\mathcal{U}}$,

$$\left|\mathrm{Median}\big(\hat{h}_1(z) - y, \ldots, \hat{h}_T(z) - y\big)\right|^p \leq \psi(z, y)^p + 4\epsilon.$$

The median is translation invariant, so we have

$$\left|\mathrm{Median}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z)\big) - y\right|^p \leq \psi(z, y)^p + 4\epsilon.$$

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

$$\sup_{z \in \mathcal{U}(x)} \left|\mathrm{Median}\big(\hat{h}_1(z) - y, \ldots, \hat{h}_T(z) - y\big)\right|^p \leq \psi(x, y)^p + 4\epsilon.$$

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,

$$\sup_{z \in \mathcal{U}(x)} \left|\mathrm{Median}\big(\hat{h}_1(z) - y, \ldots, \hat{h}_k(z) - y\big)\right|^p \leq \psi(x, y)^p + 5\epsilon.$$

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

5.5 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 Section 5.7.

Theorem 5.9 *Let $\mathcal{D}$ be a distribution that is η-uniformly realizable (see Eq. (5.4)) by a class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$. Algorithm 5 implies that the sample complexity for robust $(\eta, \beta, \epsilon, \delta)$-PAC learning a concept class $\mathcal{H}$ is*

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

for some numerical constant $c \in (0, \infty)$. Recall that $\operatorname{fat}^*(\mathcal{F}, \epsilon) \lesssim \frac{1}{\epsilon} 2^{\operatorname{fat}(\mathcal{F}, \epsilon/2)+1}$ by Eq. (2.2).

Algorithm 5 Improper Robust (η, β)-Regressor for the Realizble Setting

---

Input: $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, $S = \{(x_i, y_i)\}_{i=1}^m$.
Parameters: Approximation parameters $\eta, \beta \in (0,1)$, sparsification parameter $k \geq 1$, number of boosting rounds $T \geq 1$.
Algorithms used: ψ- $\mathrm{RERM}_{\mathcal{H}}$ (Eq. (5.5)), a variant of median boosting: MedBoost (Algorithm 7), sparsification method (Algorithm 8).

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 ψ- RERM optimizer on $\tilde{\mathcal{O}}(\operatorname{fat}(\mathcal{H}, \mathcal{O}(\beta)))$ points from S. The input cutoff parameters to the optimizer are fixed η for all points.
   (ii) Let $\tilde{\mathcal{H}} = \hat{\mathcal{H}} \cup \{h^\star\}$. Each $(z,y) \in S_{\mathcal{U}}$ defines a function in the dual space, $f_{(z,y)} : \tilde{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|$. Define $\bar{S}_{\mathcal{U}}$ to be the minimal cover of $S_{\mathcal{U}}$ under $\|\cdot\|_\infty$ norm at scale $\mathcal{O}(\beta)$.
3. Compute robust 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: $\hat{h} = \text{Median}\left(\hat{h}_1, \ldots, \hat{h}_k\right)$.

We explain the main differences from Algorithm 4. 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 (Eq. (5.4)).

In this model, the cutoff parameter is given to us as a parameter and is *fixed* for all points. This is different from Algorithms 3 and 4, 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 ψ- 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 [40] for the 0-1 loss and holds for the η-ball robust loss as well. The proof is in Section 5.7.

Theorem 5.10 *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 Eq. (2.2).

Remark 5.11 *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 Section 5.7 we show that by using a fixed cutoff for all points we can obtain an error of only $\sqrt{OPT_{\mathcal{H}}} + \epsilon$.*

5.6 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 loss (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. [2] 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.

5.7 Deferred Proofs

Theorem 5.12 (Generalization from approximate interpolation) *[33, 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 $\eta, \beta, \epsilon, \delta \in (0,1)$, 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

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

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

Proof of Theorem 5.6: Generalization from Approximate Interpolation with Changing Cutoffs

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

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

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

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

Attias and Kontorovich [4] 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 5.12 to the class $\mathcal{F}_2$ with $t(x) = 0$ and $\eta = \beta$.

Proofs for Section 5.3: Robust Regression for ℓ_p Losses

Proof (of Theorem 5.3). Fix $\epsilon, \delta \in (0,1)$ and $p \in [1, \infty]$. 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 first prove for $p \in \{1, \infty\}$, and generalize for $p \in (1, \infty)$ by using the Lipschitzness of the ℓ_p loss. We follow the steps described in Algorithm 3.

1. Compute $h^\star \leftarrow \ell_p\text{-}\mathrm{RERM}_{\mathcal{H}}(S)$ in order to get the set of cutoffs $\psi(x, y) = \sup_{z \in \mathcal{U}(x)} |h^\star(z) - y|$ for $(x, y) \in S$. Let $\psi|_S = (\psi(x_1, y_1), \ldots, \psi(x_m, y_m))$. Our goal is to construct a predictor with an empirical robust loss of $\psi(x, y)^p + \epsilon$, for $p \in (1, \infty)$, and $\psi(x, y) + \epsilon$ for $p \in \{1, \infty\}$, 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 $\psi(z, y)$ be the $\psi(x, y)$ 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 follows.

   (a) Define a set of functions, such that each function is defined by an ϵ-approximate $\psi\text{-}\mathrm{RERM}_{\mathcal{H}}$ optimizer on

$d = \mathcal{O}\Big(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H}, \epsilon/8)\log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\epsilon/8)}{\epsilon^2}\Big)\Big)$ points from S,

$$\hat{\mathcal{H}} = \left\{\psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S', \psi|_{S'}, \epsilon) : S' \subseteq S, |S'| = d\right\}.$$

Recall the definition of ψ- $\mathrm{RERM}_{\mathcal{H}}$, see Eq. (5.5). The cardinality of this class is bounded as follows

$$|\hat{\mathcal{H}}| \approx \binom{m}{d} \lesssim \left(\frac{m}{d}\right)^d. \tag{5.9}$$

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

$$\sup_{(z,y)\in S_{\mathcal{U}}} \inf_{(\bar{z},\bar{y})\in\bar{S}_{\mathcal{U}}} \big\|f_{(z,y)} - f_{(\bar{z},\bar{y})}\big\|_\infty \le \epsilon. \tag{5.10}$$

Let $\mathrm{fat}^*\big(L^1_{\tilde{\mathcal{H}}}, \epsilon\big)$ be the dual ϵ-fat-shattering of $L^1_{\tilde{\mathcal{H}}}$. Applying a covering number argument from Theorem 2.16 on the dual space and upper bounding the dual fat-shattering of the L_1 loss class with the dual fat-shattering of $\tilde{H}$, we have the following bound

$$\begin{aligned}
\big|\bar{S}_{\mathcal{U}}\big| &= \mathcal{N}\big(\epsilon, S_{\mathcal{U}}, \|\cdot\|_\infty\big) \\
&\lesssim \exp\left(\mathrm{fat}^*\big(L^1_{\tilde{\mathcal{H}}}, c\epsilon\big)\log^2\left(\frac{|\tilde{\mathcal{H}}|}{\epsilon}\right)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*\Big(\tilde{\mathcal{H}}, c\epsilon\Big)\log^2\left(\frac{|\tilde{\mathcal{H}}|}{\epsilon}\right)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon)\log^2\left(\frac{|\tilde{\mathcal{H}}|}{\epsilon}\right)\right),
\end{aligned} \tag{5.11}$$

where $c \in (0,\infty)$ is a numerical constant, derived from the covering argument in Theorem 2.16.

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

From Theorem 5.6, taking $\delta = 1/3$, $\beta = \epsilon$, 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 $\forall(\bar{z},\bar{y}) \in S'' : |h(\bar{z}) - \bar{y}| \le \psi(\bar{z},\bar{y}) + \epsilon$, it holds that

$\mathbb{P}_{(\bar{z},\bar{y})\sim\mathcal{P}}\{(\bar{z},\bar{y}) : |h(\bar{z}) - \bar{y}| \le \psi(\bar{z},\bar{y}) + 2\epsilon\} \ge 1-\epsilon$. 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, $S'' \subseteq \bigcup_{(x,y)\in S'}\bigcup\{(\bar{z},\bar{y}) : \bar{z} \in \mathcal{U}(x)\}$. Take $\hat{\mathcal{H}} \ni \hat{h} = \psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S', \psi|_{S'}, \epsilon)$, it holds that $\forall(\bar{z},\bar{y}) \in S'' : \Big|\hat{h}(\bar{z}) - \bar{y}\Big| \le \psi(\bar{z},\bar{y}) + \epsilon$, as a result we get $\mathbb{P}_{(\bar{z},\bar{y})\sim\mathcal{P}}\Big\{(\bar{z},\bar{y}) : \Big|\hat{h}(\bar{z}) - \bar{y}\Big| \le \psi(\bar{z},\bar{y})$ $1-\epsilon$.

Algorithm 6 Robust Multiplicative Weights

Input: $\mathcal{H}, S, \bar{S}_{\mathcal{U}}$.
Parameters: Approximation parameter $\epsilon \in (0,1)$, weights update parameter $\xi \in (0,1)$, number of boosting rounds $T \geq 1$, base learner sample size $d \geq 1$, loss parameter $p \in [1,\infty]$ cutoff parameters $\psi|_S = (\psi(x_1,y_1),\ldots,\psi(x_m,y_m))$ for $(x_i,y_i), \in S$ and $\psi(\bar{z},\bar{y})$ is the $\psi(x,y)$ for which $\bar{z} \in \mathcal{U}(x)$.
Algorithms used: ϵ-approximate ψ-robust empirical risk minimizer ψ- $\mathrm{RERM}_{\mathcal{H}}$ (Eq. (5.5)).
Initialize P_1 = Uniform($\bar{S}_{\mathcal{U}}$).
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 ψ- $\mathrm{RERM}_{\mathcal{H}}$ on them.

(a) Find d points $S''_t \subseteq \bar{S}_{\mathcal{U}}$ such that any $h \in \mathcal{H}$ satisfying: $\forall(\bar{z},\bar{y}) \in S''_t : \ |h(\bar{z}) - \bar{y}|^p \leq \psi(\bar{z},\bar{y})^p + \epsilon$, it holds that $\mathbb{E}_{(\bar{z},\bar{y})\sim\mathcal{P}_t}\left[\mathbb{I}\left\{|h(\bar{z}) - \bar{y}|^p \leq \psi(\bar{z},\bar{y})^p + 2\epsilon\right\}\right] \geq 1-\epsilon$. (See the analysis for why this set exists).

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

(c) Compute $\hat{h}_t = \psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S'_t, \psi|_{S'_t}, \epsilon)$.

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

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

$$P_{t+1}(\bar{z},\bar{y}) \propto P_t(\bar{z},\bar{y}) e^{-\xi \mathbb{I}\left\{\left|\hat{h}_t(\bar{z}) - \bar{y}\right|^p \leq \psi(\bar{z},\bar{y})^p + 2\epsilon\right\}}$$

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

A uniformly 5ϵ-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, $\frac{1}{T}\sum_{t=1}^{T}\hat{h}_t(\cdot)$. For $T \approx \log\left|\bar{S}_{\mathcal{U}}\right|$, we show that

$$\forall(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}} : \ \frac{1}{T}\sum_{t=1}^{T}\left|\hat{h}_t(\bar{z}) - \bar{y}\right| \leq \psi(\bar{z},\bar{y}) + 3\epsilon. \tag{5.12}$$

For any distribution $\mathcal{P}_t$ over $\bar{S}_{\mathcal{U}}$, we have a base learner $\hat{h}_t$, satisfying $\mathbb{E}_{(\bar{z},\bar{y})\sim\mathcal{P}_t}\left[\mathbb{I}\left\{\left|\hat{h}_t(\bar{z}) - \bar{y}\right| \leq \psi(\bar{z},\bar{y}) + 2\epsilon\right\}\right] \geq 1-\epsilon$, due to Theorem 5.6. Following standard analysis of MW / α-Boost (see Schapire and Freund [130, Section 6]), for any $(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}}$, $1-\epsilon$ fraction of the base learners have an error within $\psi(\bar{z},\bar{y}) + 2\epsilon$. 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 $\psi(\bar{z},\bar{y}) + 3\epsilon$. 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 (Eq. (5.27)), we have

$$\forall(z,y) \in S_{\mathcal{U}} : \ \frac{1}{T}\sum_{t=1}^{T}\left|\hat{h}_t(z) - y\right| \leq \psi(z,y) + 5\epsilon. \tag{5.13}$$

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

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\left|\hat{h}_t(z)-y\right| \leq \frac{1}{T}\sum_{t=1}^{T}\left|\hat{h}_t(\bar{z})-\bar{y}\right| + \epsilon, \tag{5.14}$$

and

$$\psi(\bar{z},\bar{y}) = \left|h^\star(\bar{z})-\bar{y}\right| \leq \left|h^\star(z)-y\right| + \epsilon = \psi(z,y)+\epsilon. \tag{5.15}$$

Combining Eqs. (5.14) and (5.15) we get Eq. (5.13). By using the convexity of the ℓ_1 loss, we have

$$\left|\frac{1}{T}\sum_{t=1}^{T}\hat{h}_t(z)-y\right| \leq \frac{1}{T}\sum_{t=1}^{T}\left|\hat{h}_t(z)-y\right|. \tag{5.16}$$

Finally, from Eqs. (5.13) and (5.16) we conclude a uniformly 5ϵ-approximate *adversarially robust* sample compression scheme for S,

$$\forall (z,y)\in S_{\mathcal{U}}: \ \left|\frac{1}{T}\sum_{t=1}^{T}\hat{h}_t(z)-y\right| \leq \psi(z,y)+5\epsilon, \tag{5.17}$$

which implies that

$$\forall (x,y)\in S: \ \sup_{z\in\mathcal{U}(x)}\left|\frac{1}{T}\sum_{t=1}^{T}\hat{h}_t(x)-y\right| \leq \psi(x,y)+5\epsilon.$$

Bounding the compression size. We have $T = \mathcal{O}\big(\log\big|\bar{S}_{\mathcal{U}}\big|\big)$ hypotheses, where each one is representable by $d = \mathcal{O}\Big(\frac{1}{\epsilon}\mathrm{fat}(\mathcal{H},\epsilon/8)\log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\epsilon/8)}{\epsilon^2}\Big)\Big)$ points. By counting the number of predictors using Eq. (5.20), we get for $m \geq 2d$

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

We get a uniformly 5ϵ-approximate *adversarially robust* sample compression scheme for S of size

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

Each weak learner is encoded by a multiset $S' \subseteq S$ of size d and is constructed by computing some $\hat{h} \in \mathcal{H}$ that solves the constrained optimization

$$\sup_{z \in \mathcal{U}(x)} \left| \hat{h}(z) - y \right| \leq \psi(x,y) + \epsilon, \ \forall (x,y) \in S'.$$

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

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

Encoding base learners. We encode each $\psi(x,y)$ by some approximation $\tilde{\psi}(x,y)$, such that $\left|\tilde{\psi}(x,y) - \psi(x,y)\right| \leq \epsilon$, by discretizing $[0,1]$ to $1/\epsilon$ buckets of size ϵ, and each $\psi(x,y)$ is rounded down to the closest value $\tilde{\psi}(x,y)$. Each approximation requires to encode $\log(1/\epsilon)$ bits, and so each learner encodes $d\log(1/\epsilon)$ bits and d samples. We have k weak learners, and the compression size is

$$k(d + d\log(1/\epsilon)) \leq 2kd\log(1/\epsilon)\,.$$

Therefore, we have a uniform 5ϵ-approximate compression for ℓ_1 and ℓ_∞ losses of size

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

Generalizing for $p \in (1,\infty)$. We rely on the Lipschitzness of the ℓ_p loss and rescaling the approximation parameter ϵ to ϵ/p. Recall the covering of $S_{\mathcal{U}}$ in step 3(b). Note that an (ϵ/p)-cover for the L_1 loss class is an ϵ-cover for the L_p loss class due to the Lipschitzness of the ℓ_p loss

$$\begin{aligned}\Big| |h(z)-y|^p - |h(\bar{z})-\bar{y}|^p \Big| &\leq p\Big||h(z)-y| - |h(\bar{z})-\bar{y}|\Big| \\ &\leq p\epsilon.\end{aligned}$$

Moreover, we constructed a function $f = \frac{1}{T}\sum_{t=1}^{T}\hat{h}_t(x)$ with $\sup_{z\in\mathcal{U}(x)}|f(z)-y| \leq \psi(x,y)+\epsilon$ for any $(x,y) \in S$. Note that since $|f(\cdot)-y| \in [0,1]$, the same z that maximizes the ℓ_1 loss also maximizes for any ℓ_p. This implies that

$$\sup_{z\in\mathcal{U}(x)}|f(z)-y|^p \overset{(i)}{\leq} ((\psi(x,y))+\epsilon)^p \overset{(ii)}{\leq} \psi(x,y)^p + p\epsilon,$$

where (i) follows by just raising both sides to the power of p and (ii) follows since the function $x \mapsto |x-y|^p$ is

p-Lipschitz for $(x-y) \in [0,1]$, and so

$$|(\psi(x,y)+\epsilon)^p - \psi(x,y)^p| \le p|\psi(x,y)+\epsilon-\psi(x,y)|$$
$$\le p\epsilon.$$

By rescaling $p\epsilon$ to ϵ, we get

$$\sup_{z\in\mathcal{U}(x)}\left|\frac{1}{T}\sum_{t=1}^{T}\hat{h}_t(x)-y\right|^p \le \psi(x,y)^p+\epsilon.$$

Therefore, we have an approximate compression for ℓ_p of size

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

Note that the $1/\epsilon^3$ term is not affected by the rescaling. The instances of ϵ that should be rescaled are the ones that arise from the covering argument in step 3(b), and the approximation parameter for the sample compression.

Generalization bound. 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,

$$\begin{aligned}
\mathrm{Err}_{\ell_p}(\rho(\kappa(S));\mathcal{D}) &\overset{(i)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_p}(\rho(\kappa(S));S)+\sqrt{\frac{|\kappa(S)|\log(m)+\log\frac{1}{\delta}}{m}}\\
&\overset{(ii)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_p}(h^\star;S)+5\epsilon+\sqrt{\frac{|\kappa(S)|\log(m)+\log\frac{1}{\delta}}{m}}\\
&\overset{(iii)}{\lesssim} \mathrm{Err}_{\ell_p}(h^\star;\mathcal{D})+5\epsilon+\sqrt{\frac{|\kappa(S)|\log(m)+\log\frac{1}{\delta}}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{m}}\\
&\lesssim \mathrm{Err}_{\ell_p}(h^\star;\mathcal{D})+5\epsilon+\sqrt{\frac{|\kappa(S)|\log(m)+\log\frac{1}{\delta}}{m}},
\end{aligned}$$

(i) follows from a generalization of sample compression scheme in the agnostic case, see Theorem 2.19, (ii) follows Eq. (5.17), (iii) follows from Hoeffding's inequality. We 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

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

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 $c \in (0, \infty)$. □

Proofs for Section 5.4: Improved Sample Complexity via Median Boosting and Sparsification

Proof (of Theorem 5.7). Fix $\epsilon, \delta \in (0,1)$ and $p \in [1, \infty]$. 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 first prove for $p \in \{1, \infty\}$, and generalize for $p \in (1, \infty)$ by using the Lipschitzness of the ℓ_p loss. We follow the steps as described in Algorithm 4.

1. Compute $h^\star \leftarrow \ell_p\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S)$ in order to get the set of cutoffs $\psi(x,y) = \sup_{z \in \mathcal{U}(x)} |h^\star(z) - y|$ for $(x,y) \in S$. Let $\psi|_S = (\psi(x_1,y_1), \ldots, \psi(x_m,y_m))$. Our goal is to construct a predictor with an empirical robust loss of $\psi(x,y)^p + \epsilon$, for $p \in (1,\infty)$, and $\psi(x,y) + \epsilon$ for $p \in \{1,\infty\}$, 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 $\psi(z,y)$ be the $\psi(x,y)$ 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 follows.

   (a) Define a set of functions, such that each function is defined by an ϵ-approximate $\psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}$ 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,

$$\hat{\mathcal{H}} = \left\{\psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S', \psi|_{S'}, \epsilon) : S' \subseteq S, |S'| = d\right\}.$$

   Recall the definition of $\psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}$, see Eq. (5.5).

   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

$$|\hat{\mathcal{H}}| \approx \binom{m}{d} \lesssim \left(\frac{m}{d}\right)^d. \tag{5.18}$$

   (b) A discretization $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ will be defined by covering of the dual class in $\|\cdot\|_\infty$. Define $\tilde{\mathcal{H}} = \hat{\mathcal{H}} \cup \{h^\star\}$. Let $L^1_{\tilde{\mathcal{H}}}$ be the L_1 loss class of $\tilde{\mathcal{H}}$, namely, $L^1_{\tilde{\mathcal{H}}} = \left\{\mathcal{Z} \times \mathcal{Y} \ni (z,y) \mapsto |h(z) - y| : h \in \tilde{\mathcal{H}}\right\}$. The *dual class* of $L^1_{\tilde{\mathcal{H}}}$, ${L^1_{\tilde{\mathcal{H}}}}^* \subseteq [0,1]^{\tilde{\mathcal{H}}}$, is defined as the set of all functions $f_{(z,y)} : \tilde{\mathcal{H}} \to [0,1]$ such that $f_{(z,y)}(h) = \left|h(z) - y\right|$, for

any $(z,y) \in S_{\mathcal{U}}$. Formally, ${L^1_{\tilde{\mathcal{H}}}}^* = \{f_{(z,y)} : (z,y) \in S_{\mathcal{U}}\}$, where $f_{(z,y)} = \Big(f_{(z,y)}(h_1), \ldots, f_{(z,y)}(h_{|\tilde{\mathcal{H}}|})\Big)$. We take $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ to be a minimal ϵ-cover for $S_{\mathcal{U}}$ in $\|\cdot\|_\infty$,

$$\sup_{(z,y)\in S_{\mathcal{U}}} \inf_{(\bar{z},\bar{y})\in \bar{S}_{\mathcal{U}}} \left\| f_{(z,y)} - f_{(\bar{z},\bar{y})} \right\|_\infty \leq \epsilon. \tag{5.19}$$

Let $\mathrm{fat}^*\left(L^1_{\tilde{\mathcal{H}}}, \epsilon\right)$ be the dual ϵ-fat-shattering of $L^1_{\tilde{\mathcal{H}}}$. Applying a covering number argument from Theorem 2.16 on the dual space and upper bounding the dual fat-shattering of the L_1 loss class with the dual fat-shattering of $\tilde{H}$, we have the following bound

$$\begin{aligned}
\left|\bar{S}_{\mathcal{U}}\right| &= \mathcal{N}\left(\epsilon, S_{\mathcal{U}}, \|\cdot\|_\infty\right) \\
&\lesssim \exp\left(\mathrm{fat}^*\left(L^1_{\tilde{\mathcal{H}}}, c\epsilon\right) \log^2\left(\frac{|\tilde{\mathcal{H}}|}{\epsilon}\right)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*\left(\tilde{\mathcal{H}}, c\epsilon\right) \log^2\left(\frac{|\tilde{\mathcal{H}}|}{\epsilon}\right)\right) \\
&\lesssim \exp\left(\mathrm{fat}^*(\mathcal{H}, c\epsilon) \log^2\left(\frac{|\tilde{\mathcal{H}}|}{\epsilon}\right)\right),
\end{aligned} \tag{5.20}$$

where $c \in (0,\infty)$ is a numerical constant, derived from the covering argument in Theorem 2.16.

4. Compute a robust variant of the real-valued boosting algorithm $\mathrm{MedBoost}$ [131, 140] on the discretized set $\bar{S}_{\mathcal{U}}$. The output of the algorithm is a uniformly 2ϵ-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. [131, 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 $d = \mathcal{O}\Big(\mathrm{fat}(\mathcal{H}, \epsilon/8) \log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\epsilon/8)}{\epsilon^2}\Big)\Big)$ and let $\psi(\bar{z},\bar{y})$ be the $\psi(x,y)$ for which $\bar{z} \in \mathcal{U}(x)$. 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 $\forall(\bar{z},\bar{y}) \in S'' : |h(\bar{z}) - \bar{y}| \leq \psi(\bar{z},\bar{y}) + \epsilon$, it holds that $\mathbb{P}_{(\bar{z},\bar{y})\sim\mathcal{P}}((\bar{z},\bar{y}) : |h(\bar{z}) - \bar{y}| > \psi(\bar{z},\bar{y}) + 2\epsilon) \leq 1/3$. That is, such a function is a $(2\epsilon, 1/6)$-weak learner for $\mathcal{P}$ and $h^\star$. 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, $S'' \subseteq \bigcup_{(x,y)\in S'} \bigcup\{(\bar{z},\bar{y}) : \bar{z} \in \mathcal{U}(x)\}$. Therefore, we can conclude that $\hat{h} = \psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S', \psi|_{S'}, \epsilon)$ 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 7 Robust MedBoost

Input: $\mathcal{H}, S, \bar{S}_{\mathcal{U}}$.
Parameters: Approximation parameter $\epsilon \in (0,1)$, number of boosting rounds $T \geq 1$, weak learner sample size $d \geq 1$, loss parameter $p \in [1,\infty]$, cutoff parameters $\psi|_S = (\psi(x_1,y_1),\ldots,\psi(x_m,y_m))$ for $(x_i,y_i), \in S$ and $\psi(\bar{z},\bar{y})$ is the $\psi(x,y)$ for which $\bar{z} \in \mathcal{U}(x)$.
Algorithms used: ϵ-approximate ψ-robust empirical risk minimizer ψ- $\mathrm{RERM}_{\mathcal{H}}$ (Eq. (5.5)).
Initialize $\mathcal{P}_1$ = Uniform($\bar{S}_{\mathcal{U}}$).
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 ψ- $\mathrm{RERM}_{\mathcal{H}}$ on them.

(a) Find d points $S''_t \subseteq \bar{S}_{\mathcal{U}}$ such that any $h \in \mathcal{H}$ satisfying: $\forall(\bar{z},\bar{y}) \in S''_t : \ |h(\bar{z}) - \bar{y}|^p \leq \psi(\bar{z},\bar{y})^p + \epsilon$, it holds that $\mathbb{E}_{(\bar{z},\bar{y})\sim\mathcal{P}_t}\left[\mathbb{I}\left\{|h(\bar{z}) - \bar{y}|^p \geq \psi(\bar{z},\bar{y})^p + 2\epsilon\right\}\right] \leq 1/3$. (See the analysis for why this set exists).

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

(c) Compute $\hat{h}_t = \psi\text{-}\mathrm{RERM}_{\mathcal{H}}(S'_t, \psi|_{S'_t}, \epsilon)$. From steps (a) and (b), it follows that $\hat{h}_t$ is a $(2\epsilon, 1/6)$-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 $i = 1,\ldots,n = \left|\bar{S}_{\mathcal{U}}\right|$:
Set
$$w_i^{(t)} = 1 - 2\mathbb{I}\left[\left|\hat{h}_t(\bar{z}_i) - \bar{y}_i\right|^p > \psi(\bar{z}_i,\bar{y}_i)^p + 2\epsilon\right].$$

(e) Set
$$\alpha_t = \frac{1}{2}\log\left(\frac{(1-1/6)\sum_{i=1}^n \mathcal{P}_t(\bar{z}_i,\bar{y}_i)\,\mathbb{I}\left[w_i^{(t)} = 1\right]}{(1+1/6)\sum_{i=1}^n \mathcal{P}_t(\bar{z}_i,\bar{y}_i)\,\mathbb{I}\left[w_i^{(t)} = -1\right]}\right).$$

(f) If $\alpha_t = \infty$: return T copies of h_t, $(\alpha_1 = 1,\ldots,\alpha_T = 1)$, and S'_t.
Else:
$$P_{t+1}(\bar{z}_i,\bar{y}_i) = P_t(\bar{z}_i,\bar{y}_i)\frac{\exp(-\alpha_t w_i^t)}{\sum_{j=1}^n \mathcal{P}_t(\bar{z}_j,\bar{y}_j)\exp\left(-\alpha_t w_j^t\right)}.$$

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

A uniformly 4ϵ-approximate adversarially robust sample compression scheme for S. The output of MedBoost is a uniformly 2ϵ-approximate sample compression scheme for the set $\bar{S}_{\mathcal{U}}$. We show that this is a uniformly 4ϵ-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. [131, Corollary 6] that the output of the algorithm satisfy

$$\forall(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}} : \ \left|\mathrm{Median}\left(\hat{h}_1(\bar{z}),\ldots,\hat{h}_T(\bar{z});\alpha_1,\ldots,\alpha_T\right) - \bar{y}\right| \leq \psi(\bar{z},\bar{y}) + 2\epsilon, \tag{5.21}$$

$\mathrm{Median}\left(\hat{h}_1(\bar{z}),\ldots,\hat{h}_T(\bar{z});\alpha_1,\ldots,\alpha_T\right)$ is the weighted median of $\hat{h}_1,\ldots,\hat{h}_T$ with weights $\alpha_1,\ldots,\alpha_T$. From the covering argument (Eq. (5.20)), this implies that

$$\forall(z,y) \in S_{\mathcal{U}} : \ \left|\mathrm{Median}\left(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\right) - y\right| \leq \psi(z,y) + 4\epsilon. \tag{5.22}$$

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

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

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

$$\begin{aligned}
\Big|\mathrm{Median}\big(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\big) - y\Big| &\overset{(a)}{=} \Big|\mathrm{Median}\big(\hat{h}_1(z)-y,\ldots,\hat{h}_T(z)-y;\alpha_1,\ldots,\alpha_T\big)\Big| \\
&\overset{(b)}{\leq} \Big|\mathrm{Median}\big(\hat{h}_1(\bar{z})-\bar{y},\ldots,\hat{h}_T(\bar{z})-\bar{y};\alpha_1,\ldots,\alpha_T\big)\Big| + \epsilon \\
&\overset{(c)}{=} \Big|\mathrm{Median}\big(\hat{h}_1(\bar{z}),\ldots,\hat{h}_T(\bar{z});\alpha_1,\ldots,\alpha_T\big) - \bar{y}\Big| + \epsilon \\
&\overset{(d)}{\leq} |h^\star(\bar{z})-\bar{y}| + 3\epsilon \\
&\overset{(e)}{\leq} |h^\star(z)-y| + 4\epsilon \\
&\overset{(f)}{=} \psi(z,y) + 4\epsilon,
\end{aligned} \tag{5.23}$$

(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 2ϵ-approximate sample compression for $\bar{S}_{\mathcal{U}}$, (f) follows from the definition in step 2 of the algorithm.

Finally, from Eq. (5.23) we conclude a uniformly 4ϵ-approximate *adversarially robust* sample compression scheme for S,

$$\forall (x,y) \in S: \quad \sup_{z \in \mathcal{U}(x)} \Big|\mathrm{Median}\big(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\big) - y\Big| \leq \psi(x,y) + 4\epsilon. \tag{5.24}$$

Bounding the compression size. We have $T = \mathcal{O}\big(\log\big|\bar{S}_{\mathcal{U}}\big|\big)$ hypotheses, where each one is representable by $d = \mathcal{O}\Big(\mathrm{fat}(\mathcal{H},\epsilon/8)\log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\epsilon/8)}{\epsilon^2}\Big)\Big)$ points. By counting the number of predictors using Eq. (5.20), we get

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

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. [131]. The idea is that by sampling functions from the ensemble, we can guarantee via a uniform convergence in the dual space, that it is sufficient to have roughly $\approx \mathrm{fat}^*(\mathcal{H},c\epsilon)$ 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 $\tau(x,y) = \psi(x,y)^p + 5\epsilon$ and $k = \mathcal{O}\big(\mathrm{fat}^*(\mathcal{H},c\epsilon)\log^2\big(\mathrm{fat}^*(\mathcal{H},c\epsilon)/\epsilon\big)\big)$, where $c \in (0,\infty)$ is a numerical constant.

Algorithm 8 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: Number of functions to sample $k \geq 1$, cutoff parameters $(\tau(x_1, y_1), \ldots, \tau(x_m, y_m))$.

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

Output: Hypotheses $\hat{h}_{J_1}, \ldots, \hat{h}_{J_k}$.

The sparsification method returns set of functions $\left\{ \hat{h}_{J_1}, \ldots, \hat{h}_{J_k} \right\}$, such that

$$\forall (x, y) \in S : \sup_{z \in \mathcal{U}(x)} \left| \mathrm{Median}\big(\hat{h}_{J_1}(x), \ldots, \hat{h}_{J_k}(x)\big) - y \right| \leq \psi(x, y) + 5\epsilon. \tag{5.25}$$

We get a uniformly 5ϵ-approximate *adversarially robust* sample compression scheme for S, where we have $k = \mathcal{O}\big(\mathrm{fat}^*(\mathcal{H}, c\epsilon) \log^2(\mathrm{fat}^*(\mathcal{H}, c\epsilon)\,/\epsilon)\big)$ functions, and each function is representable by $d = \mathcal{O}\big(\mathrm{fat}(\mathcal{H}, \epsilon/8) \log^2\big(\mathrm{fat}(\mathcal{H}, \epsilon/8)\,/\epsilon^2\big)\big)$ points.

Encoding weak learners. Each weak learner is encoded by a multiset $S' \subseteq S$ of size d and is constructed by computing some $\hat{h} \in \mathcal{H}$ that solves the constrained optimization

$$\sup_{z \in \mathcal{U}(x)} \left| \hat{h}(z) - y \right| \leq \psi(x, y) + \epsilon, \ \forall (x, y) \in S'.$$

We encode each $\psi(x, y)$ by some approximation $\tilde{\psi}(x, y)$, such that $\left| \tilde{\psi}(x, y) - \psi(x, y) \right| \leq \epsilon$, by discretizing $[0, 1]$ to $1/\epsilon$ buckets of size ϵ, and each $\psi(x, y)$ is rounded down to the closest value $\tilde{\psi}(x, y)$. Each approximation requires to encode $\log(1/\epsilon)$ bits, and so each learner encodes $d \log(1/\epsilon)$ bits and d samples. We have k weak learners, and the compression size is

$$k(d + d \log(1/\epsilon)) \leq 2kd \log(1/\epsilon)\,.$$

Therefore, we have a uniform 6ϵ-approximate compression for ℓ_1 and ℓ_∞ losses of size

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

Generalizing for $p \in (1, \infty)$. We rely on the Lipschitzness of the ℓ_p loss and rescaling the approximation parameter ϵ to ϵ/p.

Recall the covering of $S_{\mathcal{U}}$ in step 3(b). Note that an (ϵ/p)-cover for the L_1 loss class is an ϵ-cover for the L_p loss

class due to the Lipschitzness of the ℓ_p loss

$$\begin{aligned}\Big||h(z)-y|^p-|h(\bar{z})-\bar{y}|^p\Big| &\le p\Big||h(z)-y|-|h(\bar{z})-\bar{y}|\Big| \\ &\le p\epsilon.\end{aligned}$$

Moreover, we constructed a function $f=\text{Median}\big(\hat{h}_{J_1},\dots,\hat{h}_{J_k}\big)$ with $\sup_{z\in\mathcal{U}(x)}|f(z)-y|\le\psi(x,y)+\epsilon$ for any $(x,y)\in S$. Note that since $|f(\cdot)-y|\in[0,1]$, the same z that maximizes the ℓ_1 loss also maximizes for any ℓ_p. This implies that

$$\sup_{z\in\mathcal{U}(x)}|f(z)-y|^p \overset{(i)}{\le} ((\psi(x,y))+\epsilon)^p \overset{(ii)}{\le} \psi(x,y)^p+p\epsilon,$$

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

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

By rescaling $p\epsilon$ to ϵ, we get

$$\sup_{z\in\mathcal{U}(x)}\Big|\text{Median}\Big(\hat{h}_{J_1}(z),\dots,\hat{h}_{J_k}(z)\Big)-y\Big|^p\le\psi(x,y)^p+\epsilon,$$

where the number of functions in the ensemble is

$$k=\mathcal{O}\left(\text{fat}^*(\mathcal{H},c\epsilon/p)\log^2\left(\frac{p\,\text{fat}^*(\mathcal{H},c\epsilon/p)}{\epsilon}\right)\right),$$

and each function is represented by a set of samples of size

$$d=\mathcal{O}\left(\text{fat}(\mathcal{H},c\epsilon/p)\log^2\left(\frac{p^2\,\text{fat}(\mathcal{H},c\epsilon/p)}{\epsilon^2}\right)\right).$$

Therefore, we have an approximate compression for ℓ_p of size

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

Generalization bound. 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,

$$\begin{aligned}
\mathrm{Err}_{\ell_p}(\rho(\kappa(S)); \mathcal{D}) &\overset{(i)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_p}(\rho(\kappa(S)); S) + \sqrt{\frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m}} \\
&\overset{(ii)}{\lesssim} \widehat{\mathrm{Err}}_{\ell_p}(h^\star; S) + 6\epsilon + \sqrt{\frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m}} \\
&\overset{(iii)}{\lesssim} \mathrm{Err}_{\ell_p}(h^\star; \mathcal{D}) + 6\epsilon + \sqrt{\frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{m}} \\
&\lesssim \mathrm{Err}_{\ell_p}(h^\star; \mathcal{D}) + 6\epsilon + \sqrt{\frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m}},
\end{aligned}$$

(i) follows from a generalization of sample compression scheme in the agnostic case, see Theorem 2.19, (ii) follows from the approximate sample compression we proved above, (iii) follows from Hoeffding's inequality. We take m sufficiently large such that

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

Re-scale $\epsilon = \epsilon/6$ and plug in the compression size, we get sample complexity of size

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

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

Proofs for Section 5.5: Robust (η, β)-Regression

Realizable Setting

Proof (of Theorem 5.9). Fix $\epsilon, \delta, \beta, \eta \in (0, 1)$. Let $\mathcal{H} \subseteq [0, 1]^{\mathcal{X}}$. Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is η-uniformly realizable by $\mathcal{H}$, 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 5.

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 follows.

   (a) Define a set of functions, such that each function is defined by an β-approximate ψ–$\mathrm{RERM}_{\mathcal{H}}$ optimizer 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 from S, with fixed cutoff parameters $\psi(x, y) = \eta$ for each

$(x, y) \in S$,

$$\hat{\mathcal{H}} = \left\{\psi\text{-}\operatorname{RERM}_{\mathcal{H}}(S', \psi|_{S'}, \beta) : S' \subseteq S, |S'| = d\right\}.$$

Recall the definition of $\psi\text{-}\operatorname{RERM}_{\mathcal{H}}$, see Eq. (5.5). The cardinality of this class is bounded as follows

$$|\hat{\mathcal{H}}| \approx \binom{m}{d} \lesssim \left(\frac{m}{d}\right)^d. \tag{5.26}$$

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

$$\sup_{(z,y)\in S_{\mathcal{U}}} \inf_{(\bar{z},\bar{y})\in \bar{S}_{\mathcal{U}}} \left\|f_{(z,y)} - f_{(\bar{z},\bar{y})}\right\|_\infty \leq \beta. \tag{5.27}$$

Let $\operatorname{fat}^*\left(L^1_{\tilde{\mathcal{H}}}, \beta\right)$ be the dual β-fat-shattering of $L^1_{\tilde{\mathcal{H}}}$. Applying a covering number argument from Theorem 2.16 on the dual space and upper bounding the dual fat-shattering of the L_1 loss class with the dual fat-shattering of $\tilde{H}$, we have the following bound

$$\begin{aligned}
\left|\bar{S}_{\mathcal{U}}\right| &= \mathcal{N}\left(\beta, S_{\mathcal{U}}, \|\cdot\|_\infty\right) \\
&\lesssim \exp\left(\operatorname{fat}^*\left(L^1_{\tilde{\mathcal{H}}}, c\beta\right) \log^2\left(\frac{|\tilde{\mathcal{H}}|}{\beta}\right)\right) \\
&\lesssim \exp\left(\operatorname{fat}^*\left(\tilde{\mathcal{H}}, c\beta\right) \log^2\left(\frac{|\tilde{\mathcal{H}}|}{\beta}\right)\right) \\
&\lesssim \exp\left(\operatorname{fat}^*\left(\mathcal{H}, c\beta\right) \log^2\left(\frac{|\tilde{\mathcal{H}}|}{\beta}\right)\right),
\end{aligned} \tag{5.28}$$

where $c \in (0, \infty)$ is a numerical constant, derived from the covering argument in Theorem 2.16.

3. Compute a robust version of the real-valued boosting algorithm MedBoost (Algorithm 7) on the discretized set $\bar{S}_{\mathcal{U}}$. The inputs to the algorithm are as follows. Set $\epsilon = \beta$, $\psi|_S = (\eta, \ldots, \eta)$, $p = 1$, and $T \approx \log\left(\left|\bar{S}_{\mathcal{U}}\right|\right)$ rounds of boosting.

   The output of the algorithm is a uniform β-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. [131, 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 5.12, 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 $\mathcal{O}\left(\operatorname{fat}(\mathcal{H}, \beta/8) \log^2\left(\frac{\operatorname{fat}(\mathcal{H},\beta/8)}{\beta}\right)\right)$, 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 $(\beta, 1/6)$-weak learner for $\mathcal{P}$ (see Theorem 7.5). 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 $\mathcal{O}\left(\operatorname{fat}(\mathcal{H}, \beta/8) \log^2\left(\frac{\operatorname{fat}(\mathcal{H},\beta/8)}{\beta}\right)\right)$ 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 $\mathcal{O}\Big(\mathrm{fat}(\mathcal{H},\beta/8)\log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\beta/8)}{\beta}\Big)\Big)$ 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 $\hat{h} = \psi\text{-}\,\mathrm{RERM}_{\mathcal{H}}(S', \psi|_{S'}, \beta)$ 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 3β-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. [131, Corollary 6] that the output of the algorithm satisfy

$$\forall(\bar{z},\bar{y}) \in \bar{S}_{\mathcal{U}} : \ \Big|\mathrm{Median}\big(\hat{h}_1(\bar{z}),\ldots,\hat{h}_T(\bar{z});\alpha_1,\ldots,\alpha_T\big) - \bar{y}\Big| \leq \eta + \beta, \tag{5.29}$$

$\mathrm{Median}\big(\hat{h}_1(\bar{z}),\ldots,\hat{h}_T(\bar{z});\alpha_1,\ldots,\alpha_T\big)$ is the weighted median of $\hat{h}_1,\ldots,\hat{h}_T$ with weights $\alpha_1,\ldots,\alpha_T$. From the covering argument (Eq. (5.28)), this implies that

$$\forall(z,y) \in S_{\mathcal{U}} : \ \Big|\mathrm{Median}\big(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\big) - y\Big| \leq \eta + 3\beta. \tag{5.30}$$

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

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

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

$$\begin{aligned}
\Big|\mathrm{Median}\big(\hat{h}_1(z),\ldots,\hat{h}_T(z);\alpha_1,\ldots,\alpha_T\big) - y\Big| &\overset{(a)}{=} \Big|\mathrm{Median}\big(\hat{h}_1(z)-y,\ldots,\hat{h}_T(z)-y;\alpha_1,\ldots,\alpha_T\big)\Big| \\
&\overset{(b)}{\leq} \Big|\mathrm{Median}\big(\hat{h}_1(\bar{z})-\bar{y},\ldots,\hat{h}_T(\bar{z})-\bar{y};\alpha_1,\ldots,\alpha_T\big)\Big| + \beta \\
&\overset{(c)}{=} \Big|\mathrm{Median}\big(\hat{h}_1(\bar{z}),\ldots,\hat{h}_T(\bar{z});\alpha_1,\ldots,\alpha_T\big) - \bar{y}\Big| + \beta \\
&\overset{(d)}{\leq} |h^\star(\bar{z})-\bar{y}| + 2\beta \\
&\overset{(e)}{\leq} |h^\star(z)-y| + 3\beta \\
&\overset{(f)}{=} \eta + 3\beta,
\end{aligned} \tag{5.31}$$

(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 assumption of η-realizability.

Finally, from Eq. (5.31) we conclude a uniformly 3β-approximate *adversarially robust* sample compression scheme

for S,

$$\forall (x,y) \in S : \sup_{z \in \mathcal{U}(x)} \left| \mathrm{Median}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z); \alpha_1, \ldots, \alpha_T\big) - y \right| \leq \eta + 3\beta. \tag{5.32}$$

Bounding the compression size. We have $T = \mathcal{O}\big(\log\big|\bar{S}_{\mathcal{U}}\big|\big)$ hypotheses, where each one is representable by $d = \mathcal{O}\Big(\mathrm{fat}(\mathcal{H}, \beta/8) \log^2\Big(\frac{\mathrm{fat}(\mathcal{H},\beta/8)}{\beta^2}\Big)\Big)$ points. By counting the number of predictors using Eq. (5.28), we get

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

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 8). 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 $\approx \mathrm{fat}^*(\mathcal{H}, \mathcal{O}(\beta))$ predictors. Applying Hanneke et al. [131, Theorem 10] with the parameters $\tau(x,y) = \eta + 4\beta$ and $k = \mathcal{O}\big(\mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\big(\mathrm{fat}^*(\mathcal{H}, c\beta)/\beta\big)\big)$, where $c \in (0, \infty)$ is a numerical constant. The sparsification method returns set of functions $\left\{\hat{h}_{J_1}, \ldots, \hat{h}_{J_k}\right\}$, such that

$$\forall (x,y) \in S : \sup_{z \in \mathcal{U}(x)} \left| \mathrm{Median}\big(\hat{h}_{J_1}(x), \ldots, \hat{h}_{J_k}(x)\big) - y \right| \leq \eta + 4\beta.$$

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.

Encoding weak learners. Each weak learner is encoded by a multiset $S' \subseteq S$ of size d and is constructed by computing some $\hat{h} \in \mathcal{H}$ that solves the constrained optimization

$$\sup_{z \in \mathcal{U}(x)} \left| \hat{h}(z) - y \right| \leq \eta, \ \forall (x,y) \in S'.$$

We encode η by some approximation $\tilde{\eta}$, such that $|\tilde{\eta} - \eta| \leq \beta$, by discretizing $[0,1]$ to $1/\beta$ buckets of size β, and η is rounded down to the closest value $\tilde{\eta}$. The approximation requires to encode $\log(1/\beta)$ bits, and so each learner encodes $d \log(1/\beta)$ bits and d samples. We have k weak learners, and the compression size is

$$k(d + d\log(1/\beta)) \leq 2kd \log(1/\beta).$$

Therefore, we have a uniform 5β-approximate compression of size

$$\mathcal{O}\left(\mathrm{fat}(\mathcal{H}, \beta/8)\, \mathrm{fat}^*(\mathcal{H}, c\beta) \log^2\left(\frac{\mathrm{fat}(\mathcal{H}, \beta/8)}{\beta^2}\right) \log^2\left(\frac{\mathrm{fat}^*(\mathcal{H}, c\beta)}{\beta}\right) \log\frac{1}{\beta}\right).$$

Generalization bound. Let (κ, ρ) be the compression scheme and $|\kappa(S)|$ the compression size. Let $\widehat{\mathrm{Err}}_\eta(h; S)$ be the empirical loss of h on S with the η-ball robust loss. We can upper bound the error as follows,

$$\begin{aligned}
\mathrm{Err}_\eta(\rho(\kappa(S)); \mathcal{D}) &\overset{(i)}{\lesssim} \widehat{\mathrm{Err}}_\eta(\rho(\kappa(S)); S) + \frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m} \\
&\overset{(ii)}{\lesssim} \widehat{\mathrm{Err}}_\eta(h^\star; S) + 5\beta + \frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m} \\
&\overset{(iii)}{\lesssim} \mathrm{Err}_\eta(h^\star; \mathcal{D}) + 5\beta + \frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m} + \frac{\log \frac{1}{\delta}}{m} \\
&\lesssim \mathrm{Err}_\eta(h^\star; \mathcal{D}) + 5\beta + \frac{|\kappa(S)| \log(m) + \log \frac{1}{\delta}}{m},
\end{aligned}$$

(i) follows from a generalization of a sample compression scheme in the realizable case, see Theorem 2.19, (ii) follows from the approximate sample compression we proved above, (iii) follows from Hoeffding's inequality. We take m sufficiently large such that

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

By plugging it in the compression size and re-scaling β, we get a sample complexity of size

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

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

Agnostic Setting

Proof (of Theorem 5.10). The construction follows a reduction to the realizable case similar to [40], which is for the non-robust zero-one loss. Moreover, we use a margin-based analysis of MedBoost algorithm (see Kégl [140, Theorem 1]).

Denote $\Lambda_{\mathrm{RE}} = \Lambda_{\mathrm{RE}}(\eta, \beta, 1/3, 1/3, \mathcal{H}, \mathcal{U})$, 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 in order to find a maximal subset $S' \subseteq S$ with zero empirical robust loss (for the η-ball loss), such that $\inf_{h \in \mathcal{H}} \widehat{\mathrm{Err}}_\eta(h; S') = 0$. Now, Λ_{RE} samples suffice for weak robust learning for any distribution $\mathcal{D}$ on S'.

Compute the 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 $\hat{h} = \mathrm{Median}\big(\hat{h}_1(z), \ldots, \hat{h}_T(z); \alpha_1, \ldots, \alpha_T\big)$ satisfies $\widehat{\mathrm{Err}}_{\eta+\beta}\Big(\hat{h}; S'\Big) = 0$, 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 $\Lambda_{\mathrm{RE}} T$.

By plugging it into an agnostic sample compression bound Theorem 2.19, we have a sample complexity of $\tilde{\mathcal{O}}\Big(\frac{\Lambda_{\mathrm{RE}}}{\epsilon^2}\Big)$, which translates into $\tilde{\mathcal{O}}\Big(\frac{\mathrm{fat}(\mathcal{H}, c\beta)\mathrm{fat}^*(\mathcal{H}, c\beta)}{\epsilon^2}\Big)$, for some numerical constant $c \in (0, \infty)$. □

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 $\sqrt{\mathrm{OPT}_{\mathcal{H}}} + \epsilon$, where

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

Theorem 5.13 *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 $\eta, \beta, \epsilon, \delta \in (0,1)$, for some numerical constant $c \in (0, \infty)$, with probability $1 - \delta$, Section 5.7 outputs a function with error at most $\sqrt{\mathrm{OPT}_{\mathcal{H}}} + \epsilon$, by using a sample of size*

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

Fixed Cutoff Approach for Agnostic Robust Regression

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, ϵ)-regression (see Theorem 5.10): Agnostic-$(\eta + \epsilon)$-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 5.13). 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 $\eta = \sqrt{\mathrm{OPT}_{\mathcal{H}}}$,

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

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

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

Chapter 6

Fat-Shattering Dimension of k-fold Aggregations

We provide estimates on the fat-shattering dimension of aggregation rules of real-valued function classes. The latter consists of all ways of choosing k functions, one from each of the k classes, and computing pointwise an "aggregate" function of these, such as the median, mean, and maximum. The bounds are stated in terms of the fat-shattering dimensions of the component classes. For linear and affine function classes, we provide a considerably sharper upper bound and a matching lower bound, achieving, in particular, an optimal dependence on k. Along the way, we improve several known results in addition to pointing out and correcting a number of erroneous claims in the literature.

6.1 Introduction

The *fat-shattering dimension*, also known as "scale-sensitive" and the "parametrized variant of the P-dimension", was first defined by Kearns and Schapire [157]; its key role in learning theory lies in characterizing the PAC learnability of real-valued function classes [158, 159].

In this paper, we study the behavior of the fat-shattering dimension under various k-fold aggregations. Let $F_1, \ldots, F_k \subseteq \mathbb{R}^{\mathcal{X}}$ be real-valued function classes, and $G : \mathbb{R}^k \to \mathbb{R}$ be an aggregation rule. We consider the *aggregate* function class $G(F_1, \ldots, F_k)$, which consists of all mappings $x \mapsto G(f_1(x), \ldots, f_k(x))$, for any $f_1 \in F_1, \ldots, f_k \in F_k$. Some natural aggregation rules include the pointwise k-fold maximum, median, mean, and max-min. We seek to bound the fat-shattering complexity of $G(F_1, \ldots, F_k)$ in terms of the fat-shattering dimensions of the constituent F_is. This question naturally arises in the context of ensemble methods, such as boosting and bagging, where the learner's prediction consists of an aggregation of base learners.

The analogous question regarding aggregations of VC classes (VC dimension being the combinatorial complexity controlling the learnability of Boolean function classes) have been studied in detail and largely resolved [98, 160–

163]. Furthermore, closure properties were also studied in the context of online classification and private PAC learning [164, 165] for the Littlestone and Threshold dimensions. However, for real-valued functions, this question remained largely uninvestigated.

Our Contributions

- For a natural class of aggregation rules that commute with shifts (see definition (6.7)) and commute with truncation (see definition (6.20)), assuming $\mathrm{fat}_\gamma(F_i) \leq d$, for $1 \leq i \leq k$, we show that

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \leq O\left(dk \log^2(dk)\right), \qquad \gamma > 0.$$

  In particular, this result holds for the maximum, minimum, median, and max-min aggregations. The formal statement is given in Theorem 6.1.

- By using an entirely different approach, for aggregations that are L-Lipschitz ($L \geq 1$) in supremum norm (see definition (6.8)) and for bounded function classes $F_1, \ldots, F_k \subset [-R, R]^\Omega$ with $\mathrm{fat}_{\epsilon\gamma}(F_i) \leq d$, for $1 \leq i \leq k$, we show that

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \leq O\left(dk \log^{1+\epsilon}\frac{LRk}{\gamma}\right), \qquad 0 < \gamma/L < R \text{ and } 0 < \epsilon < \log 2.$$

  In particular, this result holds for the maximum, minimum, median, mean, and max-min aggregations. The formal statement is given in Theorem 6.3.

- For R-bounded affine functions and for aggregations that are L-Lipschitz in supremum norm, we show the following dimension-free bound,

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \leq \quad O\left(\frac{L^2R^2k\log(k)}{\gamma^2}\right), \qquad 0 < \gamma/L < R.$$

  This result also extends to the hinge-loss class of affine functions. In particular, this result holds for the maximum, minimum, median, mean, and max-min aggregations. We improve by a log factor the estimate of Fefferman et al. [166, Lemma 6] on the fat-shattering dimension of max-min aggregation of linear functions. The formal statement is given in Theorem 6.5

  Furthermore, in Corollary 6.7 we show an upper bound on the Rademacher complexity of the k-fold maximum aggregation of affine functions and hinge-loss affine functions. Our bound scales with $\sqrt{k}$, improving upon Raviv et al. [167] where the dependence on k is linear.

- For affine functions and the k-fold maximum aggregation, we show tight dimension-dependent bounds (up to constants),

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad = \quad \Theta\left(dk \log k\right), \qquad \gamma > 0,$$

  where d is the Euclidean dimension. For the formal statements, see Theorems 6.9 and 6.10.

Applications

The need to analyze the combinatorial complexity of a k-fold maximum of function classes (see (6.4) for the formal definition) arises in a number of diverse settings. One natural example is adversarially robust PAC learning to test time attacks for real-valued functions [1, 3]. In this setting, the learner observes an i.i.d. labeled sample from an unknown distribution, and the goal is to output a hypothesis with a small error on unseen examples from the same distribution, with high probability. The difference from the standard PAC learning model is that at test time, the learner only observes a corrupted example, while the prediction is tested on the original label. Formally, (x, y) is drawn from the unknown distribution, and there is an adversary that can map x to k possible corruptions z that are known to the learner. The learner observes only z while its loss is with respect to the original label y. This scenario is naturally captured by the k-fold max: the learner aims to learn the maximum aggregation of the loss classes. Attias et al. [1] showed that uniform convergence holds in this case, and so the sample complexity of an empirical risk minimization algorithm is determined by the complexity measure of the k-fold maximum aggregation.

Analyzing the k-fold maximum arises also in a setting of learning polyhedra with a margin. Gottlieb et al. [168] provided a learning algorithm that represents polyhedra as intersections of bounded affine functions. The sample complexity of the algorithm is determined by the complexity measure of the maximum aggregation of affine function classes.

Another natural example of where the k-fold maximum and k-fold max-min play a role is in analyzing the convergence of k-means clustering. Fefferman et al. [166] bounded the max-min aggregation and Klochkov et al. [169], Biau et al. [170], Appert and Catoni [171], Zhivotovskiy [172] bounded the max aggregation. The main challenge in this setting is bounding the covering numbers of the aggregation over k function classes which can be obtained by bounding the Rademacher complexity or the fat-shattering dimension.

Finally, there are numerous ensemble methods for regression that output some aggregation of base learners, such as the median or mean. Examples of these methods include boosting (e.g., Kégl [140], Freund and Schapire [156]), bagging (bootstrap aggregation) by Breiman [173], and its extension to the random forest algorithm [174].

Related Work

It was claimed in Attias et al. [175, Theorem 12] that

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \leq \quad 2\log(3k) \sum_{j=1}^{k} \mathrm{fat}_\gamma(F_i),$$

but the proof had a mistake (see Section 6.5); our Open Problem (6.27) asks if the general form of the bound does hold (we conjecture it does at least for the max aggregation). Using the recent disambiguation result of Alon et al. [176] presented in Lemma 6.11 here, Attias et al. [1, Lemma 15] obtained the bound

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \leq O\left(\log(k)\log^2(|\mathcal{X}|)\sum_{j=1}^{k} \mathrm{fat}_\gamma(F_i)\right), \tag{6.1}$$

where Ω is the domain of the function classes $F_1, \ldots, F_k$. The latter is, in general, incomparable to Theorem 6.1. However, for large or infinite $\mathcal{X}$, Theorem 6.1 is clearly a considerable improvement over (6.1).

Using the covering number results of Mendelson and Vershynin [35], Talagrand [177] (see Section 6.6), Duan [178, Theorem 6.2] obtained a general result, which, when specialized to k-fold maxima, yields

$$\mathrm{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \leq \quad O\left(\log \frac{k}{\gamma} \cdot \sum_{i=1}^{k} \mathrm{fat}_{c\gamma/\sqrt{k}}(F_i)\right) \tag{6.2}$$

for a universal constant $c > 0$; (6.2) is an immediate consequence of Theorem 6.12 (with $p = 2$), Lemma 6.21, and Lemma 6.23 in this paper. Our results improve over (6.2) by removing the dependence on k in the scale of the fat-shattering dimensions; however, Duan's general method is applicable to a wider class of uniformly continuous k-fold aggregations.

Srebro et al. [179, Lemma A.2] bounded the fat-shattering dimension in terms of the Rademacher complexity. Foster and Rakhlin [180] bounded the Rademacher complexity of a smooth k-fold aggregate, see also references therein. Inspired by Appert and Catoni [171], Zhivotovskiy [172] has obtained the best known upper bound on the Rademacher complexity of k-fold maxima over linear function classes. Raviv et al. [167] upper bounded the Rademacher complexity of the k-fold maximum aggregation of affine functions and hinge-loss affine functions.

6.2 Preliminaries

Aggregation Rules

A k-fold *aggregation* rule is any mapping $G : \mathbb{R}^k \to \mathbb{R}$. Just as G maps k-tuples of reals into reals, it naturally aggregates k-tuples of functions into a single one: for $f_1, \ldots, f_k : \Omega \to \mathbb{R}$, we define $G(f_1, \ldots, f_k) : \Omega \to \mathbb{R}$ as the mapping $x \mapsto G(f_1(x), \ldots, f_k(x))$. Finally, the aggregation extends to k-tuples of function classes: for $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$, we define

$$G(F_1, \ldots, F_k) := \{x \mapsto G(f_1(x), \ldots, f_k(x)) : f_i \in F_i, i \in [k]\}. \tag{6.3}$$

Examples. A canonical example of an aggregation rule is the k-fold max, induced by the mapping

$$G_{\max}(x_1, \ldots, x_k) := \max_{i \in [k]} x_i. \tag{6.4}$$

The minimum is defined analogously as

$$G_{\min}(x_1, \ldots, x_k) := \min_{i \in [k]} x_i.$$

The mean aggregation is defined as

$$G_{\mathrm{mean}}(x_1, \ldots, x_k) := \frac{1}{k} \sum_{i=0}^{k} x_i.$$

Denoting by $x_{(1)}, \ldots, x_{(k)}$ the ascending order of a sequence $x_1, \ldots, x_k$, that is, $x_{(1)} \leq \cdots \leq x_{(k)}$, the (lower[1]) median is defined as

$$G_{\text{Median}}(x_1, \ldots, x_k) := x_{(\lceil k/2 \rceil)}. \tag{6.5}$$

We also define $G_{\text{max-min}} : \mathbb{R}^{k \times \ell} \to \mathbb{R}$ as

$$G_{\text{max-min}}(x_{11}, \ldots, x_{k\ell}) := \max_{j \in [\ell]} \min_{i \in [k]} x_{ij}; \tag{6.6}$$

Next, we consider some properties that an aggregation rule might possess.

Commuting with shifts. We say that an aggregation rule G commutes with shifts if

$$G(x) - r = G(x - r), \qquad x \in \mathbb{R}^k, r \in \mathbb{R}, \tag{6.7}$$

where $x - r$ is defined as $(x_1 - r, \ldots, x_k - r)$ for $x = (x_1, \ldots, x_k)$. It is readily verified that the maximum, minimum, max-min, mean, and median commute with shifts.

Lipschitz continuity. The mapping $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz with respect to $\|\cdot\|_\infty$ if

$$\left|G(x) - G(x')\right| \leq L \left\|x - x'\right\|_\infty = L \max_{i \in [k]} |x_i - x'_i|, \qquad x, x' \in \mathbb{R}^k. \tag{6.8}$$

In Section 6.6, we show that maximum, median, and max-min aggregations are 1-Lipschitz (Lemmas 6.16, 6.17, 6.18 respectively). Showing it for the mean is a simple exercise. The proof for the minimum is similar to the one for the maximum.

We also consider aggregations that *commute with truncation*; see Section 6.4 for the formal definition.

Complexity Measures

Fat-shattering dimension at zero. As in Gottlieb et al. [181], we also define the notion of γ-shattering at 0, where the "shift" r in (2.1) is constrained to be 0. Formally, the shattering condition is

$$\min_{y \in \{-1,1\}^m} \sup_{f \in F} \min_{i \in [m]} y_i f(x_i) \geq \gamma,$$

and we denote the corresponding dimension by $\text{fat}_\gamma(F)$.

[1] Ordinarily, for even k, any $m \in [x_{(k/2)}, x_{(k/2+1)}]$ is a median of x. For the proof of Theorem 6.1, the median must be a value actually occurring in x.

Attias et al. [175, Lemma 13] showed that for all $F \subset \mathbb{R}^{\mathcal{X}}$,

$$\mathrm{fat}_\gamma(F) = \max_{r \in \mathbb{R}^{\mathcal{X}}} \mathrm{f\mathring{a}t}_\gamma(F - r), \qquad \gamma > 0, \tag{6.9}$$

where $F - r = \{f - r; f \in F\}$ is the r-shifted class (the maximum is always achieved). Lemma 6.28 presents another, apparently novel, connection between fat and fåt.

Covering numbers. We start with some background on covering numbers. Whenever $\mathcal{X}$ is endowed with a probability measure μ, this induces, for $p \in [1, \infty)$ and $f : \Omega \to \mathbb{R}^k$, the norm

$$\|f\|^p_{L_p^{(k)}(\mu)} = \mathbb{E}_{X \sim \mu} \|f(X)\|_p^p = \int_{\mathcal{X}} \|f(x)\|_p^p \, \mathrm{d}\mu(x)$$

on $L_p^{(k)}(\mu) := \left\{ f \in (\mathbb{R}^k)^{\mathcal{X}} : \|f\|_{L_p^{(k)}(\mu)} < \infty \right\}$. When $k = 1$, we write $L_p(\mu) := L_p^{(1)}(\mu)$. For $p = \infty$, $\|f\|_{L_\infty^{(k)}(\mu)}$ is the essential supremum of f with respect to μ. For $t > 0$ and $H \subset F \subset L_p(\mu)$, we say that H is a t-cover of F under $\|\cdot\|_{L_p(\mu)}$ if $\sup_{f \in F} \inf_{h \in H} \|f - h\|_{L_p(\mu)} \leq t$. The t-covering number of F, denoted by $\mathcal{N}(F, L_p(\mu), t)$, is the cardinality of the smallest t-cover of F (possibly, ∞). We note the obvious relation

$$p > q \quad \Longrightarrow \quad \mathcal{N}(F, L_p(\mu), t) \geq \mathcal{N}(F, L_q(\mu), t), \tag{6.10}$$

which holds for all probability measures μ and all $t > 0$.

We sometimes overload the notation about aggregations by defining G on k-tuples of functions (instead of k-tuples of reals), $G : (\mathbb{R}^\Omega)^k \to \mathbb{R}^\Omega$. We say that G is L-Lipschitz with respect to $\|\cdot\|_{L_p^{(k)}(\mu)}$, if

$$\left\| G(f_{1:k}) - G(f'_{1:k}) \right\|_{L_p(\mu)} \leq L \left\| (f_{1:k}) - (f'_{1:k}) \right\|_{L_p^{(k)}(\mu)}, \qquad f_{1:k}, f'_{1:k} \in (\mathbb{R}^k)^{\mathcal{X}}.$$

Notation

We write $\mathbb{N} = \{0, 1, \ldots\}$ to denote the natural numbers. For $n \in \mathbb{N}$, we write $[n] := \{1, 2, \ldots, n\}$. All of our logarithms are base e, unless explicitly denoted otherwise. We use $\max\{u, v\}$ and $u \vee v$ interchangeably, and write $\mathrm{Log}(x) := \log(\mathrm{e} \vee x)$. For any function class F over a set $\mathcal{X}$ and $E \subset \mathcal{X}$, $F(E) = F|_E$ denotes the projection on (restriction to) E. In line with the common convention in functional analysis, absolute numerical constants will be denoted by letters such as C, c, whose value may change from line to line. Any transformation $\varphi : \mathbb{R} \to \mathbb{R}$ may be applied to a function $f \in \mathbb{R}^{\mathcal{X}}$ via $\varphi(f) := \varphi \circ f$, as well as to $F \subset \mathbb{R}^{\mathcal{X}}$ via $\varphi(F) := \{\varphi(f); f \in F\}$. The sign function thresholds at 0: $\mathrm{sign}(t) = \mathbb{1}[t \geq 0]$.

6.3 Main Results

Our main results involve upper-bounding the fat-shattering dimension of aggregation rules in terms of the dimensions of the component classes. We begin with the simplest (to present):

Theorem 6.1 (General F and G that commutes with shifts and truncation) *For $F_1, \ldots, F_k \subseteq \mathbb{R}^{\mathcal{X}}$, and an aggregation rule G that commutes with shifts, (see definition (6.7)) and commutes with truncation (see definition (6.20)), we have*

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \leq \quad 35 D_\gamma \log^2(126 D_\gamma), \qquad \gamma > 0,$$

where $D_\gamma := \sum_{i=1}^k \mathrm{fat}_\gamma(F_i) > 0$. In the degenerate case where $D_\gamma = 0$, $\mathrm{fat}_\gamma(G) = 0$.

In particular, this result holds for the maximum, minimum, max-min, and median aggregation rules.

Remark 6.2 *We made no attempt to optimize the constants; these are only provided to give a rough order-of-magnitude sense. In the sequel, we forgo numerical estimates and state the results in terms of unspecified universal constants.*

The next result provides an alternative bound based on an entirely different technique:

Theorem 6.3 (Bounded function classes and Lipschitz aggregations) *For $0 < \epsilon < \log 2$, $F_1, \ldots, F_k \subseteq [-R, R]^{\mathcal{X}}$, and an aggregation rule G that is L-Lipschitz ($L \geq 1$) in supremum norm (see definition (6.8)), we have*

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \leq \quad C D \, \mathrm{Log}^{1+\epsilon} \frac{LRk}{\gamma}, \qquad 0 < \gamma/L < R,$$

where

$$D \quad = \quad \sum_{i=1}^k \mathrm{fat}_{c\epsilon\gamma}(F_i)$$

and $C, c > 0$ are universal constants. In particular, this result holds for natural aggregation rules, such as maximum, minimum, max-min, mean, and median.

Remark 6.4 *The bounds in Theorems 6.1 and 6.3 are, in general, incomparable—and not just because of the unspecified constants in the latter. One notable difference is that Theorem 6.1 only depends on the shattering scale γ, while Theorem 6.3 additionally features a (weak) explicit dependence on the aspect ratio R/γ. In particular, Theorem 6.1 is applicable to semi-bounded affine classes (see Section 6.6), while Theorem 6.3 is not. Still, for fixed R, γ and large k, the latter presents a significant asymptotic improvement over the former.*

For the special case of affine functions and hinge-loss affine functions, the technique of Theorem 6.3 yields a considerably sharper estimate:

Theorem 6.5 (Dimension-free bound for Lipschitz aggregations of affine functions) *Let $B \subset \mathbb{R}^d$ be the d-dimensional Euclidean unit ball and*

$$F_i = \left\{ x \mapsto w \cdot x + b; \|w\| \vee |b| \leq R_i, w \in \mathbb{R}^d, b \in \mathbb{R} \right\}, \qquad R_i \in \mathbb{R},\ i \in [k], \tag{6.11}$$

be k collections of R_i-bounded affine functions on $\mathcal{X} = B$ and G be an aggregation rule that is L-Lipschitz in supremum

norm (see definition (6.8)). Then

$$\mathrm{fat}_\gamma(G(F_1,\dots,F_k)) \quad \leq \quad \frac{CL^2\,\mathrm{Log}(k)}{\gamma^2}\sum_{i=1}^{k} R_i^2, \qquad 0<\gamma/L<\min_{i\in[k]} R_i, \tag{6.12}$$

where $C>0$ is a universal constant. Further, if

$$F_i^{\mathrm{Hinge}} = \{(x,y)\mapsto \max\{0,1-yf(x)\}\,;\, f\in F_i\} \tag{6.13}$$

is a family of R_i-bounded hinge-loss affine functions for $i\in[k]$ and $G_{\mathrm{Hinge}}\equiv G(F_1^{\mathrm{Hinge}},\dots,F_k^{\mathrm{Hinge}})$ is an aggregation rule that is L-Lipschitz in supremum norm, then the same bound as in (6.12) holds for $\mathrm{fat}_\gamma(G_{\mathrm{Hinge}})$.

In particular, this result holds for the maximum, minimum, max-min, mean, and median aggregation rules.

Theorem 6.5 improves by a log factor the estimate of Fefferman et al. [166], on the fat-shattering dimension of max-min aggregation (defined in Section 6.2) of linear functions:[2]

Lemma 6.6 (Fefferman et al. [166], Lemma 6) *Let $B\subset\mathbb{R}^d$ be the d-dimensional Euclidean unit ball and*

$$F_{ij} = \left\{x\mapsto w\cdot x; \|w\|\leq\|1\|, w\in\mathbb{R}^d\right\}, \qquad i\in[k], j\in[\ell],$$

be $k\ell$ (identical) linear function classes defined on $\mathcal{X}=B$. If $G_{\text{max-min}}$ is the max-min aggregation rule (6.6), then

$$\mathrm{fat}_\gamma(G_{\text{max-min}}(F_{11},\dots,F_{k\ell})) \quad \leq \quad C\frac{k\ell}{\gamma^2}\,\mathrm{Log}^2\left(\frac{k\ell}{\gamma^2}\right),$$

where $C>0$ is a universal constant.

Our Theorem 6.5 improves the latter by a log factor:

$$\mathrm{fat}_\gamma(G_{\text{max-min}}(F_{11},\dots,F_{k\ell})) \quad \leq \quad C\frac{k\ell\log(k\ell)}{\gamma^2}.$$

Corollary 6.7 (Rademacher complexity for k-Fold Maximum of Affine Functions) *Let F_i be an R_i-bounded affine function class as in (6.11) or a hinge loss affine function class as in (6.13), let $G_{\max}$ be the maximum aggregation rule, and let $\tilde{R}=\max_i R_i$, then*

$$\mathcal{R}_n(G_{\max}(F_1,\dots,F_k)) \leq C\sqrt{\frac{\mathrm{Log}(k)\,\mathrm{Log}^3(\tilde{R}n)\tilde{R}^2\sum_{i=1}^k R_i^2}{n}}.$$

where $\mathcal{R}_n$ is the Rademacher complexity and $C>0$ is a universal constant.

Corollary 6.7 improves upon Raviv et al. [167, Theorem 7]. Their upper bound scales linearly with k, whereas ours scales as $\sqrt{k\log k}$.

[2]The max-min aggregation is shown to be 1-Lipschitz in supremum norm in Lemma 6.18 of Section 6.6.

Note, however, that for linear classes a better bound is known:

Theorem 6.8 (Zhivotovskiy [172]) *Let $B \subset \mathbb{R}^d$ be the d-dimensional Euclidean unit ball and*

$$F_i = \left\{ x \mapsto w \cdot x; \|w\| \leq 1, w \in \mathbb{R}^d \right\}, \qquad i \in [k]$$

be k (identical) linear function classes defined on $\mathcal{X} = B$. If $G_{\max}$ is the maximum aggregation rule, then

$$\mathcal{R}_n(G_{\max}(F_1, \ldots, F_k)) \leq C \log\left(\frac{n}{k}\right) \sqrt{\frac{k \log k}{n}},$$

where $\mathcal{R}_n$ is the Rademacher complexity and $C > 0$ is a universal constant.

The estimate in Theorem 6.5 is *dimension-free* in the sense of being independent of d. In applications where a dependence on d is admissible, an optimal bound can be obtained:

Theorem 6.9 (Dimension-dependent bound for k-fold maximum of affine functions) *Let $\mathcal{X} = \mathbb{R}^d$ and $F_i \subset \mathbb{R}^{\mathcal{X}}$ be k (identical) function classes consisting of all real-valued affine functions:*

$$F_i = \left\{ x \mapsto w \cdot x + b; w \in \mathbb{R}^d, b \in \mathbb{R} \right\}, \qquad i \in [k]$$

and let $G_{\max}$ be the k-fold maximum (see definition (6.4)). Then

$$\operatorname{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \leq \quad Cdk \operatorname{Log} k, \qquad \gamma > 0,$$

where $C > 0$ is a universal constant.

The optimality of the upper bound in Theorem 6.9 is witnessed by the matching lower bound:

Theorem 6.10 (Dimension-dependent lower bound for k-fold maximum of affine functions) *For $k \geq 1$ and $d \geq 4$, let $F_1 = F_2 = \ldots = F_k$ be the collection of all affine functions over $\mathcal{X} = \mathbb{R}^d$ and let $G_{\max}$ be the k-fold maximum (see definition (6.4)). Then*

$$\operatorname{fat}_\gamma(G_{\max}(F_1, \ldots, F_k)) \quad \geq \quad C \log(k) \sum_{i=1}^{k} \operatorname{fat}_\gamma(F_i) = Cdk \log k, \qquad \gamma > 0,$$

where $C > 0$ is a universal constant.

The scaling argument employed in the proof of Theorem 6.10 can be invoked to show that the claim continues to hold for $\mathcal{X} = B$.

Together, Theorems 6.9 and 6.10 show that the dependence on k is optimal.

6.4 Proofs

We start with upper-bounding the fat-shattering dimension of aggregation rules that commute with shifts (definition (6.7)) and commute with truncation (defined below), in terms of the dimensions of the component classes.

Proof of Theorem 6.1

Partial concept classes and disambiguation. We say that $F^\star \subseteq \{0,1,\star\}^{\mathcal{X}}$ is a *partial* concept class over Ω; this usage is consistent with Alon et al. [176], while Attias et al. [1, 175] used the descriptor *ambiguous*. Define the *disambiguation operator* $\mathscr{D} : \{0,1,\star\} \to 2^{\{0,1\}}$ as

$$\mathscr{D}(0) = \{0\}\,; \quad \mathscr{D}(1) = \{1\}\,; \quad \mathscr{D}(\star) = \{0,1\}\,. \tag{6.14}$$

For $f^\star \in F^\star$, define its *disambiguation set* $\mathscr{D}(f^\star) \subseteq \{0,1\}^{\mathcal{X}}$ as

$$\mathscr{D}(f^\star) = \Big\{ g \in \{0,1\}^{\mathcal{X}} : \forall x \in \mathcal{X},\ g(x) \in \mathscr{D}(f^\star(x)) \Big\}\,; \tag{6.15}$$

in words, $\mathscr{D}(f^\star)$ consists of the *total* concepts $g : \mathcal{X} \to \{0,1\}$ that agree pointwise with $f^\star$, whenever the latter takes a value in $\{0,1\}$. We say that $\bar{F} \subseteq \{0,1\}^{\mathcal{X}}$ disambiguates $F^\star$ if for all $f^\star \in F^\star$, we have $\bar{F} \cap \mathscr{D}(f^\star) \neq \emptyset$; in words, every $f^\star \in F^\star$ must have a disambiguated representative in $\bar{F}$.[3]

As in Attias et al. [1], Alon et al. [176], we say[4] that $S \subset \mathcal{X}$ is VC-shattered by $F^\star$ if $F^\star(S) \supseteq \{0,1\}^S$. We write $\mathrm{VC}(F^\star)$ to denote the size of the largest VC-shattered set (possibly, ∞). The obvious relation $\mathrm{VC}(F^\star) \leq \mathrm{VC}(\bar{F})$ always holds between a partial concept class and any of its disambiguations. Alon et al. [176, Theorem 13] proved the following variant of the Sauer-Shelah-Perles Lemma for partial concept classes:

Lemma 6.11 (Alon et al. [176]) *For every $F^\star \subseteq \{0,1,\star\}^{\mathcal{X}}$ with $d = \mathrm{VC}(F^\star) < \infty$ and $|\Omega| < \infty$, there is an $\bar{F}$ disambiguating $F^\star$ such that*

$$|\bar{F}(\mathcal{X})| \quad \leq \quad (|\mathcal{X}|+1)^{(d+1)\log_2 |\mathcal{X}|+2}. \tag{6.16}$$

For $d > 0$ and $|\mathcal{X}| > 1$, this implies the somewhat more wieldy estimate[5]

$$|\bar{F}(\mathcal{X})| \quad \leq \quad |\mathcal{X}|^{7d\log_2 |\mathcal{X}|}. \tag{6.17}$$

We will make use of the elementary fact

$$x \leq A\log_2 x \quad \Longrightarrow \quad x \leq 3A\log(3A), \qquad x, A \geq 1$$

[3] Attias et al. [1] additionally required that $\bar{F} \subseteq \bigcup_{f^\star \in F^\star} \mathscr{D}(f^\star)$, but this is an unnecessary restriction, and does not affect any of the results.

[4] Attias et al. [175] had incorrectly given $F^\star(S) = \{0,1\}^S$ as the shattering condition.

[5] The estimate (6.17) does not appear in Alon et al. [176], but is an elementary consequence of (6.16).

and its corollary

$$y \le A(\log_2 y)^2 \quad \Longrightarrow \quad y \le 5A\log^2(18A), \qquad y, A \ge 1. \tag{6.18}$$

Aggregation rules commuting with truncation. Fix $\gamma > 0$ and define the truncation operator $[\cdot]^\star_\gamma : \mathbb{R} \to \{0, 1, \star\}$ as

$$[t]^\star_\gamma \quad = \quad \begin{cases} 0, & t \le -\gamma \\ 1, & t \ge \gamma \\ \star, & \text{else.} \end{cases} \tag{6.19}$$

Let $x_i \in \mathbb{R}$, $i \in [k]$. Let the γ-truncation $[x_i]^\star_\gamma \in \{0, 1, \star\}$, and $\bar{x}_i \in \mathscr{D}([x_i]^\star_\gamma) \subseteq \{0, 1\}$ be a disambiguation. We say that an aggregation rule $G : \mathbb{R}^k \to \mathbb{R}$ *commutes with truncation* if for any $\gamma > 0$,

$$G(\bar{x}_1, \dots, \bar{x}_k) \in \mathscr{D}([G(x_1, \dots, x_k)]^\star_\gamma) \tag{6.20}$$

for *all* disambiguations $\bar{x}_i$, $i \in [k]$ (see definitions in (6.14) and (6.15)). In Section 6.6, we show that median and max-min aggregations commute with truncations (Lemmas 6.19, 6.20 respectively). Showing it for the maximum and minimum is a simple exercise. We note that the mean aggregation does not satisfy this property.

Proof (of Theorem 6.1). We follow the basic techniques of discretization and r-shifting, employed in Attias et al. [1, 175].

Fix $\gamma > 0$, recall the truncation operator $[\cdot]^\star_\gamma : \mathbb{R} \to \{0, 1, \star\}$ defined in (6.19). We also define the truncation operator over functions $[\cdot]^\star_\gamma : \mathbb{R}^\Omega \to \{0, 1, \star\}^\Omega$, as $[f]^\star_\gamma = f^\star$ where $f^\star(x) = [f(x)]^\star_\gamma$, for $x \in \Omega$. Observe that for all $F \subseteq \mathbb{R}^\mathcal{X}$ and $[F]^\star_\gamma := \big\{[f]^\star_\gamma; f \in F\big\}$, we have $\mathring{\text{fat}}_\gamma(F) = \text{VC}([F]^\star_\gamma)$. Let $G : \mathbb{R}^k \to \mathbb{R}$ be a k-fold aggregation rule and $F_1, \dots, F_k \subseteq \mathbb{R}^\Omega$ be real-valued function classes. Suppose that some $S = \{x_1, \dots, x_\ell\} \subset \mathcal{X}$ is γ-shattered by $G \equiv G(F_1, \dots, F_k)$. Proving the claim amounts to upper-bounding ℓ appropriately. By (6.9), there is an $r \in \mathbb{R}^\mathcal{X}$ such that $\text{fat}_\gamma(G) = \mathring{\text{fat}}_\gamma(G - r) = \text{VC}([G - r]^\star_\gamma)$. Put $F'_i := F_i - r$ and since G commutes with r-shift, as defined in (6.7), we have

$$G' := G(F'_1, \dots, F'_k) = G(F_1 - r, \dots, F_k - r) = G(F_1 \dots, F_k) - r. \tag{6.21}$$

Hence, S is VC-shattered by $[G']^\star_\gamma$ and

$$v_i := \text{VC}([F'_i]^\star_\gamma) = \mathring{\text{fat}}_\gamma(F'_i) \le \text{fat}_\gamma(F'_i) = \text{fat}_\gamma(F_i), \qquad i \in [k]. \tag{6.22}$$

Let us assume for now that each $v_i > 0$; in this case, there is no loss of generality in assuming $\ell > 1$. Let $\bar{F}_i$ be a "good" disambiguation of $[F'_i]^\star_\gamma$ on S, as furnished by Lemma 6.11:

$$|\bar{F}_i(S)| \quad \le \quad \ell^{7v_i \log_2 \ell}.$$

Observe that $\bar{G} := G(\bar{F}_1, \ldots, \bar{F}_k)$ is a valid disambiguation of $[G']^\star_\gamma$ since we assume that G commutes with truncation. It follows that

$$2^\ell \;=\; |\bar{G}(S)| \;\leq\; \prod_{i=1}^{k} |\bar{F}_i(S)| \;\leq\; \ell^{7\log_2 \ell \sum_{i=1}^k v_i}. \tag{6.23}$$

Thus, (6.18) implies that $\ell \leq 35(\sum v_i)\log^2(126\sum v_i)$, and the latter is an upper bound on $\mathrm{VC}(\bar{G})$ — and hence, also on $\mathrm{VC}([G']^\star_\gamma) = \mathrm{fat}_\gamma(G)$. The claim now follows from (6.22).

If any one given $v_i = 0$, we claim that (6.23) is unaffected. This is because any $C^\star \subset \{0, 1, \star\}^{\mathcal{X}}$ with $\mathrm{VC}(C^\star) = 0$ has a singleton disambiguation $\bar{C} = \{c\}$. Indeed, any given $x \in \mathcal{X}$ can receive at most one of $\{0, 1\}$ as a label from the members of C (otherwise, it would be shattered, forcing $\mathrm{VC}(C^\star) \geq 1$). If *any* $c^\star \in C^\star$ labels x with 0, then *all* members of $C^\star$ are disambiguated to label x with 0 (and, *mutatis mutandis*, 1). Any x labeled with $\star$ by *every* $c^\star \in C^\star_i$ can be disambiguated arbitrarily (say, to 0). Disambiguating the degenerate $[F'_i]^\star_\gamma$ to the singleton $\bar{F}_i(S)$ has no effect on the product in (6.23).

The foregoing argument continues to hold if more than one $v_i = 0$. In particular, in the degenerate case where $\mathrm{fat}_\gamma(F_1) = \mathrm{fat}_\gamma(F_2) = \ldots = \mathrm{fat}_\gamma(F_k) = 0$, we have $\prod |\bar{F}_i(S)| = 1$, which forces $\ell = 0$. □

Proof of Theorem 6.3

First, we upper bound the covering numbers of Lipschitz aggregations as a function of the covering numbers of the component classes.

Theorem 6.12 (Covering number of L-Lipschitz aggregations) *Let $t > 0$, $p \in [1, \infty]$, and $F_1, \ldots, F_k \subset L_p(\mu)$. Let G be an aggregation rule that is L-Lipschitz. Then, for all probability measures μ on $\mathcal{X}$,*

$$\mathcal{N}(G(F_1, \ldots, F_k), L_p(\mu), t) \quad\leq\quad \begin{cases} \prod_{i=1}^k \mathcal{N}(F_i, L_p(\mu), t/Lk^{1/p}), & p < \infty \\ \prod_{i=1}^k \mathcal{N}(F_i, L_p(\mu), t/L), & p = \infty. \end{cases}$$

We proceed to the main proof.

Proof (of Theorem 6.3). . Let $G : \mathbb{R}^k \to \mathbb{R}$ be an aggregation rule that is L-Lipschitz ($L \geq 1$) in supremum norm, as defined in (6.8), and let $F_1, \ldots, F_k \subseteq [-R, R]^\Omega$ be real-valued function classes. Suppose that some $\mathcal{X}_\ell = \{x_1, \ldots, x_\ell\} \subset \mathcal{X} = B$ is a maximal set that is γ-shattered by G, let $F_i(\mathcal{X}_\ell) = F_i|_{\mathcal{X}_\ell}$, and μ_ℓ be the uniform distribution on $\mathcal{X}_\ell$. We upper bound the covering number with the fat-shattering dimension as in Lemma 6.25 (see Section 6.6), with $n = \ell$ and $p = \infty$,

$$\log \mathcal{N}(F_i(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma) \quad\leq\quad Cv_i \log(R\ell/v_i\gamma)\log^\epsilon(\ell/v_i), \qquad 0 < \gamma < R,$$

where $v_i = \mathrm{fat}_{c\epsilon\gamma}(F_i)$. Then Theorem 6.12 implies that

$$\begin{aligned}
\log \mathcal{N}(G(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma/2) &\leq \sum_{i=1}^{k} \log \mathcal{N}(F_i(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma/2L) \\
&\leq C \sum_{i=1}^{k} v_i \log(LR\ell/v_i\gamma) \log^{\epsilon}(\ell/v_i) \\
&\overset{(a)}{\leq} C \sum_{i=1}^{k} v_i \log^{1+\epsilon}(LR\ell/v_i\gamma) \\
&\overset{(b)}{\leq} CD \log^{1+\epsilon} \frac{LR\ell k}{D\gamma},
\end{aligned}$$

where $D := \sum_{i=1}^{k} v_i$, (a) follows since $R/\gamma > 1$ and assuming $L \geq 1$, and (b) follows by the concavity of $x \log^{1+\epsilon}(u/x)$ (see Lemma 6.32 in Section 6.6). We can assume $\ell \geq 2$ without loss of generality. Combining the monotonicity of the covering number (see (6.10)), a lower bound on the covering number in terms of the fat-shattering dimension (see Lemma 6.21 in Section 6.6), and the fact the $\mathcal{X}_\ell$ is a maximal set that is γ-shattered by G yields

$$\log \mathcal{N}(G(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma/2) \geq C\mathrm{fat}_\gamma(G) = C\ell,$$

whence

$$\ell \leq CD \log^{1+\epsilon} \frac{LR\ell k}{D\gamma}.$$

Using the elementary fact

$$x \leq A \operatorname{Log}^{1+\epsilon} x \implies x \leq cA \operatorname{Log}^{1+\epsilon} A \qquad x, A \geq 1$$

(with $x = LR\ell k/D\gamma$ and $A = cLRk/\gamma$), we get

$$\ell \leq CD \operatorname{Log}^{1+\epsilon} \frac{LRk}{\gamma},$$

which implies the claim. □

Proof of Theorem 6.5

We use the notation and results from the Appendix, and in particular, from Section 6.6.

Proof (of Theorem 6.5). A bound of this form for the k-fold maximum aggregation was claimed in Kontorovich [182], however the argument there was flawed, see Section 6.5.

Let $G : \mathbb{R}^k \to \mathbb{R}$ be an aggregation rule that is L-Lipschitz in supremum norm, as defined in (6.8), and let $F_1, \ldots, F_k$ be bounded affine function classes, as defined in (6.11). Suppose that some $\mathcal{X}_\ell = \{x_1, \ldots, x_\ell\} \subset \mathcal{X} = B$ is a maximal set that is γ-shattered by G, let $F_i(\mathcal{X}_\ell) = F_i|_{\mathcal{X}_\ell}$, and μ_ℓ be the uniform distribution on $\mathcal{X}_\ell$. We upper bound the covering

number as in Lemma 6.27 (with $m = \ell$),

$$\log \mathcal{N}(F_i(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma) \quad \leq \quad C \frac{R_i^2}{\gamma^2} \operatorname{Log} \frac{\ell \gamma^2}{R_i^2}, \qquad 0 < \gamma < R_i.$$

Denote $v_i := L^2 R_i^2 / \gamma^2$, and consider the L_∞ covering number of $F_i(\mathcal{X}_\ell)$ at scale γ / L:

$$\log \mathcal{N}(F_i(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma / L) \quad \leq \quad C v_i \operatorname{Log} \frac{\ell}{v_i}.$$

Then Theorem 6.12 implies that

$$\begin{aligned} \log \mathcal{N}(G(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma/2) \quad &\leq \quad \sum_{i=1}^{k} \log \mathcal{N}(F_i(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma/2L) \\ &\leq \quad C \sum_{i=1}^{k} v_i \operatorname{Log} \frac{\ell}{v_i} \\ &\overset{\text{(a)}}{\leq} \quad C D \operatorname{Log} \frac{k\ell}{D}, \end{aligned}$$

where $D := \sum_{i=1}^{k} v_i$ and (a) follows by the concavity of $x \log(u/x)$ (see Corollary 6.31 in Section 6.6). Combining the monotonicity of the covering number (see (6.10)), a lower bound on the covering number in terms of the fat-shattering dimension (see Lemma 6.21 in Section 6.6), and the fact the $\mathcal{X}_\ell$ is a maximal set that is γ-shattered by G yields

$$\log \mathcal{N}(G(\mathcal{X}_\ell), L_\infty(\mu_\ell), \gamma/2) \quad \geq \quad C \mathrm{fat}_\gamma(G) = C\ell,$$

whence

$$\ell \quad \leq \quad C D \operatorname{Log} \frac{k\ell}{D}.$$

Using the elementary fact

$$x \leq A \operatorname{Log} x \quad \Longrightarrow \quad x \leq cA \operatorname{Log} A, \qquad x, A \geq 1$$

(with $x = k\ell/D$ and $A = ck$) we get $\ell \leq cD \log k$, which implies the claim.

The result can easily be generalized to hinge-loss affine classes. Let F_i be an affine function class as in (6.11), define F_i' as the function class on $B \times \{-1, 1\}$ given by $F_i' = \{(x, y) \mapsto y f(x); f \in F_i\}$, and the *hinge-loss affine class* F_i^{Hinge} as the function class on $B \times \{-1, 1\}$ given by $F_i^{\mathrm{Hinge}} = \{(x, y) \mapsto \max\{0, 1 - f(x, y)\}; f \in F_i'\}$. One first observes that the restriction of F_i' to any $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, as a body in $\mathbb{R}^n$, is identical to the restriction of F_i to $\{x_1, \ldots, x_n\}$. Interpreting F_i^{Hinge} as a 2-fold maximum over the singleton class $H = \{h \equiv 0\}$ and the bounded affine class F_i' lets us invoke Theorem 6.12 to argue that F_i and F_i^{Hinge} have the same L_∞ covering numbers. Hence, the argument we deployed here to establish (6.12) for affine classes also applies to k-fold L-Lipschitz aggregations hinge-loss classes. □

Proof of Corollary 6.7

Proof (of Corollary 6.7). Raviv et al. [167, Theorem 7] upper-bounded the Rademacher complexity of the maximum aggregation of k hinge loss affine functions by $k/\sqrt{n}$.

For R_i-bounded affine functions or hinge loss affine functions, the analysis above, combined with the calculation in Kontorovich [182] yields a bound of $O\left(\sqrt{\frac{\mathrm{Log}(k)\,\mathrm{Log}^3(n)\sum_{i=1}^k R_i^2}{n}}\right)$. For completeness, we provide the full proof.

Let $G_{\max} : \mathbb{R}^k \to \mathbb{R}$ be the k-fold maximum aggregation rule, as defined in (6.4), and let $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$ be R_i-bounded affine function classes as in (6.11) or hinge loss affine function classes as in (6.13). Since this aggregation is 1-Lipschitz in the supremum norm, Theorem 6.5 implies that

$$\mathrm{fat}_\gamma(G_{\max}) \quad \leq \quad \frac{C\,\mathrm{Log}(k)}{\gamma^2}\sum_{i=1}^k R_i^2, \qquad 0 < \gamma < \min_{i\in[k]} R_i,$$

where $C > 0$ is a universal constant.

From fat-shattering to Rademacher. The fat-shattering estimate above can be used to upper-bound the Rademacher complexity by converting the former to a covering number bound and plugging it into Dudley's chaining integral [34]:

$$\mathcal{R}_n(F) \leq \inf_{\alpha\geq 0}\left(4\alpha + 12\int_\alpha^\infty \sqrt{\frac{\log \mathcal{N}(F, \|\cdot\|_2, t)}{n}}dt\right), \tag{6.24}$$

where $\mathcal{N}(\cdot)$ are the L_2 covering numbers.

It remains to bound the covering numbers. A simple way of doing so is to invoke Lemmas 2.6, 3.2, and 3.3 in Alon et al. [158] — but this incurs superfluous logarithmic factors in n. Instead, we use the sharper estimate of Mendelson and Vershynin [35], stated here in Lemma 6.23. Putting $\tilde{R} = \max_i R_i$, the latter yields

$$\begin{aligned}
\mathcal{R}_n(G_{\max}) &\leq \inf_{\alpha\geq 0}\left(4\alpha + 12\int_\alpha^1 \sqrt{\frac{\log \mathcal{N}(G_{\max}, \|\cdot\|_2, t)}{n}}dt\right)\\
&\leq \inf_{\alpha\geq 0}\left(4\alpha + 12c'\int_\alpha^1 \sqrt{\frac{\mathrm{fat}_{ct/\tilde{R}}(G_{\max})\log\frac{2\tilde{R}}{t}}{n}}dt\right)\\
&\leq \inf_{\alpha\geq 0}\left(4\alpha + 12c''\sqrt{\frac{\mathrm{Log}(k)\sum_{i=1}^k R_i^2}{n}}\int_\alpha^1 \frac{\tilde{R}}{t}\sqrt{\log\frac{2\tilde{R}}{t}}dt\right).
\end{aligned}$$

Now

$$\int_\alpha^1 \frac{\tilde{R}}{t}\sqrt{\log\frac{2\tilde{R}}{t}}dt = \frac{2\tilde{R}}{3}\left(\log(2\tilde{R}/\alpha)^{3/2} - (\log 2\tilde{R})^{3/2}\right)$$

and choosing $\alpha = 1/\sqrt{n}$ yields

$$\begin{aligned}
\mathcal{R}_n(G_{\max}) &\leq \frac{4}{\sqrt{n}} + 12c''\sqrt{\frac{\mathrm{Log}(k)\sum_{i=1}^k R_i^2}{n}}\frac{2\tilde{R}}{3}\left(\log(2\tilde{R}\sqrt{n})^{3/2} - (\log 2\tilde{R})^{3/2}\right)\\
&= O\left(\sqrt{\frac{\mathrm{Log}(k)\log^3(\tilde{R}n)\tilde{R}^2\sum_{i=1}^k R_i^2}{n}}\right).
\end{aligned}$$

□

Proof of Theorem 6.9

Proof (of Theorem 6.9). Let $G_{\max} : \mathbb{R}^k \to \mathbb{R}$ be the k-fold maximum aggregation rule, as defined in (6.4), and let $F_1, \ldots, F_k \subseteq \mathbb{R}^\Omega$ be identical function classes consisting of all real-valued affine functions. Note that $G_{\max}$ is an aggregation that commutes with shift, as defined in (6.7).

By (6.9), there is an $r \in \mathbb{R}^\mathcal{X}$ such that $\mathrm{fat}_\gamma(G_{\max}) = \mathrm{f\mathring{a}t}_\gamma(G_{\max} - r)$. As in (6.21), put $F_i' := F_i - r$ and $G'_{\max} := G_{\max} - r = G_{\max}(F_1', \ldots, F_k')$. Define $\bar{G}_{\max} = \mathrm{sign}(G'_{\max})$ and $\bar{F}_i = \mathrm{sign}(F_i')$.

Since sign and max commute, we have $\bar{G}_{\max} = \max_{i\in[k]}(\bar{F}_i)$. We claim that

$$\mathrm{f\mathring{a}t}_\gamma(G'_{\max}) \quad \leq \quad \mathrm{VC}(\bar{G}_{\max}). \tag{6.25}$$

Indeed, any $S \subset \mathcal{X}$ that is γ-shattered with shift $r = 0$ by any $G \subset \mathbb{R}^\mathcal{X}$ is also VC-shattered by $\mathrm{sign}(G)$. (See Section 6.4, and notice that the converse implication—and the reverse inequality—do not hold.) It holds that

$$d + 1 \overset{(a)}{=} \mathrm{VC}(\bar{F}_i) \overset{(b)}{=} \mathrm{f\mathring{a}t}_\gamma(F_i') \overset{(c)}{=} \mathrm{fat}_\gamma(F_i') \overset{(d)}{=} \mathrm{fat}_\gamma(F_i),$$

where (a) follows from a standard argument (e.g., Mohri et al. [183, Example 3.2]), (b) holds because any $S \subset \mathbb{R}^d$ that is VC-shattered by $\mathrm{sign}(F_i')$ is also γ-shattered by F_i' with shift $r = 0$, (c) follows from Lemma 6.28, since the class satisfies the closure property (6.33), and (d) holds since the shattering remains the same for the shifted class.

Now the argument of Blumer et al. [98, Lemma 3.2.3] applies:

$$\mathrm{VC}(\bar{G}_{\max}) \quad \leq \quad 2(d+1)k\log(3k) \tag{6.26}$$

(this holds for any k-fold aggregation function, not just the maximum). Combining (6.25) with (6.26) proves the claim.□

Proof of Theorem 6.10

Proof (of Theorem 6.10). It follows from Mohri et al. [183, Example 3.2] that $\mathrm{VC}(\mathrm{sign}(F_i)) = d + 1$. Since F_i is closed under scalar multiplication, a scaling argument shows that any $S \subset \mathbb{R}^d$ that is VC-shattered by $\mathrm{sign}(F_i)$ is also γ-shattered by F_i with shift $r = 0$, whence $\mathrm{f\mathring{a}t}_\gamma(F_i) = d+1$ for all $\gamma > 0$; invoking Lemma 6.28 extends this to $\mathrm{fat}_\gamma(F_i)$ as well. Now Csikós et al. [163, Theorem 1] shows that the k-fold unions of half-spaces necessarily shatter some set $S \subset \mathbb{R}^d$ of size at least $cdk\log k$. Since union is a special case of the max operator, and the latter commutes with sign, the scaling argument shows that this S is γ-shattered by $G_{\max}$ with shift $r = 0$. Hence, $\mathrm{fat}_\gamma(G_{\max}) \geq \mathrm{f\mathring{a}t}_\gamma(G_{\max}) \geq |S|$, which proves the claim. □

6.5 Discussion

In this paper, we proved upper and lower bounds on the fat-shattering dimension of aggregation rules as a function of the fat-shattering dimension of the component classes. We leave some remaining gaps for future work. First, for aggregation

rules that commute with shifts and commute with truncation, assuming $\mathrm{fat}_\gamma(F_i) \leq d$, for $1 \leq i \leq k$, we show in Theorem 6.1 that

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \leq Cdk \operatorname{Log}^2(dk), \qquad \gamma > 0,$$

$C > 0$ is a universal constant. We pose the following

Open Problem 6.13 *Let G be an aggregation rule with the properties as in Theorem 6.1. Is it the case that for all $F_i \subseteq \mathbb{R}^{\mathcal{X}}$ with $\mathrm{fat}_\gamma(F_i) \leq d$, $i \in [k]$, we have*

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \leq \quad Cdk \operatorname{Log}(k), \qquad \gamma > 0, \tag{6.27}$$

for some universal $C > 0$?

In light of Theorem 6.10, this is the best one could hope for in general. We pose also the following conjecture about bounded affine functions.

Conjecture 6.14 *Theorem 6.5 is tight up to constants. For R_i-bounded affine functions and an aggregation rule G that is 1-Lipschitz in supremum norm,*

$$\mathrm{fat}_\gamma(G(F_1, \ldots, F_k)) \quad \geq \quad \frac{C \operatorname{Log}(k)}{\gamma^2} \sum_{i=1}^{k} R_i^2, \qquad 0 < \gamma < \min_{i \in [k]} R_i, \tag{6.28}$$

where $C > 0$ is a universal constant.

Throughout the paper, we mentioned several mistaken claims in the literature. In this section, we briefly discuss the nature of these mistakes—which are, in a sense, variations on the same kind of error. We begin with Attias et al. [175, Lemma 14], which incorrectly claimed that any partial function class $F^\star$ has a disambiguation $\bar{F}$ such that $\mathrm{VC}(\bar{F}) \leq \mathrm{VC}(F^\star)$ (see Section 6.4 for the definitions). The mistake was pointed out to us by Yann Guermeur, and later, Alon et al. [176, Theorem 11] showed that there exist partial classes, $F^\star$ with $\mathrm{VC}(F^\star) = 1$ for which every disambiguation $\bar{F}$ has $\mathrm{VC}(\bar{F}) = \infty$.

Kontorovich [182] attempted to prove the bound stated in our Theorem 6.5 (up to constants, and only for linear classes). The argument proceeded via a reduction to the Boolean case, as in our proof of Theorem 6.9. It was correctly observed that if, say, some finite $S \subset \mathcal{X}$ is 1-shattered by F_i with shift $r = 0$, then it is also VC-shattered by $\mathrm{sign}(F_i)$. Neglected was the fact that $\mathrm{sign}(F_i)$ might shatter additional points in $\mathcal{X} \setminus S$—and, in sufficiently high dimension, it necessarily will. The crux of the matter is that (6.25) holds in the dimension-dependent but not the dimension-free setting; again, this may be seen as a variant of the disambiguation mistake.

Finally, the proof of Hanneke and Kontorovich [184, Lemma 6] claims, in the first display, that the shattered set can be classified with a large margin, which is incorrect — yet another variant of mistaken disambiguation.

6.6 Auxiliary results

Properties of Aggregation Rules

Lemma 6.15 *If $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz under $\|\cdot\|_p$, then $G : (\mathbb{R}^\Omega)^k \to \mathbb{R}^\Omega$ is L-Lipschitz in $\|\cdot\|_{L_p^{(k)}(\mu)}$.*

Proof.

$$\begin{aligned}\left\|G(f_1,\dots,f_k) - G(f_1',\dots,f_k')\right\|_{L_p(\mu)}^p &= \int_{\mathcal{X}} |G(f_1,\dots,f_k)(x) - G(f_1',\dots,f_k')(x)|^p \mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} |G(f_1(x),\dots,f_k(x)) - G(f_1'(x),\dots,f_k'(x))|^p \mathrm{d}\mu(x) \\ &\leq \int_{\mathcal{X}} L^p \left\|(f_1(x),\dots,f_k(x)) - (f_1'(x),\dots,f_k'(x))\right\|_p^p \mathrm{d}\mu(x),\end{aligned}$$

where the inequality follows from the assumption that $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz in $\|\cdot\|_p$. This proves

$$\left\|G(f_1,\dots,f_k) - G(f_1',\dots,f_k')\right\|_{L_p(\mu)} \leq L \left\|(f_1,\dots,f_k) - (f_1',\dots,f_k')\right\|_{L_p^{(k)}(\mu)},$$

and hence the claim. □

Proof (of Theorem 6.12). Suppose $p < \infty$, and let $g = G(f_1,\dots,f_k) \in G(F_1,\dots,F_k)$. For each $i \in [k]$, let $\hat{F}_i \subset F_i$ be a $t/Lk^{1/p}$-cover of F_i. Let each f_i be "$t/Lk^{1/p}$-covered" by some $\hat{f}_i \in \hat{F}_i$, in the sense that $\left\|f_i - \hat{f}_i\right\|_{L_p(\mu)} \leq t/Lk^{1/p}$. Assuming that $G : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz in $\|\cdot\|_p$, Lemma 6.15 implies that $G : (\mathbb{R}^\Omega)^k \to \mathbb{R}^\Omega$ is L-Lipschitz in $\|\cdot\|_{L_p^{(k)}(\mu)}$. Then it follows that g is t-covered by $G(\hat{f}_1,\dots,\hat{f}_k)$, since

$$\begin{aligned}\left\|G(f_1,\dots,f_k) - G(\hat{f}_1,\dots,\hat{f}_k)\right\|_{L_p(\mu)}^p &\leq L^p \left\|(f_1,\dots,f_k) - (\hat{f}_1,\dots,\hat{f}_k)\right\|_{L_p^{(k)}(\mu)}^p \\ &= L^p \int_{\mathcal{X}} \left\|(f_1(x),\dots,f_k(x)) - (\hat{f}_1(x),\dots,\hat{f}_k(x))\right\|_p^p \mathrm{d}\mu(x) \\ &= L^p \int_{\mathcal{X}} \sum_{i=1}^k \left|f_i(x) - \hat{f}_i(x)\right|^p \mathrm{d}\mu(x) \\ &= L^p \sum_{i=1}^k \int_{\mathcal{X}} \left|f_i(x) - \hat{f}_i(x)\right|^p \mathrm{d}\mu(x) \\ &= L^p \sum_{i=1}^k \left\|f_i - \hat{f}_i\right\|_{L_p(\mu)}^p \\ &\leq L^p k \left(\frac{t}{Lk^{1/p}}\right)^p \\ &= t^p,\end{aligned}$$

and so $\left\|G(f_1,\dots,f_k) - G(\hat{f}_1,\dots,\hat{f}_k)\right\|_{L_p(\mu)} \leq t$.

We conclude that $G(F_1,\dots,F_k)$ has a t-cover of size $|\hat{F}_1 \times \hat{F}_2 \times \dots \times \hat{F}_k|$, which proves the claim. The case $p = \infty$ is proved analogously (or, alternatively, as a limiting case of $p < \infty$). □

We show that natural aggregations are Lipschitz in $\|\cdot\|_p$ norms, $p \in [1,\infty)$, and in supremum norm. The following facts

are elementary:

$$|a \vee b - c \vee d| \quad \leq \quad |a - c| \vee |b - d|, \qquad a, b, c, d \in \mathbb{R}; \tag{6.29}$$

$$|a \wedge b - c \wedge d| \quad \leq \quad |a - c| \vee |b - d|, \qquad a, b, c, d \in \mathbb{R}, \tag{6.30}$$

where $s \vee t := \max\{s, t\}$ and $s \wedge t := \min\{s, t\}$.

Lemma 6.16 (Maximum aggregation is 1-Lipschitz) *Let $G_{\max} : \mathbb{R}^k \to \mathbb{R}$ be the maximum aggregation, then for any $x, x' \in \mathbb{R}^k$ and $p \in [1, \infty]$,*

$$\left|G_{\max}(x) - G_{\max}(x')\right| \leq \left\|x - x'\right\|_p .$$

Proof. For $k = 2$ and $p = \infty$, the claim follows from the stronger, pointwise inequality (6.29). The proof follows by simple induction on k. Since $\|\cdot\|_\infty \leq \|\cdot\|_p$, we conclude the proof for $p \in [1, \infty]$. □

Lemma 6.17 (Median aggregation is 1-Lipschitz) *Let $G_{\text{Median}} : \mathbb{R}^k \to \mathbb{R}$ be the median aggregation, then for any $x, x' \in \mathbb{R}^k$ and $p \in [1, \infty]$,*

$$\left|G_{\text{Median}}(x) - G_{\text{Median}}(x')\right| \leq \left\|x - x'\right\|_p .$$

Proof. Denote by $x_{(1)}, \ldots, x_{(k)}$ the ascending order of a sequence $x_1, \ldots, x_k$, that is, $x_{(1)} \leq \cdots \leq x_{(k)}$. For all $x, x' \in \mathbb{R}^k$ we have

$$\begin{aligned} \left|G_{\text{Median}}(x_1, \ldots, x_k) - G_{\text{Median}}(x'_1, \ldots, x'_k)\right| \quad &= \quad \left|x_{(\lceil k/2 \rceil)} - x'_{(\lceil k/2 \rceil)}\right| \\ &\leq \quad \max_{i \in [k]} \left|x_{(i)} - x'_{(i)}\right| \leq \max_{i \in [k]} \left|x_i - x'_i\right|, \end{aligned}$$

where the last inequality follows from Cohen et al. [185, Eq. (16)][6] Since $\|\cdot\|_\infty \leq \|\cdot\|_p$, we conclude the proof for $p \in [1, \infty]$. □

Lemma 6.18 (Max-Min aggregation is 1-Lipschitz) *Let $G_{\text{max-min}} : \mathbb{R}^{k \times \ell} \to \mathbb{R}$ be the max-min aggregation, then for any $x, x' \in \mathbb{R}^{k \times \ell}$ and $p \in [1, \infty]$,*

$$\left|G_{\text{max-min}}(x) - G_{\text{max-min}}(x')\right| \leq \left\|x - x'\right\|_p .$$

Proof. The inequalities (6.29), (6.30) imply that the k-fold max and min aggregations are both 1-Lipschitz with respect to $\|\cdot\|_\infty$. Hence, for all $x, y \in \mathbb{R}^{k \times \ell}$, we have

$$\left|\min_{i \in [k]} x_{ij} - \min_{i \in [k]} y_{ij}\right| \leq \max_{i \in [k]} |x_{ij} - y_{ij}| , \qquad j \in [\ell]$$

[6] stated there for distributions but true for all vectors, by the same argument

and further,

$$\left|\max_{j\in[\ell]}\min_{i\in[k]} x_{ij} - \max_{j\in[\ell]}\min_{i\in[k]} y_{ij}\right| \le \max_{j\in[\ell]}\max_{i\in[k]} |x_{ij} - y_{ij}|\,.$$

This proves that $|G_{\text{max-min}}(x) - G_{\text{max-min}}(x')| \le \|x - x'\|_\infty$. Since $\|\cdot\|_\infty \le \|\cdot\|_p$, the claim holds for all $p \in [1,\infty]$. □

Lemma 6.19 (Median aggregation commutes with truncation) *Let $G_{\text{Median}} : \mathbb{R}^k \to \mathbb{R}$ be the median aggregation, then G_{Median} commutes with truncation. That is, for any $\gamma > 0$ and $x \in \mathbb{R}^d$,*

$$G_{\text{Median}}(\bar{x}_1,\dots,\bar{x}_k) \in \mathscr{D}([G_{\text{Median}}(x_1,\dots,x_k)]^\star_\gamma)$$

for all disambiguations $\bar{x}_i \in \mathscr{D}([x_i]^\star_\gamma)$, $i \in [k]$.

Proof. Fix $\gamma > 0$ and denote by $x_{(1)},\dots,x_{(k)}$ the ascending order of a sequence $x_1,\dots,x_k$. Now for $\bar{x}_i \in \mathscr{D}([x_i]^\star_\gamma) \subseteq \{0,1\}$, our definition of the median (6.5) implies that $G_{\text{Median}}(\bar{x}_1,\dots,\bar{x}_k) \in \{0,1\}$. It remains to perform an exhaustive verification of the possible cases.

If $[G_{\text{Median}}(x_1,\dots,x_k)]^\star_\gamma = \star$ then any value in $\{0,1\}$ is valid. If $[G_{\text{Median}}(x_1,\dots,x_k)]^\star_\gamma = 0$ it means that $G_{\text{Median}}(x_1,\dots,x_k)$ outputs a value smaller than $-\gamma$, which means that at least half of inputs $x_1,\dots,x_k$ have a value smaller than $-\gamma$. Let these values be $x_{(1)},\dots,x_{(m)}$ where $m \ge k/2$. We have $[x_{(j)}]^\star_\gamma = 0$ for $j \in [m]$ and $[x_{(\ell)}]^\star_\gamma \subseteq \{\star,1\}$ for $\ell \in \{m+1,\dots,k\}$. For any disambiguation $\bar{x}_{(\ell)}$, $G_{\text{Median}}(\bar{x}_1,\dots,\bar{x}_k)$ would still output 0 and is a valid disambiguation of $[G_{\text{Median}}(x_1,\dots,x_k)]^\star_\gamma$. The case $[G_{\text{Median}}(x_1,\dots,x_k)]^\star_\gamma = 1$ follows from the same argument. □

Lemma 6.20 (Max-Min aggregation commutes with truncation) *Let $G_{\text{max-min}} : \mathbb{R}^{k\times\ell} \to \mathbb{R}$ be the max-min aggregation, then $G_{\text{max-min}}$ commutes with truncation. That is, for any $\gamma > 0$ and $x \in \mathbb{R}^{k\times\ell}$,*

$$G_{\text{max-min}}(\bar{x}_{11},\dots,\bar{x}_{k\ell}) \in \mathscr{D}([G_{\text{max-min}}(x_{11},\dots,x_{k\ell})]^\star_\gamma),$$

for all disambiguations $\bar{x}_{ij} \in \mathscr{D}([x_{ij}]^\star_\gamma)$, $i \in [k], j \in [\ell]$.

Proof. Fix $\gamma > 0$. For any $i \in [k]$ denote by $x_{i(1)},\dots,x_{i(\ell)}$ the ascending order of a sequence $x_{i1},\dots,x_{i\ell}$. We assume $\bar{x}_{ij} \in \mathscr{D}([x_{ij}]^\star_\gamma) \subseteq \{0,1\}$ and $G_{\text{max-min}}(\bar{x}_{11},\dots,\bar{x}_{k\ell})$ outputs a value in $\{0,1\}$ by our definition of the max-min. We check all possible outputs of $[G_{\text{max-min}}(x_{11},\dots,x_{k\ell})]^\star_\gamma$ and verify that $G_{\text{max-min}}(\bar{x}_{11},\dots,\bar{x}_{k\ell})$ is a valid disambiguation.

If $[G_{\text{max-min}}(x_{11},\dots,x_{k\ell})]^\star_\gamma = \star$ then any value in $\{0,1\}$ is valid. If $[G_{\text{max-min}}(x_{11},\dots,x_{k\ell})]^\star_\gamma = 0$ it means that $G_{\text{max-min}}(x_{11},\dots,x_{k\ell})$ outputs a value smaller than $-\gamma$. This means that all values that minimize each row $x_{1(1)},\dots,x_{k(1)}$ are smaller than $-\gamma$ since the maximum of them is smaller than $-\gamma$. We have $[x_{i(1)}]^\star_\gamma = 0$ for $i \in [k]$. For any disambiguation $\bar{x}_{ij}$ $G_{\text{max-min}}(\bar{x}_{11},\dots,\bar{x}_{k\ell})$ would still output 0 and is a valid disambiguation of $[G_{\text{max-min}}(x_{11},\dots,x_{k\ell})]^\star_\gamma$. The case$[G_{\text{max-min}}(x_{11},\dots,x_{k\ell})]^\star_\gamma = 1$ follows from the same argument. □

Covering Numbers and the Fat-Shattering Dimension

In this section, we summarize some known results connecting the covering numbers of a bounded function class to its fat-shattering dimension.

Lemma 6.21 (Talagrand [177], Proposition 1.4) *For any $F \subseteq [-R, R]^{\mathcal{X}}$, there exists a probability measure μ on $\mathcal{X}$ such that*

$$\mathcal{N}(F, L_2(\mu), t) \quad \geq \quad 2^{C\mathrm{fat}_{2t}(F)}, \qquad 0 < t < R, \tag{6.31}$$

where $C > 0$ is a universal constant. Moreover, μ may be taken to be the uniform distribution on any $2t$-shattered subset of $\mathcal{X}$.

Remark 6.22 *The tightness of (6.31) is trivially demonstrated by the example $F = \{-\gamma, \gamma\}^n$.*

Lemma 6.23 (Mendelson and Vershynin [35], Theorem 1) *For all $F \subseteq [-1, 1]^{\mathcal{X}}$ and all probability measures μ,*

$$\mathcal{N}(F, L_2(\mu), t) \quad \leq \quad \left(\frac{2}{t}\right)^{C\mathrm{fat}_{ct}(F)}, \qquad 0 < t < 1, \tag{6.32}$$

where $C, c > 0$ are universal constants.

Remark 6.24 *The following example due to Vershynin [186] shows that (6.32) is tight. Take $\mathcal{X} = [n]$ and $F = [-1, 1]^{\mathcal{X}}$. Then, for all sufficiently small $t > 0$, we have $\mathrm{fat}_t(F) = n$. However, a simple volumetric calculation shows that $\mathcal{N}(F, L_2(\mu), t)$ behaves as $(C/t)^n$ for small t, where $C > 0$ is a constant.*

Lemma 6.25 (Rudelson and Vershynin [187]) *Suppose that $p \in [2, \infty)$, μ is a probability measure on $\mathcal{X}$, and $R > 0$. If $F \subset L_p(\mathcal{X}, \mu)$ satisfies $\sup_{f \in F} \|f\|_{L_{2p}(\mu)} \leq R$, then*

$$\log \mathcal{N}(F, L_p(\mu), t) \quad \leq \quad Cp^2 \mathrm{fat}_{ct}(F) \log \frac{R}{ct}, \qquad 0 < t < R;$$

furthermore, for all $\epsilon > 0$, if $\sup_{f \in F} \|f\|_{L_\infty(\mu)} \leq R$, then

$$\log \mathcal{N}(F, L_\infty(\mu), t) \quad \leq \quad Cv \log(Rn/vt) \log^{\epsilon}(n/v), \qquad 0 < t < R,$$

where $n = |\mathcal{X}|$, $v = \mathrm{fat}_{c\epsilon t}(F)$, and $C, c > 0$ are universal constants.

Covering Numbers of Linear and Affine Classes

Let $B \subset \mathbb{R}^d$ be the d-dimensional Euclidean unit ball and

$$F = \{x \mapsto w \cdot x + b; \|w\| \vee |b| \leq R\}$$

be the collection of R-bounded affine functions on $\mathcal{X} = B$.

Remark 6.26 *There is a trivial reduction from an R-bounded affine class in d dimensions to a $2R$-bounded* linear *class in $d+1$ dimensions, via the standard trick of adding an extra dummy dimension. This only affects the covering number bounds up to constants.*

For $\mathcal{X}_n \subset B$, $|\mathcal{X}_n| = n$, define $F(\mathcal{X}_n) = F|_{\mathcal{X}_n}$, and endow $\mathcal{X}_n$ with the uniform measure μ_n. Zhang [188, Theorem 4] implies the covering number estimate

$$\log \mathcal{N}(F(\mathcal{X}_n), L_\infty(\mu_n), t) \quad \leq \quad C\frac{R^2}{t^2}\operatorname{Log}\frac{nR}{t}, \qquad t > 0,$$

where $C > 0$ is a universal constant (Zhang's result is more general and allows to compute explicit constants). We will use the following sharper bound:

Lemma 6.27

$$\log \mathcal{N}(F(\mathcal{X}_n), L_\infty(\mu_n), t) \quad \leq \quad C\frac{R^2}{t^2}\operatorname{Log}\frac{mt^2}{R^2}, \qquad 0 < t < R,$$

where $m = \min\{n, d\}$ and $C > 0$ is a universal constant.

Proof. The result is folklore knowledge, but we provide proof for completeness.

Let $B = B_2^d$ be the Euclidean unit ball and $X = \{x_1, \ldots, x_n\} \subset B$. This induces the set $F = \{(w \cdot x_1, w \cdot x_2, \ldots, w \cdot x_n); w \in B\} \subset \mathbb{R}^n$. We argue that there is no loss of generality in assuming $d \geq n$. Indeed, if $n > d$, then X is spanned by some $X' = \{x'_1, \ldots, x'_d\} \subset B$ and $F \subset \operatorname{span}(X')$ is also a d-dimensional set. Thus, we assume $d \geq n$ henceforth.

Via a standard infinitesimal perturbation, we can assume that X is a linearly independent set (i.e., spans $\mathbb{R}^n$). If we treat X as an $n \times d$ matrix, then $F = XB$, which means that F is an ellipsoid. We are interested in estimating the ℓ_∞ covering numbers of F.

Let $K \subset \mathbb{R}^d$ be such that $XK = L$, where $L = B_\infty^n$ is the unit cube. (The existence of a K such that $XK \subset L$ is obvious, but because we assumed that X spans $\mathbb{R}^n$, every point in $[-1, 1]^n$ has a pre-image under X.) Let us compute the polar body K°, defined as

$$K^\circ = \left\{ u \in \mathbb{R}^d : \sup_{v \in K} v \cdot u \leq 1 \right\}.$$

We claim that

$$K^\circ = \operatorname{absconv}(X) =: \left\{ \sum_{i=1}^n \alpha_i x_i; \sum |\alpha_i| \leq 1 \right\}.$$

Indeed, consider a $z = \sum_{i=1}^n \alpha_i x_i \in \mathrm{absconv}(X)$. Then, for any $v \in K$, we have

$$\begin{aligned} v \cdot z &= v \cdot \sum_{i=1}^n \alpha_i x_i \\ &= \sum_{i=1}^n \alpha_i (v \cdot x_i) \\ &\leq \sum_{i=1}^n |\alpha_i| \leq 1 \qquad \Longrightarrow z \in K^\circ, \end{aligned}$$

where we have used $|v \cdot x_i| \leq 1$, since $XK = L = B_\infty^n = [-1,1]^n$. This shows that $\mathrm{absconv}(X) \subseteq K^\circ$. On the other hand, consider any $u \in K^\circ$. There is no loss of generality in assuming that u is in the span of X, that is, $u = \sum_{i=1}^m \alpha_i x_i$, for $\alpha_i \in \mathbb{R}$. By definition of $u \in K^\circ$, we have

$$\sup_{v \in K} v \cdot u = \sup_{v \in K} v \cdot \sum_{i=1}^n \alpha_i x_i = \sup_{v \in K} \sum_{i=1}^n \alpha_i (v \cdot x_i) \leq 1.$$

Now because $XK = [-1,1]^n$, for each choice of $\alpha \in \mathbb{R}^n$, there is a $v \in K$ such that $|v \cdot x_i| = \mathrm{sign}(\alpha_i)$ for all $i \in [n]$. This shows that we must have $\sum_{i=1}^n |\alpha_i| \leq 1$, and proves $K^\circ \subseteq \mathrm{absconv}(X)$.

It is well-known (and easy to verify) that covering numbers enjoy an affine invariance:

$$N(F, L) := N(XB, XK) = N(B, K),$$

where $N(A, B)$, for two sets A, B, is the smallest number of copies of B necessary to cover A. Now the seminal result of Artstein et al. [189] applies: for all $t > 0$,

$$\log N(B, tK) \leq a \log N(K^\circ, btB),$$

where $a, b > 0$ are universal constants.

This reduces the problem to estimating the ℓ_2-covering numbers of $\mathrm{absconv}(X)$. The latter may be achieved via Maurey's method [190, Corollary 0.0.4 and Exercise 0.0.6]: the t-covering number of $\mathrm{absconv}(rX)$ under ℓ_2 is at most

$$(c + cmt^2/r^2)^{\lceil r^2/t^2 \rceil}, \qquad \square$$

where $c > 0$ is a universal constant.

Fat-Shattering Dimension of Linear and Affine Classes

In this section, $\mathcal{X} = \mathbb{R}^d$ and $B \subset \mathbb{R}^d$ denotes the Euclidean unit ball. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be *affine* if it is of the form $f(x) = w \cdot x + b$, for some $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, where $\cdot$ denotes the Euclidean inner product.

Throughout the paper, we have referred to R-bounded affine function classes as those for which $\|w\| \vee |b| \leq R$. In this section, we define the larger class of *R-semi-bounded* affine functions, as those for which $\|w\| \leq R$, but b may be unbounded. In particular, the covering-number results (and the reduction to linear classes spelled out in Remark 6.26)

do not apply to semi-bounded affine classes.

The following simple result may be of independent interest.

Lemma 6.28 *Let $F \subset \mathbb{R}^{\mathcal{X}}$ be some collection of functions with the closure property*

$$f, g \in F \quad \Longrightarrow \; (f-g)/2 \in F. \tag{6.33}$$

Then, for all $\gamma > 0$, we have $\mathrm{fat}_\gamma(F) = \mathrm{f\mathring{a}t}_\gamma(F)$.

Proof. Suppose that some set $\{x_1, \ldots, x_k\}$ is γ-shattered by F. That means that there is an $r \in \mathbb{R}^k$ such that for all $y \in \{-1,1\}^k$, there is an $f = f_y \in F$ for which

$$\gamma \;\leq\; y_i(f(x_i) - r_i), \qquad i \in [k]. \tag{6.34}$$

Now for any $y \in \{-1,1\}^k$, let $\hat{f} = f_y$ and $\check{f} = f_{-y}$. Then, for each $i \in [k]$, we have

$$\begin{aligned} \gamma &\leq \;\; y_i(\hat{f}(x_i) - r_i), \\ \gamma &\leq -y_i(\check{f}(x_i) - r_i). \end{aligned}$$

It follows that $f = (\hat{f} - \check{f})/2$ achieves (6.34), for the given y, with $r \equiv 0$. Now (6.33) implies that the function defined by f belongs to F, which completes the proof. □

Now it is well-known [100, Theorem 4.6] that bounded *linear* functions — i.e., function classes on B of the form $F = \{x \mapsto w \cdot x; \|w\| \leq R\}$, also known as *homogeneous hyperplanes* — satisfy $\mathrm{fat}_\gamma(F) \leq (R/\gamma)^2$. The discussion in Hanneke and Kontorovich [184, p. 102] shows that the common approach of reducing of the general (affine) case to the linear (homogeneous, $b = 0$) case, via the addition of a "dummy" coordinate, incurs a large suboptimal factor in the bound. Hanneke and Kontorovich [184, Lemma 6] is essentially an analysis of the fat-shattering dimension of bounded affine functions. Although this result contains a mistake (see Section 6.5), much of the proof technique can be salvaged:

Lemma 6.29 *The semi-bounded affine function class on B defined by $F = \{x \mapsto w \cdot x + b; \|w\| \leq R\}$ in d dimensions satisfies*

$$\mathrm{fat}_\gamma(F) \;\leq\; \min\left\{ d+1, \left(\frac{\left(1 + \sqrt{\frac{8}{\pi}}\right) R}{\gamma} \right)^2 \right\}, \qquad 0 < \gamma \leq R.$$

Proof. Since F satisfies (6.33), it suffices to consider $\mathrm{f\mathring{a}t}_\gamma(F)$, and so the shattering condition simplifies to

$$\gamma \;\leq\; y_i(w \cdot x_i + b), \qquad i \in [k]. \tag{6.35}$$

Now $\mathrm{f\mathring{a}t}_\gamma(F)$ is always upper-bounded by the VC-dimension of the corresponding class thresholded at zero, i.e., $\mathrm{sign}(F)$. For d-dimensional inhomogeneous hyperplanes, the latter is exactly $d+1$ [183, Example 3.2]. Having dispensed with the dimension-dependent part in the bound, we now focus on the R-dependent one.

Let us observe, as in Hanneke and Kontorovich [184, Lemma 6], that for $\|x_i\| \leq 1$ and $\|w\|, \gamma \leq R$, one can always realize (6.35) with $|b| \leq 2R$; which is what we shall assume, without loss of generality, henceforth. Summing up the k inequalities in (6.35) yields

$$k\gamma \quad \leq \quad w \cdot \sum_{i=1}^{k} y_i x_i + b \sum_{i=1}^{k} y_i \leq R \left\| \sum_{i=1}^{k} y_i x_i \right\| + 2R \left| \sum_{i=1}^{k} y_i \right|.$$

Letting y be drawn uniformly from $\{-1, 1\}^k$ and taking expectations, we have

$$\begin{aligned} k\gamma \quad &\leq \quad R\mathbb{E} \left\| \sum_{i=1}^{k} y_i x_i \right\| + 2R\mathbb{E} \left| \sum_{i=1}^{k} y_i \right| \leq R\sqrt{\mathbb{E} \left\| \sum_{i=1}^{k} y_i x_i \right\|^2} + 2R\sqrt{\mathbb{E} \left(\sum_{i=1}^{k} y_i \right)^2} \\ &= \quad R\sqrt{\sum_{i=1}^{k} \|x_i\|^2} + 2R\sqrt{\sum_{i=1}^{k} \mathbb{E} y_i^2} \leq 3R\sqrt{k}. \end{aligned}$$

Isolating k on the left-hand side of the inequality proves the claim $k \leq \left(\frac{3R}{\gamma}\right)^2$.

Following a referee's suggestion, we improve the constant as follows. Note that

$$\mathbb{E} \left| \sum_{i=1}^{k} y_i \right| = \frac{1}{2^k} \sum_{i=0}^{k} \binom{k}{i} |k - 2i| = \frac{k}{2^{k-1}} \binom{k-1}{\lfloor \frac{k}{2} \rfloor} \leq \sqrt{\frac{2}{\pi}} \frac{k}{\sqrt{k + \frac{1}{2}}},$$

where the inequality follows from a binomial coefficient estimate via Stirling's approximation. Thus,

$$k\gamma \quad \leq \quad R\sqrt{k} + 2R\sqrt{\frac{2}{\pi}} \frac{k}{\sqrt{k + \frac{1}{2}}} \leq R\sqrt{k} + 2R\sqrt{\frac{2}{\pi}}\sqrt{k},$$

which proves that $k \leq \left(\frac{\left(1 + \sqrt{\frac{8}{\pi}}\right) R}{\gamma} \right)^2$. □

Concavity Miscellanea

The results below are routine exercises in differentiation and Jensen's inequality.

Lemma 6.30 *For $u > 0$, the function $x \mapsto x \log(u/x)$ is concave on $(0, \infty)$.*

Corollary 6.31 *For all $u > 0$ and $v_i > 0$, $i \in [k]$,*

$$\sum_{i=1}^{k} v_i \log(u/v_i) \quad \leq \quad \left(\sum v_i \right) \log \frac{uk}{\sum v_i}.$$

Lemma 6.32 *For $0 \leq \epsilon \leq \log 2$ and $u \geq 2$, the function $x \mapsto x \log^{1+\epsilon}(u/x)$ is concave on $[1, \infty)$. It follows that for ϵ, u as above and $v_i \geq 1$, $i \in [k]$,*

$$\sum_{i=1}^{k} v_i \log^{1+\epsilon}(u/v_i) \quad \leq \quad \left(\sum v_i \right) \log^{1+\epsilon} \frac{uk}{\sum v_i}.$$

Chapter 7

Agnostic Sample Compression Schemes for Regression

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 [40] 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 [191].

7.1 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 [192] and its positive resolution by Moran and Yehudayoff [132], with various intermediate steps in between [193–202]. A stronger form of this problem, where *small* means $\mathcal{O}(d)$ (or even exactly d), remains open [191].

David, Moran, and Yehudayoff [40] 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 Theorem 7.2, 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. [40] 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. [40] 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 7.1 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 Eqs. (7.3) and (7.4) for formal definitions.

Our approach begins with proposing a boosting method (Algorithm 9) 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 7.3. 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 Eq. (2.2). 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 [191]. While such an approximate compression has been previously acknowledged in realizable regression [131], and exact compression in agnostic binary classification [40], in Section 7.3 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. [40] 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 7.8 and 7.9. 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 polynomial (in fact, linear) size linear program.

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

We then propose Algorithm 10 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 7.7. 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 7.11), generalizing the argument of David et al. [40].

These appear to be the first positive results for a 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 and Warmuth [13], Floyd and Warmuth [14]. 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 [41, 132, 203], multiclass classification [40, 154, 204, 205], regression [20, 131], active learning [206], density estimation [207], adversarially robust learning [2, 3, 9, 55, 105, 208], learning with partial concepts [97], and showing Bayes-consistency for nearest-neighbor methods [209, 210]. As a matter of fact, compressibility and learnability are known to be equivalent for general learning problems [40]. A remarkable result by Moran and Yehudayoff [132] showed that VC classes enjoy a sample compression that is independent of the sample size.

David et al. [40] 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. [131] 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. [207] 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.

Problem Setup	Compression Type	Compression Size	Reference
Realizable/Agnostic Binary Classification	Exact	$\mathcal{O}(\mathrm{VC} \cdot \mathrm{VC}^*)$	[40, 132]
Realizable/Agnostic Multiclass Classification	Exact	$\mathcal{O}\big(\mathrm{d_G} \cdot \mathrm{d_G^*}\big)$	[40]
		$\mathcal{O}\big(\mathrm{DS}^{1.5} \cdot \mathrm{polylog}(m)\big)$	[205]
		$\Omega\Big(\log(m)^{1-o(1)}\Big)$	[211]
Realizable ℓ_∞ Regression	α-Approximate	$\mathcal{O}\big(\mathrm{fat}_{c\alpha} \cdot \mathrm{fat}^*_{c\alpha} \cdot \mathrm{polylog}\big(\mathrm{fat}_{c\alpha}, \mathrm{fat}^*_{c\alpha}, \frac{1}{\alpha}\big)\big)$	[131]
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	$\Omega(m)$	[40]
Agnostic ℓ_p Linear Regression: $p \in [1, \infty]$	Exact	$\Omega(\log(m))$	This work

Table 7.1: **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 [131] 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 7.5 for open problems.

7.2 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

$$L_p(f, S) := \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - y_i|^p, \tag{7.1}$$

while for $p = \infty$,

$$L_\infty(f, S) := \max_{1 \le i \le m} |f(x_i) - y_i|. \tag{7.2}$$

Remark 7.1 *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. Recall Definition 2.18 of a sample compression scheme, which we use here for the ℓ_p loss. We say that a compression scheme (κ, ρ) is a k-size agnostic *exact sample compression scheme* for $\mathcal{F}$ if κ is a k-selection and for all $S = \{(x_i, y_i) : i \in [m]\}$, and $f_S := \rho(\kappa(S))$ achieves $\mathcal{F}$-competitive empirical loss:

$$L_p(f_S, S) \leq \inf_{f \in \mathcal{F}} L_p(f, S). \tag{7.3}$$

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

$$L_p(f_S, S) \leq \inf_{f \in \mathcal{F}} L_p(f, S) + \alpha. \tag{7.4}$$

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, $k(m) \leq k$ 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 $1 \leq p \leq \infty$.

Remark 7.2 *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 are 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 [192, 195, 212]. Also, as mentioned, in this work we are primarily interested in sample compression schemes that have* bounded size*: $k(m) \leq k$ 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 $1 \leq p < \infty$, 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 $1 \leq p < \infty$, 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 [212]. 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., 213 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. [131, 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.

7.3 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 [31, 32]. We prove the following main result.

Theorem 7.3 (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 9 with $T \leftarrow \mathcal{O}\left(\frac{1}{\beta^2}\log(m)\right)$ and*

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

we get an α-approximate sample compression scheme of size

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

for some universal constant $c > 0$. Recall that the dual fat-shattering is at most exponential in the primal dimension, see Eq. (2.2). $\tilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors of $(\mathrm{fat}, \mathrm{fat}^*, p, 1/\alpha, 1/\beta)$.

Remark 7.4 *Note that having an α-approximate compression of size k implies the following bound on the generalization error: $\alpha + \sqrt{\frac{k\log(m/k)}{m}}$ [40, 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 7.5 (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. [131, Section 4]. On the other hand, by making the requirement too strong (for example, $\mathrm{AdaBoost.R}$ in Freund and Schapire [156]), 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 Hanneke et al. [131], Kégl [140] 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 5.6. A similar generalization result was used in the context of adversarially robust learning [3].

The compression approach for agnostic binary classification, as discussed in [40], encounters a similar 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 9 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 Eqs. (7.1) and (7.2). 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 Theorem 7.5 of weak learners. By using Theorem 5.6 with

Algorithm 9 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: $P_1 \leftarrow \text{Uniform}(S)$.

▷ 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) $f^\star \leftarrow \text{argmin}_{f \in \mathcal{F}} L_p(f, S)$ (defined in Eqs. (7.1) and (7.2)).
   (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 5.6).
   (b) For $i = 1, \ldots, m$:
   Set $w_i^{(t)} \leftarrow 1 - 2\mathbb{I}\Big[\Big|\hat{f}_t(x_i) - y_i\Big| > \psi(x_i, y_i) + 2\alpha\Big]$.
   (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:
   $(J_1, \ldots, J_n) \sim \text{Categorial}\left(\frac{\alpha_1}{\sum_{s=1}^T \alpha_s}, \ldots, \frac{\alpha_T}{\sum_{s=1}^T \alpha_s}\right)^n$.
   (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 $\text{Median}(f_{J_1}, \ldots, f_{J_n})$.

$\delta = 1/3$ and $\epsilon = 1/2 - \beta$, for any distribution P_t on S, upon receiving an i.i.d. sample $S_t \subseteq S$ from P_t of size

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

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

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

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 $\forall (x,y) \in S_t : \ |f(x) - y| \leq \psi(x,y) + \alpha$, 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, Theorem 7.15 guarantees that for all $(x,y) \in S$ the output of the boosting algorithm satisfies

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

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. Theorem 7.16 implies that the sparsification method in Step 3 ensures that we can sample

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

such that 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 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 $|h(x) - y| \leq \psi(x,y) + \alpha$ for any $(x,y) \in S$, which implies

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

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

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

By rescaling $p\alpha$ to α, we get

$$|\mathrm{Median}(f_{J_1}(x), \ldots, f_{J_n}(x)) - y|^p \leq \psi(x,y)^p + \alpha,$$

where

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

and

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

We proved the correctness of an α-approximate compression

$$L_p(\mathrm{Median}(f_{J_1}, \ldots, f_{J_n}), S) \leq \inf_{f\in\mathcal{F}} L_p(f, S) + \alpha.$$

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

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

We encode each $\psi(x, y)$ by some approximation $\tilde{\psi}(x, y)$, such that $\left|\tilde{\psi}(x, y) - \psi(x, y)\right| \leq \alpha$, 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

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

7.4 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.)

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 7.3 we proved an approximate compression for general function classes with ℓ_p losses of size $\mathcal{O}\big(\mathrm{fat}_{c\alpha/p} \cdot \mathrm{fat}^*_{c\alpha/p} \cdot \mathrm{polylog}\big(\mathrm{fat}_{c\alpha/p}, \mathrm{fat}^*_{c\alpha/p}, p, 1/\alpha\big)\big)$. We have an immediate corollary for linear functions. Let $\mathrm{Pdim}(\mathcal{F})$ be the pseudo-dimension of a function class $\mathcal{F}$ [25, 214], that can be defined as $\mathrm{Pdim}(\mathcal{F}) = \lim_{\gamma \to 0} \mathrm{fat}_\gamma(\mathcal{F})$. 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 [33], and the dimension of the dual vector space is equal to the one of the primal space. This implies the following.

Corollary 7.6 *Algorithm 9 is a sample compression scheme of size $\mathcal{O}\big(d^2 \cdot \mathrm{polylog}\big(d, p, \frac{1}{\alpha}\big)\big)$ 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 7.7 to 7.9 improve upon these bounds by using a dedicated algorithm for linear functions. We start with the following result:

Theorem 7.7 (Approximate compression for agnostic linear regression) *Let $\mathcal{F} = \big\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq 1\big\}$, $S = \{(\mathbf{x}_i, y_i) : \|\mathbf{x}_i\|_2 \leq 1, \forall i \in [m]\} \subseteq \mathcal{X} \times [0,1]$, and an approximation parameter $\alpha \in (0,1)$. Algorithm 10 is an α-approximate sample compression scheme for the ℓ_p loss of size*

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

Algorithm 10 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 α

2. Define $S_\alpha = A \cup B$, where

$$A = \left\{(\mathbf{x}_i, j\alpha) : i \in [m], j \in \{-1/\alpha, \ldots, -1, 0, 1, \ldots, 1/\alpha\}\right\},\ B = \left\{(\mathbf{x}_i, j(1+\alpha)) : i \in [m], j \in \{-1, +1\}\right\}.$$

▷ Label by ± 1 the discretized dataset with $f^\star$

3. Define

$$S_\alpha(f^\star) = \{((\mathbf{x}_i, \tilde{y}), z) : \text{ for any } (\mathbf{x}_i, \tilde{y}) \in S_\alpha : z = +1 \text{ if } f^\star(\mathbf{x}_i) - \tilde{y} \leq 0, \text{ otherwise } z = -1\}.$$

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 $f_{\text{SVM}} = \rho(\kappa(S)) = (\mathbf{w}_{\text{SVM}}, b_{\text{SVM}})$, 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 $\langle \mathbf{w}_{\text{SVM}}, (\mathbf{x}, y) \rangle + b_{\text{SVM}} = 0$,

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

where $\mathbf{w}^d_{\text{SVM}} = (\mathbf{w}_{\text{SVM}}(1), \ldots, \mathbf{w}_{\text{SVM}}(d))$. It holds that $\mathbf{w}_{\text{SVM}}(d+1) \neq 0$, 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 $\mathbf{w}_{\text{SVM}}(d+1) = 0$ 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 $d + 2 + (d+2)\log(1/\alpha) + d + 2 = \mathcal{O}(d\log(1/\alpha))$. □

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 7.4.

Theorem 7.8 *There exists an efficiently computable (see the linear program in Eq. (7.7)) 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 [215]). We derive a compression scheme from this method.

Proof. 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

$$L(w, b) := \sum_{i\in[m]} |(wx_i + b) - y_i|, \tag{7.5}$$

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_i, y_i) 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[2] $\hat{w} = (y_{\hat{\jmath}} - y_{\hat{\imath}})/(x_{\hat{\jmath}} - x_{\hat{\imath}})$ 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 (7.5) as a linear program.

$$\begin{aligned} \min_{(\epsilon_1,\ldots,\epsilon_m,w,b)\in\mathbb{R}^{m+2}} \quad & \sum_{i=1}^{m} \epsilon_i \quad \text{s.t.} \\ \forall i \in [m] \quad & \epsilon_i \geq 0 \\ \forall i \in [m] \quad & wx_i + b - y_i \leq \epsilon_i \\ \forall i \in [m] \quad & -wx_i - b + y_i \leq \epsilon_i. \end{aligned} \tag{7.6}$$

We observe that the linear program in (7.6) is feasible with a finite solution (and actually, the constraints $\epsilon_i \geq 0$ 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 $\epsilon_i = 0$ 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 7.1.)

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

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

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 $S^-_{a,b} := \{i : |wx_i + b > y_i|\}$

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

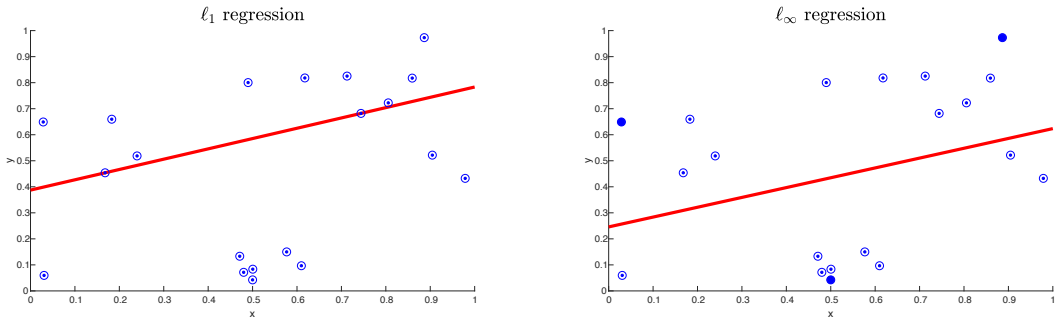

Figure 7.1: Sample compression schemes for ℓ_1 and ℓ_∞ linear regression. A sample S of $m = 20$ points (x_i, y_i) was drawn iid uniformly from $[0,1]^2$. On this sample, ℓ_1 regression was performed by solving the LP in (7.6), shown on the left, and ℓ_∞ regression was performed by solving the LP in (7.8), 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 ●.

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_i, y_i). If the line $(w^\star, b^\pm)$ contains other data points then we are done, so assume to the contrary that $\epsilon_{\hat{\imath}}$ 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 $\epsilon_{\hat{\imath}} = 0$, 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 $\mathcal{P}^\star_{\hat{\imath}} \subset \mathcal{P}_i$ 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 (7.6),

Linear programming for ℓ_1 regression:

$$\begin{aligned} \min_{(\epsilon_1, \ldots, \epsilon_m, \mathbf{w}, b) \in \mathbb{R}^{m+d+1}} \quad & \sum_{i=1}^{m} \epsilon_i \quad \text{s.t.} \\ \forall i \in [m] \quad & \epsilon_i \geq 0 \\ \forall i \in [m] \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \leq \epsilon_i \\ \forall i \in [m] \quad & -\langle \mathbf{w}, \mathbf{x}_i \rangle - b + y_i \leq \epsilon_i. \end{aligned} \tag{7.7}$$

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. □

Similarly, we can obtain a compression scheme for ℓ_∞ loss via linear programming.

Theorem 7.9 *There exists an efficiently computable (see the linear program in Eq. (7.8)) exact compression scheme for agnostic ℓ_∞ linear regression of size $d + 2$.*

Proof. 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

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

We cast the risk minimization problem as a linear program.

Linear programming for ℓ_∞ regression:

$$\begin{aligned} \min_{(\epsilon, \mathbf{w}, b) \in \mathbb{R}^{d+2}} : \quad & \epsilon \\ s.t. \quad \forall i : \quad & \epsilon - \langle \mathbf{w}, \mathbf{x}_i \rangle - b + y_i \geq 0 \\ & \epsilon + \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \geq 0. \end{aligned} \tag{7.8}$$

(As before, the constraint $\epsilon \geq 0$ 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:

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

The KKT conditions imply, in particular, that

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

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

$$\langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i = \epsilon. \tag{7.9}$$

□

In analogy with SVM, let us refer to the datapoints satisfying (7.9) as the *support vectors*; clearly, the remaining sample points may be discarded without affecting the solution. Solutions to (7.8) lie in $\mathbb{R}^{d+2}$ and hence $d+2$ linearly independent datapoints suffice to uniquely pin down an optimal $(\epsilon, \mathbf{w}, b)$ via the equations (7.9).

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. [40, Theorem 4.1].

Theorem 7.10 (David et al. [40]) *There is no exact agnostic sample compression scheme for zero-dimensional linear regression with size $k(m) \leq m/2$.*

Theorem 7.11 *There is no exact agnostic sample compression scheme for zero-dimensional linear regression under ℓ_p loss, $1 < p < \infty$, with size $k(m) < \log(m)$.*

Proof. 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

$$F'(s) = pN_0 s^{p-1} - pN_1(1-s)^{p-1}$$

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 7.12 *A more refined analysis, along the lines of David et al. [40, 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.*

7.5 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 [33, 216, 217] of linear functions in $\mathbb{R}^d$ is precisely $d+1$ [33], 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 7.13 *: **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 and Yehudayoff [132] 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 Warmuth [191], Floyd and Warmuth [195], 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.[3] Noting that, for $\mathcal{F}$ containing $\{0,1\}$-valued functions,

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

$\mathrm{Pdim}(\mathcal{F})$ is equal to 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 7.14 : ***Compressing to a 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 $\mathrm{fat}(\mathcal{F}, c\alpha)$?*

7.6 Deferred Proofs

Our proof for Theorem 7.3 relies on several auxiliary results.

Existence of approximate weak learners. 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 [131], 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

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

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 7.15 *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 9). 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 7.16 *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 9, 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.*

Bibliography

[1] Attias, Idan, Kontorovich, Aryeh, and Mansour, Yishay. Improved generalization bounds for adversarially robust learning. *Journal of Machine Learning Research*, 23(175):1–31, 2022.

[2] Attias, Idan, Hanneke, Steve, and Mansour, Yishay. A characterization of semi-supervised adversarially robust pac learnability. *Advances in Neural Information Processing Systems*, 35:23646–23659, 2022.

[3] Attias, Idan and Hanneke, Steve. Adversarially robust pac learnability of real-valued functions. In *International Conference on Machine Learning*, pages 1172–1199. PMLR, 2023.

[4] Attias, Idan and Kontorovich, Aryeh. Fat-shattering dimension of k-fold aggregations. *Journal of Machine Learning Research*, 25(144):1–29, 2024.

[5] Attias, Idan, Hanneke, Steve, Kontorovich, Aryeh, and Sadigurschi, Menachem. Agnostic sample compression schemes for regression. In *Proceedings of the 41th International Conference on Machine Learning*, 2024.

[6] Vapnik, Vladimir and Chervonenkis, Alexey. Theory of pattern recognition. 1974.

[7] Valiant, Leslie G. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[8] Feige, Uriel, Mansour, Yishay, and Schapire, Robert. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657, 2015.

[9] Montasser, Omar, Hanneke, Steve, and Srebro, Nathan. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530, 2019.

[10] Long, Philip M. On agnostic learning with {0,*, 1}-valued and real-valued hypotheses. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, pages 289–302. Springer, 2001.

[11] Alon, Noga, Hanneke, Steve, Holzman, Ron, and Moran, Shay. A theory of pac learnability of partial concept classes. *arXiv preprint arXiv:2107.08444*, 2021.

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

[13] Littlestone, Nick and Warmuth, Manfred. Relating data compression and learnability. 1986.

[14] Floyd, Sally and Warmuth, Manfred. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[15] Amir, Idan, Attias, Idan, Koren, Tomer, Mansour, Yishay, and Livni, Roi. Prediction with corrupted expert advice. *Advances in Neural Information Processing Systems*, 33:14315–14325, 2020.

[16] Levi, Matan, Attias, Idan, and Kontorovich, Aryeh. Domain invariant adversarial learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[17] Attias, Idan, Cohen, Edith, Shechner, Moshe, and Stemmer, Uri. A framework for adversarial streaming via differential privacy and difference estimators. *Algorithmica*, pages 1–56, 2024.

[18] Mashiah, Eitan-Hai, Attias, Idan, and Mansour, Yishay. Learning revenue maximization using posted prices for stochastic strategic patient buyers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9090–9098, 2023.

[19] Assos, Angelos, Attias, Idan, Dagan, Yuval, Daskalakis, Constantinos, and Fishelson, Maxwell K. Online learning and solving infinite games with an erm oracle. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 274–324. PMLR, 2023.

[20] Attias, Idan, Hanneke, Steve, Kalavasis, Alkis, Karbasi, Amin, and Velegkas, Grigoris. Optimal learners for realizable regression: Pac learning and online learning. *Advances in Neural Information Processing Systems*, 36, 2023.

[21] Attias, Idan, Dziugaite, Gintare Karolina, Haghifam, Mahdi, Livni, Roi, and Roy, Daniel M. Information complexity of stochastic convex optimization: Applications to generalization and memorization. *arXiv preprint arXiv:2402.09327*, 2024.

[22] Liu, Ziyi, Attias, Idan, and Roy, Daniel M. Causal bandits: The pareto optimal frontier of adaptivity, a reduction to linear bandits, and limitations around unknown marginals. *arXiv preprint arXiv:2407.00950*, 2024.

[23] Attias, Idan, Hanneke, Steve, Kalavasis, Alkis, Karbasi, Amin, and Velegkas, Grigoris. Universal rates for regression: Separations between cut-off and absolute loss. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 359–405. PMLR, 2024.

[24] Liu, Ziyi, Attias, Idan, and Roy, Daniel M. Sequential probability assignment with contexts: Minimax regret, contextual shtarkov sums, and contextual normalized maximum likelihood. *arXiv preprint arXiv:2410.03849*, 2024.

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

[26] Vapnik, Vladimir N and Chervonenkis, A Ya. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[27] Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[28] Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations of machine learning*. MIT press, 2018.

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

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

[31] Alon, Noga, Ben-David, Shai, Cesa-Bianchi, Nicolo, and Haussler, David. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

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

[33] Anthony, Martin, Bartlett, Peter L, Bartlett, Peter L, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

[34] Dudley, Richard M. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[35] Mendelson, S. and Vershynin, R. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003.

[36] Bartlett, Peter L and Mendelson, Shahar. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[37] Rudelson, Mark and Vershynin, Roman. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

[38] Assouad, Patrick. Densité et dimension. In *Annales de l'institut Fourier*, volume 33, pages 233–282, 1983.

[39] Kleer, Pieter and Simon, Hans. Primal and dual combinatorial dimensions. *arXiv preprint arXiv:2108.10037*, 2021.

[40] David, Ofir, Moran, Shay, and Yehudayoff, Amir. Supervised learning through the lens of compression. *Advances in Neural Information Processing Systems*, 29:2784–2792, 2016.

[41] Graepel, Thore, Herbrich, Ralf, and Shawe-Taylor, John. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

[42] Maurer, Andreas and Pontil, Massimiliano. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[43] Roth, Aaron. Lecture notes: The algorithmic foundations of adaptive data analysis. 2017.

[44] Schmidt, Ludwig, Santurkar, Shibani, Tsipras, Dimitris, Talwar, Kunal, and Madry, Aleksander. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[45] Madry, Aleksander, Makelov, Aleksandar, Schmidt, Ludwig, Tsipras, Dimitris, and Vladu, Adrian. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[46] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A.S. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[47] Cesa-Bianchi, Nicolo, Mansour, Yishay, and Stoltz, Gilles. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

[48] Freund, Yoav and Schapire, Robert E. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

[49] Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, and Fergus, Rob. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[50] Biggio, Battista, Corona, Igino, Maiorca, Davide, Nelson, Blaine, Šrndić, Nedim, Laskov, Pavel, Giacinto, Giorgio, and Roli, Fabio. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[51] Goodfellow, Ian J, Shlens, Jonathon, and Szegedy, Christian. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[52] Kurakin, Alexey, Goodfellow, Ian, and Bengio, Samy. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[53] Moosavi-Dezfooli, Seyed-Mohsen, Fawzi, Alhussein, and Frossard, Pascal. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[54] Tramèr, Florian, Papernot, Nicolas, Goodfellow, Ian, Boneh, Dan, and McDaniel, Patrick. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

[55] Montasser, Omar, Hanneke, Steve, and Srebro, Nati. Reducing adversarially robust learning to non-robust pac learning. *Advances in Neural Information Processing Systems*, 33:14626–14637, 2020.

[56] Attias, Idan, Kontorovich, Aryeh, and Mansour, Yishay. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183, 2019.

[57] Yin, Dong, Kannan, Ramchandran, and Bartlett, Peter. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.

[58] Awasthi, Pranjal, Frank, Natalie, and Mohri, Mehryar. On the rademacher complexity of linear hypothesis sets. *arXiv preprint arXiv:2007.11045*, 2020.

[59] Cullina, Daniel, Bhagoji, Arjun Nitin, and Mittal, Prateek. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

[60] Khim, Justin and Loh, Po-Ling. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

[61] Raghunathan, Aditi, Xie, Sang Michael, Yang, Fanny, Duchi, John C, and Liang, Percy. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

[62] Diochnos, Dimitrios, Mahloujifar, Saeed, and Mahmoody, Mohammad. Adversarial risk and robustness: General definitions and implications for the uniform distribution. *Advances in Neural Information Processing Systems*, 31, 2018.

[63] Balda, Emilio Rafael, Behboodi, Arash, Koep, Niklas, and Mathar, Rudolf. Adversarial risk bounds for neural networks through sparsity based compression. *arXiv preprint arXiv:1906.00698*, 2019.

[64] Pydi, Muni Sreenivas and Jog, Varun. Adversarial risk via optimal transport and optimal couplings. *arXiv preprint arXiv:1912.02794*, 2019.

[65] Tu, Zhuozhuo, Zhang, Jingwei, and Tao, Dacheng. Theoretical analysis of adversarial learning: A minimax approach. In *Advances in Neural Information Processing Systems*, pages 12259–12269, 2019.

[66] Chen, Lin, Min, Yifei, Zhang, Mingrui, and Karbasi, Amin. More data can expand the generalization gap between adversarially robust and standard models. *arXiv preprint arXiv:2002.04725*, 2020.

[67] Carmon, Yair, Raghunathan, Aditi, Schmidt, Ludwig, Duchi, John C, and Liang, Percy S. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

[68] Alayrac, Jean-Baptiste, Uesato, Jonathan, Huang, Po-Sen, Fawzi, Alhussein, Stanforth, Robert, and Kohli, Pushmeet. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019.

[69] Zhai, Runtian, Cai, Tianle, He, Di, Dan, Chen, He, Kun, Hopcroft, John, and Wang, Liwei. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

[70] Najafi, Amir, Maeda, Shin-ichi, Koyama, Masanori, and Miyato, Takeru. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pages 5542–5552, 2019.

[71] Xu, Huan and Mannor, Shie. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

[72] Bubeck, Sébastien, Price, Eric, and Razenshteyn, Ilya. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

[73] Mahloujifar, Saeed, Diochnos, Dimitrios I, and Mahmoody, Mohammad. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4536–4543, 2019.

[74] Mahloujifar, Saeed and Mahmoody, Mohammad. Can adversarially robust learning leveragecomputational hardness? In *Algorithmic Learning Theory*, pages 581–609. PMLR, 2019.

[75] Chen, Robert S, Lucier, Brendan, Singer, Yaron, and Syrgkanis, Vasilis. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, pages 4705–4714, 2017.

[76] Awasthi, Pranjal, Dutta, Abhratanu, and Vijayaraghavan, Aravindan. On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems*, pages 13737–13747, 2019.

[77] Awasthi, Pranjal, Chatziafratis, Vaggos, Chen, Xue, and Vijayaraghavan, Aravindan. Adversarially robust low dimensional representations. *arXiv preprint arXiv:1911.13268*, 2019.

[78] Sinha, Aman, Namkoong, Hongseok, and Duchi, John. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[79] Diakonikolas, Ilias, Kane, Daniel, and Manurangsi, Pasin. Nearly tight bounds for robust proper learning of halfspaces with a margin. In *Advances in Neural Information Processing Systems*, pages 10473–10484, 2019.

[80] Diakonikolas, Ilias, Kane, Daniel M, and Manurangsi, Pasin. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *Advances in Neural Information Processing Systems*, 33, 2020.

[81] Montasser, Omar, Goel, Surbhi, Diakonikolas, Ilias, and Srebro, Nathan. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning*, pages 7010–7021. PMLR, 2020.

[82] Gourdeau, Pascale, Kanade, Varun, Kwiatkowska, Marta, and Worrell, James. On the hardness of robust classification. In *Advances in Neural Information Processing Systems*, pages 7446–7455, 2019.

[83] Ashtiani, Hassan, Pathak, Vinayak, and Urner, Ruth. Black-box certification and learning under adversarial perturbations. In *International Conference on Machine Learning*, pages 388–398. PMLR, 2020.

[84] Mansour, Yishay, Rubinstein, Aviad, and Tennenholtz, Moshe. Robust probabilistic inference. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 449–460. SIAM, 2014.

[85] Feige, Uriel, Mansour, Yishay, and Schapire, Robert E. Robust inference for multiclass classification. In *Algorithmic Learning Theory*, pages 368–386, 2018.

[86] Globerson, Amir and Roweis, Sam. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pages 353–360, 2006.

[87] Teo, Choon H, Globerson, Amir, Roweis, Sam T, and Smola, Alex J. Convex learning with invariances. In *Advances in neural information processing systems*, pages 1489–1496, 2008.

[88] Dekel, Ofer, Shamir, Ohad, and Xiao, Lin. Learning to classify with missing and corrupted features. *Machine learning*, 81(2):149–178, 2010.

[89] Angluin, Dana and Laird, Philip. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[90] Valiant, Leslie G. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566. Citeseer, 1985.

[91] Kearns, Michael and Li, Ming. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4): 807–837, 1993.

[92] Raviv, Dolev, Hazan, Tamir, and Osadchy, Margarita. Hinge-minimax learner for the ensemble of hyperplanes. *The Journal of Machine Learning Research*, 19(1):2517–2546, 2018.

[93] Foster, Dylan J and Rakhlin, Alexander. ℓ_∞vector contraction for rademacher complexity. *arXiv preprint arXiv:1911.06468*, 2019.

[94] Boucheron, Stéphane, Bousquet, Olivier, and Lugosi, Gábor. Theory of classification: A survey of recent advances. *ESAIM Probab. Statist.*, 9:323–375, 2005.

[95] Ben-David, Shai, Cesa-Bianchi, Nicolò, Haussler, David, and Long, Philip M. Characterizations of learnability for classes of {0, ..., n}-valued functions. *J. Comput. Syst. Sci.*, 50(1):74–86, 1995.

[96] Gottlieb, Lee-Ad, Kontorovich, Aryeh, and Krauthgamer, Robert. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

[97] Alon, Noga, Hanneke, Steve, Holzman, Ron, and Moran, Shay. A theory of pac learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671. IEEE, 2022.

[98] Blumer, Anselm, Ehrenfeucht, Andrzej, Haussler, David, and Warmuth, Manfred K. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989.

[99] Mendelson, Shahar. On the size of convex hulls of small sets. *Journal of Machine Learning Research*, 2(Oct): 1–18, 2001.

[100] Bartlett, Peter and Shawe-Taylor, John. *Generalization performance of support vector machines and other pattern classifiers*, pages 43–54. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3.

[101] Uesato, Jonathan, Alayrac, Jean-Baptiste, Huang, Po-Sen, Stanforth, Robert, Fawzi, Alhussein, and Kohli, Pushmeet. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

[102] Wei, Colin, Shen, Kendrick, Chen, Yining, and Ma, Tengyu. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

[103] Awasthi, Pranjal, Frank, Natalie, Mao, Anqi, Mohri, Mehryar, and Zhong, Yutao. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, 34, 2021.

[104] Montasser, Omar, Hanneke, Steve, and Srebro, Nathan. Transductive robust learning guarantees. *arXiv preprint arXiv:2110.10602*, 2021.

[105] Montasser, Omar, Hanneke, Steve, and Srebro, Nathan. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory*, pages 3452–3482. PMLR, 2021.

[106] Awasthi, Pranjal, Frank, Natalie, and Mohri, Mehryar. On the existence of the adversarial bayes classifier. *Advances in Neural Information Processing Systems*, 34, 2021.

[107] Dan, Chen, Wei, Yuting, and Ravikumar, Pradeep. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

[108] Awasthi, Pranjal, Frank, Natalie, and Mohri, Mehryar. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR, 2020.

[109] Bhattacharjee, Robi, Jha, Somesh, and Chaudhuri, Kamalika. Sample complexity of robust linear classification on separated data. In *International Conference on Machine Learning*, pages 884–893. PMLR, 2021.

[110] Xing, Yue, Zhang, Ruizhi, and Cheng, Guang. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 514–522. PMLR, 2021.

[111] Ashtiani, Hassan, Pathak, Vinayak, and Urner, Ruth. Adversarially robust learning with tolerance. *arXiv preprint arXiv:2203.00849*, 2022.

[112] Blum, Avrim. Semi-supervised learning. In *Encyclopedia of Algorithms*, pages 1936–1941. 2016.

[113] Chapelle, Olivier, Scholkopf, Bernhard, and Zien, Alexander. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[114] Zhu, Xiaojin and Goldberg, Andrew B. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

[115] Urner, Ruth, Shalev-Shwartz, Shai, and Ben-David, Shai. Access to unlabeled data can speed up prediction time. In *ICML*, 2011.

[116] Göpfert, Christina, Ben-David, Shai, Bousquet, Olivier, Gelly, Sylvain, Tolstikhin, Ilya, and Urner, Ruth. When can unlabeled data improve the learning rate? In *Conference on Learning Theory*, pages 1500–1518. PMLR, 2019.

[117] Darnstädt, Malte, Simon, Hans Ulrich, and Szörényi, Balázs. Unlabeled data does provably help. 2013.

[118] Balcan, Maria-Florina and Blum, Avrim. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

[119] Balcan, Maria-Florina and Blum, Avrim. 21 an augmented pac model for semi-supervised learning. 2006.

[120] Auer, Peter and Ortner, Ronald. A new pac bound for intersection-closed concept classes. *Machine Learning*, 66 (2):151–163, 2007.

[121] Darnstädt, Malte. The optimal pac bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2015.

[122] Hanneke, Steve. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17(1):4667–4721, 2016.

[123] Hanneke, Steve. *Theoretical foundations of active learning*. Carnegie Mellon University, 2009.

[124] Long, Philip M. An upper bound on the sample complexity of pac-learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003.

[125] Giné, Evarist and Koltchinskii, Vladimir. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

[126] Bshouty, Nader H, Li, Yi, and Long, Philip M. Using the doubling dimension to analyze the generalization of learning algorithms. *Journal of Computer and System Sciences*, 75(6):323–335, 2009.

[127] Balcan, Maria-Florina and Long, Phil. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316. PMLR, 2013.

[128] Haussler, David, Littlestone, Nick, and Warmuth, Manfred K. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

[129] Warmuth, Manfred K. The optimal pac algorithm. In *International Conference on Computational Learning Theory*, pages 641–642. Springer, 2004.

[130] Schapire, Robert E and Freund, Yoav. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

[131] Hanneke, Steve, Kontorovich, Aryeh, and Sadigurschi, Menachem. Sample compression for real-valued learners. In *Algorithmic Learning Theory*, pages 466–488. PMLR, 2019.

[132] Moran, Shay and Yehudayoff, Amir. Sample compression schemes for vc classes. *Journal of the ACM (JACM)*, 63(3):1–10, 2016.

[133] Sauer, Norbert. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[134] Awasthi, Pranjal, Mao, Anqi, Mohri, Mehryar, and Zhong, Yutao. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174. PMLR, 2022.

[135] Awasthi, Pranjal, Mao, Anqi, Mohri, Mehryar, and Zhong, Yutao. Multi-class h-consistency bounds. *Advances in Neural Information Processing Systems*, 35:782–795, 2022.

[136] Awasthi, Pranjal, Mao, Anqi, Mohri, Mehryar, and Zhong, Yutao. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 10077–10094. PMLR, 2023.

[137] Mao, Anqi, Mohri, Mehryar, and Zhong, Yutao. Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*, 2023.

[138] Anthony, Martin and Bartlett, Peter L. Function learning from interpolation. *Combinatorics, Probability and Computing*, 9(3):213–225, 2000.

[139] Simon, Hans Ulrich. Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, 26(3):751–763, 1997.

[140] Kégl, Balázs. Robust regression by boosting the median. In *Learning Theory and Kernel Machines*, pages 258–272. Springer, 2003.

[141] Duffy, Nigel and Helmbold, David. Boosting methods for regression. *Machine Learning*, 47(2):153–200, 2002.

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

[143] Pollard, David. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

[144] Gourdeau, Pascale, Kanade, Varun, Kwiatkowska, Marta, and Worrell, James. On the hardness of robust classification. *The Journal of Machine Learning Research*, 22(1):12521–12549, 2021.

[145] Bubeck, Sébastien, Lee, Yin Tat, Price, Eric, and Razenshteyn, Ilya. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840. PMLR, 2019.

[146] Srebro, Nathan, Rennie, Jason, and Jaakkola, Tommi. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

[147] Candes, Emmanuel and Recht, Benjamin. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[148] Anava, Oren, Hazan, Elad, Mannor, Shie, and Shamir, Ohad. Online learning for time series prediction. In *Conference on learning theory*, pages 172–184. PMLR, 2013.

[149] Hazan, Elad, Livni, Roi, and Mansour, Yishay. Classification with low rank and missing data. In *International conference on machine learning*, pages 257–266. PMLR, 2015.

[150] Hazan, Elad and Ma, Tengyu. A non-generative framework and convex relaxations for unsupervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[151] Hazan, Elad, Kale, Satyen, and Shalev-Shwartz, Shai. Near-optimal algorithms for online matrix prediction. In *Conference on Learning Theory*, pages 38–1. JMLR Workshop and Conference Proceedings, 2012.

[152] Agarwal, Naman, Bullins, Brian, Hazan, Elad, Kakade, Sham, and Singh, Karan. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

[153] Daniely, Amit, Sabato, Sivan, Ben-David, Shai, and Shalev-Shwartz, Shai. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 207–232. JMLR Workshop and Conference Proceedings, 2011.

[154] Daniely, Amit and Shalev-Shwartz, Shai. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

[155] Angluin, Dana. Queries and concept learning. *Machine learning*, 2:319–342, 1988.

[156] Freund, Yoav and Schapire, Robert 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.

[157] Kearns, Michael J. and Schapire, Robert E. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994. ISSN 0022-0000. doi: http://dx.doi.org/10.1016/S0022-0000(05)80062-5.

[158] Alon, Noga, Ben-David, Shai, Cesa-Bianchi, Nicolò, and Haussler, David. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997. URL citeseer.ist.psu.edu/alon97scalesensitive.html.

[159] Bartlett, Peter L. and Long, Philip M. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *J. Comput. Syst. Sci.*, 56(2):174–190, 1998.

[160] Baum, Eric B. and Haussler, David. What size net gives valid generalization? *Neural Comput.*, 1(1):151–160, 1989. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1989.1.1.151.

[161] Eisenstat, David and Angluin, Dana. The VC dimension of k-fold union. *Inf. Process. Lett.*, 101(5):181–184, 2007.

[162] Eisenstat, David. k-fold unions of low-dimensional concept classes. *Inf. Process. Lett.*, 109(23-24):1232–1234, 2009.

[163] Csikós, Mónika, Mustafa, Nabil H., and Kupavskii, Andrey. Tight lower bounds on the vc-dimension of geometric set systems. *J. Mach. Learn. Res.*, 20:81:1–81:8, 2019.

[164] Alon, Noga, Beimel, Amos, Moran, Shay, and Stemmer, Uri. Closure properties for private classification and online prediction. In *Conference on Learning Theory*, pages 119–152. PMLR, 2020.

[165] Ghazi, Badih, Golowich, Noah, Kumar, Ravi, and Manurangsi, Pasin. Near-tight closure bounds for the littlestone and threshold dimensions. In *Algorithmic Learning Theory*, pages 686–696. PMLR, 2021.

[166] Fefferman, Charles, Mitter, Sanjoy, and Narayanan, Hariharan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[167] Raviv, Dolev, Hazan, Tamir, and Osadchy, Margarita. Hinge-minimax learner for the ensemble of hyperplanes. *J. Mach. Learn. Res.*, 19:62:1–62:30, 2018.

[168] Gottlieb, Lee-Ad, Kaufman, Eran, Kontorovich, Aryeh, and Nivasch, Gabriel. Learning convex polytopes with margin. In *Neural Information Processing Systems (NIPS)*, 2018.

[169] Klochkov, Yegor, Kroshnin, Alexey, and Zhivotovskiy, Nikita. Robust k-means clustering for distributions with two moments. *The Annals of Statistics*, 49(4):2206–2230, 2021.

[170] Biau, Gérard, Devroye, Luc, and Lugosi, Gábor. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

[171] Appert, Gautier and Catoni, Olivier. New bounds for k-means and information k-means. *arXiv preprint arXiv:2101.05728*, 2021.

[172] Zhivotovskiy, Nikita. A bound for k-fold maximum. 2022.

[173] Breiman, Leo. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[174] Breiman, Leo. Random forests. *Machine learning*, 45:5–32, 2001.

[175] Attias, Idan, Kontorovich, Aryeh, and Mansour, Yishay. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory, ALT 2019*, volume 98 of *Proceedings of Machine Learning Research*, pages 162–183. PMLR, 2019.

[176] Alon, Noga, Hanneke, Steve, Holzman, Ron, and Moran, Shay. A theory of PAC learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671, 2022.

[177] Talagrand, Michel. Vapnik–chervonenkis type conditions and uniform donsker classes of functions. *The Annals of Probability*, 31(3):1565–1582, 2003.

[178] Duan, Hubert Haoyang. *Bounding the Fat Shattering Dimension of a Composition Function Class Built Using a Continuous Logic Connective*. PhD thesis, University of Waterloo, 2012.

[179] Srebro, Nathan, Sridharan, Karthik, and Tewari, Ambuj. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

[180] Foster, Dylan J. and Rakhlin, Alexander. ℓ_∞ vector contraction for rademacher complexity. *CoRR*, abs/1911.06468, 2019.

[181] Gottlieb, Lee-Ad, Kontorovich, Aryeh, and Krauthgamer, Robert. Efficient classification for metric data (extended abstract: COLT 2010). *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

[182] Kontorovich, Aryeh. Rademacher complexity of k-fold maxima of hyperplanes. 2018.

[183] Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations Of Machine Learning*. The MIT Press, 2012.

[184] Hanneke, Steve and Kontorovich, Aryeh. Optimality of SVM: novel proofs and tighter bounds. *Theor. Comput. Sci.*, 796:99–113, 2019.

[185] Cohen, Doron, Kontorovich, Aryeh, Koolyk, Aaron, and Wolfer, Geoffrey. Dimension-free empirical entropy estimation. *IEEE Transactions on Information Theory*, 69(5):3190–3202, 2023. doi: 10.1109/TIT.2022.3232739.

[186] Vershynin, Roman, 2021. Private communication.

[187] Rudelson, M. and Vershynin, R. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 164(2):603–648, 2006.

[188] Zhang, Tong. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.

[189] Artstein, S., Milman, V., and Szarek, S. J. Duality of metric entropy. *Annals of Mathematics*, 159(3):1313–1328, 2004.

[190] Vershynin, Roman. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

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

[192] Littlestone, Nick and Warmuth, Manfred K. Relating data compression and learnability. Technical report, Department of Computer and Information Sciences, Santa Cruz, CA, Ju, 1986.

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

[194] Helmbold, David, Sloan, Robert, and Warmuth, Manfred K. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

[195] Floyd, Sally and Warmuth, Manfred K. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

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

[197] Kuzmin, Dima and Warmuth, Manfred K. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007.

[198] Rubinstein, Benjamin IP, Bartlett, Peter L, and Rubinstein, J Hyam. Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59, 2009.

[199] Rubinstein, Benjamin IP and Rubinstein, J Hyam. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13(4), 2012.

[200] Chernikov, Artem and Simon, Pierre. Externally definable sets and dependent pairs. *Israel J. Math.*, 194(1): 409–425, 2013.

[201] Livni, Roi and Simon, Pierre. Honest compressions and their application to compression schemes. In *Conference on Learning Theory*, pages 77–92, 2013.

[202] Moran, Shay, Shpilka, Amir, Wigderson, Avi, and Yehudayoff, Amir. Teaching and compressing for low vc-dimension. In *A Journey Through Discrete Mathematics*, pages 633–656. Springer, 2017.

[203] Bousquet, Olivier, Hanneke, Steve, Moran, Shay, and Zhivotovskiy, Nikita. Proper learning, helly number, and an optimal svm bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020.

[204] Daniely, Amit, Sabato, Sivan, Ben-David, Shai, and Shalev-Shwartz, Shai. Multiclass learnability and the erm principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

[205] Brukhim, Nataly, Carmon, Daniel, Dinur, Irit, Moran, Shay, and Yehudayoff, Amir. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

[206] Wiener, Yair, Hanneke, Steve, and El-Yaniv, Ran. A compression technique for analyzing disagreement-based active learning. *J. Mach. Learn. Res.*, 16:713–745, 2015.

[207] Ashtiani, Hassan, Ben-David, Shai, Harvey, Nicholas JA, Liaw, Christopher, Mehrabian, Abbas, and Plan, Yaniv. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

[208] Montasser, Omar, Hanneke, Steve, and Srebro, Nati. Adversarially robust learning: A generic minimax optimal learner and characterization. *Advances in Neural Information Processing Systems*, 35:37458–37470, 2022.

[209] Gottlieb, Lee-Ad, Kontorovich, Aryeh, and Nisnevitch, Pinhas. Near-optimal sample compression for nearest neighbors. *Advances in Neural Information Processing Systems*, 27, 2014.

[210] Kontorovich, Aryeh, Sabato, Sivan, and Weiss, Roi. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. *Advances in Neural Information Processing Systems*, 30, 2017.

[211] Pabbaraju, Chirag. Multiclass learnability does not imply sample compression. *arXiv preprint arXiv:2308.06424*, 2023.

[212] Graepel, Thore, Herbrich, Ralf, and Shawe-Taylor, John. PAC-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

[213] Mendelson, Shahar. Learning without concentration. *J. ACM*, 62(3):21:1–21:25, 2015.

[214] Pollard, David. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

[215] Dodge, Y. Least absolute deviation regression. *The Concise Encyclopedia of Statistics*, pages 299–302, 2008.

[216] Pollard, David. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

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