Journal of Machine Learning Research 24 (2023) 1-27 Submitted 8/22; Revised 5/23; Published 6/23

Sample Complexity for Distributionally Robust Learning under χ^2-divergence

Zhengyu Zhou ZZYSINCE1999@GMAIL.COM
School of Computer Science
National Engineering Research Center for Multimedia Software
Institute of Artificial Intelligence
Hubei Key Laboratory of Multimedia and Network Communication Engineering
Wuhan University
Wuhan, China

Weiwei Liu* LIUWEIWEI863@GMAIL.COM
School of Computer Science
National Engineering Research Center for Multimedia Software
Institute of Artificial Intelligence
Hubei Key Laboratory of Multimedia and Network Communication Engineering
Wuhan University
Wuhan, China

Editor: Mehryar Mohri

Abstract

This paper investigates the sample complexity of learning a distributionally robust predictor under a particular distributional shift based on χ^2-divergence, which is well known for its computational feasibility and statistical properties. We demonstrate that any hypothesis class $\mathcal{H}$ with finite VC dimension is distributionally robustly learnable. Moreover, we show that when the perturbation size is smaller than a constant, finite VC dimension is also necessary for distributionally robust learning by deriving a lower bound of sample complexity in terms of VC dimension.

Keywords: distributionally robustness, PAC learning, sample complexity, χ^2-divergence

1. Introduction

Due to the prevalence of heterogeneous but often latent subpopulations in modern datasets (Meinshausen and Bühlmann, 2015; Rothenhäusler et al., 2016), many applications in statistics and machine learning are prone to distributional shifts, leading to significant performance disparities across different demographic groupings, such as race, gender, or age. Examples of such applications include speech recognition systems for people with minority accents, facial recognition, automatic video captioning, language identification, and academic recommender systems (Grother et al., 2011; Hovy and Søgaard, 2015; Blodgett et al., 2016; Sapiezynski et al., 2017; Tatman, 2017).

∗. Corresponding Author.

Learning models that can perform well against the distributional shift, such as latent heterogeneous subpopulations, unknown covariate shift (Ben-David et al., 2006; Shimodaira, 2000), or unobserved confounding variables (Hand, 2006), remains a challenging task in contemporary machine learning. Based on the samples drawn independently and identically distributed (i.i.d.) from the data-generating distribution P, this paper considers the problem of learning a predictor that is robust to distributional shift at test time.

Concretely, let $\mathcal{X}$ be the *instance space*, while $\mathcal{Y} = \{+1, -1\}$ denotes the *label space*. We formalize the distributional shift that we would like to protect against as an uncertainty set $\mathcal{U}(P)$ containing distributions with certain constraints, such as moment condition (Delage and Ye, 2010), Wasserstein distance (Gao, 2020) and f-divergence (Duchi and Namkoong, 2021). For a distribution P over $\mathcal{X} \times \mathcal{Y} = \{(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\}$, we observe m i.i.d. samples $S \sim P^m$, and distributionally robust learning attempts to learn a predictor $h : \mathcal{X} \to \mathcal{Y}$ having small *distributionally robust risk*,

$$R_{\mathcal{U}}(h; P) := \sup_{Q \in \mathcal{U}(P)} \mathop{\mathbb{E}}_{(x,y)\sim Q}[\mathbb{1}[h(x) \neq y]]. \tag{1}$$

The common approach to distributionally robust learning involves selecting a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and learning a predictor $\hat{h} : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{H}$ through Distributionally Robust Empirical Risk Minimization:

$$\hat{h} \in \mathrm{DRERM}_{\mathcal{H}}(S) := \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_{\mathcal{U}}(h; S),$$

where $\hat{R}_{\mathcal{U}}(h; S) := \sup_{Q \in \mathcal{U}(P_m)} \mathbb{E}_{(x,y)\sim Q}[\mathbb{1}[h(x) \neq y]]$ and P_m denotes the empirical distribution over samples S.

One line of research has focused on bounding the excess risk $R_{\mathcal{U}}(\hat{h}; P) - \inf_{h \in \mathcal{H}} R_{\mathcal{U}}(h; P)$. For example, Duchi and Namkoong (2021) study the excess risk based on χ^2-divergence through the lens of the covering number argument. Lee and Raginsky (2018) derive a bound of the excess risk by means of the Rademacher complexity under the Wasserstein distance regime. However, these approaches do not consider VC dimension, which is a fundamental tool in learning theory. Moreover, the lower bound of the sample complexity for distributionally robust learning remains unknown. This paper attempts to address these issues.

It has been shown that finite VC dimension (Vapnik, 1998) is a necessary and sufficient condition for the learnability of classical statistical learning (Shalev-Shwartz and Ben-David, 2014, Theorem 6.7, Theorem 6.8), which prompts us to ask the following question:

> *Is finite VC dimension a necessary and sufficient condition for the distributionally robust learnability?*

This paper answers the above question in the affirmative. More specifically, for a given hypothesis $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and distributional shift $\mathcal{U}$, we study how many i.i.d. samples are necessary and sufficient for learning a predictor h with distributionally robust risk which is as good as any predictor in $\mathcal{H}$ (see Definition 1 in §2). We focus on the χ^2-divergence, which is a special case in the Cressie-Read family of f-divergence (Cressie and Read, 1984). Namely, we consider the distributional shift as follows:

$$\mathcal{U}(P) = \{Q \ll P : D_2(Q \| P) \leq \rho\},$$

where $D_2(Q\|P) := \frac{1}{2}\int\left(\frac{dQ}{dP}-1\right)^2 dP$ and $Q \ll P$ indicates that distribution Q is absolutely continuous with respect to P. The χ^2-divergence is a commonly explored concept in the distributionally robust optimization (DRO) literature (Duchi and Namkoong, 2019). Moreover, it is also a crucial concept in a variety of fields such as information theory, statistics, learning, signal processing, and various branches of mathematics (Park et al., 2011; Saraswat, 2014; Nishiyama and Sason, 2020). The χ^2-divergence plays a fundamental role in problems related to source and channel coding, combinatorics, large deviation theory, goodness-of-fit, and independence tests in statistics, as demonstrated by Csiszár et al. (2004). Additionally, it is widely recognized for its computational feasibility and statistical properties, as noted in Tsybakov (2009).

Our main contributions are as below:

- We show that under χ^2-divergence regime, a hypothesis class $\mathcal{H}$ with finite VC dimension can be distributionally robustly PAC-learnable with DRERM.
- Under χ^2-divergence, we prove that, when the perturbation size ρ is smaller than a constant, finite VC dimension is necessary for distributionally robust learning. We further show that without a sufficient amount of samples (depending on the VC dimension of $\mathcal{H}$), any hypothesis class $\mathcal{H}$ is not distributionally robustly PAC-learnable.

The remainder of the paper is organized as follows. In §2, we begin by providing definitions of distributionally robust learnability. In §3 and §4, we present our main results of agnostic and realizable case, respectively. We provide proof overviews of upper bound in realizable case and lower bound in agnostic case in §5. We proof the lower bound in agnostic case in §6, with certain more technical aspects deferred to the appendices. Finally, we compare our results to previous work and conclude our theoretical results in §7.

2. Problem Setup

Given a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, our goal is to design a learning rule $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$ such that for any distribution P over $\mathcal{X}\times\mathcal{Y}$, the rule $\mathcal{A}$ will find a predictor that can compete with the best predictor $h^* \in \mathcal{H}$ in terms of the distributionally robust risk using a number of samples that is independent of the distribution P. In this paper, we use $(\mathcal{X}\times\mathcal{Y})^*$ to denote the set of all sequences in the space $\mathcal{X}\times\mathcal{Y}$. The following definitions formalize the notion of distributionally robust PAC learning under the realizable case and agnostic settings.

Definition 1 (Agnostic Distributionally Robust PAC Learnability) *For any $\varepsilon, \delta \in (0,1)$, the sample complexity of agnostic distributionally robust (ε,δ)-PAC learning of $\mathcal{H}$ with respect to the distributional shift $\mathcal{U}$, denoted by $\mathcal{M}_{AG}(\varepsilon,\delta;\mathcal{H},\mathcal{U})$, is defined as the smallest $m \in \mathbb{N}\cup\{0\}$ for which there exists a learning rule $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$ such that, for every data distribution P over $\mathcal{X}\times\mathcal{Y}$, with probability of at least $1-\delta$ over $S \sim P^m$,*

$$R_{\mathcal{U}}(\mathcal{A}(S);P) \le \inf_{h\in\mathcal{H}} R_{\mathcal{U}}(h;P) + \varepsilon.$$

If no such m exists, define $\mathcal{M}_{AG}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) = \infty$. We say that $\mathcal{H}$ is distributionally robustly PAC-learnable in the agnostic setting with respect to the distributional shift $\mathcal{U}$ if $\forall \varepsilon,\delta \in (0,1)$, $\mathcal{M}_{AG}(\varepsilon,\delta;\mathcal{H},\mathcal{U})$ scales polynomially with $1/\varepsilon$ and $1/\delta$.

Definition 2 (Realizable Distributionally Robust PAC Learnability) *For any $\varepsilon, \delta \in (0,1)$, the sample complexity of realizable distributionally robust (ε,δ)-PAC learning of $\mathcal{H}$ with respect to the distributional shift $\mathcal{U}$, denoted by $\mathcal{M}_{RE}(\varepsilon,\delta;\mathcal{H},\mathcal{U})$, is defined as the smallest $m \in \mathbb{N} \cup \{0\}$ for which there exists a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$ such that, for every data distribution P over $\mathcal{X} \times \mathcal{Y}$ where there exists a predictor $h^* \in \mathcal{H}$ with zero distributionally robust risk, $R_{\mathcal{U}}(h^*;P) = 0$, with probability of at least $1-\delta$ over $S \sim P^m$,*

$$R_{\mathcal{U}}(\mathcal{A}(S);P) \leq \varepsilon.$$

If no such m exists, define $\mathcal{M}_{RE}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) = \infty$. We say that $\mathcal{H}$ is distributionally robustly PAC-learnable in the realizable setting with respect to the distributional shift $\mathcal{U}$ if $\forall \varepsilon, \delta \in (0,1)$, $\mathcal{M}_{RE}(\varepsilon,\delta;\mathcal{H},\mathcal{U})$ scales polynomially with $1/\varepsilon$ and $1/\delta$.

We also denote $\mathrm{er}(h;P) = P(h(x) \neq y)$, the (non-robust) error rate under $0-1$ loss, and $\hat{\mathrm{er}}(h;S) = \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{1}[h(x) \neq y]$, the empirical error rate; here, $|S|$ denotes the cardinality of set S, while $\mathbb{1}[\cdot]$ is the indicator function that takes 1 when the statement in the square brackets is true and 0 otherwise. The definition of Vapnik-Chervonenkis dimension (VC dimension) is provided below.

Definition 3 (VC dimension) *We say that a sequence $\{x_1,\ldots,x_k\} \subseteq \mathcal{X}$ is shattered by $\mathcal{H}$ if $\forall y_1,\ldots,y_k \in \mathcal{Y}, \exists h \in \mathcal{H}$ such that $\forall i \in [k], h(x_i) = y_i$. The VC dimension of $\mathcal{H}$ (denoted as $vc(\mathcal{H})$) is then defined as the largest integer k for which there exists $\{x_1,\ldots,x_k\} \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$. If no such k exists, then $vc(\mathcal{H})$ is said to be infinite.*

In the standard PAC learning framework, we know that a hypothesis class $\mathcal{H}$ is PAC-learnable if and only if the VC dimension of $\mathcal{H}$ is finite (Vapnik and Chervonenkis, 2015). The question then naturally arises as to whether the finite VC dimension of $\mathcal{H}$ is a necessary and sufficient condition for distributionally robust PAC learnability. In the following sections, we arrive at an affirmative answer to this question.

Denote the loss class of $\mathcal{H}$ by $\mathcal{L}_{\mathcal{H}}$, where

$$\mathcal{L}_{\mathcal{H}} = \Big\{(x,y) \mapsto \mathbb{1}[h(x) \neq y] : h \in \mathcal{H}\Big\}.$$

3. Agnostic Case

We use $R_2(h;P)$ to denote the distributionally robust risk under distributional shift

$$\mathcal{U}(P) = \{Q \ll P : D_2(Q\|P) \leq \rho\}.$$

We recall the following duality formulation (Shapiro, 2017, Section 3.2) for distributionally robust risk, which is essential in our derivation.

Proposition 4 (Duality Formulation) *For any probability P on $\mathcal{X} \times \mathcal{Y}$, any $\rho > 0$, and $c_2(\rho) := \sqrt{1+2\rho}$, for all $h \in \mathcal{H}$, we have*

$$R_2(h;P) = \inf_{\eta\in\mathbb{R}} \Big\{ c_2(\rho)\mathbb{E}_P \left[(\mathbb{1}[h(x) \neq y] - \eta)_+^2 \right]^{1/2} + \eta \Big\}, \tag{2}$$

where $a_+ := \max(a,0)$.

To bring the above proposition into effect, we need the following lemma. In the interests of simplicity, for any fixed $h \in \mathcal{H}$, let

$$g_2(\eta, P) \coloneqq c_2(\rho)\mathbb{E}_P\left[(\mathbb{1}[h(x) \neq y] - \eta)_+^2\right]^{1/2} + \eta.$$

Lemma 5 *For any distribution P,*

$$\inf_{\eta\in\mathbb{R}} g_2(\eta, P) = \inf_{\eta}\left\{g_2(\eta, P) : \eta \in \left[-\frac{1}{c_2(\rho)-1}, 1\right]\right\}.$$

Remark 6 *The above lemma restricts the domain of η to a compact set, which is crucial to our uniform convergence result.*

Theorem 7 *For any $\mathcal{H}$ with $vc(\mathcal{H}) = d$ and $\mathcal{U}(P) = \{Q \ll P : D_2(Q\|P) \leq \rho\}$, $\forall \varepsilon, \delta \in (0, 1)$,*

$$\mathcal{M}_{AG}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) = O\left(\frac{d}{\varepsilon^4}\log\left(\frac{d}{\varepsilon^4}\right) + \frac{1}{\varepsilon^4}\left(\frac{1}{c_2(\rho)-1}\vee 1\right) + \frac{d}{\varepsilon^4}\log\left(\frac{e}{d}\right) + \frac{\log(2/\delta)}{\varepsilon^4}\right). \quad (3)$$

The proof of Lemma 5 and Theorem 7 can be found in §A.1.

Remark 8 *The dependence on $\frac{1}{c_2(\rho)-1}$ is due to the lower bound for η in Lemma 5. While more advanced techniques could potentially yield a better dependence on ρ, it is worth noting that the upper bound shows that the finite VC dimension is sufficient for distributionally robust PAC learnability for finite non-zero values of ρ.*

Theorem 9 *For any $\mathcal{H}$ with $vc(\mathcal{H}) = d$ and $\mathcal{U}(P) = \{Q \ll P : D_2(Q\|P) \leq \rho\}$ with $\rho \in \left(0, \frac{3-2\sqrt{2}}{2}\right)$, $\forall \varepsilon, \delta \in (0, 1)$,*

$$\mathcal{M}_{AG}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) = \Omega\left(\frac{\left(\frac{1}{2} - \frac{\sqrt{2}}{16}\rho^{1/2}\right)^2 d + \log(1/\delta)}{\varepsilon^2}\right). \quad (4)$$

The proof of Theorem 9 can be found in §6.

Remark 10 *We derive the result under the assumption that the perturbation size $\rho \in \left(0, \frac{3-2\sqrt{2}}{2}\right)$. Intuitively, if the perturbation size is prohibitively large, the learning problem can become "easier", since the benchmark $\inf_{h\in\mathcal{H}} R_2(h; P)$ may increase too much. When ρ approaches 0, the lower bound recovers that of the classical statistical learning (Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014).*

4. Realizable Case

To study the upper bound of the sample complexity under the realizable case, it is necessary to introduce the definition of *Distributionally Robust ε-net*, which is similar to the definition in (Shalev-Shwartz and Ben-David, 2014, Definition 28.2). In the realizable case, we have a target hypothesis h^* that generates the label. We will frequently refer $\mathcal{C}_h$ to the set $\{x \in \mathcal{X} : h(x) \neq h^*(x)\}$, where h is a predictor in hypothesis class $\mathcal{H}$. The distributionally robust risk has the following form:

$$R_2(h; P) = \sup_{Q\ll P}\left\{\mathbb{E}_{x\sim Q}[\mathbb{1}[h(x) \neq h^*(x)]] : D_2(Q\|P) \leq \rho\right\}.$$

Definition 11 (Distributionally Robust ε-net) *Let $\mathcal{X}$ be a domain. $S \subseteq \mathcal{X}$ is a Distributionally Robust ε-net for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to a distribution P over $\mathcal{X}$ if:*

$$\forall h \in \mathcal{H} : R_2(h; P) \geq \varepsilon \implies \mathcal{C}_h \cap S \neq \emptyset. \tag{5}$$

Theorem 12 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $vc(\mathcal{H}) = d$ and $\mathcal{U}(P) = \{Q \ll P : D_2(Q\|P) \leq \rho\}$. $\forall \varepsilon \in (0,1), \forall \delta \in (0,1/4)$, we have*

$$\begin{aligned}\mathcal{M}_{RE}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = O\Bigg(&\frac{16(1+2\rho)d}{\varepsilon^2} \log\left(\frac{8(1+2\rho)d}{\varepsilon^2}\right) \\ &+ \frac{8(1+2\rho)}{\varepsilon^2}\left(d \log\left(\frac{2e}{d}\right) + \log\left(\frac{2}{\delta}\right)\right)\Bigg).\end{aligned} \tag{6}$$

The proof of Theorem 12 can be found in §A.2.

Remark 13 *In contrast to the upper bound derived in the agnostic case, the upper bound in the realizable case is proportional to $1 + 2\rho$. The relationship between distributionally robust risk and standard risk, which is highlighted in Lemma 28, accounts for this dependence on $1 + 2\rho$. As ρ approaches 0, the upper bound remains valid. However, it is important to note that the upper bound scales quadratically with $1/\varepsilon$, which is distinct from the scaling observed in classical statistical learning.*

Theorem 14 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $vc(\mathcal{H}) = d$ and $\mathcal{U}(P) = \{Q \ll P : D_2(Q\|P) \leq \rho\}$, $\forall \varepsilon \in (0, 1/8), \forall \delta \in (0, 1/100)$, we have*

$$\mathcal{M}_{RE}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = \Omega\left(\frac{d-1}{\varepsilon}\right). \tag{7}$$

Furthermore, if $\mathcal{H}$ contains at least three functions, $\forall \varepsilon \in (0, 3/4), \forall \delta \in (0,1)$, we have

$$\mathcal{M}_{RE}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) \geq \frac{\log(1/\delta)}{2\varepsilon}. \tag{8}$$

The proof of Theorem 14 can be found in §B.1.

Remark 15 *In the proof of the lower bound in the realizable case, we leverage the fact that distributionally robust risk exceeds the standard risk. The absence of the parameter ρ in the lower bound is due to the specific inequality we use in the proof. This inequality allows us to derive the lower bound in terms of the VC dimension.*

5. Proof overviews

We highlight the proof overviews of upper bound in realizable case and lower bound in agnostic case, which, we believe, may bring us some new insights.

Upper bound in realizable case. We first show that, for a hypothesis class $\mathcal{H}$ with finite VC dimension, given sufficient samples, the samples form a *distributionally robust ε-net* for $\mathcal{H}$ with high probability over the random draw of samples, namely Proposition 18; subsequently, we prove that such samples are sufficient for distributionally robust learning.

We decompose the first step into two subroutines. Firstly, we denote the set of sample sequence which is not ε-net by B. We draw another m sample points. We bound the probability $\mathbb{P}[S \in B]$ by $2\mathbb{P}[(S,T) \in B']$, where $(S,T) \in B'$ denotes the event where there exists a hypothesis h which has 0 empirical error on sample S, but has true error larger than ε and errs on at least $\frac{m\varepsilon^2}{2(1+2\rho)}$-fraction of the points in T. The constant $\frac{m\varepsilon^2}{2(1+2\rho)}$ is carefully chosen, where we use our Lemma 28 in our main context. The key idea here is that conditioning on the event $S \in B$, given an hypothesis h_S which has 0 empirical error on sample S, but has true error larger than ε, a sufficient condition for event B' is that h_S errs on at least $\frac{2m\varepsilon^2}{2(1+2\rho)}$-fraction of the points in T; its probability can provide a lower bound of the probability of B' conditioned on $S \in B$. The event $\left\{|T \cap \mathcal{C}_{h_S}| > \frac{m\varepsilon^2}{2(1+2\rho)}\right\}$ can be viewed as m repeated Bernoulli test with a success rate larger than the given constant. Next, we bound the probability $\mathbb{P}[(S,T) \in B']$ with a symmetrization argument. The key idea here is that the probability can be bounded by the probability of the randomness over the draw of $2m$ samples which satisfy: there exists a hypothesis $h \in \mathcal{H}$ such that it only errs on the last m sample points with an error rate $\frac{m\varepsilon^2}{2(1+2\rho)}$. The existence of the hypothesis can be turned into a maximization over the hypothesis class. However, the hypothesis class is often infinite, so we need to focus on the effective number of hypotheses on A. Now, we can bound it with Sauer's lemma in terms of VC dimension.

In the second step, we show that, any DRERM hypothesis has a true error of at most ε, with high probability over a choice of m i.i.d. instances. The key idea of the proof is that for any distributionally robust ε-net, by its definition, for any hypothesis h with $R_{2,\rho}(h; P) \geq \varepsilon$, the hypothesis h will err on the sample S; thus, h cannot be a DRERM hypothesis.

Lower bound in agnostic case. The main argument that lies in the heart of the proof is a probability method argument. With every labeling $b \in \{-1, 1\}^m$, we associate a distribution $\mathcal{D}_b$ over $\mathcal{X} \times \{-1, +1\}$. We then show with some positive probability if we sample $b \in \{-1, +1\}^m$, $\mathcal{D}_b$ satisfies the requiment that without sufficient samples, no hypothesis in the class can have excess risk smaller than ε. **Constructing a Family of Distributions.** We start by first describing the construction of $\mathcal{D}_b$ for $b \in \{-1, +1\}^m$. Our construction follows previous studied distribution construction patterns in a subtle manner. Anthony and Bartlett (2002, Chapter 5) observed that for a distribution $\mathcal{D}$ that assigns each point in $\mathcal{X}$ a random label, if S does not sample a point x enough times, any classifier f, that is constructed using only information supplied by S, cannot determine with good probability the Bayes label of x, that is the label of x that minimizes the error probability. To follow the above construction, we need to show that which classifier in $\mathcal{H}$ has the best distributionally robust error. It seems not obvious whether the same labeling rule as above will have the lowest distributionlly robust error. We show that the Bayes labeling also has the lowest distributionally robust risk making us think more about the relation between ERM and DRERM. Next, to carefully study the difference between the risk of the output hypothesis and the lower risk, we derive a explicit form of the distributionally robsut risk under the χ^2-divergence setting (see Lemma 23 in the appendix). Then, we turn the existence argument to an maximization argument, and use the fact that average is smaller than the maximum to lower bound the expected (with respect to the random choice of samples and random labeling b) excess risk of a given algorithm $\mathcal{A}$. Subsequently, we minimizes the lower bound by chosing the Maximum-Likelihood learning rule. Finally, using some probabilistic method, we can give an explicit form of the minimized lower bound, thus showing that the expected excess risk is lager than ε with a positive probability.

6. Lower Bound in Agnostic Case

Proof [Proof of Theorem 9] We will prove the lower bound in two parts. First, we will show that $\mathcal{M}_{\mathrm{AG}}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) \geq \log(\frac{1}{4\delta})/(2\varepsilon^2)$; second, we will show that for every $\delta \leq 1/8$, we have that $\mathcal{M}_{\mathrm{AG}}(\varepsilon,1/8;\mathcal{H},\mathcal{U}) \geq \frac{\left(\frac{1}{2}-\frac{\sqrt{2}}{16}\rho^{1/2}\right)^2 d}{128\varepsilon^2}$. These two bounds will conclude the proof.

We first demonstrate that $\mathcal{M}_{\mathrm{AG}}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) \geq \log(\frac{1}{4\delta})/(2\varepsilon^2)$.

To do so, we show that for a sample with size $m \leq \log(\frac{1}{4\delta})/(2\varepsilon^2)$, $\mathcal{H}$ is not learnable for any $\varepsilon \in \left(0,\frac{1}{\sqrt{2}}\right)$ and $\delta \in (0,1)$.

Let us choose one example that is shattered by $\mathcal{H}$. That is, let c be an example such that there are $h_+, h_- \in \mathcal{H}$ for which $h_+(c)=1$ and $h_-(c)=-1$. Define two distributions, P_+ and P_-, such that for $b \in \{+1,-1\}$, we have

$$P_b(x,y) = \begin{cases} \dfrac{1+yb\varepsilon}{2}, & \text{if} \quad x=c,\\ 0 \quad, & \text{otherwise.}\end{cases}$$

Any training set sampled from P_b has the form $S=((c,y_1),\dots,(c,y_m))$. Let $\mathcal{A}$ be an arbitrary algorithm. Therefore, the hypothesis that $\mathcal{A}$ outputs receiving sample S is fully characterized by the vector $\mathbf{y}=(y_1,\dots,y_m)\in\{+1,-1\}^m$. Upon receiving a training set S, the algorithm $\mathcal{A}$ returns a hypothesis $h_S:\mathcal{X}\to\{+1,-1\}$. Since the error of h_S w.r.t. P_b depends only on $h(c)$, we can think of $\mathcal{A}$ as a mapping from $\{+1,-1\}^m$ into $\{+1,-1\}$.

Therefore, we denote by $\mathcal{A}(\mathbf{y})$ the value in $\{+1,-1\}$ corresponding to the prediction $h_S(c)$; here, h_S is the hypothesis that $\mathcal{A}$ outputs upon receiving the training set $S=((c,y_1),\dots,(c,y_m))$. **Claim 1**. The hypothesis $h_b(c)=b$ has optimal distributionally robust risk on P_b.

Note that for any hypothesis h, we have

$$\begin{aligned}
R_2(h;P_b) &= \sup\left\{\mathbb{E}_P\big[\mathbb{1}[h(x)\neq y]\big] : P \ll P_b, D_2(P\|P_b)\leq\rho\right\}\\
&= \inf_{\eta\in\mathbb{R}}\left\{c_2(\rho)\mathbb{E}_{P_b}\left[(\mathbb{1}[h(x)\neq y]-\eta)_+^2\right]^{1/2}+\eta\right\}\\
&= \inf_{\eta\in\mathbb{R}}\left\{c_2(\rho)\left(\frac{1+\varepsilon}{2}(\mathbb{1}[h(c)\neq b]-\eta)_+^2+\frac{1-\varepsilon}{2}(\mathbb{1}[h(c)\neq -b]-\eta)^2\right)^{1/2}+\eta\right\}.
\end{aligned}$$

Substituting h_b into the above formulation, we obtain:

$$R_2(h_b;P_b) = \inf_{\eta\in\mathbb{R}}\left\{c_2(\rho)\left(\frac{1+\varepsilon}{2}(-\eta)_+^2+\frac{1-\varepsilon}{2}(1-\eta)_+^2\right)^{1/2}+\eta\right\}.$$

Noting that $(-\eta)_+^2 \leq (1-\eta)_+^2$, we get $R_2(h_b;P_b)\leq R_2(h;P_b)$ for any $h\in\mathcal{H}$.

Invoking Lemma 23, we have

$$R_2(\mathcal{A}(\mathbf{y});P_b) - \inf_{h\in\mathcal{H}} R_2(h;P_b) = \begin{cases}\varepsilon, & \text{if } \mathcal{A}(\mathbf{y})(c)\neq b,\\ 0, & \text{otherwise.}\end{cases}$$

Fix $\mathcal{A}$. For $b\in\{+1,-1\}$, let $Y^b=\{\mathbf{y}\in\{+1,-1\}^m : \mathcal{A}(\mathbf{y})\neq b\}$. The distribution P_b induces a probability $\mathcal{D}_b$ over $\{+1,-1\}^m$. Hence,

$$\mathbb{P}\left[R_2(\mathcal{A}(\mathbf{y});P_b)-R_2(h_b;P_b)=\varepsilon\right] = P_b(Y^b) = \sum_{\mathbf{y}}\mathcal{D}_b[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y})\neq b].$$

Denote $N^+ = \{\mathbf{y} : |\{i : y_i = +1\}| \geq m/2\}$ and $N^- = \{+1,-1\}^m \setminus N^+$. Note that for any $\mathbf{y} \in N^+$, we have $\mathcal{D}_+[\mathbf{y}] \geq \mathcal{D}_-[\mathbf{y}]$, while for any $\mathbf{y} \in N^-$, we have $\mathcal{D}_-[\mathbf{y}] \geq \mathcal{D}_+[\mathbf{y}]$. Therefore,

$$
\begin{aligned}
\max_{b\in\{+1,-1\}} & \mathbb{P}\left[R_2(\mathcal{A}(\mathbf{y}); P_b) - R_2(h_b; P_b) = \varepsilon\right] \\
&= \max_{b\in\{+1,-1\}} \sum_{\mathbf{y}} \mathcal{D}_b[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq b] \\
&\geq \frac{1}{2}\sum_{\mathbf{y}} \mathcal{D}_+[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq +1] + \frac{1}{2}\sum_{\mathbf{y}} \mathcal{D}_-[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq -1] \\
&= \frac{1}{2}\sum_{\mathbf{y}\in N^+} \left(\mathcal{D}_+[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq +1] + \mathcal{D}_-[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq -1]\right) \\
&\quad + \frac{1}{2}\sum_{\mathbf{y}\in N^-} \left(\mathcal{D}_+[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq +1] + \mathcal{D}_-[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq -1]\right) \\
&\geq \frac{1}{2}\sum_{\mathbf{y}\in N^+} \left(\mathcal{D}_-[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq +1] + \mathcal{D}_-[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq -1]\right) \\
&\quad + \frac{1}{2}\sum_{\mathbf{y}\in N^-} \left(\mathcal{D}_+[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq +1] + \mathcal{D}_+[\mathbf{y}]\mathbb{1}[\mathcal{A}(\mathbf{y}) \neq -1]\right) \\
&= \frac{1}{2}\sum_{\mathbf{y}\in N^+} \mathcal{D}_-[\mathbf{y}] + \frac{1}{2}\sum_{\mathbf{y}\in N^-} \mathcal{D}_+[\mathbf{y}].
\end{aligned}
$$

Next, note that $\sum_{\mathbf{y}\in N^+} \mathcal{D}_-[\mathbf{y}] = \sum_{\mathbf{y}\in N^-} \mathcal{D}_+[\mathbf{y}]$, and both values are the probability that a Binomial $(m,(1-\varepsilon)/2)$ random variable will have a value greater than $m/2$. Using Lemma 24, this probability is lower bounded by

$$
\frac{1}{2}\left(1 - \sqrt{1-\exp(-m\varepsilon^2/(1-\varepsilon^2))}\right) \geq \frac{1}{2}\left(1 - \sqrt{1-\exp(-2m\varepsilon^2)}\right),
$$

where we derive under the assumption that $\varepsilon^2 \leq 1/2$. It follows that if $m \leq 0.5\log(1/(4\delta))/\varepsilon^2$, then there exists b such that

$$
\mathbb{P}[R_2(\mathcal{A}(\mathbf{y}); P_b) - R_2(h_b; P_b) = \varepsilon] \geq \frac{1}{2}\left(1 - \sqrt{1-\sqrt{4\delta}}\right) \geq \delta,
$$

where the last inequality can be obtained through standard algebraic manipulations. This concludes our proof.

Next, we demonstrate that $\mathcal{M}_{\mathrm{AG}}(\varepsilon, 1/8; \mathcal{H}, \mathcal{U}) \geq \frac{\left(\frac{1}{2}-\frac{\sqrt{2}}{16}\rho^{1/2}\right)^2 d}{128\varepsilon^2}$.

We shall now prove that for every $\varepsilon < \frac{1}{8\sqrt{2}}$ we have that $\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) \geq \frac{\left(\frac{1}{2}-\frac{\sqrt{2}}{16}\rho^{1/2}\right)^2 d}{128\varepsilon^2}$.

Let $r = 8\varepsilon$, and note that $r \in (0, 1/\sqrt{2})$. We will construct a family of distributions as follows. First, let $C = \{c_1, \ldots, c_d\}$ be a set of d instances that are shattered by $\mathcal{H}$. Second, for each vector $\mathbf{b} = (b_1, \ldots, b_d) \in \{+1,-1\}^d$, define a distribution $P_{\mathbf{b}}$ such that

$$
P_{\mathbf{b}}(x,y) = \begin{cases} \frac{1}{d}\cdot\frac{1+yb_i r}{2} & , \quad \text{if } \exists i : x = c_i \\ 0 & , \quad \text{otherwise.} \end{cases}
$$

That is, to sample an example according to $P_{\mathbf{b}}$, we first sample an element $c_i \in C$ uniformly at random, then set the label to be b_i with probability $(1+r)/2$ or $-b_i$ with probability $(1-r)/2$.

Claim 2. The hypothesis $h_{\mathbf{b}}$ satisfying $h(c_i) = b_i, \forall i \in [d]$, has the optimal distributionally robust risk on $P_{\mathbf{b}}$.

Recalling the dual formulation (2), $R_2(h; P_{\mathbf{b}})$ can be rewritten as follows:

$$R_2(h; P_{\mathbf{b}}) = \inf_{\eta \in \mathbb{R}} \left\{ c_2 \left[\frac{1}{d} \sum_{i=1}^{d} \left(\frac{1+r}{2} (\mathbb{1}[h(c_i) \neq b_i] - \eta)_+^2 + \frac{1-r}{2} (\mathbb{1}[h(c_i) \neq -b_i] - \eta)_+^2 \right) \right]^{1/2} + \eta \right\}.$$

For each $i \in [d]$, $h_{\mathbf{b}}(c_i) = b_i$, the summand above can be written as $\frac{1+r}{2}(-\eta)_+^2 + \frac{1-r}{2}(1-\eta)_+^2$; for $h(c_i) \neq b_i$, the summand is $\frac{1+r}{2}(1-\eta)_+^2 + \frac{1-r}{2}(-\eta)_+^2$. Combining $r > 0$ and $(-\eta)_+^2 \leq (1-\eta)_+^2$, we have $\frac{1+r}{2}(-\eta)_+^2 + \frac{1-r}{2}(1-\eta)_+^2 \leq \frac{1+r}{2}(1-\eta)_+^2 + \frac{1-r}{2}(-\eta)_+^2$, which concludes our claim.

We denote $d_+ = |\{i \in [d] : \mathcal{A}(S)(c_i) = b_i\}|$ and $d_- = |\{i \in [d] : \mathcal{A}(S)(c_i) \neq b_i\}|$, therefore $d_+ + d_- = d$. Next, we will simplify the distributionally robust risk as follows:

$$\begin{aligned} R_2(\mathcal{A}(S); P_{\mathbf{b}}) &= \inf_{\eta \in \mathbb{R}} \left\{ c_2 \left[\frac{d_+}{d} \left(\frac{1+r}{2} (-\eta)_+^2 + \frac{1-r}{2} (1-\eta)_+^2 \right) + \right. \right. \\ &\qquad \left. \left. \frac{d_-}{d} \left(\frac{1+r}{2} (1-\eta)_+^2 + \frac{1-r}{2} (-\eta)_+^2 \right) \right]^{1/2} + \eta \right\} \\ &= \inf_{\eta \in \mathbb{R}} \left\{ c_2 \left[\left(\frac{1}{2} + \frac{r}{2} \cdot \frac{d_+ - d_-}{d} \right) (-\eta)_+^2 + \left(\frac{1}{2} - \frac{r}{2} \cdot \frac{d_+ - d_-}{d} \right) (1-\eta)_+^2 \right]^{1/2} + \eta \right\}. \end{aligned}$$

Therefore, $R_2(\mathcal{A}(S); P_{\mathbf{b}})$ can be viewed as the distributionally robust risk of the classifier $h'(x) \equiv 1$ on distribution Q, with

$$Q(x, y) = \begin{cases} \dfrac{1 + yr(d_+ - d_-)/d}{2} & \text{, if } x = c \\ 0 & \text{, otherwise.} \end{cases}$$

Invoking Lemma 23, we obtain

$$R_2(\mathcal{A}(S); P_{\mathbf{b}}) = \frac{1}{2} \left(1 - r \cdot \frac{d_+ - d_-}{d} + \sqrt{2\rho \left(1 - r^2 \cdot \frac{(d_+ - d_-)^2}{d^2} \right)} \right).$$

Following the same logic, we have $R_2(h_{\mathbf{b}}; P_{\mathbf{b}}) = \frac{1}{2}\left(1 - r + \sqrt{2\rho\,(1-r^2)}\right)$.

Then, after some algebraic manipulations, we have:

$$\begin{aligned} R_2(\mathcal{A}(S); P_{\mathbf{b}}) - \inf_{h \in \mathcal{H}} R_2(h; P_{\mathbf{b}}) &= R_2(\mathcal{A}(S); P_{\mathbf{b}}) - R_2(h_{\mathbf{b}}; P_{\mathbf{b}}) \\ &= r \cdot \frac{d_-}{d} + \frac{1}{2} \left(\sqrt{2\rho\,(1 - r^2 (d_+ - d_-)^2/d^2)} - \sqrt{2\rho(1-r^2)} \right) \\ &\geq r \cdot \frac{d_-}{d}, \end{aligned} \tag{9}$$

where the last line follows from $(d_+ - d_-)^2/d^2 \leq 1$.

Next, fix some learning algorithm $\mathcal{A}$, we have that:

$$\max_{P_{\mathbf{b}}:\mathbf{b}\in\{+1,-1\}^d} \mathop{\mathbb{E}}_{S\sim P_{\mathbf{b}}^m}\left[R_2(\mathcal{A}(S);P_{\mathbf{b}}) - \inf_{h\in\mathcal{H}} R_2(h;P_{\mathbf{b}})\right] \tag{10}$$

$$\geq \mathop{\mathbb{E}}_{P_{\mathbf{b}}:\mathbf{b}\sim U(\{+1,-1\}^d)} \mathop{\mathbb{E}}_{S\sim P_{\mathbf{b}}^m}\left[R_2(\mathcal{A}(S);P_{\mathbf{b}}) - \inf_{h\in\mathcal{H}} R_2(h;P_{\mathbf{b}})\right] \tag{11}$$

$$\geq \mathop{\mathbb{E}}_{P_{\mathbf{b}}:\mathbf{b}\sim U(\{+1,-1\}^d)} \mathop{\mathbb{E}}_{S\sim P_{\mathbf{b}}^m}\left[r\cdot\frac{|\{i\in[d]:\mathcal{A}(S)(c_i)\neq b_i\}|}{d}\right] \tag{12}$$

$$= \frac{r}{d}\sum_{i=1}^d \mathop{\mathbb{E}}_{P_{\mathbf{b}}:\mathbf{b}\sim U(\{+1,-1\}^d)} \mathop{\mathbb{E}}_{S\sim P_{\mathbf{b}}^m} \mathbb{1}[\mathcal{A}(S)(c_i)\neq b_i], \tag{13}$$

where the second inequality follows from (9). In addition, using the definition of $P_{\mathbf{b}}$, in order to sample $S\sim P_{\mathbf{b}}$, we can first sample $(j_1,\ldots,j_m)\sim U([d])^m$, set $x_i = c_{j_i}$, and finally sample y_i such that $\mathbb{P}[y_i = b_{j_i}] = (1+r)/2$. Let us simplify the notation and use $y\sim b$ to denote sampling according to $\mathbb{P}[y=b] = (1+r)/2$. Therefore, the right-hand side of (13) equals

$$\frac{r}{d}\sum_{i=1}^d \mathop{\mathbb{E}}_{j\sim U([d])^m} \mathop{\mathbb{E}}_{\mathbf{b}\sim U(\{+1,-1\}^d)} \mathop{\mathbb{E}}_{\forall k, y_k\sim b_{j_k}} \mathbb{1}[\mathcal{A}(S)(c_i)\neq b_i]. \tag{14}$$

We now proceed in two steps. First, in Lemma 25, we show that among all learning algorithms, $\mathcal{A}$, the one that minimizes (14) is the Maximum-Likelihood learning rule, denoted as $\mathcal{A}_{\mathrm{ML}}$. Formally, for each i, $\mathcal{A}_{\mathrm{ML}}(S)(c_i)$ is the majority vote among the set $\{y_k : k\in[m], x_k = c_i\}$. Second, we lower bound (14) for $\mathcal{A}_{\mathrm{ML}}$.

Fix i. For every $\mathbf{j}\in[d]^m$, let $n_i(\mathbf{j}) = |\{k:\mathbf{j}_k = i\}|$ be the number of instances in which the instance is c_i. For the Maximum-Likelihood rule, we have that the quantity

$$\mathop{\mathbb{E}}_{\mathbf{b}\sim U(\{+1,-1\}^d)} \mathop{\mathbb{E}}_{\forall k, y_k\sim b_{j_k}} \mathbb{1}[\mathcal{A}(S)(c_i)\neq b_i]$$

is exactly equal to the probability that a binomial $(n_i(\mathbf{j}),(1-r)/2)$ random variable will be larger than $n_i(\mathbf{j})/2$. Using Lemma 24, and the assumption $r^2\leq 1/2$, we have that

$$\mathbb{P}[B\geq n_i(\mathbf{j})/2] \geq \frac{1}{2}\left(1-\sqrt{1-\exp\big(-2n_i(\mathbf{j})r^2\big)}\right).$$

We have thus demonstrated that

$$\frac{r}{d}\sum_{i=1}^d \mathop{\mathbb{E}}_{j\sim U([d])^m} \mathop{\mathbb{E}}_{\mathbf{b}\sim U(\{+1,-1\}^d)} \mathop{\mathbb{E}}_{\forall k, y_k\sim b_{j_k}} \mathbb{1}[\mathcal{A}(S)(c_i)\neq b_i]$$
$$\geq \frac{r}{2d}\sum_{i=1}^d \mathop{\mathbb{E}}_{\mathbf{j}\sim U([d])^m}\left(1-\sqrt{1-\exp\big(-2n_i(\mathbf{j})r^2\big)}\right)$$
$$\geq \frac{r}{2d}\sum_{i=1}^d \mathop{\mathbb{E}}_{\mathbf{j}\sim U([d])^m}\left(1-\sqrt{2n_i(\mathbf{j})r^2}\right),$$

where the last inequality follows from the fact that $1-e^{-a}\leq a$.

Since the square root function is concave, we can apply Jensen's inequality to obtain that the above is lower bounded by

$$\begin{aligned}
&\frac{r}{2d}\sum_{i=1}^{d}\left(1-\sqrt{2r^2\mathop{\mathbb{E}}_{\mathbf{j}\sim U([d])^m} n_i(\mathbf{j})}\right)\\
&=\frac{r}{2d}\sum_{i=1}^{d}\left(1-\sqrt{2r^2m/d}\right)\\
&=\frac{r}{2}\left(1-\sqrt{2r^2m/d}\right).
\end{aligned}$$

As long as $m<\frac{\left(\frac{1}{2}-\frac{\sqrt{2}}{16}\rho^{1/2}\right)^2 d}{2r^2}$, this term will be larger than $\frac{r}{4}+\frac{\sqrt{2}r\rho^{1/2}}{32}$.

In summary, we have shown that if $m<\frac{\left(\frac{1}{2}-\frac{\sqrt{2}}{16}\rho^{1/2}\right)^2 d}{2r^2}$ then for any algorithm, there exists a distribution $P_{\mathbf{b}}$ such that

$$\mathop{\mathbb{E}}_{S\sim P^m}\left[R_2(\mathcal{A}(S);P_{\mathbf{b}})-\inf_{h\in\mathcal{H}}R_2(h;P_{\mathbf{b}})\right]\geq\frac{r}{4}+\frac{\sqrt{2}r\rho^{1/2}}{32}.$$

Finally, let $\Delta=\frac{1}{r}\left(R_2(\mathcal{A}(S);P_{\mathbf{b}})-\inf_{h\in\mathcal{H}}R_2(h;P_{\mathbf{b}})\right)$; we proof that $\Delta\in[0,1+\sqrt{2}\rho^{1/2}/4]$ in Lemma 26. Therefore, using Lemma 27, we obtain that

$$\begin{aligned}
\mathbb{P}\left[R_2(\mathcal{A}(S);P_{\mathbf{b}})-\inf_{h\in\mathcal{H}}R_2(h;P_{\mathbf{b}})\geq\epsilon\right]&=\mathbb{P}\left[\Delta>\frac{\varepsilon}{r}\right]\geq\left(1+\sqrt{2}\rho^{1/2}/4\right)^{-1}\left(\mathbb{E}[\Delta]-\frac{\varepsilon}{r}\right)\\
&\geq\left(1+\sqrt{2}\rho^{1/2}/4\right)^{-1}\left(\frac{1}{4}+\frac{\sqrt{2}\rho^{1/2}}{32}-\frac{\varepsilon}{r}\right).
\end{aligned}$$

Choosing $r=8\varepsilon$, we conclude that if $m<\frac{\left(\frac{1}{2}-\frac{\sqrt{2}}{16}\rho^{1/2}\right)^2 d}{2r^2}$, then with probability of at least $1/8$, we will have that $R_2(\mathcal{A}(S);P_{\mathbf{b}})-\inf_{h\in\mathcal{H}}R_2(h;P_{\mathbf{b}})\geq\epsilon$. ■

7. Discussion

In this paper, 1) we provide lower bounds on the sample complexity of distributionally robust learning based on VC dimension both in agnostic and realizable case, which has not been studied before to our knowledge; 2) we also provide upper bounds both in agnostic and realizable cases; moreover, we provide a new analysis of the excess risk, which is different from the covering argument with respect to L^∞-norm used in Duchi and Namkoong (2021).

Comparison with Duchi and Namkoong (2021) We study the $0-1$ loss of the VC classes. There is a situation, where the VC dimension is finite, while the covering number of the $0-1$ loss class with respect to L^∞-norm is infinite. Specifically, for any hypothesis class with finite VC dimension and infinite elements, given two different hypotheses h_1 and h_2, there exists x such that $h_1(x)\neq h_2(x)$, thus $\sup_{x,y}|\mathbb{1}[h_1(x)\neq y]-\mathbb{1}[h_2(x)\neq y]|=1$. Then, for any $\delta<1/2$, the δ-packing number of the $0-1$ loss class is infinite. Using the relation between the covering number and packing number (Wainwright, 2019, Lemma 5.5), we deduce that the $\delta/2$-covering number of the $0-1$ loss class is infinite, for any $\delta < 1/2$. It is well known that the covering number with L^r-norm can be controlled by the VC dimension (Wellner et al., 2013, Theorem 2.6.4). Since this is valid only for finite r, our work significantly extends the results of Duchi and Namkoong (2021) in the χ^2-divergence setting. Duchi and Namkoong (2021) provide a minimax lower bound showing that the rate they obtain is optimal. However, what role does VC dimension play in distributionally robust learning is still unknown. Our paper takes a step forward and studies distributionally robust learnability through the lens of VC dimension. We show that the finite VC dimension is necessary and sufficient for distributionally robust learnability under certain assumptions.

Appendix A. Proofs of Upper Bound

A.1 Upper Bound in Agnostic Case

To prove (3), it suffices to show that applying the DRERM with a sample size m of the same order as in (3) yields an ε, δ-learner for $\mathcal{H}$.

We use $\varphi_{\eta,h}(x, y)$ to denote $\mathbb{1}[h(x) \neq y] - \eta$ and c_2 as the shorthand of $c_2(\rho)$ when there is no ambiguity. First, we present the proof of Lemma 5.

Moreover, we introduce the definition of *Rademacher Complexity* and *Growth Function* in order to bound the excess risk in terms of VC dimension.

Definition 16 (Rademacher Complexity) *We define the empirical Rademacher Complexity of a hypothesis class $\mathcal{F}$ for a given sample $z_i = (x_i, y_i), i = 1, \ldots, m$ as follows :*

$$\hat{\mathcal{R}}_m(\mathcal{F}) \coloneqq \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right],$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ is a vector of i.i.d. Rademacher variables. The Rademacher Complexity is defined as the expectation of this quantity:

$$\mathcal{R}_m(\mathcal{F}) \coloneqq \mathbb{E}_{(z_1,\ldots,z_m)\sim P^m} \left[\hat{\mathcal{R}}_m(\mathcal{F}) \right].$$

Definition 17 (Growth Function) *Let $\mathcal{H}$ be a hypothesis class. The growth function of $\mathcal{H}$, denoted as $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as follows:*

$$\tau_{\mathcal{H}}(m) \coloneqq \max_{C \subset \mathcal{X}: |C| = m} |\mathcal{H}_C|,$$

where $\mathcal{H}_C \coloneqq \left\{ (h(x_1), \cdots, h(x_{|C|})) : h \in \mathcal{H} \right\}$ for $C = \{x_i : 1 \leq i \leq |C|\}$.

Proof [Proof of Lemma 5] By definition, $g_2(\eta, P) = \eta$ for $\eta \geq 1$, and

$$\begin{aligned} g_2\left(-\frac{1}{c_2 - 1}, P\right) &\geq \frac{c_2}{c_2 - 1} - \frac{1}{c_2 - 1} \\ &= 1 = g_2(1, P). \end{aligned}$$

Since $\eta \mapsto g_2(\eta, P)$ is convex, this implies the result. ■

Proof [Proof of Theorem 7] The proof is a modification of the techniques utilized by Koltchinskii and Panchenko (2002). Let $(x_1, y_1), \ldots, (x_m, y_m)$ be a classification training set, P_m denote the corresponding empirical distribution and $h^* \in \underset{h\in\mathcal{H}}{\operatorname{argmin}} R_2(h; P)$. We begin by decomposing the excess risk:

$$R_2(\hat{h}; P) - R_2(h^*; P) \leq R_2(\hat{h}; P) - R_2(\hat{h}; P_m) + R_2(h^*; P_m) - R_2(h^*; P),$$

where the last step follows from the definition of $\hat{h}$. Define

$$\hat{\eta} := \underset{\eta\in\mathbb{R}}{\operatorname{argmin}} \left\{ c_2 \mathbb{E}_{P_m} \left[\left(\varphi_{\eta,\hat{h}}(x,y) \right)_+^2 \right]^{1/2} + \eta \right\},$$
$$\eta^* := \underset{\eta\in\mathbb{R}}{\operatorname{argmin}} \left\{ c_2 \mathbb{E}_{P} \left[\left(\varphi_{\eta,h^*}(x,y) \right)_+^2 \right]^{1/2} + \eta \right\}.$$

We can then write

$$\begin{aligned} R_2(\hat{h}; P) - R_2(\hat{h}; P_m) &= \min_{\eta\in\mathbb{R}} \left\{ c_2 \mathbb{E}_P \left[\left(\varphi_{\eta,\hat{h}}(x,y) \right)_+^2 \right]^{\frac{1}{2}} + \eta \right\} - \left(c_2 \mathbb{E}_{P_m} \left[\left(\varphi_{\hat{\eta},\hat{h}}(x,y) \right)_+^2 \right]^{\frac{1}{2}} + \hat{\eta} \right) \\ &\leq c_2 \left(\mathbb{E}_P \left[\left(\varphi_{\hat{\eta},\hat{h}}(x,y) \right)_+^2 \right]^{\frac{1}{2}} - \mathbb{E}_{P_m} \left[\left(\varphi_{\hat{\eta},\hat{h}}(x,y) \right)_+^2 \right]^{\frac{1}{2}} \right) \\ &\leq c_2 \left| \int \left(\varphi_{\hat{\eta},\hat{h}}(x,y) \right)_+^2 (P_m - P)(dxdy) \right|^{\frac{1}{2}}. \end{aligned}$$

Following the same logic,

$$R_2(h^*; P_m) - R_2(h^*; P) \leq c_2 \left| \int \left(\varphi_{\eta^*,h^*}(x,y) \right)_+^2 (P_m - P)(dxdy) \right|^{\frac{1}{2}}. \tag{15}$$

By Lemma 5, $\hat{\eta} \in \left[-\frac{1}{c_2-1}, 1 \right]$. We now define $\Phi = \left\{ \varphi_{\eta,h} : h \in \mathcal{H}, \eta \in \left[-\frac{1}{c_2-1}, 1 \right] \right\}, \psi(t) = t_+^2$, and $\psi \circ \Phi = \{ \psi \circ \varphi : \varphi \in \Phi \}$, where $\circ$ denotes the composition of functions. Thus, we can write

$$R_2(\hat{h}; P) - R_2(\hat{h}; P_m) \leq c_2 \left(\sup_{\varphi\in\psi\circ\Phi} \left[\int \varphi(x,y) (P - P_m)(dxdy) \right] \right)^{\frac{1}{2}}.$$

Since $\mathbb{1}[h(x) \neq y] \in [0, 1]$ for any $h \in \mathcal{H}$ and $\eta \in \left[-\frac{1}{c_k-1}, 1 \right]$, then for any $\varphi \in \psi \circ \Phi$, we have $\|\varphi\|_\infty \leq \left(\frac{c_2}{c_2-1} \right)^2$.

By a standard symmetrization argument, with probability of at least $1 - \delta/2$,

$$R_2(\hat{h}; P) - R_2(\hat{h}; P_m) \leq c_2 \left(2\mathcal{R}_m(\psi \circ \Phi) + \left(\frac{c_2}{c_2 - 1} \right)^2 \sqrt{\frac{2\log(2/\delta)}{m}} \right)^{\frac{1}{2}}$$

Moreover, from (15) and Hoeffding's inequality, it follows that

$$R_2(h^*; P_m) - R_2(h^*; P) \leq \frac{c_2}{c_2 - 1} \left(\frac{\log(2/\delta)}{2m} \right)^{\frac{1}{4}}, \tag{16}$$

with probability of at least $1-\delta/2$.

Combining these results, with probability of at least $1-\delta$,

$$
\begin{aligned}
R_2(\hat{h};P)-R_2(h^*;P) &\leq c_2\left[\left(2\mathcal{R}_m(\psi\circ\Phi)+\left(\frac{c_2}{c_2-1}\right)^2\sqrt{\frac{2\log(2/\delta)}{m}}\right)^{\frac{1}{2}}+\frac{c_2}{c_2-1}\left(\frac{\log(2/\delta)}{2m}\right)^{\frac{1}{4}}\right] \\
&\leq c_2\left[2\left(\mathcal{R}_m(\psi\circ\Phi)\right)^{\frac{1}{2}}+\frac{3c_2}{c_2-1}\left(\frac{\log(2/\delta)}{2m}\right)^{\frac{1}{4}}\right].
\end{aligned} \tag{17}
$$

Therefore, it suffices to bound $\mathcal{R}_m(\psi\circ\Phi)$. It can be readily observed that $t\mapsto t_+^2$ is $\frac{2c_2}{c_2-1}$-Lipschitz on $\left[-1,\frac{c_2}{c_2-1}\right]$; thus, by invoking Lemma 21, we get:

$$
\mathcal{R}_m(\psi\circ\Phi)\leq\frac{2c_2}{c_2-1}\mathcal{R}_m(\Phi). \tag{18}
$$

More specifically,

$$
\begin{aligned}
\mathcal{R}_m(\Phi) &= \mathbb{E}\left[\frac{1}{m}\sup_{h\in\mathcal{H},\eta\in\left[-\frac{1}{c_2-1},1\right]}\sum_{i=1}^m\sigma_i(\mathbb{1}[h(x)\neq y]-\eta)\right] \\
&\leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m\sigma_i\mathbb{1}[h(x)\neq y]\right]+\mathbb{E}\left[\sup_{\eta\in\left[-\frac{1}{c_2-1},1\right]}\frac{1}{m}\left|\sum_{i=1}^m\eta\sigma_i\right|\right] \\
&\leq \mathcal{R}_m(\mathcal{L}_{\mathcal{H}})+\frac{1}{m}\left(\frac{1}{c_2-1}\vee 1\right)\mathbb{E}\left[\left|\sum_{i=1}^m\sigma_i\right|\right] \\
&\leq \mathcal{R}_m(\mathcal{L}_{\mathcal{H}})+\frac{1}{m}\left(\frac{1}{c_2-1}\vee 1\right)\mathbb{E}\left[\left(\sum_{i=1}^m\sigma_i\right)^2\right]^{1/2} \\
&\leq \mathcal{R}_m(\mathcal{L}_{\mathcal{H}})+\frac{1}{\sqrt{m}}\left(\frac{1}{c_2-1}\vee 1\right).
\end{aligned} \tag{19}
$$

To bound $\mathcal{R}_m(\mathcal{L}_{\mathcal{H}})$, we define $R(A)=\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m\sigma_i\mathbb{1}[h(x_i)\neq y_i]\right]$;

Recall that the Sauer-Shelah lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 6.10) tells us that if $\mathrm{vc}(\mathcal{H})=d$, then

$$
|\{(h(x_1),\ldots,h(x_m)):h\in\mathcal{H}\}|\leq\left(\frac{em}{d}\right)^d.
$$

By the Massart Lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 26.8), we have:

$$
\mathcal{R}_m(\mathcal{L}_{\mathcal{H}})\leq\sqrt{\frac{2d\log(em/d)}{m}} \tag{20}
$$

Combining (17), (18), (19) and (20), the following holds:

$$R_2(\hat{h}; P) - R_2(h^*; P) \leq C \left(\frac{2d\log(em/d) + \left(\frac{1}{c_2-1} \vee 1\right)^2 + \log(2/\delta)}{m} \right)^{1/4} \tag{21}$$

for some constant C. To ensure that the right-hand side of (21) is smaller than ε, we need:

$$m \geq \frac{C}{\varepsilon^4} \left(d\log(m) + d\log(e/d) + \left(\frac{1}{c_2 - 1} \vee 1\right)^2 + \log(2/\delta) \right).$$

Using Lemma 22, a sufficient condition for the inequality to hold is that

$$m \geq \frac{4Cd}{\varepsilon^4}\log(\frac{2Cd}{\varepsilon^4}) + \frac{2C}{\varepsilon^4}\left(\left(\frac{1}{c_2-1} \vee 1\right)^2 + d\log(e/d) + \log(2/\delta) \right),$$

which concludes our proof. ■

A.2 Proof of Theorem 12

Our proof is organized as follows: we first show that, for a hypothesis class $\mathcal{H}$ with finite VC dimension, given sufficient samples (the magnitude is provided in Theorem 12), the samples form a Distributionally Robust ε-net for $\mathcal{H}$; subsequently, we prove that such samples is sufficient for distributionally robust learning.

Proposition 18 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $vc(\mathcal{H}) = d$. Fix $\varepsilon \in (0,1), \delta \in (0, 1/4)$ and let*

$$m \geq \frac{16(1+2\rho)d}{\varepsilon^2}\log\left(\frac{8(1+2\rho)d}{\varepsilon^2}\right) + \frac{8(1+2\rho)}{\varepsilon^2}\left(d\log\left(\frac{2e}{d}\right) + \log\left(\frac{2}{\delta}\right)\right). \tag{22}$$

Then, with probability of at least $1-\delta$ over a choice of $S \sim P^m$, we can conclude that S is a Distributionally Robust ε-net for $\mathcal{H}$.

Proof Let $B \coloneqq \{S \subseteq \mathcal{X} : |S| = m, \exists h \in \mathcal{H}, R_2(h; P) \geq \varepsilon, \mathcal{C}_h \cap S = \emptyset\}$ be the set of sets that are not a Distributionally Robust ε-net. We need to bound $\mathbb{P}[S \in B]$.

Define

$$B' \coloneqq \left\{ (S,T) \subseteq \mathcal{X} : |S| = |T| = m, \exists h \in \mathcal{H}, R_2(h; P) \geq \varepsilon, \mathcal{C}_h \cap S = \emptyset, |T \cap \mathcal{C}_h| > \frac{m\varepsilon^2}{2(1+2\rho)} \right\}.$$

Claim 1. $\mathbb{P}[S \in B] \leq 2 \cdot \mathbb{P}[(S,T) \in B']$.

Since S and T are chosen independently, we can write

$$\begin{aligned}\mathbb{P}[(S,T) \in B'] &= \mathop{\mathbb{E}}_{(S,T)\sim P^{2m}} \mathbb{1}[(S,T) \in B'] \\ &= \mathop{\mathbb{E}}_{S\sim P^m} \mathop{\mathbb{E}}_{T\sim P^m} \mathbb{1}[(S,T) \in B'].\end{aligned}$$

Note that $(S,T) \in B'$ implies $S \in B$, and therefore $\mathbb{1}[(S,T) \in B'] = \mathbb{1}[(S,T) \in B'] \cdot \mathbb{1}[S \in B]$, which yields

$$\begin{aligned}\mathbb{P}[(S,T) \in B'] = \mathop{\mathbb{E}}_{(S,T)\sim P^{2m}} \mathbb{1}[(S,T) \in B'] \cdot \mathbb{1}[S \in B] \\ \mathop{\mathbb{E}}_{S\sim P^m} \mathbb{1}[S \in B] \mathop{\mathbb{E}}_{T\sim P^m} \mathbb{1}[(S,T) \in B'].\end{aligned} \tag{23}$$

Fix some S. Then, either $\mathbb{1}[S \in B] = 0$, or $S \in B$ and then $\exists h_S$ such that $R_2(h_S; P) \geq \varepsilon$ and $|\mathcal{C}_{h_S} \cap S| = 0$. It follows that a sufficient condition for $(S,T) \in B'$ is that $|T \cap \mathcal{C}_{h_S}| > \frac{m\varepsilon^2}{2(1+2\rho)}$. Therefore, whenever $S \in B$, we have

$$\mathop{\mathbb{E}}_{T\sim P^m} \mathbb{1}[(S,T) \in B'] \geq \mathop{\mathbb{P}}_{T\sim P^m} \left[|T \cap \mathcal{C}_{h_S}| > \frac{m\varepsilon^2}{2(1+2\rho)} \right]. \tag{24}$$

However, since we now assume $S \in B$, we know that $R_2(h_S; P) \geq \varepsilon$; accordingly, by Lemma 28, we have $\mathrm{er}(h_S; P) = p \geq \frac{\varepsilon^2}{1+2\rho}$.

Therefore, $|T \cap \mathcal{C}_{h_S}|$ is a binomial random variable with parameters p (probability of success for a single try) and m (number of tries). Chernoff's inequality implies

$$\begin{aligned}\mathbb{P}\left[|T \cap \mathcal{C}_{h_S}| \leq \frac{pm}{2}\right] &= \mathbb{P}\left[|T \cap \mathcal{C}_{h_S}| - pm \leq -pm/2\right] \\ &\leq \exp(-mp/2) \leq \exp\left(-\frac{m\varepsilon^2}{2(1+2\rho)}\right) \\ &\leq \exp\left(-d\log(1/\delta)/2\right) = \delta^{d/2} \leq 1/2,\end{aligned}$$

where the first inequality is obtained via Chernoff's inequality and the penultimate inequality follows from our choice of m.

Thus,

$$\begin{aligned}\mathbb{P}\left[|T \cap \mathcal{C}_{h_S}| > \frac{m\varepsilon^2}{2(1+2\rho)}\right] &= 1 - \mathbb{P}\left[|T \cap \mathcal{C}_{h_S}| \leq \frac{m\varepsilon^2}{2(1+2\rho)}\right] \\ &\geq 1 - \mathbb{P}\left[|T \cap \mathcal{C}_{h_S}| \leq \frac{mp}{2}\right] \geq 1/2.\end{aligned} \tag{25}$$

Combining (23), (24) and (25), we conclude our proof of Claim 1.

Claim 2.(Symmetrization) $\mathbb{P}\left[(S,T) \in B'\right] \leq \exp\left(-\frac{m\varepsilon^2}{4(1+2\rho)}\right) \cdot \tau_{\mathcal{H}}(2m)$.

For ease of notation, let $\alpha = \frac{m\varepsilon^2}{2(1+2\rho)}$, and for a sequence $A = (x_1, \ldots, x_{2m})$, let $A_0 = (x_1, \ldots, x_m)$. Using the definition of B', we get

$$\begin{aligned}\mathbb{P}\left[A \in B'\right] &= \mathop{\mathbb{E}}_{A\sim P^{2m}} \max_{h\in\mathcal{H}} \left\{\mathbb{1}\left[R_2(h;P) \geq \varepsilon\right] \cdot \mathbb{1}\left[|\mathcal{C}_h \cap A_0| = 0\right] \cdot \mathbb{1}\left[|\mathcal{C}_h \cap A| \geq \alpha\right]\right\} \\ &\leq \mathop{\mathbb{E}}_{A\sim P^{2m}} \max_{h\in\mathcal{H}} \left\{\mathbb{1}\left[|\mathcal{C}_h \cap A_0| = 0\right] \cdot \mathbb{1}\left[|\mathcal{C}_h \cap A| \geq \alpha\right]\right\}.\end{aligned}$$

Now let us denote by $\mathcal{H}_A$ the effective number of different hypotheses on A, namely, $\mathcal{H}_A \coloneqq \{\mathcal{C}_h \cap A : h \in \mathcal{H}\}$. It follows that

$$\mathbb{P}[A \in B'] \leq \mathop{\mathbb{E}}_{A\sim P^{2m}} \max_{\mathcal{C}_h\in\mathcal{H}_A} \left\{\mathbb{1}[|\mathcal{C}_h \cap A_0| = 0] \cdot \mathbb{1}[|\mathcal{C}_h \cap A| \geq \alpha]\right\}.$$

Let $J = \{\mathbf{j} \subseteq [2m] : |\mathbf{j}| = m\}$. For any $\mathbf{j} \in J$ and $A = (x_1, \ldots, x_{2m})$, define $A_{\mathbf{j}} = (x_{\mathbf{j}_1}, \ldots, x_{\mathbf{j}_m})$. Since the elements of A are chosen i.i.d., for any $\mathbf{j} \in J$ and any function $f(A, A_0)$, it holds that $\mathbb{E}_{A\sim P^{2m}}[f(A, A_0)] = \mathbb{E}_{A\sim P^{2m}}[f(A, A_{\mathbf{j}})]$. Since this holds for any $\mathbf{j}$, it also holds for the expectation of $\mathbf{j}$ chosen at random from J. In particular, it holds for the function $f(A, A_0) = \sum_{\mathcal{C}_h\in\mathcal{H}_A} \mathbb{1}[|\mathcal{C}_h \cap A_0| = 0] \cdot \mathbb{1}[|\mathcal{C}_h \cap A| \geq \alpha]$. We therefore obtain that

$$\begin{aligned}\mathbb{P}[A \in B'] &\leq \mathop{\mathbb{E}}_{A\sim P^{2m}} \mathop{\mathbb{E}}_{\mathbf{j}\sim U(J)} \sum_{\mathcal{C}_h\in\mathcal{H}_A} \mathbb{1}[|\mathcal{C}_h \cap A_{\mathbf{j}}| = 0] \cdot \mathbb{1}[|\mathcal{C}_h \cap A| \geq \alpha] \\ &= \mathop{\mathbb{E}}_{A\sim P^{2m}} \sum_{\mathcal{C}_h\in\mathcal{H}_A} \mathbb{1}[|\mathcal{C}_h \cap A| \geq \alpha] \mathop{\mathbb{E}}_{\mathbf{j}\sim U(J)} \mathbb{1}[|\mathcal{C}_h \cap A_{\mathbf{j}}| = 0].\end{aligned}$$

Now fix some A, such that $|\mathcal{C}_h \cap A| \geq \alpha$. Thus, $\mathbb{E}_{j\sim U(J)} \mathbb{1}[|\mathcal{C}_h \cap A_{\mathbf{j}}| = 0]$ represents the probability that when choosing m balls from a bag containing at least α red balls, we will never choose a red ball. This probability is at most

$$\left(1 - \frac{\alpha}{2m}\right)^2 = \left(1 - \frac{\varepsilon^2}{4(1+2\rho)}\right)^m \leq \exp\left(-\frac{m\varepsilon^2}{4(1+2\rho)}\right).$$

We therefore have

$$\begin{aligned}\mathbb{P}\left[A \in B'\right] &\leq \mathop{\mathbb{E}}_{A\sim P^{2m}} \sum_{\mathcal{C}_h\in\mathcal{H}_A} \exp\left(-\frac{m\varepsilon^2}{4(1+2\rho)}\right) \\ &\leq \exp\left(-\frac{m\varepsilon^2}{4(1+2\rho)}\right) \mathop{\mathbb{E}}_{A\sim P^{2m}} |\mathcal{H}_A| \\ &\leq \exp\left(-\frac{m\varepsilon^2}{4(1+2\rho)}\right) \cdot \tau_{\mathcal{H}}(2m).\end{aligned}$$

By Sauer's lemma, we know that $\tau_{\mathcal{H}} \leq (2em/d)^d$; combining this with the above two claims, we obtain that

$$\mathbb{P}[S \in B] \leq 2(2em/d)^d \exp\left(-\frac{m\varepsilon^2}{4(1+2\rho)}\right).$$

We would like the right-hand side of the inequality to be at most δ; that is,

$$2(2em/d)^d \exp\left(-\frac{m\varepsilon^2}{4(1+2\rho)}\right) \leq \delta.$$

Through rearrangement, we arrive at

$$m \geq \frac{4(1+2\rho)d}{\varepsilon^2} \log(m) + \frac{4(1+2\rho)d}{\varepsilon^2} \log\left(\frac{2e}{d}\right) + \frac{4(1+2\rho)}{\varepsilon^2} \log\left(\frac{2}{\delta}\right).$$

Using Lemma 22, a sufficient condition for the preceding to hold is that

$$m \geq \frac{16(1+2\rho)d}{\varepsilon^2} \log\left(\frac{8(1+2\rho)d}{\varepsilon^2}\right) + \frac{8(1+2\rho)}{\varepsilon^2} \left(d \log\left(\frac{2e}{d}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

■

Next, we derive distributionally robust PAC learnability from the definition of distributionally robust ε-net.

Proposition 19 *Let $\mathcal{H}$ be a hypothesis class over $\mathcal{X}$ with $vc(\mathcal{H})$. Let P be a distribution over $\mathcal{X}$ and let h^* be a target hypothesis. Fix $\varepsilon, \delta \in (0,1)$ and let m be as defined in Proposition 18; then, with probability of at least $1-\delta$ over a choice of m i.i.d. instances from $\mathcal{X}$ with labels according to h^*, we have that any DRERM hypothesis has a true error of at most ε.*

Proof Define the class $\mathcal{H}^{h^*} = \{\mathcal{C}_{h^*} \triangle \mathcal{C}_h : h \in \mathcal{H}\}$, where $\mathcal{C}_{h^*} \triangle \mathcal{C}_h = (\mathcal{C}_h \backslash \mathcal{C}_{h^*}) \cup (\mathcal{C}_{h^*} \backslash \mathcal{C}_h)$. It can be readily verified that if some $A \subseteq \mathcal{X}$ is shattered by $\mathcal{H}$ then it is also shattered by $\mathcal{H}^{h^*}$, and vice versa. Hence, $\mathrm{vc}(\mathcal{H}) = \mathrm{vc}\left(\mathcal{H}^{h^*}\right)$. Therefore, using Proposition 18, we can determine that with probability of at least $1-\delta$, the sample S is a distributionally robust ε-net for $\mathcal{H}^{h^*}$, note that

$$R_2(h;P) = \sup\left\{\mathbb{E}_Q[\mathbb{1}[h(x) \neq y]] : Q \ll P, D_2(Q\|P)\right\}.$$

Since h^* is the target hypothesis, we have that

$$\mathbb{E}_P[\mathbb{1}[h^*(x) \neq y]] = 0.$$

Thus, $R_2(h^*;P) = \inf_{\eta\in\mathbb{R}}\left\{c_2\mathbb{E}_P[\left(\mathbb{1}[h^(x) \neq y] - \eta\right)_+^2]^{1/2} + \eta\right\} \leq c_2\mathbb{E}_P[(\mathbb{1}[h^*(x) \neq y])_+^2]^{1/2} = 0$, which means that for any $Q \ll P$ with $D_2(Q\|P) \leq \rho$, we have $Q[y = h^*(x)] = 1$.

Therefore,

$$\begin{aligned} R_2(h;P) &= \sup\left\{\mathbb{E}_Q[\mathbb{1}[h(x) \neq y] : Q \ll P, D_2(Q\|P) \leq \rho\right\} \\ &= \sup\left\{\mathbb{E}_Q[\mathbb{1}[h(x) \neq h^*(x)] : Q \ll P, D_2(Q\|P) \leq \rho\right\} \\ &= R_2(h \triangle h^*;P), \end{aligned}$$

where $h \triangle h^*$ is the hypothesis that satisfies $\mathcal{C}_{h\triangle h^*} = \mathcal{C}_{h^*} \triangle \mathcal{C}_h$. Therefore, for any $h \in \mathcal{H}$ with $R_2(h;P) \geq \varepsilon$, we have that $|\mathcal{C}_{h^*\triangle h} \cap S| > 0$; this which implies that h cannot be a DRERM hypothesis, which concludes our proof. ■

Appendix B. Proof of Lower Bound

B.1 Lower Bound in Realizable Case

Proof [Proof of Theorem 14] Suppose that $S = \{x_0, x_1, \ldots, x_{d-1}\}$ is shattered by $\mathcal{H}$. Let P be the probability distribution on the domain $\mathcal{X}$ of $\mathcal{H}$ such that $P(x) = 0$ if $x \notin S$, $P(x_0) = 1 - 8\varepsilon$, and for $i = 1, \ldots, d-1$, $P(x_i) = 8\varepsilon/(d-1)$. With probability 1, for any m, a P^m-random sample lies in S^m; henceforth, to simplify our analysis, we assume without loss of generality that $\mathcal{X} = S$ and that $\mathcal{H}$ consists precisely of all 2^d functions from S to $\{0,1\}$. For convenience, and to be explicit, if a training sample $\mathbf{z} = (z_1, \ldots, z_m) \in (S \times \{0,1\})^m$ corresponds to a sample $\mathbf{x} \in \mathcal{X}^m$ and a labeling function $t \in \mathcal{H}$, we shall denote $\mathcal{A}(\mathbf{z})$ by $\mathcal{A}(\mathbf{x}, t)$.

Let $S' = \{x_1, \ldots, x_{d-1}\}$ and let $\mathcal{H}'$ be the set of all 2^{d-1} functions $h \in \mathcal{H}$ such that $h(x_0) = 0$. We shall employ the probabilistic method, with target function t drawn at random according to the uniform distribution U on $\mathcal{H}'$. Let $\mathcal{A}$ be any algorithm for $\mathcal{H}$. We obtain a lower bound on the sample complexity of $\mathcal{A}$ under the assumption that $\mathcal{A}$ always returns a hypothesis in $\mathcal{H}'$; that is, we assume that whatever sample $\mathbf{z}$ is given, $\mathcal{A}$ will classify x_0 correctly. (This assumption causes no loss of generality: if the output hypothesis of $\mathcal{A}$ does not always belong to $\mathcal{H}'$, we can consider the "better" learning algorithm derived from $\mathcal{A}$ whose output hypotheses are forced to classify x_0 correctly. Clearly, a lower bound on the sample complexity of this latter algorithm is also a lower bound on the sample complexity of $\mathcal{A}$.) Let m be any fixed positive integer and, for $\mathbf{x} \in S^m$, denote by $l(\mathbf{x})$ the number of distinct elements of S' occurring in the sample $\mathbf{x}$. It is evident that for any $x \in S'$, exactly half of the functions $h' \in \mathcal{H}'$ satisfy $h'(x) = 1$. It follows that for any fixed $\mathbf{x} \in S^m$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{t\sim U(\mathcal{H}')} \mathrm{er}(\mathcal{A}(\mathbf{x}, t); P) &= \sum_{t'\in\mathcal{H}} \mathbb{P}_{t\sim U(\mathcal{H}')}(t = t')\mathrm{er}(\mathcal{A}(\mathbf{x}, t); P) \\
&= \sum_{t'\in\mathcal{H}} \mathbb{P}_{t\sim U(\mathcal{H}')}(t = t') \sum_{y\in S'} \mathbb{P}_{y'\sim P}(y' = y)\mathbb{1}[\mathcal{A}(\mathbf{x}, t) \neq t(y)] \\
&\geq \sum_{t'\in\mathcal{H}} \mathbb{P}_{t\sim U(\mathcal{H}')}(t = t') \sum_{y\in S'\setminus\mathbf{x}} \mathbb{P}_{y'\sim P}(y' = y)\mathbb{1}[\mathcal{A}(\mathbf{x}, t) \neq t(y)] \\
&= \sum_{y\in S'\setminus\mathbf{x}} \frac{1}{2} \cdot \mathbb{P}_{y'\sim P}(y' = y) = \frac{1}{2} \cdot \frac{8\varepsilon}{d-1} \cdot (d - 1 - l(\mathbf{x})).
\end{aligned} \tag{26}$$

The penultimate equality can be obtained from the following: for a fixed y, we can divide $\mathcal{H}'$ into two groups, namely $\mathcal{H}'_i = \{t \in \mathcal{H}' : t(y) = i\}$, $i \in \{0, 1\}$, and for any $h_1 \in \mathcal{H}'_1$, we can always find a hypothesis $h_2 \in \mathcal{H}'_2$, such that for any $x \neq y$ and $x \in S$, $h_1(x) = h_2(x)$. Since $y \in S'\setminus\mathbf{x}$, we have $\mathcal{A}(\mathbf{x}, h_1) = \mathcal{A}(\mathbf{x}, h_2)$. Therefore, it can be seen that $\mathbb{1}[\mathcal{A}(\mathbf{x}, h_1) \neq h_1(x)] + \mathbb{1}[\mathcal{A}(\mathbf{x}, h_2) \neq h_2(x)] = 1$.

We now focus on a special subset $\mathcal{S}$ of S^m, consisting of all $\mathbf{x}$ for which $l(\mathbf{x}) < \frac{d-1}{2}$. If $\mathbf{x} \in \mathcal{S}$, then by (26),

$$\mathop{\mathbb{E}}_{t\sim U(\mathcal{H}')} \mathrm{er}(\mathcal{A}(\mathbf{x}, t); P) > 2\varepsilon. \tag{27}$$

Now let Q denote the restriction of P^m to $\mathcal{S}$, so that for any $A \subseteq S^m$, $Q(A) = P^m(A \cap \mathcal{S})/P^m(\mathcal{S})$. Accordingly,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim Q} \mathop{\mathbb{E}}_{t\sim U(\mathcal{H}')} \mathrm{er}(\mathcal{A}(\mathbf{x}, t); P) > 2\varepsilon,$$

since (27) holds for every $\mathbf{x} \in \mathcal{S}$. By Fubini's theorem, the two expectation operations may be interchanged. In other words,

$$\mathop{\mathbb{E}}_{t\sim U(\mathcal{H}')} \mathop{\mathbb{E}}_{\mathbf{x}\sim Q} \mathrm{er}(\mathcal{A}(\mathbf{x}, t); P) = \mathop{\mathbb{E}}_{\mathbf{x}\sim Q} \mathop{\mathbb{E}}_{t\sim U(\mathcal{H}')} \mathrm{er}(\mathcal{A}(\mathbf{x}, t); P) > 2\varepsilon.$$

This implies that for some $t' \in \mathcal{H}'$,

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim Q} \mathrm{er}(\mathcal{A}(\mathbf{x}, t); P) > 2\varepsilon.$$

Let p_ε be the probability with respect to Q that $\mathrm{er}(\mathcal{A}(\mathbf{x};t');P) \geq \varepsilon$.

Given our assumption that $\mathcal{A}$ returns a function in $\mathcal{H}'$, the error of $\mathcal{A}(\mathbf{x},t')$ with respect to P is never more than 8ε (the probability of S'). Hence, we must have

$$2\varepsilon < \mathop{\mathbb{E}}_{\mathbf{x}\sim Q} \mathrm{er}(\mathcal{A}(\mathbf{x},t');P) \leq 8\varepsilon \cdot p_\varepsilon + (1-p_\varepsilon)\varepsilon$$

from which we obtain $p_\varepsilon > 1/7$. It now follows that

$$\begin{aligned}
\mathbb{P}_{\mathbf{x}\sim P^m}\left(\mathrm{er}(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\right) &= \frac{\mathbb{P}_{\mathbf{x}\sim P^m}\left(\mathrm{er}(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\right)}{\mathbb{P}_{\mathbf{x}\sim P^m}(\mathbf{x}\in\mathcal{S})} \cdot \mathbb{P}_{\mathbf{x}\sim P^m}(\mathbf{x}\in\mathcal{S}) \\
&\geq \frac{\mathbb{P}_{\mathbf{x}\sim P^m}\left(\mathbf{x}\in\{\mathrm{er}(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\}\cap\mathcal{S}\right)}{\mathbb{P}_{\mathbf{x}\sim P^m}(\mathbf{x}\in\mathcal{S})} \cdot \mathbb{P}_{\mathbf{x}\sim P^m}(\mathbf{x}\in\mathcal{S}) \\
&= \mathbb{P}_{\mathbf{x}\sim Q}\left(\mathrm{er}(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\right) \cdot \mathbb{P}_{\mathbf{x}\sim P^m}(\mathbf{x}\in\mathcal{S}) \\
&> \frac{1}{7} \cdot \mathbb{P}_{\mathbf{x}\sim P^m}(\mathbf{x}\in\mathcal{S}).
\end{aligned} \tag{28}$$

Since $R_2(\mathcal{A}(\mathbf{x},t');P) \geq \mathrm{er}(\mathcal{A}(\mathbf{x},t');P)$, we have

$$\mathbb{P}_{\mathbf{x}\sim P^m}\left(R_2(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\right) \geq \mathbb{P}_{\mathbf{x}\sim P^m}\left(\mathrm{er}(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\right) > \frac{1}{7} \cdot \mathbb{P}_{\mathbf{x}\sim P^m}(\mathbf{x}\in\mathcal{S}).$$

Now, $P^m(\mathcal{S})$ is the probability that a P^m-random sample $\mathbf{z}$ has no more than $\frac{d-1}{2}$ distinct entries from S', and this is at least $1 - \mathrm{GE}\left(8\varepsilon, m, \frac{d-1}{2}\right)$, where

$$\mathrm{GE}\left(p, m, (1+\varepsilon)mp\right) \coloneqq \sum_{i=\lceil(1+\varepsilon)mp\rceil}^{m} \tbinom{m}{i} p^i (1-p)^{m-i}.$$

Using Lemma 29, generally, it follows that $\mathrm{GE}(p,m,k) \leq \exp\left(-(k-pm)^2/(3pm)\right)$.

If $m \leq \frac{d-1}{32m}$, then it is evident that this probability is at least $7/100$. Therefore, if $m \leq \frac{d-1}{32\varepsilon}$ and $\delta < 1/100$,

$$\begin{aligned}
\mathbb{P}_{\mathbf{x}\sim P^m}\left(R_2(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\right) &\geq \mathbb{P}_{\mathbf{x}\sim P^m}\left(\mathrm{er}(\mathcal{A}(\mathbf{x},t');P) \geq \varepsilon\right) \\
&> \frac{1}{7} \cdot \frac{7}{100} = \frac{1}{100} \geq \delta
\end{aligned}$$

and the first part of the result follows.

To prove the second part of the theorem, note that if $\mathcal{H}$ contains at least three functions, there exist examples a, b and functions $h_1, h_2 \in \mathcal{H}$ such that $h_1(a) = h_2(a)$ and $h_1(b) = 1, h_2(b) = 0$. Without loss of generality, we shall assume that $h_1(a) = h_2(a) = 1$. Let P be the probability distribution for which $P(a) = 1-\varepsilon$ and $P(b) = \varepsilon$. The probability that a sample $\mathbf{x} \in \mathcal{X}^m$ has all its entries equal to a is $(1-\varepsilon)^m$. Now, $(1-\varepsilon)^m \geq \delta$ if and only if

$$m \leq \frac{\log(1/\delta)}{-\log(1-\varepsilon)}.$$

Furthermore, $-\log(1-\varepsilon) \leq 2\varepsilon$ for $0 < \varepsilon \leq 3/4$. It follows that if m is no lager than $\frac{\log(1/\delta)}{2\varepsilon}$, then with probability greater than δ, a sample $\mathbf{x} \in \mathcal{X}^m$ has all its entries equal to a.

Let $\mathbf{a}^1$ denote the training sample $\mathbf{a}^1 = ((a,1),\ldots,(a,1))$ with length m. Note that $\mathbf{a}^1$ is a training sample corresponding to h_1 and h_2. Suppose that $\mathcal{A}$ is a learning algorithm for $\mathcal{H}$, and let $\mathcal{A}_a$ denote the output $\mathcal{A}(\mathbf{a}^1)$ of $\mathcal{A}$ on the sample $\mathbf{a}^1$.

If $\mathcal{A}_a(b) = 1$ then $\mathcal{A}_a$ has an error of at least ε (the probability of b) with respect to h_2, which implies $R_2(\mathcal{A}(\mathbf{a}^1, h_2); P) \geq \varepsilon$; while if $\mathcal{A}_a(b) = 0$, then it has error of at least with respect to h_1, which implies that $R_2(\mathcal{A}(\mathbf{a}^1, h_1); P) \geq \varepsilon$.

It follows that if $m \leq \frac{\log(1/\delta)}{2\varepsilon}$, then either

$$\mathbb{P}_{\mathbf{z}\sim P^m}\left(R_2(\mathcal{A}(\mathbf{z}, h_1); P) \geq \varepsilon\right) \geq \mathbb{P}_{\mathbf{z}\sim P^m}\left(\mathbf{z} = \mathbf{a}^1\right) > \delta$$

or

$$\mathbb{P}_{\mathbf{z}\sim P^m}\left(R_2(\mathcal{A}(\mathbf{z}, h_2); P) \geq \varepsilon\right) \geq \mathbb{P}_{\mathbf{z}\sim P^m}\left(\mathbf{z} = \mathbf{a}^1\right) > \delta$$

We therefore deduce that the learning algorithm fails for some $t \in \mathcal{H}$ if m is this small. ■

Appendix C. Auxiliary Lemma

The following Hoeffding's lemma can be found in (Shalev-Shwartz and Ben-David, 2014, Lemma 4.5).

Lemma 20 (Hoeffding's Inequality) *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables, and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right] \leq 2\exp\left(-2m\varepsilon^2/(b-a)^2\right).$$

We next recall an important lemma that is useful for bounding the Rademacher complexity and can be found in (Mohri et al., 2012, Lemma 4.2).

Lemma 21 (Talagrand's Lemma) *Let ψ be a ρ-Lipschitz function. For any function class $\mathcal{H}$, we have:*

$$\mathcal{R}_m(\psi \circ \mathcal{H}) \leq \rho\mathcal{R}_m(\mathcal{H}).$$

The following lemma is fundamental and can be found in (Shalev-Shwartz and Ben-David, 2014, Lemma A.1).

Lemma 22 *Let $a > 1$ and $b > 0$. Then: $x \geq 4a\log(2a) + 2b \implies x \geq a\log(x) + b$.*

Lemma 23 *For $P_b(x,y) = \begin{cases} \dfrac{1+yb\varepsilon}{2}, & \text{if} \quad x = c, \\ 0 \quad , & \text{otherwise.}\end{cases}$ and $\rho \in \left(0, \frac{3-2\sqrt{2}}{2}\right)$, we have $R_2(h;P_b) = \left[1 - h(c)b\varepsilon + \sqrt{2\rho(1-\varepsilon^2)}\right]/2$.*

Proof Recall the definition of $R_2(h;P_b) = \sup\left\{\mathbb{E}_P\left[\mathbb{1}[h(x) \neq y]\right] : P \ll P_b, D_2(P\|P_b) \leq \rho\right\}$.

We write P as

$$P(x,y)=\begin{cases}\dfrac{1+\xi}{2}, & \text{if } x=c \text{ and } y=b,\\ \dfrac{1-\xi}{2}, & \text{if } x=c \text{ and } y=-b,\\ \;\;0\;\;\;\;, & \text{otherwise.}\end{cases}$$

Thus, $R_2(h;P_b)$ can be rewritten as

$$R_2(h;P_b)=\sup_{\xi}\left\{\frac{1+\xi}{2}\mathbb{1}[h(c)\neq b]+\frac{1-\xi}{2}\mathbb{1}[h(c)\neq -b]:\xi^2-2\varepsilon\xi+\varepsilon^2-2\rho(1-\varepsilon^2)\leq 0,\right.$$
$$\left.-1\leq\xi\leq 1\right\}.$$

Solving the quadratic inequality in the above formulation, we get $\varepsilon-\sqrt{2\rho(1-\varepsilon^2)}\leq\xi\leq\varepsilon+\sqrt{2\rho(1-\varepsilon^2)}$. Since we assume $\rho\in\left(0,\frac{3-2\sqrt{2}}{2}\right)$ and $\rho\in\left(0,\frac{1}{\sqrt{2}}\right)$, the left-hand and right-hand sides can both be achieved.

When $h(c)=b$, we have $\mathrm{er}(h;P)=\frac{1-\xi}{2}$, and thus $R_2(h;P_b)=\frac{1-\varepsilon+\sqrt{2\rho(1-\varepsilon^2)}}{2}$; following the same logic, when $h(c)=-b$, we have $R_2(h;P_b)=\frac{1+\varepsilon+\sqrt{2\rho(1-\varepsilon^2)}}{2}$.

Thus, we obtain the desired result. ■

We frequently employ the following estimate on the binomial random variable probability (Slud, 1977):

Lemma 24 (Slud's Inequality) *Let X be a (m,p) binomial random variable and assume that $p=(1-\varepsilon)/2$. Then,*

$$\mathbb{P}[X\geq m/2]\geq\frac{1}{2}\left(1-\sqrt{1-\exp(-m\varepsilon^2/(1-\varepsilon^2))}\right).$$

Lemma 25 *Among all algorithms, (14) is minimized for $\mathcal{A}$ being the Maximum-Likelihood algorithm, $\mathcal{A}_{ML}$, defined as*

$$\forall i,\quad \mathcal{A}_{ML}(S)(c_i)=sign\left(\sum_{k:x_k=c_i}y_k\right).$$

Proof Fix some $\mathbf{j}\in[d]^m$. Note that, given $\mathbf{j}$ and $\mathbf{y}\in\{+1,-1\}^m$, the training set S is fully determined; we can therefore write $\mathcal{A}(\mathbf{j},\mathbf{y})$ instead of $\mathcal{A}(S)$. Let us also fix $i\in[d]$. Denote by $\mathbf{b}^{\neg i}$ the sequence $(b_1,\ldots,b_{i-1},b_{i+1},\ldots,b_m)$. Also, for any $\mathbf{b}\in\{+1,-1\}^m$, let $\mathbf{y}^I$ denote the elements of $\mathbf{y}$ corresponding to indices for which $j_k=i$, and let $\mathbf{y}^{\neg I}$ be the rest of the elements of $\mathbf{y}$. We therefore have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{b}\sim U(\{+1,-1\}^d)}\mathop{\mathbb{E}}_{\forall k,y_k\sim b_{j_k}}\mathbb{1}[\mathcal{A}(S)(c_i)\neq b_i]\\ &=\frac{1}{2}\sum_{b_i\in\{+1,-1\}}\mathop{\mathbb{E}}_{\mathbf{b}^{\neg i}\sim U(\{+1,-1\}^{d-1})}\sum_{\mathbf{y}}P[\mathbf{y}|\mathbf{b}^{\neg i},b_i]\mathbb{1}[\mathcal{A}(\mathbf{j},\mathbf{y})(c_i)\neq b_i]\\ &=\mathop{\mathbb{E}}_{\mathbf{b}^{\neg i}\sim U(\{+1,-1\}^{d-1})}\sum_{\mathbf{y}^{\neg I}}P[\mathbf{y}^{\neg I}|\mathbf{b}^{\neg i}]\frac{1}{2}\sum_{\mathbf{y}^I}\left(\sum_{b_i\in\{+1,-1\}}P[\mathbf{y}^I|\mathbf{b}_i]\mathbb{1}[\mathcal{A}(\mathbf{j},\mathbf{y})(c_i)\neq b_i]\right).\end{aligned}$$

The sum within the parentheses is minimized when $\mathcal{A}(\mathbf{j},\mathbf{y})(c_i)$ is the maximizer of $P[\mathbf{y}^I|b_i]$ over $b_i \in \{+1,-1\}$, which is exactly the Maximum-Likelihood rule. Repeating the same argument for all i we conclude our proof. ■

Lemma 26 *Let $\Delta = \frac{1}{r}\left(R_2(\mathcal{A}(S);P) - \inf_{h\in\mathcal{H}} R_2(h;P)\right)$, where r is defined in the proof of Theorem 9. We have $\Delta \in [0, 1+\sqrt{2}\rho^{1/2}/4]$.*

Proof Recalling (9), we have

$$\begin{aligned} R_2(\mathcal{A}(S);P_{\mathbf{b}}) - \inf_{h\in\mathcal{H}} R_2(h;P_{\mathbf{b}}) &= R_2(\mathcal{A}(S);P_{\mathbf{b}}) - R_2(h_{\mathbf{b}};P_{\mathbf{b}}) \\ &= r\cdot\frac{d_-}{d} + \frac{1}{2}\left(\sqrt{2\rho\left(1-r^2(d_+-d_-)^2/d^2\right)} - \sqrt{2\rho(1-r^2)}\right). \end{aligned} \tag{29}$$

Next, we upper bound the second term in (29),

$$\begin{aligned} &\frac{1}{2}\left(\sqrt{2\rho\left(1-r^2(d_+-d_-)^2/d^2\right)} - \sqrt{2\rho(1-r^2)}\right) \\ &\leq \frac{1}{2}\rho\left(2\rho(1-r^2)\right)^{-1/2} r^2\left(1-\frac{(d_+-d_-)^2}{d^2}\right) \\ &\leq \frac{1}{2}\rho^{1/2}r^2, \end{aligned}$$

where the first inequality follows from $\sqrt{x} \leq \frac{1}{2}(\sqrt{x}+\sqrt{y}), \forall x \leq y$, while the second follows from the fact that $r^2 \leq 1/2$. Thus, $\Delta \leq \rho^{1/2}r/2 \leq \sqrt{2}\rho^{1/2}/4$. ■

Lemma 27 *Let Z be a random variable that takes values in $[0,c], c > 1$. Assume that $\mathbb{E}[Z] = \mu$. Then, for any $a \in (0,c)$,*

$$\mathbb{P}[Z > a] \geq \frac{\mu - a}{c - a}.$$

Proof

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}\left[Z\cdot\mathbb{1}[0\leq Z\leq a]\right] + \mathbb{E}\left[Z\cdot\mathbb{1}[a < Z \leq c]\right] \\ &\leq a\left(1-\mathbb{P}[Z>a]\right) + c\mathbb{P}[Z>a]. \end{aligned}$$

Following rearrangement, we obtain the desired result. ■

Lemma 28 *For any probability distribution P and predictor $h \in \mathcal{H}$, we have*

$$R_2(h;P) \leq c_2(\rho)er(h;P)^{1/2}.$$

Proof Recalling the dual formulation of $R_2(h;P)$ (Proposition 4), we have

$$\begin{aligned} R_2(h;P) &= \inf_{\eta\in\mathbb{R}}\left\{c_2(\rho)\mathbb{E}_P\left[(\mathbb{1}[h(x)\neq y]-\eta)_+^2\right]^{1/2} + \eta\right\} \\ &\leq c_2(\rho)\mathbb{E}_P[(\mathbb{1}[h(x)\neq y])_+^2]^{1/2} \\ &= c_2(\rho)\mathrm{er}(h;P)^{1/2}, \end{aligned}$$

where the first inequality follows by setting $\eta = 0$ and the last line follows from the fact that the indicator $\mathbb{1}[h(x) \neq y]$ is a $0-1$-valued function. ■

Lemma 29 *For* $\varepsilon \in (0,1)$, $GE(p, m, (1+\varepsilon)mp) := \sum_{i=\lceil(1+\varepsilon)mp\rceil}^{m} \binom{m}{i} p^i (1-p)^{m-i} \leq \exp\left(-\varepsilon^2 pm/3\right)$.

Proof Let $Z_1, \ldots, Z_m$ be independent Bernoulli variables, where for every i, $\mathbb{P}[Z_i = 1] = p$ and $\mathbb{P}[Z_i = 0] = 1 - p$. Thus, $\mathrm{GE}(p, m, (1+\varepsilon)mp)$ can be written as follows:

$$\begin{aligned} \mathrm{GE}(p, m, (1+\varepsilon)mp) &= \sum_{i=\lceil(1+\varepsilon)mp\rceil}^{m} \binom{m}{i} p^i (1-p)^{m-i} \\ &= \mathbb{P}\left[\sum_{i=1}^{m} X_i \geq (1+\varepsilon)mp\right] \end{aligned}$$

Using (Shalev-Shwartz and Ben-David, 2014, Lemma B.4), we obtain

$$\begin{aligned} \mathrm{GE}(p, m, (1+\varepsilon)mp) = \sum_{i=\lceil(1+\varepsilon)mp\rceil}^{m} \binom{m}{i} p^i (1-p)^{m-i} &\leq \exp\left(-p\frac{\varepsilon^2}{2+2\varepsilon/3}\right) \\ &\leq \exp\left(-\varepsilon^2 pm/3\right), \end{aligned}$$

where the last inequality holds, since we assume $\varepsilon \in (0,1)$. ■

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61976161, the Fundamental Research Funds for the Central Universities under Grant 2042022rc0016.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, pages 137–144, 2006.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.

Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, 46(3):440–464, 1984.

Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operation Research*, 58(3):595–612, 2010.

John C. Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20:68:1–68:55, 2019.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *CoRR*, abs/2009.04382, 2020.

Patrick J Grother, Patrick J Grother, P Jonathon Phillips, and George W Quinn. *Report on the evaluation of 2D still-image face recognition algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2011.

David J. Hand. Classifier technology and the illusion of progress. *Statistical science*, 21(1): 1–14, 2006.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 483–488, 2015.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *NeurIPS*, pages 2692–2701, 2018.

Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.

Tomohiro Nishiyama and Igal Sason. On relations between the relative entropy and χ 2-divergence, generalizations and applications. *Entropy*, 22(5):563, 2020.

Il Park, Sohan Seth, Murali Rao, and José C Príncipe. Estimation of symmetric chi-square divergence for point processes. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2016–2019. IEEE, 2011.

Dominik Rothenhäusler, Nicolai Meinshausen, and Peter Bühlmann. Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, pages 255–277. Springer, 2016.

Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson. Academic performance prediction in a gender-imbalanced environment. 2017.

Ram Naresh Saraswat. Chi square divergence measure and their bounds. In *3rd International Conference on "Innovative Approach in Applied Physical, Mathematical/Statistical", Chemical Sciences and Emerging Energy Technology for Sustainable Development*, page 55, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. 2014.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Eric Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, 5(3): 404–412, 1977.

Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59, 2017.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

Vladimir Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. 2015.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.