Journal of Machine Learning Research 23 (2022) 1-31 Submitted 11/20; Revised 4/22; Published 6/22

Improved Generalization Bounds for Adversarially Robust Learning

Idan Attias IDANATTI@POST.BGU.AC.IL
Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Department of Computer Science
Ben-Gurion University of the Negev, Beer Sheva, Israel

Yishay Mansour MANSOUR.YISHAY@GMAIL.COM
Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv Israel
and Google Research, Tel Aviv

Editor: Sebastien Bubeck

Abstract

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. (2018); Madry et al. (2017).

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. (2015), 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. (2015) 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.

Keywords: Adversarial Robustness, PAC Learning, Sample Complexity, Zero-Sum Game.

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. (2015). 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

©2022 Idan Attias, Aryeh Kontorovich, and Yishay Mansour..

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/20-1353.html.

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. (2018); Madry et al. (2017) and other common robust optimization approaches (Ben-Tal et al., 2009), 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. (2015). 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. (2015) 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. (2007) and based on the ideas of Freund and Schapire (1999). 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 (Szegedy et al., 2013; Biggio et al., 2013; Goodfellow et al., 2014; Kurakin et al., 2016; Moosavi-Dezfooli et al., 2016; Tramèr et al., 2017).

1.1 Subsequent Work: Montasser, Hanneke, and Srebro (2019, 2020b)

Following the conference version (Attias et al., 2019) of this work, Montasser, Hanneke, and Srebro (2019) 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-

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

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

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

which we refer to as the $[0, 1]$-*robust loss*. Montasser et al. (2019, 2020b) defined a loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ as follows

$$\ell_2(h, x, y) = \max_{z \in \rho(x)} \mathbb{I}\left[h(z) \neq y\right], \tag{2}$$

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

$$\ell_3(h_1, \dots, h_T, x, y) = \max_{z \in \rho(x)} \mathbb{I}\left[\text{Majority}(h_1(z), \dots, h_T(z)) \neq y\right], \tag{3}$$

where Majority takes the majority of its the Boolean inputs (and assume that T is odd). Clearly, if $\ell_1(h_1, \dots, h_T, x, y) < 1/2$ then $\ell_3(h_1, \dots, h_T, x, y) = 0$. However, if $\ell_3(h_1, \dots, h_T, x, y) = 0$ it only guarantees that $\ell_1(h_1, \dots, h_T, x, y) < 1/2$ but can be very far from zero. This is why an upper bound on 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 1.3.

The work of Montasser et al. (2019), 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 infinite set of corruptions. The learner competes with the single optimal hypothesis in the class, and outputs a mixture of hypothesis to do so. In this work, considering the $[0, 1]$-robust loss, we would like to guarantee and ϵ-optimal value for the learner in a zero-sum game, via a mixed strategy, and so we find an ϵ-optimal mixture of hypothesis. That is, we compete with the optimal mixture of hypothesis from the function class. In that sense, we are having a proper learning algorithm, with respect to the convex hull of the hypothesis class.

In another closely related work from the computational perspective, Montasser, Hanneke, and Srebro (2020b) reduced the problem of robust learning to non-robust learning. Namely, their algorithm using access to only a black-box PAC learner, similar to the algorithm of Feige et al. (2015) 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.

1.2 Uniform Convergence of the Zero-One Robust Loss Class

For the case of 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 Appendix A).

Lemma 1 *For any class $\mathcal{H}$ of* VC *dimension d, and any adversary $\rho : \mathcal{X} \to 2^{\mathcal{X}}$ such that $|\rho(x)| \leq k$, the* VC*-dimension of the zero-one robust loss class $L^{\rho}_{\mathcal{H}} = \left\{ (x, y) \mapsto \max_{z \in \rho(x)} \mathbb{I}\left[h(z) \neq y\right] : h \in \mathcal{H} \right\}$ is at most $\mathcal{O}(d \log k)$.*

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

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

1.3 Main Results

We provide a summary of the results for the $[0, 1]$-robust loss and the zero-one robust loss (see Eqs. (1) and (2) for the definitions) for robust (ϵ, δ)-PAC learning with finite set of possible 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 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. (2015)
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. (2019)
Uniform Convergence	$\tilde{\mathcal{O}}\left(\frac{kd}{\epsilon^2} + \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$	This work
	REGRESSION	
Uniform Convergence	$\tilde{\mathcal{O}}\left(\inf_{\beta \geq 0} \left\{\beta + \sqrt{\frac{k}{n}} \int_\beta^1 \sqrt{\mathrm{fat}_\gamma(\mathcal{H})} d\gamma\right\} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right)$	This work

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. (2019)
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

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. (2019) 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 $\frac{d\log(k)}{\epsilon^4}$.

1.4 Other Related Work

The most closley related works studying robust learning with adversarial examples are Schmidt et al. (2018); Madry et al. (2017). 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 Montasser et al. (2019); Yin et al. (2019); Awasthi et al. (2020); Cullina et al. (2018); Khim and Loh (2018); Raghunathan et al. (2019); Diochnos et al. (2018); Balda et al. (2019); Pydi and Jog (2019); Tu et al. (2019); Chen et al. (2020); Carmon et al. (2019); Alayrac et al. (2019); Zhai et al. (2019); Najafi et al. (2019); Levi et al. (2021); Attias et al. (2022); Attias and Hanneke (2022). A different notion of robustness by Xu and Mannor (2012) 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 (Bubeck et al., 2018; Mahloujifar et al., 2019; Mahloujifar and Mahmoody, 2019; Chen et al., 2017; Awasthi et al., 2019b,a; Sinha et al., 2017; Diakonikolas et al., 2019, 2020; Montasser et al., 2020a; Gourdeau et al., 2019; Ashtiani et al., 2020).

All of our results based on a robust learning model for binary classification suggested by Feige et al. (2015). The works of Mansour et al. (2014); Feige et al. (2015, 2018) 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, and the main task is given a corrupted input, derive in a computationally efficient way a classification which 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 (2006) and its extensions Teo et al. (2008); Dekel et al. (2010) 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 (Feige et al., 2015) 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 in 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 (Angluin and Laird, 1988) and malicious noise (Valiant, 1985; Kearns and Li, 1993). 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.

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 is able to corrupt an input by mapping an uncorrupted input $x \in \mathcal{X}$ to a corrupted one $z \in \mathcal{Z}$. There is a mapping ρ which for every $x \in \mathcal{X}$ defines a set $\rho(x) \subseteq \mathcal{Z}$, such that $|\rho(x)| \leq k$. The adversary can map an uncorrupted input x to any corrupted input $z \in \rho(x)$. We assume that the learner has an access to $\rho(\cdot)$ 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 $\sum_{i=1}^{T} \alpha_i \mathbb{I}\left[h_i(z) = y\right]$ 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., $\ell(h(z), y) = \mathbb{I}\left[h(z) \neq y\right]$. 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 $z \in \rho(x)$, 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 $z \in \rho(x)$. 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

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

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

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

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 $\rho(\cdot)$,

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

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

$$\mathcal{F}^{(j)}_{\mathcal{H}} := \{\mathcal{X} \times \mathcal{Y} \ni (x, y) \mapsto \mathbb{I}[h(z_j) \neq y] : h \in \mathcal{H},\ \rho(x) = \{z_1, \ldots, z_k\}\}, \tag{4}$$

where we treat the set-valued output of $\rho(x)$ 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

$$\mathrm{conv}^{\rho}(L_{\mathcal{H}}) \subseteq \max\left((\mathrm{conv}(\mathcal{F}^{(j)}_{\mathcal{H}}))_{j\in[k]} \right). \tag{5}$$

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

$$\mathrm{Risk}(h) = \mathbb{E}_{x\sim D}[\max_{z\in\rho(x)} \ell(h(z), c(x))],$$

and the empirical error on sample S under corruption of $\rho(\cdot)$ by

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

3.1 Combinatorial Dimensions and Capacity Measures

Rademacher Complexity. Let $\mathcal{H}$ be of real valued function class on the domain space $\mathcal{W}$. Define the empirical Rademacher complexity on a given sequence $\mathbf{w} = (w_1, \ldots, w_n) = w_{1:n} \in \mathcal{W}^n$:

$$R_n(\mathcal{H}|\mathbf{w}) = E_\sigma \sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(w_i).$$

Fat-Shattering Dimension. 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 $\mathrm{fat}_\gamma(\mathcal{F})$, is the size of the largest γ-shattered set (possibly ∞).

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 $\mathcal{H}$.

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 $\mathcal{H}$.

Growth Function. The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \mapsto \mathbb{N}$ for a binary function class $\mathcal{H} : \mathcal{X} \mapsto \{0, 1\}$ is defined by

$$\forall m \in \mathbb{N},\ \Pi_{\mathcal{H}}(m) = \max_{\{x_1, \ldots, x_m\} \subseteq \mathcal{X}} |\{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\}|$$

And the VC-dimension of $\mathcal{H}$ is defined by

$$\mathrm{VC}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}.$$

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 corruptions by the adversary. We employ the regret minimization algorithm proposed by Feige et al. (2015) 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 $|\rho(x)| \times \mathcal{H}$ 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 $P_x \in \Delta(\rho(x))$, for a given $x \in \mathcal{X}$, is a distribution over the corrupted inputs $\rho(x)$. 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

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

Hence, the learner selects

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

A hypotheses 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 $P_x \in \Delta(\rho(x))$ over corrupted inputs.

Algorithm 1

parameter: $\eta > 0$

for all $(x, y) \in S, z \in \rho(x)$ **do** ▷ initialize weights and distributions vector

  $w_1(z, (x, y)) \leftarrow 1, \forall (x, y) \in S, \forall z \in \rho(x)$

  $P^1(z, (x, y)) \leftarrow \frac{w_1(z,(x,y))}{\sum_{z' \in \rho(x)} w_1(z',(x,y))}$ ▷ for each $(x, y) \in S$ we have a distribution over $\rho(x)$

end for

for `t = 1:T` **do**

  $h_t \leftarrow \arg\min_{h \in \mathcal{H}} E_{(z,y) \sim D^{P^t}}[\ell(h(z), y]$ ▷ using the ERM oracle for $\mathcal{H}$

  for all $(x, y) \in S, z \in \rho(x)$ **do** ▷ update weights for P^{t+1}

    $w_{t+1}(z, (x, y)) \leftarrow (1 + \eta \cdot [\ell(h_t(z), y)]) \cdot w_t(z, (x, y))$

    $P^{t+1}(z, (x, y)) \leftarrow \frac{w_{t+1}(z,(x,y))}{\sum_{z' \in \rho(x)} w_{t+1}(z',(x,y))}$

  end for

end for

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

Similar to Feige et al. (2015, Theorem 1), for the binary classification case and zero-one loss we have:

Theorem 3 *(Feige, Mansour, and Schapire, 2015, 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.

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. (2015), 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 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 $n_0 = \frac{C_\alpha}{\epsilon^2}\Big(k\,\mathrm{VC}(\mathcal{H})\log^{\frac{3}{2}+\alpha}(k\,\mathrm{VC}(\mathcal{H})) + \log(\frac{1}{\delta})\Big)$, such that for $|S| \geq n_0$, for every $\tilde{h} \in \Delta(\mathcal{H})$*

$$|\operatorname{Risk}(\tilde{h}) - \widehat{\operatorname{Risk}}(\tilde{h})| \leq \epsilon$$

with probability at least $1 - \delta$.

Theorem 5 *(Mohri et al., 2018, Theorem 3.3) 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 $S = (w_1, \cdots, w_n) = \mathbf{w}$ from distribution D, for all $g \in \mathcal{G}$:*

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

Remark. The corresponding result in the conference version of this paper, Attias et al. (2019), Theorem 2, was proved via Lemma 3 therein. The latter contained a mistake, as pointed out to us by Digvijay Pravin Boob and Praneeth Netrapalli. The current proof relies on a recent result of Foster and Rakhlin (2019).

Theorem 6 *(Foster and Rakhlin, 2019) Let $\mathcal{F}$ be a $\mathbb{R}^k$-valued function class, such that the coordinate projection class is denoted by $\mathcal{F}_j = \{w \mapsto f(w)_j \mid f \in \mathcal{F}\}$, 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 $\mathbf{w} = (w_1, \cdots, w_n)$,*

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

Proof [Proof of Theorem 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. (5), $\mathrm{conv}^\rho(L_\mathcal{H}) \subseteq \max(\mathrm{conv}(\mathcal{F}_\mathcal{H}^{(j)}))_{j\in[k]})$. Recall that functions contained in $\mathcal{F}_\mathcal{H}^{(j)}$ are loss functions of the learner when the adversary corrupts input x to $z_j \in \rho(x)$. We are left to bound the Rademacher complexity of the function class $\max((\mathrm{conv}(\mathcal{F}_\mathcal{H}^{(j)}))_{j\in[k]})$. Formally,

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

where the inequality stems from applying Theorem 5 on the function class $\mathrm{conv}^\rho(L_\mathcal{H})$ and Eq. (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 $\mathcal{F} = \{(x,y) \mapsto (f_1(x,y), \cdots, f_k(x,y)) \,|\, f_j \in \mathrm{conv}(\mathcal{F}_\mathcal{H}^{(j)}), 1 \leq j \leq k\}$ we can apply Theorem 6, for any $\alpha > 0$, there exists a constant $C_\alpha > 0$ such that

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

where the last equality follows from the well-known identity $R_n(\mathcal{F}|\mathbf{w}) = R_n(\mathrm{conv}(\mathcal{F})|\mathbf{w})$, (see, e.g., Boucheron et al. (2005, 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 (Bartlett and Mendelson, 2002): $R_n(\mathcal{F}|\mathbf{w}) \leq C\sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}}$, and for $\left(C\sqrt{\mathrm{VC}(\mathcal{F}_\mathcal{H})}e^{3/2+\alpha}\right)^{2/3} \leq n$, by the monotonicity of the function $x\log^{3/2+\alpha}(n/x)$ we get an upper bound of

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

where the inequality follows from Lemma 8. We require that

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

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

$$
R_n\left(\max((\mathrm{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})|\mathbf{x}\times\mathbf{y}\right)\leq\mathcal{O}\left(\sqrt{\frac{\mathrm{VC}(\mathcal{H})\log^2(\mathrm{VC}(\mathcal{H}))k\log k\log^9(n)}{n}}\right)\tag{6}
$$

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

Remark. Theorem 4 provides an improvement to Theorem 7 in Raviv, Hazan, and Osadchy (2018), where they considered learning with intersection of hyperplanes for imbalanced binary classification problem.

5.1 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.

Theorem 7 (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 $\mathrm{d}_G(\mathcal{H})$. 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 $\tilde{h}\in\Delta(\mathcal{H})$,*

$$
|\operatorname{Risk}(\tilde{h})-\widehat{\operatorname{Risk}}(\tilde{h})|\leq\epsilon
$$

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. (4)).

Lemma 8 *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 $\{(x_1, y_1), \dots, (x_d, y_d)\} \subset \mathcal{X} \times \mathcal{Y}$. 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 ρ on input x. We claim that $\mathcal{H}$ is able to G-shatter $S = \{z_j(x_1), \dots, z_j(x_d)\} \subset \mathcal{Z}$. 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 7, we follow the same proof of Theorem 4 and use the Graph-dimension property of Lemma 8.

Remark. A similar bound to that of Theorem 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. (1995).

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 9 (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,

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

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 A) we derive a simplified bound for hyperplanes.

Corollary 10 *Let $\mathcal{H}$ be a function class of homogeneous hyperplanes with domain $\mathbb{R}^m$. Using the same assumptions as in Theorem 9, we have*

$$|\operatorname{Risk}(\tilde{h}) - \widehat{\operatorname{Risk}}(\tilde{h})| \leq \mathcal{O}\left(C_\alpha\sqrt{\frac{k}{n}}\log^{\frac{7}{2}}(n)\log^{\frac{3}{2}+\alpha}\left(\frac{n}{\log^{\frac{7}{2}}(n)}\right) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right),$$

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 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*

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

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

$$|\operatorname{Risk}(\tilde{h}) - \widehat{\operatorname{Risk}}(\tilde{h})| \leq \mathcal{O}\left(m_n(\mathcal{H}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right),$$

with probability at least $1-\delta$.

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

Theorem 12 (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 $d = \operatorname{fat}_{\frac{\epsilon}{4}}(\mathcal{H})$. 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 $\tilde{h} \in \Delta(\mathcal{H})$

$$|\operatorname{Risk}(\tilde{h}) - \widehat{\operatorname{Risk}}(\tilde{h})| \leq \epsilon$$

with probability at least $1-\delta$.

We would like to compare the bounds in Theorems 9, 11 and 12. In terms of dependence in the fat-shattering dimension and k, Theorem 9 would give a better bound than Theorem 11. However, the latter has a better dependence in $\log(n)$ factors. Regarding Theorem 12, 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 $\operatorname{fat}_\gamma(\mathcal{H}) = 1/\gamma^6$, Theorem 9 guarantees learning rate of $1/n^{1/6}$ and so Theorem 12 provides a sharper bound.

6.1 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 13 (Fat-shattering of k-fold maxima) *Let $S = \{x_1, \ldots, x_m\}$. For any k real valued functions classes $\mathcal{F}_1, \ldots, \mathcal{F}_k \subseteq \mathbb{R}^S$,*

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

Remark. It was pointed out to us by Yann Guermeur that the corresponding result in the conference version of this paper, Attias et al. (2019), Theorem 12, contained a mistake — the root of which was an erroneous claim in Lemma 14 therein. The corrected version of that result was proved by Alon, Hanneke, Holzman, and Moran (2022), Lemma 15 below, allows for a corrected version of Theorem 12, with an additional $\log^2(m)$ factor. In a subsequent work (Kontorovich and Attias, 2021, Theorem 1), this result was further improved,

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

Moreover, Attias and Hanneke (2022) studied the regression setting with ℓ_p losses and arbitrary perturbation sets.

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. (2014)). The analogous dimension will be denoted by $\mathrm{fat}^0_\gamma(\mathcal{F})$.

Lemma 14 *For all $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and $\gamma > 0$, we have*

$$\mathrm{fat}_\gamma(\mathcal{F}) = \max_{r\in\mathbb{R}^{\mathcal{X}}}\mathrm{fat}^0_\gamma(\mathcal{F} - r), \tag{7}$$

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

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

It was mistakenly claimed in the conference version (Attias et al., 2019, 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 Appendix A for completeness.

Lemma 15 *(Alon, Hanneke, Holzman, and Moran, 2022, 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 2 *(Alon, Hanneke, Holzman, and Moran, 2022) 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.

Remark. There exist an ambiguous function class $F^\star$, such that for any disambiguation $\bar{F}$ it holds that $\mathrm{VC}(\bar{F}) = \infty$. See Alon, Hanneke, Holzman, and Moran (2022), Theorem 1.

Lemma 16 *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. (1989, 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 that of Blumer et al.: one readily verifies that for $m = |S| = 2\bar{d}k \log(3k)$, we have $(em/\bar{d})^{\bar{d}k} < 2^m$. ∎

Proof [Proof of Theorem 13] To prove the Theorem, it suffices to show that for all $\mathcal{F}_j \subseteq \mathbb{R}^S$

$$\mathrm{fat}_\gamma^0(\max((\mathcal{F}_j)_{j\in[k]})) \leq \mathcal{O}(\log(k) \log^2(m) \sum_{j=1}^{k} \mathrm{fat}_\gamma^0(\mathcal{F}_j)). \tag{8}$$

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

$$\max((\mathcal{F}_j - r)_{j\in[k]}) = \max((\mathcal{F}_j)_{j\in[k]}) - r. \tag{9}$$

By applying Lemma 14 to the function class $\max((\mathcal{F}_j)_{j\in[k]})$ and using Eq. (9), we have

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

Applying Eq. (8) to classes $F_j - r$ obtains

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

Then,

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

where the last identity follows from Lemma 14.

Now we proceed to prove Eq. (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: $v_i^\star = \operatorname{sgn}(v_i)\gamma$ 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 $\mathcal{F}_j$ has the same γ-shattering dimension at zero.

Lemma 15 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 $j \in [k]$,

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

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 16 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

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

We have shown that,

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

this concludes the proof of Eq. (8). ∎

6.2 Shattering Dimension of L_1 and L_2 Loss Classes

Lemma 17 *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 $\mathcal{H}$,*

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

Lemma 18 *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be an arbitrary loss function. For $j \in [k]$ define*

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

and

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

Then, for all $\gamma > 0$,

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

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

Proof [Proof of Lemma 17] 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 $\mathrm{fat}_\gamma(\mathcal{H}^{\Delta}) \le \mathrm{fat}_\gamma(\mathcal{H})$. 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 13 that

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

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

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

Then

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

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

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

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

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

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

■

6.3 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 13.

Theorem 19 *(Mendelson, 2001, Theorem 1.5) There is an absolute constant C, such that for every function class F bounded by $[0,1]$ and every $\gamma > 0$,*

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

Corollary 20 *Let $S = \{x_1, \dots, x_m\}$. For any k real valued functions classes $\mathcal{F}_1, \dots, \mathcal{F}_k \subseteq [0,1]^S$,*

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

Proof

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

where (i) stems from Theorem 13 and (ii) stems from Theorem 19. ■

Theorem 21 *(Dudley, 1967; Mendelson and Vershynin, 2003) For any $\mathcal{F} \subseteq [-1,1]^X$, any $\gamma \in (0,1)$ and $S = (w_1, \dots, w_n) = \mathbf{w} \in \mathcal{W}^n$,*

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

where c and C are universal constants.

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

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

6.4 Proofs

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

Proof [Proof of Theorem 9] We follow the same steps as in the proof of Theorem 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 21,

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

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 17 and Lemma18,

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

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 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

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

We conclude that

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

■

Proof [Proof of Theorem 11] Similar to the proof for binary case, we bound the empirical Rademacher complexity of the loss class of $\tilde{h} \in \Delta(\mathcal{H})$.

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

where the inequality stems from applying Theorem 5 on the function class $\operatorname{conv}^{\rho}(L_{\mathcal{H}})$ and Eq. (5). From Theorem 21 we have

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

Using Corollary 20 we upper bound the inner term by

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

Lemmas 17 and 18 concludes the proof with

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

■

Proof [Proof of Theorem 12] Denote the sample size by $|S| = n$. We start off with a known generalization bound by Bartlett and Long (1998), showing that for any function class $\mathcal{H} : \mathcal{Z} \to [0,1]$, the sample size is at least

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

In our case, the function class we are interested in is $\max((\operatorname{conv}(\mathcal{F}_{\mathcal{H}}^{(j)}))_{j\in[k]})$. by Corollary 20 we have that

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

Thus, it suffices to solve the following

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

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,

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

■

Acknowledgments

We deeply thank the anonymous reviewers for many insightful comments and suggestions, in particular, for providing us the proof about the VC-dimension of the zero-one robust loss class (Lemma 1).

We thank Digvijay Boob, Praneeth Netrapalli, and Yann Guermeur for pointing out a mistakes in the conference version of this paper. We thank Shay Moran, Ron Holzman, and Steve Hanneke for resolving a mistake (see Lemma 15) and providing Example 2. We are also gratful to Uri Stemmer for fruitful discussions. IA is thankful to Doron Cohen and Netanel Rabinowitz for many helpful discussions.

The work of AK and YM was supported in part by grants from the Israel Science Foundation (ISF). The work of IA was supported in part by Kreitman School and the Vatat Scholarship from the Israeli Council for Higher Education.

Appendix A. Additional Proofs

Proof [of Lemma 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, $S_\rho = \bigcup_{i\in[n]} \{z : z \in \rho(x_i)\}$, the size of S_ρ is at most nk. Denote by $L^\rho_{\mathcal{H}}(S)$ the set of all possible behaviors on S using functions in $L^\rho_{\mathcal{H}}$, and by $\mathcal{H}(S_\rho)$, the set of all possible behaviors on S_ρ using functions in $\mathcal{H}$. Namely, $L^\rho_{\mathcal{H}}(S) = \left\{(\ell(x_1, y_1), \dots, \ell(x_n, y_n)) : \ell \in L^\rho_{\mathcal{H}}\right\}$ and $\mathcal{H}(S_\rho) = \{(h(z_1), \dots, h(z_m)) : h \in \mathcal{H}\}$. Observe that each pattern in the set $L^\rho_{\mathcal{H}}(S)$ will map to at least one pattern in $\mathcal{H}(S_\rho)$, implying that the size of $L^\rho_{\mathcal{H}}(S)$ is at most the size of $\mathcal{H}(S_\rho)$. Using Sauer's lemma, the size of $\mathcal{H}(S_\rho)$ is at most $(nk)^d$, solving for n such that $(nk)^d < 2^n$ yields the stated bound. ■

Proof [Proof of Corollary 10] We seek an upper bounds on the following term in the case of homogeneous hyperplanes with norm bounded by 1.

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

where the inequality stems from the bound $\mathrm{fat}_\delta(H) \leq \frac{1}{\delta^2}$ (Bartlett and Shawe-Taylor, 1999).

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

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

■

Proof [of Lemma 15] For any finite sequence $(x_1, y_1), \dots, (x_k, y_k)$ with $x_i \in X$, $y_i \in \{0, 1\}$, and $x_1 < \dots < x_k$, denote by $F^\star|_{(x_1,y_1),\dots,(x_k,y_k)}$ 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*:

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

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

$$w(F^\star) \leq \sum_{n \in \mathbb{N}} \frac{n^{d-1}}{n^{d+1}} = \sum_{n \in \mathbb{N}} \frac{1}{n^2} = \frac{\pi^2}{6}, \tag{11}$$

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:

$$M[F^\star|_{(x_1,y_1),\ldots,(x_k,y_k)}](x) \quad = \quad \left[\!\left[w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k),(x,1)}) > w(F^\star|_{(x_1,y_1),\ldots,(x_k,y_k),(x,0)}) \right]\!\right] \tag{12}$$

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

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

Additional Generalization Bound for Binary Classification. We derive the result in Eq. (6). Denote the sample size $|S| = n$ and $\mathrm{VC}(\mathcal{H}) = d$. Using Theorem 11 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

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

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

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

References

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems*, pages 12192–12202, 2019.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. 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.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Black-box certification and learning under adversarial perturbations. *arXiv preprint arXiv:2006.16520*, 2020.

Idan Attias and Steve Hanneke. Adversarially robust learning of real-valued functions. *arXiv preprint arXiv:2206.12977*, 2022.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183, 2019.

Idan Attias, Steve Hanneke, and Yishay Mansour. A characterization of semi-supervised adversarially-robust pac learnability. *arXiv preprint arXiv:2202.05420*, 2022.

Pranjal Awasthi, Vaggos Chatziafratis, Xue Chen, and Aravindan Vijayaraghavan. Adversarially robust low dimensional representations. *arXiv preprint arXiv:1911.13268*, 2019a.

Pranjal Awasthi, Abhratanu Dutta, and Aravindan Vijayaraghavan. On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems*, pages 13737–13747, 2019b.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the rademacher complexity of linear hypothesis sets. *arXiv preprint arXiv:2007.11045*, 2020.

Emilio Rafael Balda, Arash Behboodi, Niklas Koep, and Rudolf Mathar. Adversarial risk bounds for neural networks through sparsity based compression. *arXiv preprint arXiv:1906.00698*, 2019.

Peter Bartlett and John Shawe-Taylor. *Generalization performance of support vector machines and other pattern classifiers*, pages 43–54. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3.

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

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler, and Philip M. Long. Characterizations of learnability for classes of {0, ..., n}-valued functions. *J. Comput. Syst. Sci.*, 50(1):74–86, 1995.

A. Ben-Tal, L. El Ghaoui, and A.S. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 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.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of recent advances. *ESAIM Probab. Statist.*, 9:323–375, 2005.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. *arXiv preprint arXiv:2002.04725*, 2020.

Robert S Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, pages 4705–4714, 2017.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

Ofer Dekel, Ohad Shamir, and Lin Xiao. Learning to classify with missing and corrupted features. *Machine learning*, 81(2):149–178, 2010.

Ilias Diakonikolas, Daniel Kane, and Pasin Manurangsi. Nearly tight bounds for robust proper learning of halfspaces with a margin. In *Advances in Neural Information Processing Systems*, pages 10473–10484, 2019.

Ilias Diakonikolas, Daniel M. Kane, and Pasin Manurangsi. The complexity of adversarially robust proper learning of halfspaces with agnostic noise, 2020.

Dimitrios Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, pages 10359–10368, 2018.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Uriel Feige, Yishay Mansour, and Robert Schapire. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657, 2015.

Uriel Feige, Yishay Mansour, and Robert E Schapire. Robust inference for multiclass classification. In *Algorithmic Learning Theory*, pages 368–386, 2018.

Dylan J Foster and Alexander Rakhlin. ℓ_∞vector contraction for rademacher complexity. *arXiv preprint arXiv:1911.06468*, 2019.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Amir Globerson and Sam Roweis. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pages 353–360, 2006.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. On the hardness of robust classification. In *Advances in Neural Information Processing Systems*, pages 7446–7455, 2019.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2, 2018.

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

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Matan Levi, Idan Attias, and Aryeh Kontorovich. Domain invariant adversarial learning. *arXiv preprint arXiv:2104.00322*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Saeed Mahloujifar and Mohammad Mahmoody. Can adversarially robust learning leveragecomputational hardness? In *Algorithmic Learning Theory*, pages 581–609. PMLR, 2019.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. 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.

Yishay Mansour, Aviad Rubinstein, and Moshe Tennenholtz. Robust probabilistic inference. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 449–460. SIAM, 2014.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1): 37–55, 2003. ISSN 0020-9910. doi: 10.1007/s00222-002-0266-3. URL http://dx.doi.org/10.1007/s00222-002-0266-3.

Shahar Mendelson. On the size of convex hulls of small sets. *Journal of Machine Learning Research*, 2(Oct):1–18, 2001.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530, 2019.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. *arXiv preprint arXiv:2005.07652*, 2020a.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Reducing adversarially robust learning to non-robust pac learning. *arXiv preprint arXiv:2010.12039*, 2020b.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 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.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pages 5542–5552, 2019.

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. *arXiv preprint arXiv:1912.02794*, 2019.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Dolev Raviv, Tamir Hazan, and Margarita Osadchy. Hinge-minimax learner for the ensemble of hyperplanes. *Journal of Machine Learning Research*, 19(62):1–30, 2018. URL http://jmlr.org/papers/v19/17-402.html.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Choon H Teo, Amir Globerson, Sam T Roweis, and Alex J Smola. Convex learning with invariances. In *Advances in neural information processing systems*, pages 1489–1496, 2008.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Zhuozhuo Tu, Jingwei Zhang, and Dacheng Tao. Theoretical analysis of adversarial learning: A minimax approach. In *Advances in Neural Information Processing Systems*, pages 12259–12269, 2019.

Leslie G Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566. Citeseer, 1985.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.