Improved Generalization Bounds for 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

Editors: Aurélien Garivier and Satyen Kale

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 effected by the adversary during testing. The learner's goal is to build a robust classifier that would be tested on future adversarial examples. We use a zero-sum game between the learner and the adversary as our game theoretic framework. The adversary is limited to k possible corruptions for each input. Our 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 multi-class classification, as well as the real-valued case (regression). For the binary classification setting, we both tighten the generalization bound of Feige, Mansour, and Schapire (2015), 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 $\mathcal{O}(\frac{1}{\epsilon^2}(k\log(k)\,\mathrm{VC}(\mathcal{H})+\log\frac{1}{\delta}))$. 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 blackbox; we adapt it for the multi-class and regression settings. The algorithm provides us with near optimal policies for the players on a given training sample.

Keywords: Robust Learning, Adversarial Learning, Generalization Bounds, Rademacher Complexity, Fat-Shattering Dimension, 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, both at training and testing time, may be corrupted in an adversarial manner to disrupt the learner's performance. This challenge to design reliable robust models gained significant attention as standard supervised learning methods have shown vulnerability, and is named *adversarial examples*. We study the adversarially robust learning paradigm from a generalization point of view and concentrate on the case of having adversarial examples at test time.

We consider the following robust learning framework for multi-class 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 access during the training phase to all possible corruptions that the adversary might effect. The learner selects a hypothesis from a fixed hypothesis class (in our case, a mixture of hypotheses from base class $\mathcal{H}$) that gives a prediction (a distribution over predictions) for a corrupted input. The learner's accuracy is measured by predicting the true label of the uncorrupted input while they observe only the corrupted input during test time. Thus, their goal is to find a policy that is immune to those corruptions. The adversary is capable of corrupting each future input, but there are only k possible corruptions for each input. This leads to a game theoretic framework of a zero-sum game between the learner and the adversary. The model is closely related to the one suggested by Schmidt et al. (2018); Madry et al. (2017) and 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 sample space to any other space, but is limited with finite possible corruptions for each input.

Our focus is on adversarial examples during testing time. The training data is clean, but we take into consideration all possible corruptions when we build the robust classifier. 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., 2015; Kurakin et al., 2017; Moosavi-Dezfooli et al., 2016; Tramèr et al., 2017). Another interesting direction (not pursued here) is the setting where the training data is adversarially corrupted beforehand, without any direct access to uncorrupted sample.

Studying worst-case (adversarial) corruptions is interesting for a couple of reasons. First, it models situations where the corruption occurs maliciously (such as a spammer who tailors messages to avoid a spam detector) and not merely as a result of random noise. Additionally, a robust classifier that is successful against adversarial corruption extends to less adversarial settings.

Our main results are generalization bounds for all settings. For the binary classification setting, we improve the generalization bound given in Feige et al. (2015). We generalize to the case of mixture of hypotheses from $\mathcal{H}$ when $\mathcal{H}$ is not necessarily finite. The sample complexity has been improved from $\mathcal{O}(\frac{1}{\epsilon^4}\log(\frac{|\mathcal{H}|}{\delta}))$ to $\mathcal{O}(\frac{1}{\epsilon^2}(k\log(k)\,\mathrm{VC}(\mathcal{H})+\log\frac{1}{\delta}))$. 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 case we provide a tight bound on the fat shattering dimension of k-fold maximum class and bound the fat shattering dimension of L_1 and L_2 loss classes.

For our algorithm, we employ a regret minimization algorithm proposed for binary classification by Feige et al. (2015) for computing near optimal policies for the players on the training data. We adapt it multiclass classification and regression as well. The algorithm is a variant of the algorithm found in Cesa-Bianchi et al. (2005) and based on the ideas of Freund and Schapire (1999). An ERM (empirical risk minimization) oracle is used multiple times 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.

1.1. Related work

The most related work 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 closely related works that analyse the theoretical aspects of adversarial robust generalization and learning rules are (Yin et al., 2018; Khim and Loh, 2018; Cullina et al., 2018; Bubeck et al., 2018; Chen et al., 2017; Diochnos et al., 2018; Mahloujifar et al., 2018; Mahloujifar and Mahmoody, 2018). A different notion of robustness by (Xu and Mannor, 2012) is shown to be sufficient and necessary for standard generalization.

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. (2015); Feige et al. (2015, 2018) consider *robust inference* for the binary and multi-class 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 and Shamir (2008) discuss a robust learning model where an uncorrupted sample is drawn from an unknown distribution, and the goal is to learn a linear classifier that would be able to overcome 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.

Other motivating applications such as spam messages detection, web spam detection, computer intrusion detection, fraud detection, network failure detection, noisy bio-sensors measurements, and many more can be found in Laskov and Lippmann (2010).

1.2. The structure of the paper

The structure of this paper is as follows: Section 2 discusses the model in detail. Section 3 contains relevant definitions and notations. Section 4 is the learning algorithm, and Sections 5, 6 and 7 contain the generalization bounds for binary and multiclass classification and regression.

2. Model

There is some unknown distribution D over a finite domain $\mathcal{X}$ of uncorrupted examples and a finite domain of corrupted examples $\mathcal{Z}$, possibly the same as $\mathcal{X}$. We work in deterministic scenario, where there is some concept class C such that $c \in C$ has domain $\mathcal{X}$ and range $\mathcal{Y}$ that can be $\{1, \ldots, l\}$ or $\mathbb{R}$. There is some unknown target function $c^* \in C$ which maps an uncorrupted example to its label.

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 $\sum_{i=1}^{T} \alpha_i (\mathbb{I}_{\{h_i(z)=y\}})$ where $\sum_{i=1}^{T} \alpha_i$=1. For each hypothesis $h \in \mathcal{H}$ in the mixture we use the zero-one loss to measure the quality of the classification, i.e., $\mathsf{loss}(h(z), y) = \mathbb{I}_{\{h(z) \neq y\}}$. The loss of $\tilde{h} \in \Delta(\mathcal{H})$ is defined by $\mathsf{loss}(\tilde{h}(z), y) = \sum_{i=1}^{T} \alpha_i \mathsf{loss}(h_i(z), y)$. 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., $\mathsf{loss}(h(z), y) = |h(z) - y|^p$, 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 $\mathsf{loss}(\tilde{h}(z), y) = \sum_{i=1}^{T} \alpha_i \mathsf{loss}(h_i(z), y)$.

The basic scenario is as follows. First, an uncorrupted input $x \in \mathcal{X}$ is selected using 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

$$\text{Risk}(\tilde{h}) = \mathbb{E}_{x \sim D}[\max_{z \in \rho(x)} \mathsf{loss}(\tilde{h}(z), c^*(x)],$$

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

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

Uncorrupted Training Data Learning Algorithm. Uncorrupted training data (UTD) learning algorithm receives an uncorrupted sample S_u and outputs a hypothesis $h \in \mathcal{H}$ (mixture of hypotheses in our case). The UTD-learning algorithm (ϵ, δ)-learns C if for any target function $c^* \in C$, with probability $1 - \delta$, the algorithm outputs some hypothesis $h \in \mathcal{H}$, such that $\text{Risk}(h) \leq v + \epsilon$. The risk is measured by adversarially corrupted inputs as mentioned above.

3. Definitions and Notations

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 $\mathsf{loss}(h(z), y) \leq M$. Without the loss of generality we use $M = 1$, otherwise, we can rescale M.

Define the following operations on the loss class $L_{\mathcal{H}}$. The convex hull of $L_{\mathcal{H}}$ is the set of all convex combinations of hypotheses from $L_{\mathcal{H}}$:

$$\operatorname{conv}(L_{\mathcal{H}}) = \left\{ \mathcal{Z} \times \mathcal{Y} \ni (z, y) \mapsto \sum_{t=1}^{T} \alpha_t f_t(z, y) : T \in \mathbb{N}, \alpha_t \in [0, 1], \sum_{t=1}^{T} \alpha_t = 1, f_t \in L_{\mathcal{H}} \right\}.$$

The convex hull of $L_{\mathcal{H}}$, where the data is corrupted by $\rho(\cdot)$, is denoted by

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

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

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

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.

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

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

We also define a hybrid $\max$–conv operator:

$$\max\text{–}\operatorname{conv}((\mathcal{A}^{(j)})_{j \in [k]}) := \left\{ W \ni w \mapsto \max_{j \in [k]} \sum_{t=1}^{T} \alpha_t f^{(j)}_t(w) : T \in \mathbb{N}, \alpha_t \in [0, 1], \sum_{t=1}^{T} \alpha_t = 1, f^{(j)}_t \in \mathcal{A}^{(j)} \right\}.$$

Note that

$$\max\text{–}\operatorname{conv}((\mathcal{A}^{(j)})_{j \in [k]}) \subseteq \max(\operatorname{conv}((\mathcal{A}^{(j)})_{j \in [k]})),$$

and the containment will generally be strict, since the former requires the same choice of convex coefficients for all $\mathcal{A}^{(j)}$'s, while the latter allows distinct ones.

Remark. Observe that

$$\mathrm{conv}^{\rho}(L_{\mathcal{H}}) = \max\text{–}\mathrm{conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})$$

We use the notation $\max\text{–}\mathrm{conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})$ and exploit its structural properties.

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

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

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

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

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 empricial Rademacher complexity on a given sequence $\mathbf{w} = (w_1, \ldots, w_n) \in \mathcal{W}^n$:

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

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 fuction $\Pi_{\mathcal{H}} : \mathbb{N} \mapsto \mathbb{N}$ for a binary function class $\mathcal{H} : \mathcal{X} \mapsto \{0, 1\}$ is defined by

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

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 $\{1, \dots, l\}$ 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 multi-class 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

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

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** $\triangleright$ initialize weights and distributions vector

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

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

end for

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

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

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

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

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

$\quad$ **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 1 *(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., $\mathsf{loss}(h(z),y) \leq 1$, $\forall x \in \mathcal{X}, \forall z \in \mathcal{Z}, \forall h \in \mathcal{H}$, the result remains the same for the other settings.

5. Generalization Bound For Binary 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 $\mathcal{O}(\frac{1}{\epsilon^2}(k\log(k)\,\mathrm{VC}(\mathcal{H}) + \log\frac{1}{\delta}))$.

Theorem 2 *Let $\mathcal{H} : \mathcal{Z} \mapsto \{0,1\}$ be a hypothesis class with finite VC-dimension. There is a sample complexity $m_0 = \mathcal{O}(\frac{1}{\epsilon^2}(k\log(k)\,\mathrm{VC}(\mathcal{H}) + \log\frac{1}{\delta}))$, such that for $|S| \geq m_0$, for every $\tilde{h} \in \Delta(\mathcal{H})$*

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

with probability at least $1-\delta$.

Lemma 3 *For any k real valued function classes $\mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(k)}$, over a set $\mathcal{X}$ and $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we have*

$$R_n(\text{max–conv}((\mathcal{F}^{(j)})_{j\in[k]})|\mathbf{x}) \quad = \quad R_n(\max((\mathcal{F}^{(j)})_{j\in[k]})|\mathbf{x}).$$

Proof It is easily seen that $\max((\mathcal{F}^{(j)})_{j\in[k]}) \subseteq \max\text{–conv}((\mathcal{F}^{(j)})_{j\in[k]})$, since for any $f^{(j)} \in \mathcal{F}^{(j)}$, $j \in [k]$ and $T = 1$ we have that $\max_j f^{(j)} \in \max\text{–conv}((\mathcal{F}^{(j)})_{j\in[k]})$. This proves that the right-hand side is at least as large as the left-hand side. Conversely,

$$\begin{aligned} R_n(\max\text{–conv}((\mathcal{F}^{(j)})_{j\in[k]})|\mathbf{x}) &= E_\sigma \sup_{\substack{(f_t^{(j)})_{j\in[k],t\in[T]} \\ (\alpha_t)_{t\in[T]}}} \frac{1}{n}\sum_{i=1}^n \sigma_i \max_{j\in[k]} \sum_{t=1}^T \alpha_t f_t^{(j)}(x_i) \\ &\leq E_\sigma \sup_{\substack{(f_t^{(j)})_{j\in[k],t\in[T]} \\ (\alpha_t)_{t\in[T]}}} \frac{1}{n}\sum_{i=1}^n \sigma_i \sum_{t=1}^T \alpha_t \max_{j\in[k]} f_t^{(j)}(x_i) \\ &= R_n(\text{conv}(\max((\mathcal{F}^{(j)})_{j\in[k]}))|\mathbf{x}) = R_n(\max((\mathcal{F}^{(j)})_{j\in[k]})|\mathbf{x}), \end{aligned}$$

where the last equality stems from the well-known identity $R_n(\mathcal{F}|\mathbf{x}) = R_n(\text{conv}(\mathcal{F})|\mathbf{x})$ (Boucheron et al., 2005, Theorem 3.3). ■

Lemma 4 *Let $\Psi : \mathcal{F}_1 \times \ldots \times \mathcal{F}_k \to \mathcal{F}$ be an arbitrary mapping, where $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_k \subseteq \{-1,1\}^X$ and $\text{VC}(\mathcal{F}_j) = d_j$ for $j \in [k]$. Then the VC-dimension of $\Psi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)$ is less than $2k\log(3k)\bar{d}$, where $\bar{d} := \frac{1}{k}\sum_{i=1}^k d_j$.*

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 $\Psi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)(S) \subseteq \{-1,1\}^S$ to be the restriction of $\Psi(\mathcal{F}_1 \times \ldots \times \mathcal{F}_k)$ to S. The key observation is that

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

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 $(em/\bar{d})^{\bar{d}k} < 2^m$. ■

Theorem 5 *(Mohri et al., 2012, Theorem 3.1) Let $\mathcal{F}$ be a family of functions mapping from $\mathcal{W}$ to [0,1]. Denote $E(f) := E_{w\sim D}[f(w)]$ and $\hat{E}_S(f) = \frac{1}{n}\sum_{i=1}^n f(w_i)$. Then, for every $\delta > 0$ with* *probability at least* $1-\delta$, *for all* $f \in \mathcal{F}$*:*

$$E[f] - \hat{E}_S(f) \leq 2R_n(\mathcal{F}|\mathbf{w}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Theorem 6 *(Dudley, 1967; Bartlett and Mendelson, 2002) Let* $d = \mathrm{VC}(\mathcal{F})$ *and* $S = (w_1, \ldots, w_n) = \mathbf{w} \in \mathcal{W}^n$*, then for* $n \geq d \geq 1$ *and for some absolute constant* $c > 0$*:*

$$R_n(\mathcal{F}|\mathbf{w}) \leq c\sqrt{\frac{d}{n}}.$$

Proof of Theorem 2 Our strategy is to bound the empirical Rademacher complexity of the loss class of $\tilde{h} \in \Delta(\mathcal{H})$. As we mentioned in Section 3, the loss class is $\mathrm{conv}^{\rho}(L_{\mathcal{H}}) = \max\text{–}\mathrm{conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})$, we use the notation $\max\text{–}\mathrm{conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})$. Recall that functions contained in $\mathcal{F}_{\mathcal{H}}^{(j)}$ are loss functions of the learner when the adversary corrupts input x to $z_j \in \rho(x)$. Combining everything together,

$$\begin{aligned}
|\operatorname{Risk}(\tilde{h}) - \widehat{\operatorname{Risk}}(\tilde{h})| &= |E_{(x,y)\sim D} \max_{j\in[k]} \sum_{t=1}^{T} \alpha_t f_t^{(j)}(x,y) - \frac{1}{|S|}\sum_{(x,y)\in S} \max_{j\in[k]} \sum_{t=1}^{T} \alpha_t f_t^{(j)}(x,y)| \\
&\overset{(i)}{\leq} 2R_n(\max\text{–}\mathrm{conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})|(\mathbf{x}\times\mathbf{y})) + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(ii)}{=} 2R_n(\max((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})|(\mathbf{x}\times\mathbf{y})) + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(iii)}{\leq} 2c\sqrt{\frac{\mathrm{VC}(\max((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})}{|S|}} + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(iv)}{\leq} 2c\sqrt{\frac{2k\log(3k)\max_{j\in[k]}\mathrm{VC}(\mathcal{F}_{\mathcal{H}}^{(j)})}{|S|}} + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(v)}{\leq} 2c\sqrt{\frac{2k\log 3k\,\mathrm{VC}(\mathcal{H})}{|S|}} + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\leq \epsilon,
\end{aligned}$$

where (i) stems from Theorem 5 (generalization error by Rademacher complexity bound for the function class $\max\text{–}\mathrm{conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})$, (ii) stems from Lemma 3, (iii) stems from Theorem 6 (bound on Rademacher complexity using VC-dimension), (iv) stems from Lemma 4 and (v) stems from Lemma 8. ■

6. Generalization Bound For Multi-Class 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 *Let $\mathcal{H}$ be a function class with domain $\mathcal{Z}$ and range $\mathcal{Y} = [l]$ with finite Graph-dimension $d_G(\mathcal{H})$. There is a sample complexity $m_0 = \mathcal{O}(\frac{1}{\epsilon^2}(k\log(k)d_G(\mathcal{H}) + \log\frac{1}{\delta}))$, such that for $|S| \geq m_0$, for every $\tilde{h} \in \Delta(\mathcal{H})$,*

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

with probability at least $1-\delta$.

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*

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

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

For the proof of Theorem 7, we follow the same proof of Theorem 2 and use the Graph-dimension property of Lemma 8 in the *(v)* inequality.

Remark. A similar bound to that of Theorem 2 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).

7. Generalization Bound For Regression

Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ be a hypothesis class of real functions. We refer to a bounded regression problem, we assume a bounded loss function by 1.

In order to use similar arguments to the binary case, we need an analogous to Lemma 4 for the fat-shattering dimension and understand the connection between the shattering dimension of loss classes (L_1 and L_2) to the original function class.

Theorem 9 *Let $\mathcal{H}$ be a function class with domain $\mathcal{Z}$ and range $\mathbb{R}$. Assume $\mathcal{H}$ has a finite γ-fat-shattering dimension for all $\gamma > 0$. Denote $m_{\mathcal{H}}(\gamma) = \int_0^1 \sqrt{\operatorname{fat}_{c\gamma}(\mathcal{H})\log(\frac{2}{\gamma})}d\gamma$, where c¿0 is an absolute constant. For the L_1 loss function, there is a sample complexity $m_0 = \mathcal{O}(\frac{1}{\epsilon^2}(k\log(k)m_{\mathcal{H}}(\gamma) + \log\frac{1}{\delta}))$, such that for $|S| \geq m_0$, for every $\tilde{h} \in \Delta(\mathcal{H})$,*

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

with probability at least $1-\delta$.

Remark. In case of integral divergence we can use a refined version of Dudley's entropy integral as in Theorem 17.

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, the sample complexity is $m_0 = \mathcal{O}(\frac{1}{\epsilon^2}(k\log^2(k/\epsilon) + \log\frac{1}{\delta}))$.*

Corollary 11 *For the L_2 loss the same result of Theorem 9 holds when we redefine $m_{\mathcal{H}}(\gamma) = \int_0^1 \sqrt{\operatorname{fat}_{c\gamma/2}(\mathcal{H})\log(\frac{2}{\gamma})}d\gamma$.*

7.1. Shattering dimension of the class $\max((\mathcal{A}^{(j)})_{j\in[k]})$

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

Theorem 12 *For any k real valued functions classes $\mathcal{F}_1,\ldots,\mathcal{F}_k$ with finite fat-shattering dimension, we have*

$$\mathrm{fat}_\gamma(\max((\mathcal{F}_j)_{j\in[k]})) < 2\log(3k)\sum_{j=1}^{k}\mathrm{fat}_\gamma(\mathcal{F}_j)$$

for all $\gamma > 0$.

Remark. This result generalizes an analogous bound obtained in Kontorovich (2018) for maxima of linear classes. For a fixed scale $\gamma > 0$, Csikos et al. (2018) have recently shown that the $O(\log k)$ factor cannot, in general, be removed. Whether a single function class can attain the lower bound for *every* $\gamma > 0$ simultaneously is an open problem.

We begin with an auxiliary definition. We say that $\mathcal{F}$ "γ-shatters a set S at zero" if the shift (or witness) 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 $\mathrm{fat}_\gamma^0(\mathcal{F})$.

Lemma 13 *For all $\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}}$ and $\gamma > 0$, we have*

$$\mathrm{fat}_\gamma(\mathcal{F}) = \max_{r\in\mathbb{R}^{\mathcal{X}}}\mathrm{fat}_\gamma^0(\mathcal{F}-r), \tag{1}$$

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 (1) 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. ■

Lemma 14 *Suppose that $F\subseteq\{-1,1,\star\}^m$ has VC-dimension d in the sense that some $J\subseteq[m]$ of size d verifies $F(J) = \{-1,1\}^J$ and for all $J'\subseteq[m]$ with $|J'| > d$, we have $F(J')\subsetneq\{-1,1\}^{J'}$. Then there is a mapping $\varphi : F\to\{-1,1\}^m$ such that (i) for all $v\in F$ and all $i\in[m]$, we have $v_i\neq\star \implies (\varphi(v))_i = v_i$ and (ii) $\varphi(F)$ does not shatter more than d points.*

Proof The mapping φ must resolve each "ambiguity" $v_i = \star$ as $(\varphi(v))_i\in\{-1,1\}$ in such a way that the resulting set of vectors $\varphi(F)$ does not shatter more points than F does. We achieve this via an iterative procedure, which initializes $F' := F$ and modifies each $v\in F'$, element-wise, until $F'\subseteq\{-1,1\}^m$ — that is, all of the ambguities have been resolved.

Suppose that the VC-dimension of F' is d and some $v\in F'$ and $i\in[m]$ are such that $v_i = \star$; we must choose a value for $(\varphi(v))_i\in\{-1,1\}$. If one of these choices ensures the condition that the VC-dimension will not increase, then we're done. Otherwise, the VC-dimension will increase from d to $d+1$ for *both* choices of $(\varphi(v))_i = 1$ and $(\varphi(v))_i = -1$. This means, in particular, that F shatters some set $J \subseteq [m]$ of size d and $i \notin J$ — since otherwise, disambiguating v_i from $\star$ to ± 1 would not increase the VC-dimension. Since the choice $(\varphi(v))_i = 1$ increases the VC-dimension, it must be the case that

$$F(J \cup \{i\}) = \{-1, 1\}^{J \cup \{i\}} \setminus \left\{ z_{J \cup \{i\}} \right\} \tag{2}$$

for some "missing witness" $z \in \{-1, 1, \star\}^m$, which agrees with v on J and $z_i = 1$; the notation z_E indicates the restriction of Z to the index set $E \subseteq [m]$. Analogously, since the choice $(\varphi(v))_i = -1$ also increases the VC-dimension, we have

$$F(J \cup \{i\}) = \{-1, 1\}^{J \cup \{i\}} \setminus \left\{ z'_{J \cup \{i\}} \right\}, \tag{3}$$

where $z'_J = v_J$ and $z'_i = -1$. The conditions (2) and (3) are in obvious contradiction, from which we conclude that the ambiguity can be resolved for each $v_i = \star$ without increasing the VC-dimension. ■

Proof of Theorem 12 First, we observe that r-shift commutes with the max operator:

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

and so, in light of Lemma 13, we have

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

Hence, to prove Theorem 12, it suffices to show that

$$\mathrm{fat}^0_\gamma(\max((\mathcal{F}_j)_{j \in [k]})) \le 2 \log(3k) \sum_{j=1}^{k} \mathrm{fat}^0_\gamma(\mathcal{F}_j). \tag{4}$$

To prove (4), let us fix some $S = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ and convert each $\mathcal{F}_j(S) \subseteq \mathbb{R}^m$ to a finite class $\mathcal{F}^\star_j(S) \subseteq \{-\gamma, \gamma, \star\}^m$ as follows. For every vector in $v \in \mathcal{F}_j(S)$, define $v^\star \in \mathcal{F}^\star_j(S)$ by: $v^\star_i = \mathrm{sgn}(v_i)\gamma$ if $|v_i| \ge \gamma$ and $v^\star_i = \star$ else. The notion of shattering (at zero) remains the same: a set $T \subseteq S$ is shattered by $\mathcal{F}_j$ iff $\mathcal{F}^\star_j(T) = \{-\gamma, \gamma\}^T$. Lemma 14 furnishes a mapping $\varphi : \mathcal{F}^\star_j(S) \to \{-\gamma, \gamma\}^m$ such that (i) for all $v \in \mathcal{F}^\star_j(S)$ and all $i \in [m]$, we have $v_i \ne \star \implies (\varphi(v))_i = v_i$ and (ii) $\varphi(\mathcal{F}^\star_j(S))$ does not shatter more points than $\mathcal{F}^\star_j(S)$. Together, properties (i) and (ii) imply that $\mathrm{fat}^0_\gamma(\mathcal{F}_j(S)) = \mathrm{fat}^0_\gamma(\varphi(\mathcal{F}^\star_j(S)))$ for all j.

Finally, observe that any d points in S γ-shattered by $\max(\mathcal{F}_{j \in [k]})$ are also shattered by $\max(\varphi(\mathcal{F}^\star_{j \in [k]}(S))$. Applying Lemma 4 with $\Psi(f_1, \ldots, f_k)(x) = \max_{j \in [k]} f_j(x)$ shows that $\max(\varphi(\mathcal{F}^\star_{j \in [k]}(S))$ cannot shatter $2 \log(3k) \sum_{j=1}^k d_j$ points, where $d_j = \mathrm{fat}^0_\gamma(\varphi(\mathcal{F}^\star_j(S))) = \mathrm{fat}^0_\gamma(\mathcal{F}_j(S)) \le \mathrm{fat}^0_\gamma(\mathcal{F}_j)$. We have shown that, for all finite $S \subseteq \mathcal{X}$, we have $\mathrm{fat}^0_\gamma(\max(\mathcal{F}_{j \in [k]}(S))) \le 2 \log(3k) \sum_{j=1}^k \mathrm{fat}^0_\gamma(\mathcal{F}_j(S)$.

Since this latter estimate holds independently of S, it is also an upper bound on $\mathrm{fat}^0_\gamma(\max(\mathcal{F}_{j \in [k]}))$. ■

7.2. Shattering dimension of L_1 and L_2 loss classes

Lemma 15 *Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be a real valued function class. 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 $\mathcal{H}$,*

$$\operatorname{fat}_\gamma(L^1_{\mathcal{H}}) \quad \le \quad 8 \operatorname{fat}_\gamma(\mathcal{H}), \quad \textit{and} \quad \operatorname{fat}_\gamma(L^2_{\mathcal{H}}) \le 8 \operatorname{fat}_{\gamma/2M}(\mathcal{H}).$$

Lemma 16 *For $p \in \{1,2\}$ and $j \in [k]$, define*

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

Then

$$\operatorname{conv}^\rho(L^p_{\mathcal{H}}) = \max\text{–}\operatorname{conv}((\mathcal{F}^{(p,j)}_{\mathcal{H}})_{j \in [k]})$$

and, for all $\gamma > 0$,

$$\operatorname{fat}_\gamma(\mathcal{F}^{(p,j)}_{\mathcal{H}}) \le \operatorname{fat}_\gamma(L^p_{\mathcal{H}}).$$

Proof The first claim is immediate from the definitions, while the second is proved using the argument (almost verbatim) of Lemma 8. ■

Proof [of Lemma 15] 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 $\operatorname{fat}_\gamma(\mathcal{H}^\Delta) \le \operatorname{fat}_\gamma(\mathcal{H})$. Indeed, if $\mathcal{H}^\Delta$ γ-shatters some set $\{(x_1,y_1), \dots, (x_k,y_k)\} \subset \mathcal{X} \times \mathbb{R}$ with shift $r \in \mathbb{R}^k$, then $\mathcal{H}$ γ-shatters the set $\{x_1, \dots, x_k\} \subset \mathcal{X}$ with shift $r + (y_1, \dots, y_k)$.

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 $\operatorname{abs}(\mathcal{F}) := \{|f|;\ f \in \mathcal{F}\}$. Observe that $\operatorname{abs}(\mathcal{F}) \subseteq \max((F_j)_{j \in [2]})$, where $\mathcal{F}_1 = \mathcal{F}$ and $\mathcal{F}_2 = -\mathcal{F} =: \{-f; f \in \mathcal{F}\}$. It follows from Theorem 12 that

$$\operatorname{fat}_\gamma(\operatorname{abs}(\mathcal{F})) < 2 \log 6 (\operatorname{fat}_\gamma(\mathcal{F}) + \operatorname{fat}_\gamma(-\mathcal{F})) < 8 \operatorname{fat}_\gamma(\mathcal{F}). \tag{5}$$

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

$$\begin{aligned} \operatorname{fat}_\gamma(\mathcal{F}) &= \operatorname{fat}_\gamma(\operatorname{abs}(\mathcal{H}^\Delta)) \\ &\le 8 \operatorname{fat}_\gamma(\mathcal{H}^\Delta) \\ &\le 8 \operatorname{fat}_\gamma(\mathcal{H}); \end{aligned}$$

this proves the claim for L_1.

To analyze the L_2 case, consider $\mathcal{F} \subset [0, M]^{\mathcal{X}}$ and define $\mathcal{F}^{\circ 2} := \{f^2; f \in \mathcal{F}\}$. We would like to bound $\mathrm{fat}_\gamma(\mathcal{F}^{\circ 2})$ in terms of $\mathrm{fat}_\gamma(\mathcal{F})$. Suppose that $\mathcal{F}^{\circ 2}$ γ-shatters some set $\{x_1, \ldots, x_k\}$ with shift $r^2 = (r_1^2, \ldots, r_k^2) \in [0, M]^k$ (there is no loss of generality in assuming that the shift has the same range as the function class). Using the elementary inequality

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

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

To extend this result to the case where $\mathcal{F} \subset [-M, M]^{\mathcal{X}}$, we use (5). In particular, define $\mathcal{F}$ as the L_2 loss class of $\mathcal{H}$:

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

Then

$$\begin{aligned}
\mathrm{fat}_\gamma(\mathcal{F}) &= \mathrm{fat}_\gamma((\mathcal{H}^\Delta)^{\circ 2}) \\
&= \mathrm{fat}_\gamma((\mathrm{abs}(\mathcal{H}^\Delta))^{\circ 2}) \\
&\leq \mathrm{fat}_{\gamma/(2M)}(\mathrm{abs}(\mathcal{H}^\Delta)) \\
&\leq 8\,\mathrm{fat}_{\gamma/(2M)}(\mathcal{H}^\Delta) \\
&\leq 8\,\mathrm{fat}_{\gamma/(2M)}(\mathcal{H}).
\end{aligned}$$

■

7.3. Generalization bound proof

Theorem 17 *(Dudley, 1967; Mendelson and Vershynin, 2003) For any $\mathcal{F} \subseteq [-1, 1]^X$, any $\gamma \in (0, 1)$ and $S = (w_1, \ldots, w_n) = \mathbf{w} \in \mathcal{W}^n$,*

$$R_n(\mathcal{F}|\mathbf{w}) \leq 12\sqrt{\frac{\tilde{K}}{n}} \int_0^1 \sqrt{\mathrm{fat}_{c\gamma}(\mathcal{F}) \log(\frac{2}{\gamma})} d\gamma$$

where c and $\tilde{K}$ are universal constants.

Remark. *When the integral above diverges, a refined version is available:*

$$R_n(\mathcal{F}|\mathbf{w}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + 12\sqrt{\frac{\tilde{K}}{n}} \int_\alpha^1 \sqrt{\mathrm{fat}_{c\gamma}(\mathcal{F}) \log(\frac{2}{\gamma})} d\gamma \right\}$$

Proof [of Theorem 9] Similar to the proof for binary case, we bound the empirical Rademacher complexity of the loss class of $\tilde{h} \in \Delta(\mathcal{H})$.

$$
\begin{aligned}
|\operatorname{Risk}(\tilde{h}) - \widehat{\operatorname{Risk}}(\tilde{h})| &= |E_{(x,y)\sim D} \max_{j\in[k]} \sum_{t=1}^{T} \alpha_t f_t^{(j)}(x,y) - \frac{1}{|S|} \sum_{(x,y)\in S} \max_{j\in[k]} \sum_{t=1}^{T} \alpha_t f_t^{(j)}(x,y)| \\
&\overset{(i)}{\leq} 2R_n(\text{max–conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})|(\mathbf{x}\times\mathbf{y})) + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(ii)}{=} 2R_n(\max((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})|(\mathbf{x}\times\mathbf{y})) + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(iii)}{\leq} 24\sqrt{\frac{\tilde{K}}{|S|}} \int_0^1 \sqrt{\operatorname{fat}_{c\gamma}(\max((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]}) \log(\frac{2}{\gamma})} d\gamma + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(iv)}{\leq} 24\sqrt{\frac{2k\log(3k)\tilde{K}}{|S|}} \int_0^1 \sqrt{\max_{j\in[k]} \operatorname{fat}_{c\gamma}(\mathcal{F}_{\mathcal{H}}^{(j)}) \log(\frac{2}{\gamma})} d\gamma + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\overset{(v)}{\leq} 24\sqrt{\frac{2k\log(3k)\tilde{K}}{|S|}} \int_0^1 \sqrt{(8\operatorname{fat}_{c\gamma}(\mathcal{H}))\log(\frac{2}{\gamma})} d\gamma + 3\sqrt{\frac{\log(2/\delta)}{2|S|}} \\
&\leq \epsilon
\end{aligned}
$$

(i) stems from Theorem 5 (generalization error by Rademacher complexity bound for the function class $\text{max–conv}((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})$), (ii) stems from Lemma 3, (iii) stems from Theorem 17, (iv) stems from Theorem 12 (fat shattering of max operator) and (v) stems from Lemmas 15 and 16.

■

Proof [of Corollary 10] Let $\mathcal{H}$ be a function class of homogeneous hyperplanes bounded by 1 with domain $\mathbb{R}^m$.

$$
\begin{aligned}
R_n(\max((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})|(\mathbf{x}\times\mathbf{y})) &\leq \inf_{\alpha\geq 0}\left\{4\alpha + 12\sqrt{\frac{\tilde{K}}{|S|}} \int_\alpha^1 \sqrt{\operatorname{fat}_{c\gamma}(\max((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]}))\log(\frac{2}{\gamma})} d\gamma\right\} \\
&\leq \inf_{\alpha\geq 0}\left\{4\alpha + 12\sqrt{\frac{2k\log(3k)\tilde{K}}{|S|}} \int_\alpha^1 \sqrt{\max_{j\in[k]} \operatorname{fat}_{c\gamma}(\mathcal{F}_{\mathcal{H}}^{(j)})\log(\frac{2}{\gamma})} d\gamma\right\} \\
&\leq \inf_{\alpha\geq 0}\left\{4\alpha + 12\sqrt{\frac{2k\log(3k)\tilde{K}}{|S|}} \int_\alpha^1 \sqrt{(8\operatorname{fat}_{c\gamma}(\mathcal{H}))\log(\frac{2}{\gamma})} d\gamma\right\} \\
&\overset{(i)}{\leq} \inf_{\alpha\geq 0}\left\{4\alpha + 12c'\sqrt{\frac{2k\log(3k)}{|S|}} \int_\alpha^1 \frac{1}{t}\sqrt{\log\frac{2}{t}} dt\right\}
\end{aligned}
$$

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

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

and choosing $\alpha = 1/\sqrt{|S|}$ yields

$$\begin{aligned} R_n(\max((\mathcal{F}_{\mathcal{H}}^{(j)})_{j\in[k]})|(\mathbf{x}\times\mathbf{y})) &\le \frac{4}{\sqrt{|S|}} + 8c'\sqrt{\frac{2k\log(3k)}{|S|}}\left(\log(2\sqrt{|S|})^{3/2} - (\log 2)^{3/2}\right) \\ &= O\left(\sqrt{\frac{k\log k\cdot(\log|S|)^3}{|S|}}\right). \end{aligned}$$

A standard calculation yields sample complexity $m_0 = \mathcal{O}(\frac{1}{\epsilon^2}(k\log^2(k/\epsilon) + \log\frac{1}{\delta}))$. ■

Acknowledgments

The work of AK and YM was supported in part by grants from the Israel Science Foundation (ISF). This research was also supported by the Lynn and William Frankel Center for Computer Science at Ben-Gurion University.

References

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, Apr 1988. ISSN 1573-0565. doi: 10.1023/A:1022873112823. URL https://doi.org/10.1023/A:1022873112823.

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 Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 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 Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40994-3.

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.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. In *COLT*, volume 3559 of *Lecture Notes in Computer Science*, pages 217–232. Springer, 2005.

Robert S. Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4705–4714. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7056-robust-optimization-for-non-convex-objectives.pdf.

Monika Csikos, Andrey Kupavskii, and Nabil H. Mustafa. Optimal bounds on the vc-dimension. *CoRR*, abs/1807.07924, 2018.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. In *NeurIPS*, pages 228–239, 2018.

Ofer Dekel and Ohad Shamir. Learning to classify with missing and corrupted features. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 216–223, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390184. URL http://doi.acm.org/10.1145/1390156.1390184.

Dimitrios I. 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 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 10380–10389, 2018. URL http://papers.nips.cc/paper/8237-adversarial-risk-and-robustness-general-definitions-and-implications-for-

R.M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967. ISSN 0022-1236. doi: http://dx.doi.org/10.1016/0022-1236(67)90017-1. URL http://www.sciencedirect.com/science/article/pii/0022123667900171.

Uriel Feige, Yishay Mansour, and Robert E. Schapire. Learning and inference in the presence of corrupted inputs. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 637–657, 2015.

Uriel Feige, Yishay Mansour, and Robert E. Schapire. Robust inference for multiclass classification. In *Algorithmic Learning Theory, ALT 2018, 7-9 April 2018, Lanzarote, Canary Islands, Spain*, pages 368–386, 2018.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999. URL https://EconPapers.repec.org/RePEc:eee:gamebe:v:29:y:1999:i:1-2:p:79-103.

Amir Globerson and Sam Roweis. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 353–360, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143889. URL http://doi.acm.org/10.1145/1143844.1143889.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014. doi: 10.1109/TIT.2014.2339840. URL http://dx.doi.org/10.1109/TIT.2014.2339840.

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

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *CoRR*, abs/1810.09519, 2018.

Aryeh Kontorovich. Rademacher complexity of k-fold maxima of hyperplanes. 2018.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. 2017. URL https://arxiv.org/abs/1611.01236.

Pavel Laskov and Richard Lippmann. Machine learning in adversarial environments. *Mach. Learn.*, 81(2):115–119, November 2010. ISSN 0885-6125. doi: 10.1007/s10994-010-5207-6. URL https://doi.org/10.1007/s10994-010-5207-6.

Saeed Mahloujifar and Mohammad Mahmoody. Can adversarially robust learning leverage computational hardness? *CoRR*, abs/1810.01407, 2018. URL http://arxiv.org/abs/1810.01407.

Saeed Mahloujifar, Dimitrios I. Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *CoRR*, abs/1809.03063, 2018. URL http://arxiv.org/abs/1809.03063.

Yishay Mansour, Aviad Rubinstein, and Moshe Tennenholtz. Robust probabilistic inference. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 449–460, 2015.

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.

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

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.

S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, June 2016. doi: 10.1109/CVPR.2016.282.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *CoRR*, abs/1804.11285, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. URL http://arxiv.org/abs/1312.6199.

Choon H. Teo, Amir Globerson, Sam T. Roweis, and Alex J. Smola. Convex learning with invariances. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1489–1496. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3218-convex-learning-with-invariances.pdf.

Florian Tramèr, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. The space of transferable adversarial examples. *CoRR*, abs/1704.03453, 2017.

L. G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'85, pages 560–566, San Francisco, CA, USA, 1985. Morgan Kaufmann Publishers Inc. ISBN 0-934613-02-8. URL http://dl.acm.org/citation.cfm?id=1625135.1625242.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*, 86(3):391–423, Mar 2012. ISSN 1573-0565. doi: 10.1007/s10994-011-5268-1. URL https://doi.org/10.1007/s10994-011-5268-1.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *CoRR*, abs/1810.11914, 2018. URL http://arxiv.org/abs/1810.11914.

Appendix A. Proof of Lemma 8

Suppose that the binary function class $\mathcal{F}_{\mathcal{H}}^{(j)}$ shatters the points $\{(x_1, y_1), \ldots, (x_d, y_d)\} \subset \mathcal{X} \times \mathcal{Y}$. That means that for each $b \in \{0,1\}^d$, there is an $h_b \in \mathcal{H}$ such that $\mathbb{I}_{\{h_b(z_j(x_i)) \neq y_i\}} = b_i$ 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 $S = \{z_j(x_1), \ldots, z_j(x_d)\} \subset \mathcal{Z}$. Indeed, for each $T \subseteq S$, let $b = b(T) \in \{0,1\}^S$ be its characteristic function. Taking $f : S \to \mathcal{Y}$ to be $f(x_i) = y_i$, we see that the definition of G-shattering holds.