
How Many Domains Suffice for Domain Generalization? A Tight Characterization via the Domain Shattering Dimension

Cynthia Dwork
Harvard University
dwork@seas.harvard.edu

Lunjia Hu *
Northeastern University
lunjia@alumni.stanford.edu

Han Shao †
University of Maryland, College Park
hanshao@umd.edu

Abstract

We study a fundamental question of domain generalization: given a family of domains (*i.e.*, data distributions), how many randomly sampled domains do we need to collect data from in order to learn a model that performs reasonably well on every seen and unseen domain in the family? We model this problem in the PAC framework and introduce a new combinatorial measure, which we call the *domain shattering dimension*. We show that this dimension characterizes the *domain sample complexity*. Furthermore, we establish a tight quantitative relationship between the domain shattering dimension and the classic VC dimension, demonstrating that every hypothesis class that is learnable in the standard PAC setting is also learnable in our setting.

1 Introduction

The ability to generalize across domains is an important component of human intelligence and a crucial milestone in the development of increasingly powerful artificial intelligence. It is not surprising that an experienced driver in one country can reasonably, albeit imperfectly, drive in another country even without further training. It is also not surprising that a skilled chess player can outperform an average person on a totally different board game. An expert doctor who has studied a disease using data from a few large hospitals can potentially provide reasonable treatment for patients in a geographically remote and biologically different population that does not have access to those large hospitals and has not been the subject of prior study. Such domain generalization abilities are empowered by the capability of the learner (*i.e.*, the driver, chess player, or doctor) to distill and master universal laws (about driving, game playing, or medical treatment) that hold even on unseen domains, while separating them from idiosyncratic patterns that are domain-specific.

In this work, we study the theoretical foundations of domain generalization with a focus on capturing this ability of learning “universal laws” while not overfitting to the “idiosyncratic patterns”. Consider a family $\mathcal{G}$ of domains, where every domain $\mathcal{D}$ is a distribution on examples (x, y) each consisting of an instance $x \in X$ and a binary label $y \in \{0, 1\}$. On a specific domain $\mathcal{D}$, there may be a hypothesis

*Work done while at Harvard. Supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness and the Harvard Center for Research on Computation and Society (CRCS).

†Work done while at Harvard. Supported by the Harvard Center of Mathematical Sciences and Applications (CMSA).

(i.e., classifier) $h : X \rightarrow \{0, 1\}$ with low classification error $\text{err}_{\mathcal{D}}(h)$, say, below 0.01:

$$\text{err}_{\mathcal{D}}(h) := \Pr_{(x,y) \sim \mathcal{D}}[y \neq h(x)] \leq 0.01.$$

However, the strong performance of this hypothesis may rely on patterns specific to the current domain and may fail to transfer to other domains $\mathcal{D}' \in \mathcal{G}$. For the purpose of domain generalization, we would instead prefer a different hypothesis h^* that achieves reasonably low—though not necessarily minimal—error across *all* domains $\mathcal{D} \in \mathcal{G}$. Specifically, for a mild but meaningful error threshold, say 0.3, we assume

$$\max_{\mathcal{D} \in \mathcal{G}} \text{err}_{\mathcal{D}}(h^*) \leq 0.3. \quad (1)$$

Motivated by this, we model the underlying “universal laws” by assuming the existence of such a universally good hypothesis $h^* \in \mathcal{H}$ that satisfies Equation (1). We make no further assumptions on the structure of the domains in $\mathcal{G}$ or on the relationships among them.

In our domain generalization task, a learner has data access to a limited number of domains $\mathcal{D}_1, \dots, \mathcal{D}_k$ sampled i.i.d. from a meta-distribution $\mathcal{P}$ over domain space $\mathcal{G}$. Using this limited data access, the learner’s goal is to output a model $h : X \rightarrow \{0, 1\}$ that achieves reasonably good performance on new, unseen domains drawn from $\mathcal{P}$ *without any additional training*. Concretely, for performance threshold and error parameter $\tau, \gamma \in [0, 1]$, the learner’s goal is to output $h : X \rightarrow \{0, 1\}$ such that

$$\Pr_{\mathcal{D} \sim \mathcal{P}}[\text{err}_{\mathcal{D}}(h) \leq \tau] \geq 1 - \gamma.$$

In this work, we aim to answer the following quantitative question:

How many domains does the learner need to see to achieve domain generalization?

The answer to this question, which we call the *domain sample complexity*, is analogous to the notion of sample complexity in classic learning theory. Sample complexity is defined as the number of *data points* needed to learn a good model on a *single* domain, whereas our notion of domain sample complexity measures the number of *domains* the learner needs to collect data from in order to generalize to unseen domains. Thus, domain sample complexity can be viewed as a higher-level meta-variant of the notion of sample complexity.

Characterizing the sample complexity of various learning tasks is a central question in learning theory. For example, the celebrated VC theory characterizes the sample complexity of classic PAC learning [Valiant, 1984] using the VC dimension [Vapnik and Chervonenkis, 1971], a combinatorial complexity measure of the hypothesis class $\mathcal{H}$. The goal of our work is to characterize the domain sample complexity, which depends not only on the complexity of the hypothesis class $\mathcal{H}$, but also on the *interaction* between $\mathcal{H}$ and the domain family $\mathcal{G}$. For example, even if $\mathcal{H}$ is itself very complex, the domain sample complexity can still be very small if the domains in $\mathcal{G}$ are very similar. Even when both $\mathcal{H}$ and $\mathcal{G}$ are complex, the domain sample complexity can still be small if their complexities are concentrated on disjoint subsets of the input space X . For instance, suppose all hypotheses $h \in \mathcal{H}$ are identical on a subset $X_1 \subseteq X$, and $\mathcal{G}$ only contains domains $\mathcal{D}$ satisfying $\Pr_{\mathcal{D}}[x \in X_1] = 1$. In this case $\mathcal{H}$ can be complex outside of X_1 and $\mathcal{G}$ can be complex inside X_1 , but the domain sample complexity is always zero: on every fixed domain $\mathcal{D} \in \mathcal{G}$, all hypotheses $h \in \mathcal{H}$ achieve the same error. Finding a combinatorial quantity that tightly characterizes the domain sample complexity requires accurately capturing this interaction between $\mathcal{H}$ and $\mathcal{G}$.³

1.1 Our Contributions

We model the domain generalization problem in the PAC framework and introduce a new combinatorial measure, the *domain shattering dimension* (Definition 4.1), for a hypothesis class $\mathcal{H}$ with respect to a family $\mathcal{G}$ of domains.

For each hypothesis h , we define the function $\text{err}_{\cdot}(h) : \mathcal{G} \rightarrow \mathbb{R}$ that maps each domain to the error of h on that domain, thereby inducing a new function class. A natural idea is to analyze the fat-shattering dimension [Kearns and Schapire, 1994] of this class. This turns out to overestimate the domain sample complexity. For a target error threshold τ (e.g. $\tau = 0.3$), the domain sample complexity can

³A similar situation that requires capturing the *interaction* between a pair of hypothesis classes occurs in the study of *comparative learning* (hybrid of *realizable* and *agnostic* learning) [Hu and Peale, 2023, Hu et al., 2022].

be very small (e.g. when $\text{err}_{\mathcal{D}}(h) \leq \tau$ is achieved by every $h \in \mathcal{H}$ on every domain $\mathcal{D} \in \mathcal{G}$), but the size of a shattered set can be large at a different threshold τ' , causing the fat-shattering dimension to be large as well.

To address this, we propose a new dimension by modifying the fat-shattering dimension: instead of allowing different thresholds across shattered domains, we require a uniform and fixed threshold across all of them. We show that the resulting domain shattering dimension gives both upper (Theorem 4.1) and lower (Theorem 4.3) bounds on the domain sample complexity, and these bounds match up to a poly-logarithmic factor. We obtain the upper bound by a min-max variant of the empirical risk minimization (ERM) algorithm. To analyze the domain sample complexity achieved by the min-max ERM algorithm, we establish a uniform convergence bound (Lemma 4.2) for *partial concept classes* based on a generalized Sauer-Shelah-Perles lemma by Alon et al. [2022], which may be of independent interest. Our algorithm and upper bound can be directly applied to other learning tasks beyond binary classification, such as multi-class classification and regression (Remark 4.2).

The next question we address in this paper is *the relationship between domain generalization learnability and standard PAC learnability*. It is well known that PAC learnability is characterized by the VC dimension. We compare our domain shattering dimension with the VC dimension and show that, for a hypothesis class $\mathcal{H}$ with VC dimension d and a margin α , the domain shattering dimension is always upper bounded by $O(d \log(1/\alpha))$ (Theorem 5.2). Moreover, we construct a hypothesis class with VC dimension d and a domain family $\mathcal{G}$ for which the domain shattering dimension is $\Omega(d \log(1/\alpha))$ (Theorem 5.1). This establishes a tight relationship between the two measures and demonstrates that standard PAC learnability implies domain generalization learnability, but the domain sample complexity can be much smaller than the sample complexity for PAC learning.

Finally, we relate our work to the literature on domain adaptation. In domain adaptation, the goal is to generalize from a source distribution to a related target distribution, under the assumption that the two are sufficiently similar. This similarity is often quantified using measures such as the $\mathcal{H}$ -divergence introduced in Ben-David et al. [2010]. We show that if all domains in $\mathcal{G}$ are similar to each other under a metric based on a modified version of the $\mathcal{H}$ -divergence, then the domain shattering dimension is 1. Moreover, when the hypothesis class admits a finite cover under this metric, the domain shattering dimension can be upper bounded by the covering number (Theorem 6.1). These results illustrate the potential of our domain shattering dimension as a general notion for characterizing domain generalization without explicitly modeling how domains are related.

1.2 Related Work

There is a vast literature on domain generalization and related learning paradigms such as meta-learning, zero-shot learning, domain adaptation, out of distribution generalization, transfer learning, invariant risk minimization, and multi-task learning. The excellent survey of domain generalization of Wang et al. [2022] provides a helpful taxonomy of many of these paradigms and explores theoretical underpinnings of domain adaptation and domain generalization. Zhou et al. [2022] organizes a plethora of works by (1) application area; (2) method; and (3) learning paradigm, and provides pointers to many datasets commonly used for domain generalization.

Our work is distinguished from prior theoretical work in that simultaneously (1) we make only the minimalist assumption of the existence of a universally good hypothesis $h^* \in \mathcal{H}$, (e.g., one satisfying Equation (1)), with no further requirements regarding how different domains in $\mathcal{G}$ are related, making our model very general; and (2) we obtain *provable* results for the strong learning objective of requiring the output model h to achieve a reasonably low error simultaneously on essentially *all* domains (including the unobserved ones), rather than achieving low error *in expectation* over choice of domain. Achieving low error simultaneously on all domains can be substantially different from achieving low average error when τ is not very close to zero. For example, an average error of 0.25 could arise if the hypothesis incurs zero error on 3/4 of the domains but makes completely incorrect predictions (error 1) on the remaining 1/4 of the domains. Finally, we explicitly focus on the question of *domain* sample complexity – the total number of domains that need to be sampled in order to generalize well to a random new domain. This is different from the common notions of sample complexity and query complexity, and this lens yielded the key concept of the domain shattering dimension.

2 Preliminaries

We include the definition of *partial concepts* that will be very useful in our analysis. A partial concept is a function $f : Z \rightarrow \{0, 1, \perp\}$ that assigns each individual z in an input space Z a binary label (0 or 1), or the “unknown” label, denoted by $\perp$. A *total* concept $f : Z \rightarrow \{0, 1\}$ assigns a binary label (0 or 1) to each individual $z \in Z$ without using the “unknown” label $\perp$. Throughout the paper, for a positive integer n , we use $[n]$ to denote the set $\{1, \dots, n\}$.

Definition 2.1 (VC dimension of partial concept classes [Vapnik and Chervonenkis, 1971, Alon et al., 2022]). *Let $\mathcal{F}$ be a class of partial concepts $f : Z \rightarrow \{0, 1, \perp\}$ on an arbitrary input space Z . We say a subset $S \subseteq Z$ is shattered by $\mathcal{F}$ if for every subset $E \subseteq S$, there exists $f_E \in \mathcal{F}$ such that*

$$\begin{aligned} f_E(s) &= 0 \quad \text{for every } s \in E, \\ f_E(s) &= 1 \quad \text{for every } s \in S \setminus E. \end{aligned}$$

The VC dimension of $\mathcal{F}$ (denoted $\text{VCdim}(\mathcal{F})$) is the size of the largest shattered set $S \subseteq Z$.

Clearly, the classical VC dimension, defined for total concept classes, is a special case of the VC dimension for partial concept classes.

3 Problem Setup for Domain Generalization

Given input space X and label space $Y = \{0, 1\}$, a domain (or data distribution) $\mathcal{D}$ is a distribution over $X \times Y$. We consider a family $\mathcal{G}$ of domains $\mathcal{D}$. For any hypothesis/predictor $h : X \rightarrow Y$, we define the error rate of h under domain $\mathcal{D}$ as

$$\text{err}_{\mathcal{D}}(h) := \mathbb{P}_{(x,y) \sim \mathcal{D}} (y \neq h(x)) .$$

We consider an underlying meta distribution $\mathcal{P}$ over the domains $\mathcal{D} \in \mathcal{G}$. Given a threshold τ , for any hypothesis h and a threshold τ , we define the domain error of h with respect to τ as

$$\text{Er}_{\mathcal{P},\tau}(h) := \mathbb{P}_{\mathcal{D} \sim \mathcal{P}} (\text{err}_{\mathcal{D}}(h) > \tau) ,$$

which quantifies the probability mass of the domains where h incurs error greater than τ . It follows immediately that $\text{Er}_{\mathcal{P},\tau}(h)$ is monotonically decreasing in τ .

As in the standard PAC learning setting, we are given a hypothesis class $\mathcal{H}$ and aim to output a hypothesis that performs well compared to the best hypothesis in $\mathcal{H}$.

Definition 3.1 (Optimal domain error bound and optimal hypothesis). *Given a hypothesis class $\mathcal{H}$, domain class $\mathcal{G}$, and a distribution $\mathcal{P}$ over $\mathcal{G}$, the optimal domain error bound is defined as*

$$\tau_{\mathcal{P},\mathcal{H}}^* = \min\{\tau | \exists h \in \mathcal{H}, \text{Er}_{\mathcal{P},\tau}(h) = 0\} ,$$

and the optimal hypothesis h^ is defined as one hypothesis h achieving $\text{Er}_{\mathcal{P},\tau_{\mathcal{P},\mathcal{H}}^*}(h) = 0$.*

This definition ensures that there exists a hypothesis h^* in $\mathcal{H}$ that achieves an error rate below threshold $\tau_{\mathcal{P},\mathcal{H}}^*$ on every domain sampled from $\mathcal{P}$, and $\tau_{\mathcal{P},\mathcal{H}}^*$ represents the smallest such achievable threshold. Thus, $\tau_{\mathcal{P},\mathcal{H}}^*$ serves as a benchmark for our learning problem and any threshold $\tau \geq \tau_{\mathcal{P},\mathcal{H}}^*$ is achievable. We assume that $\tau_{\mathcal{P},\mathcal{H}}^*$ is reasonably small—if $\tau_{\mathcal{P},\mathcal{H}}^* = 0.5$, then no hypothesis in $\mathcal{H}$ can perform well across the domains—though we do not assume it is zero.

Then given a hypothesis class $\mathcal{H}$ and a threshold τ , a learner $\mathcal{A}$ access to a set of i.i.d. sampled domains $G = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n\}$, and i.i.d. data S_i collected from each domain $\mathcal{D}_i$, with the goal of outputting a hypothesis $h := \mathcal{A}(\{S_1, \dots, S_n\})$ that achieves error rate of at most τ under almost every domain. This can be formalized into the following learnability problem for domain generalization, where we focus on the domain sample complexity: the number n of observed domains needed to achieve domain generalization.

Definition 3.2 ($(\tau, \alpha, \gamma, \delta)$ -domain learnability). *For any $\tau, \alpha, \gamma, \delta \in (0, 1)$, we say $(\mathcal{H}, \mathcal{G})$ is $(\tau, \alpha, \gamma, \delta)$ -learnable if there exists finite integers n and m for which there exists an algorithm $\mathcal{A}$ such that for any distribution $\mathcal{P}$ over $\mathcal{G}$ with optimal error bound $\tau_{\mathcal{P},\mathcal{H}}^* \leq \tau - \alpha$, with probability at least $1 - \delta$ over $G \sim \mathcal{P}^n$ and $S_i \sim \mathcal{D}_i^m$ for all $\mathcal{D}_i \in G$,*

$$\text{Er}_{\mathcal{P},\tau}(\mathcal{A}(\{S_1, \dots, S_n\})) \leq \gamma .$$

The domain sample complexity is the smallest integer n satisfying the above constraint.

4 Characterizing Domain Sample Complexity Using Domain Shattering Dimension

To characterize the domain sample complexity, we introduce the following combinatorial measure inspired by the fat-shattering dimension [Kearns and Schapire, 1994].

Definition 4.1 (Domain shattering dimension). *Let $\mathcal{G}$ be a set of domains. We say a subset $S \subseteq \mathcal{G}$ is α -shattered by $\mathcal{H}$ at τ if for all $E \subseteq S$, there exists a hypothesis $h_E \in \mathcal{H}$ satisfying*

$$\begin{aligned} \text{err}_{\mathcal{D}}(h_E) &< \tau - \alpha && \text{for every } \mathcal{D} \in E, \\ \text{err}_{\mathcal{D}}(h_E) &> \tau && \text{for every } \mathcal{D} \in S \setminus E. \end{aligned}$$

The domain shattering dimension of $\mathcal{H}$, denoted by $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$, is defined to be the maximum size of set S that can be α -shattered by $\mathcal{H}$ at τ .

Monotonicity of domain shattering dimension. It follows directly from the definition that for any $\tau, \alpha, \tau', \alpha'$, if $\tau' \geq \tau$ and $\tau' - \alpha' \leq \tau - \alpha$, we have $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau', \alpha') \leq \text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$.

4.1 Upper Bound

We show an upper bound on the domain sample complexity using the domain shattering dimension in Theorem 4.1 below. We prove this upper bound using the following natural min-max variant of the empirical risk minimization (ERM) algorithm. In the next subsection (Section 4.2), we will prove a lower bound that matches our upper bound up to a polylogarithmic factor.

Min-Max ERM Algorithm. Given a set of i.i.d. sampled domains $G = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n\}$, we assume access to approximate error rates for any hypothesis $h \in \mathcal{H}$ on each domain. Specifically, for some $\varepsilon > 0$, for every domain $\mathcal{D} \in G$ and hypothesis $h \in \mathcal{H}$, we can access an estimate $\widehat{\text{err}}_{\mathcal{D}}(h)$ of the true error $\text{err}_{\mathcal{D}}(h)$ such that

$$|\widehat{\text{err}}_{\mathcal{D}}(h) - \text{err}_{\mathcal{D}}(h)| < \varepsilon \quad \text{for all } \mathcal{D} \in G \text{ and } h \in \mathcal{H}. \quad (2)$$

By standard uniform convergence guarantees, with success probability at least $1 - \delta$, these estimates can be obtained as the empirical error on $O((\text{VCdim}(\mathcal{H}) + \log(n/\delta))/\varepsilon^2)$ i.i.d. data points from each domain's distribution. Given access to $\widehat{\text{err}}_{\mathcal{D}}(h)$, we return the min-max predictor

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \max_{\mathcal{D} \in G} \widehat{\text{err}}_{\mathcal{D}}(h). \quad (3)$$

The following theorem upper bounds $\text{Er}_{\mathcal{P}, \tau}(\hat{h})$ in terms of the domain shattering dimension and the number of training domains. This implies a domain sample complexity upper bound.

Theorem 4.1. *Let $\mathcal{H}$ be a class of hypotheses $h : X \rightarrow \{0, 1\}$ and let $\mathcal{G}$ be a family of domains $\mathcal{D}$ each being a distribution over $X \times \{0, 1\}$. Define $d := \text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$ for $\tau, \alpha \in [0, 1]$. For every $\varepsilon, \delta \in (0, 1/2)$, for every domain distribution $\mathcal{P}$ over $\mathcal{G}$ satisfying $\tau_{\mathcal{P}, \mathcal{H}}^* \leq \tau - \alpha - 2\varepsilon$, with probability at least $1 - \delta$ over a sample G of n i.i.d. domains drawn from $\mathcal{P}$, when given access to $\widehat{\text{err}}_{\mathcal{D}}(h)$ for all $\mathcal{D} \in G$ and $h \in \mathcal{H}$ satisfying (2), the min-max predictor $\hat{h}$ in (3) satisfies*

$$\text{Er}_{\mathcal{P}, \tau}(\hat{h}) \leq O\left(\frac{d \log^2 n + \log(1/\delta)}{n}\right).$$

Theorem 4.1 guarantees good accuracy of $\hat{h}$ on new, unseen domains drawn from $\mathcal{P}$. To prove this result, we start by analyzing the performance of $\hat{h}$ on the training domains in G . By the assumption of Theorem 4.1, there exists $h^* \in \mathcal{H}$ such that $\text{err}_{\mathcal{D}}(h^*) \leq \tau - \alpha - 2\varepsilon$ for every $\mathcal{D} \in G$. Now by (2) and (3), we have

$$\text{err}_{\mathcal{D}}(\hat{h}) < \widehat{\text{err}}_{\mathcal{D}}(\hat{h}) + \varepsilon \leq \widehat{\text{err}}_{\mathcal{D}}(h^*) + \varepsilon < \text{err}_{\mathcal{D}}(h^*) + 2\varepsilon \leq \tau - \alpha \quad \text{for every } \mathcal{D} \in G. \quad (4)$$

This ensures that $\hat{h}$ achieves low error on every training domain $\mathcal{D} \in G$. To prove Theorem 4.1, we need to show that $\hat{h}$ achieves low error on new domains drawn from $\mathcal{P}$. Specifically, for some

$$\gamma = O\left(\frac{d \log^2 n + \log(1/\delta)}{n}\right),$$

our goal is to show that with probability at least $1 - \delta$, the $\hat{h}$ returned by min-max ERM satisfies

$$\Pr_{\mathcal{D} \sim \mathcal{P}} [\text{err}_{\mathcal{D}}(\hat{h}) > \tau] \leq \gamma. \quad (5)$$

To prove the guarantee (5) for $\mathcal{D} \sim \mathcal{P}$ from the guarantee (4) for $\mathcal{D} \in G$, we establish a uniform convergence bound over all $h \in \mathcal{H}$.

One natural idea for establishing a desired uniform convergence bound is by applying existing results about the fat-shattering dimension Kearns and Schapire [1994], which is similar to our domain shattering dimension. However, this idea falls short for our purpose because 1) the fat-shattering dimension is defined as a maximum over *all* thresholds τ , and 2) prior results from the fat-shattering dimension incur a constant-factor blow-up in the margin α . We instead use results about partial concept classes from Alon et al. [2022]. Specifically, for each $h \in \mathcal{H}$, we construct a partial concept $f_h : \mathcal{Z} \mapsto \{0, 1, \perp\}$ by letting

$$f_h(\mathcal{D}) = \begin{cases} 1 & \text{if } \text{err}_{\mathcal{D}}(h) > \tau, \\ 0 & \text{if } \text{err}_{\mathcal{D}}(h) < \tau - \alpha, \\ \perp & \text{o.w.} \end{cases}$$

This allows us to construct a new partial concept class $\mathcal{F} = \{f_h | h \in \mathcal{H}\}$. Our assumption in Theorem 4.1 that the domain shattering dimension of $\mathcal{H}$ is d ensures that the VC dimension of $\mathcal{F}$ is d (see Definitions 2.1 and 4.1). Now (4) can be equivalently written as

$$f_{\hat{h}}(\mathcal{D}) = 0 \quad \text{for every } \mathcal{D} \in G,$$

and similarly, our goal (5) is equivalent to

$$\Pr_{\mathcal{D} \sim \mathcal{P}} [f_{\hat{h}}(\mathcal{D}) = 1] \leq \gamma.$$

Thus, to prove Theorem 4.1, it suffices to establish the following uniform convergence bound:

$$\Pr_{G \sim \mathcal{P}^n} [\exists f \in \mathcal{F} \text{ s.t. } \Pr_{\mathcal{D} \sim \mathcal{P}} [f(\mathcal{D}) = 1] > \gamma \text{ and } \forall \mathcal{D} \in G, f(\mathcal{D}) = 0] \leq \delta.$$

We formally state this uniform convergence bound in the following general lemma:⁴

Lemma 4.2 (Uniform convergence for partial concepts). *Let $\mathcal{F}$ be a class of partial concepts $f : Z \rightarrow \{0, 1, \perp\}$ on an arbitrary input space Z . Assume that $\mathcal{F}$ has VC dimension d . For every $n \in \mathbf{Z}_{>0}$ and $\delta \in (0, 1/2)$, there exists*

$$\gamma = O\left(\frac{d \log^2 n + \log(1/\delta)}{n}\right) \quad (6)$$

such that for every distribution $\mathcal{P}$ over Z , for n i.i.d. data points $z_1, \dots, z_n$ drawn from $\mathcal{P}$,

$$\Pr_{z_1, \dots, z_n} [\exists f \in \mathcal{F} \text{ s.t. } \Pr_{z \sim \mathcal{P}} [f(z) = 1] > \gamma \text{ and } \forall i \in [n], f(z_i) = 0] \leq \delta. \quad (7)$$

We establish this uniform convergence bound using a standard symmetrization trick combined with a generalized Sauer-Shelah-Perles lemma for partial concept classes (Theorem B.1) by Alon et al. [2022]. We defer the full proof of Lemma 4.2 to Appendix B.1.

Remark 4.1 (Choice of the threshold τ). *It is important to note that while our upper bound depends on the choice of τ and α , the algorithm itself does not. Let us fix the relationship $\tau_{\mathcal{P}, \mathcal{H}}^* = \tau - \alpha - 2\varepsilon$. As we increase τ and α at the same rate—i.e., relax the target error threshold τ —the value of $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$ decreases monotonically. This means that a larger proportion of domains can meet the relaxed threshold τ . Therefore, the choice of τ captures a trade-off between the strictness of the generalization goal and the fraction of domains that are able to satisfy it.*

Remark 4.2 (Beyond binary classification). *We emphasize that neither the algorithm nor the analysis relies on binary labels, and both can be directly applied to other learning tasks such as multi-class classification and regression.*

⁴We remark that a key message from the work of Alon et al. [2022] is that uniform convergence and the ERM algorithm both fail for learning partial concepts. This, however, does not contradict our Lemma 4.2. Roughly speaking, the uniform convergence needed in the setting of Alon et al. [2022] requires replacing $\Pr_{z \sim \mathcal{P}} [f(z) = 1]$ in (7) with $\Pr_{z \sim \mathcal{P}} [f(z) = 1 \text{ or } \perp]$. This stronger form of uniform convergence does not hold.

Remark 4.3 (Standard ERM fails). *Note that standard ERM, which selects a hypothesis minimizing the empirical error over the entire pool of training data, may fail in our setting. Consider a toy example with two hypotheses: one incurs an error of 0.3 on every domain, while the other achieves 0 error on half of the domains and 0.5 on the other half. Standard ERM would choose the latter, despite its poor worst-case performance across domains.*

4.2 Lower Bound

We prove a lower bound (Theorem 4.3) showing that the error upper bound in Theorem 4.1 is essentially information-theoretically optimal up to an $O(\log^2 n)$ factor. It would be ideal to show that the error bound is *instance-wise* optimal: for every fixed choice of $\mathcal{H}, \mathcal{G}$ and parameters $\tau, \alpha, \varepsilon, \delta$, the error bound in Theorem 4.1 cannot be improved even if one uses a learning algorithm specifically designed for those fixed choices. However, this perfect instance-wise optimality does not hold for the following trivial reason. For every domain $\mathcal{D} \in \mathcal{G}$, let $\mathcal{D}_X$ denote the marginal distribution of $x \in X$ where $(x, y) \sim \mathcal{D}$. Consider a fixed choice of $\mathcal{G}$ and suppose that the marginal distributions $\mathcal{D}_X$ for $\mathcal{D} \in \mathcal{G}$ are supported on disjoint subsets of X . In this case, the learner can simply always output the hypothesis h^* that simultaneously minimizes the error on every domain $\mathcal{D} \in \mathcal{G}$. This solves our domain generalization task without observing any training domains drawn from $\mathcal{P}$. The domain sample complexity is zero, but the domain shattering dimension can be very large, implying that our upper bound Theorem 4.1 is far from optimal in this setting.

We thus make a slight compromise in the level of instance-wise optimality. We still consider arbitrary fixed choices of $\mathcal{H}$ and $\mathcal{G}$. However, in our learning task, the domains do not solely come from $\mathcal{G}$, but instead come from a mildly extended family $\mathcal{G}'$. Importantly, for some domains $\mathcal{D}$ in the extended family $\mathcal{G}'$, their marginal distributions $\mathcal{D}_X$ will be supported on overlapping subsets of X . In this case, we are able to prove a domain sample complexity lower bound (Theorem 4.3) that nearly matches our upper bound (Theorem 4.1).

Concretely, we assume that there exists a distribution $\mathcal{D}_0$ of data points $(x, y) \in X \times \{0, 1\}$ such that $\text{err}_{\mathcal{D}_0}(h) = 0$ for every $h \in \mathcal{H}$.⁵ Let d be the domain shattering dimension of $\mathcal{H}$ on $\mathcal{G}$, and let $\mathcal{D}_1, \dots, \mathcal{D}_d \in \mathcal{G}$ be d domains shattered by $\mathcal{H}$. For every $i = 1, \dots, d$, we define a new domain $\mathcal{D}'_i$ as follows:

$$\mathcal{D}'_i = (1 - \lambda)\mathcal{D}_0 + \lambda(\neg\mathcal{D}_i). \quad (8)$$

Here, $\lambda \in [0, 1]$ is a fixed parameter, and $\neg\mathcal{D}_i$ is obtained by flipping the labels in $\mathcal{D}_i$. That is, $\neg\mathcal{D}_i$ is the distribution of $(x, \neg y)$ for $(x, y) \sim \mathcal{D}_i$. Equation (8) defines $\mathcal{D}'_i$ as a mixture of $\mathcal{D}_0$ and $\neg\mathcal{D}_i$.

We define $\mathcal{G}'$ to be $\mathcal{G} \cup \{\mathcal{D}_0, \mathcal{D}'_1, \dots, \mathcal{D}'_d\}$. We view $\mathcal{G}'$ as a mild extension of $\mathcal{G}$. In particular, as we show in Theorem 4.3 below, $\mathcal{H}$ has the same domain shattering dimension on $\mathcal{G}'$ as on $\mathcal{G}$ for an appropriate choice of λ . Moreover, in Theorem 4.3 we show a domain sample complexity lower bound on $\mathcal{G}'$ that nearly matches our upper bound in Theorem 4.1.

Theorem 4.3 (Lower bound). *Consider an arbitrary class $\mathcal{H}$ of hypotheses $h : X \rightarrow \{0, 1\}$ and an arbitrary family $\mathcal{G}$ of domains $\mathcal{D}$ (i.e. distributions over $X \times \{0, 1\}$). Assume that $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha) = d$ for some $\tau, \alpha \in \mathbb{R}$ and $d \in \mathbb{Z}_{>0}$, where $0 \leq \alpha < \tau \leq 1/2$. Let $\mathcal{D}_0$ be a distribution of $(x, y) \in X \times \{0, 1\}$ satisfying $\text{err}_{\mathcal{D}_0}(h) = 0$ for every $h \in \mathcal{H}$. Let $\mathcal{D}_1, \dots, \mathcal{D}_d \in \mathcal{G}$ be d domains shattered by $\mathcal{H}$, and define $\mathcal{G}' = \mathcal{G} \cup \{\mathcal{D}_0, \mathcal{D}'_1, \dots, \mathcal{D}'_d\}$ as in (8) for*

$$\lambda = \frac{\tau - \alpha}{1 - \tau} \in (0, 1]. \quad (9)$$

Then $\text{Gdim}(\mathcal{H}, \mathcal{G}', \tau, \alpha) = d$. Moreover, for some $\gamma > 0, \delta \in (0, 1/4), n \in \mathbb{Z}_{>0}$, and an error threshold $\tau' < \tau - \frac{1-\tau}{1-\alpha} \cdot \alpha \in (\tau - \alpha, \tau]$, suppose there is an algorithm A with the following property on every distribution $\mathcal{P}$ over $\mathcal{G}'$ satisfying $\tau_{\mathcal{P}, \mathcal{H}}^ < \tau - \alpha$: it takes n domains drawn i.i.d. from $\mathcal{P}$ as input, and with probability at least $1 - \delta$, outputs a hypothesis $\hat{h}$ such that*

$$\text{Er}_{\mathcal{P}, \tau'}(\hat{h}) \leq \gamma.$$

Then

$$\gamma = \Omega\left(\min\left\{1, \frac{d + \log(1/\delta)}{n}\right\}\right). \quad (10)$$

⁵We view this as a mild assumption. Note that $\mathcal{D}_0$ does not need to be a member of $\mathcal{G}$. This assumption is satisfied as long as there is an input point $x_0 \in X$ receiving the same label $h(x_0) = b \in \{0, 1\}$ for all $h \in \mathcal{H}$, in which case we choose $\mathcal{D}_0$ as the degenerate distribution supported on the single point (x_0, b) .

We remark that the lower bound in Theorem 4.3 is for a lower (i.e., more challenging) error threshold τ' instead of the original error threshold τ in the definition of the domain shattering dimension. Nonetheless, the lower threshold τ' is always allowed to be above the optimal error $\tau_{\mathcal{P},\mathcal{H}}^*$ (i.e. $\tau' > \tau - \alpha > \tau_{\mathcal{P},\mathcal{H}}^*$).

We defer the proof of Theorem 4.3 to Appendix B.2.

5 Relationship between Domain Shattering Dimension and VC Dimension

In this section, we study the relationship between the VC dimension and the domain shattering dimension of a hypothesis class $\mathcal{H}$. It is easy to see that the domain shattering dimension can be much smaller than the VC dimension when the domains in $\mathcal{G}$ are similar—in the extreme case where $\mathcal{G}$ contains only a single domain, the domain shattering dimension cannot be more than one. We thus focus on the other direction: can the domain shattering dimension of a hypothesis class exceed its VC dimension? How much larger can it be?

We give an accurate answer to this question in Theorems 5.1 and 5.2 below: for an error margin α (see Definition 4.1), the domain shattering dimension can be as large as $\Omega(d \log(1/\alpha))$, which is also the largest possible (up to a constant factor). Therefore, the domain shattering dimension can be arbitrarily larger than the VC dimension as α approaches zero, but for any fixed $\alpha > 0$, the domain shattering dimension is upper bounded linearly by the VC dimension. This implies that a PAC learnable hypothesis class is also learnable for our domain generalization task in Definition 3.2.

Theorem 5.1. *For every positive integer d , there exists a hypothesis class $\mathcal{H}$ with $\text{VCdim}(\mathcal{H}) = d$ satisfying the following property. For any $\alpha \in (0, 1/12)$, there exist $k = \Omega(d \log(1/\alpha))$ domains $\mathcal{D}_1, \dots, \mathcal{D}_k$ such that*

$$\text{Gdim}(\mathcal{H}, \{\mathcal{D}_1, \dots, \mathcal{D}_k\}, 0.3, \alpha) = k.$$

Theorem 5.2. *Let $\mathcal{H}$ be an arbitrary hypothesis class with VC dimension d . For any set $\mathcal{G}$ of domains, any threshold $\tau \in \mathbb{R}$, and any margin $\alpha \in (0, 1/2)$,*

$$\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha) = O(d \log(1/\alpha)). \quad (11)$$

We defer the proofs of these two theorems to Appendices B.3 and B.4. Our proof of Theorem 5.2 uses a dimension reduction argument combined with the standard Sauer-Shelah-Perles lemma, which is inspired by the proof of a classic covering number upper bound in terms of the VC dimension [see e.g. Vershynin, 2018, Theorem 8.3.18].

6 Connection to Domain Adaptation

In the domain adaptation literature, a common setting assumes that the training data are drawn from a source distribution, while the test data come from a different but related target distribution. When the source and target distributions are sufficiently similar, a model trained on the source data can generalize well to the target distribution. To quantify this similarity, various notions have been proposed, such as the $\mathcal{H}$ -divergence [Ben-David et al., 2010] and the propensity scoring function class [Kim et al., 2022].

Our work makes no assumptions about the inter-domain structure, except that the optimal domain error bound $\tau_{\mathcal{P},\mathcal{H}}^*$ is reasonably small. We then upper bound the domain generalization error using the domain shattering dimension. In this section, we show that when the covering number of the domain space $\mathcal{G}$ w.r.t. the $\mathcal{H}$ -divergence is small, the domain shattering dimension is also small. In fact, we show a stronger result (Theorem 6.1) for a refined variant of the $\mathcal{H}$ -divergence. This illustrates that our domain shattering dimension can capture the similarity between domains without requiring an explicit model of their relationships.

Inspired by the $\mathcal{H}$ -divergence introduced by Ben-David et al. [2010], where

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') := \sup_{h \in \mathcal{H}} |\text{err}_{\mathcal{D}}(h) - \text{err}_{\mathcal{D}'}(h)|,$$

we define a refined notion of divergence. Given $\mathcal{H}, \tau$, the $(\mathcal{H}, \tau)$ -divergence between two domains $\mathcal{D}, \mathcal{D}'$ is

$$d_{\mathcal{H},\tau}(\mathcal{D}, \mathcal{D}') := \sup_{h \in \mathcal{H}: \min\{\text{err}_{\mathcal{D}}(h), \text{err}_{\mathcal{D}'}(h)\} \leq \tau} |\text{err}_{\mathcal{D}}(h) - \text{err}_{\mathcal{D}'}(h)|.$$

Note that $d_{\mathcal{H},\tau}(\mathcal{D}, \mathcal{D}') \leq d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}')$ as we only take supremum over a subset of hypotheses $\{h \in \mathcal{H} : \min\{\text{err}_{\mathcal{D}}(h), \text{err}_{\mathcal{D}'}(h)\} \leq \tau\}$. Given domain space $\mathcal{G}$, we say $\mathcal{G}'$ is an α -cover of $\mathcal{G}$ w.r.t. $d_{\mathcal{H},\tau}$ if for every $\mathcal{D} \in \mathcal{G}$, there exists $\mathcal{D}' \in \mathcal{G}'$ such that $d_{\mathcal{H},\tau}(\mathcal{D}, \mathcal{D}') \leq \alpha$. Any α -cover $\mathcal{G}'$ w.r.t. $d_{\mathcal{H}}$ is also an α -cover w.r.t. $d_{\mathcal{H},\tau}$.

Theorem 6.1. *For every hypothesis class $\mathcal{H}$, domain space $\mathcal{G}$ and $\tau, \alpha \in (0, 1)$, let $\mathcal{G}'$ be an $\frac{\alpha}{2}$ -cover of $\mathcal{G}$ w.r.t. $d_{\mathcal{H},\tau}$. We have $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha) \leq |\mathcal{G}'|$.*

Proof of Theorem 6.1. For every $\mathcal{D}' \in \mathcal{G}'$, define $\mathcal{G}(\mathcal{D}') := \{\mathcal{D} \in \mathcal{G} : d_{\mathcal{H},\tau}(\mathcal{D}, \mathcal{D}') \leq \frac{\alpha}{2}\}$. Since $\mathcal{G}'$ is an $\frac{\alpha}{2}$ -cover, the union of $\mathcal{G}(\mathcal{D}')$ over $\mathcal{D}' \in \mathcal{G}'$ is $\mathcal{G}$.

Fix an arbitrary $\mathcal{D}' \in \mathcal{G}'$ and let $\mathcal{D}_1, \mathcal{D}_2$ be two domains in $\mathcal{G}(\mathcal{D}')$. We show that $\mathcal{D}_1, \mathcal{D}_2$ cannot be α -shattered at τ . In particular, we show that for every $h \in \mathcal{H}$, the two conditions $\text{err}_{\mathcal{D}_1}(h) < \tau - \alpha$ and $\text{err}_{\mathcal{D}_2}(h) > \tau$ cannot hold simultaneously. Indeed, if $\text{err}_{\mathcal{D}_1}(h) < \tau - \alpha$, then we have

$$\text{err}_{\mathcal{D}'}(h) \leq \text{err}_{\mathcal{D}_1}(h) + \frac{\alpha}{2} < \tau - \frac{\alpha}{2},$$

and consequently,

$$\text{err}_{\mathcal{D}_2}(h) \leq \text{err}_{\mathcal{D}'}(h) + \frac{\alpha}{2} < \tau.$$

Therefore, $\text{err}_{\mathcal{D}_2}(h) > \tau$ cannot hold.

We have now shown that every subset of domains that is α -shattered at τ contains at most one domain from each $\mathcal{G}(\mathcal{D}')$ for $\mathcal{D}' \in \mathcal{G}'$. Thus the domain shattering dimension $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$ does not exceed $|\mathcal{G}'|$. $\square$

When all the domains are similar, the cover size is 1, then sampling one domain is sufficient. Below are two examples of domain spaces with cover size 1.

Example 6.1 (Small $\mathcal{H}$ -divergences). *For any domain space $\mathcal{G}$, if for every pair of distributions $\mathcal{D}, \mathcal{D}' \in \mathcal{G}$, $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') \leq \alpha$. Then the α -cover size of $\mathcal{G}$ is 1.*

Example 6.2 (Smooth Distributions). *Given a marginal data distribution μ and a labeling function p^* , we define $\mathcal{D}_{\mu,p^*}$ as the distribution over labeled data where features are drawn from μ and labels are assigned according to p^* . For a parameter $\gamma \in (0, 1)$, a labeling function p^* , and a base marginal distribution μ_0 , define the smooth domain space as*

$$\mathcal{G}_{p^*,\mu_0,\gamma} = \left\{ \mathcal{D}_{\mu,p^*} \left| \frac{\mu(x)}{\mu_0(x)} \in [\gamma, \frac{1}{\gamma}], \forall x \in \text{supp}(\mu_0) \right. \right\}.$$

For domain space $\mathcal{G}_{p^,\mu_0,\gamma}$, we can find a $\left(\frac{1}{\gamma^2} - 1\right)$ τ -cover of size 1.*

For any two distributions $\mathcal{D}_{\mu_1,p^*}, \mathcal{D}_{\mu_2,p^*} \in \mathcal{G}_{p^*,\mu_0,\gamma}$, it holds that $\frac{\mu_1(x)}{\mu_2(x)} \in [\gamma^2, \frac{1}{\gamma^2}]$ for all $x \in \text{supp}(\mu_0)$. Consequently, for any hypothesis h , we have

$$\gamma^2 \cdot \text{err}_{\mathcal{D}_{\mu_2,p^*}}(h) \leq \text{err}_{\mathcal{D}_{\mu_1,p^*}}(h) \leq \frac{1}{\gamma^2} \cdot \text{err}_{\mathcal{D}_{\mu_2,p^*}}(h).$$

Therefore, by the definition of the $(\mathcal{H}, \tau)$ -divergence, it follows that

$$d_{\mathcal{H},\tau}(\mathcal{D}_{\mu_1,p^*}, \mathcal{D}_{\mu_2,p^*}) \leq \left(\frac{1}{\gamma^2} - 1\right) \tau.$$

Hence, for domain space $\mathcal{G}_{p^*,\mu_0,\gamma}$, we can find a $\left(\frac{1}{\gamma^2} - 1\right)$ τ -cover of size 1.

7 Discussion and Limitations

This work was inspired by the very real and widely acknowledged problem of transferring research from well-funded flagship medical research institutions to essentially all communities, even those with no direct access to these facilities, "from preeminent bench to geographically remote bedside"

(Deng et al. [2020]). Our succinct characterization and algorithm close some facets of this central question, while opening others.

How to compute the shattering dimension. As discussed in Remark 4.1, the choice of τ reflects a trade-off between the strictness of the generalization goal and the fraction of domains that can satisfy it, as captured by $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$. However, choosing an appropriate τ depends on understanding the value of $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$, which—like the VC dimension—is generally difficult to compute. Developing tools or approximation techniques to estimate Gdim is an important direction for future work.

Choice of hypothesis class $\mathcal{H}$. We focus on the setting where a hypothesis class $\mathcal{H}$ is given and assumed to contain a reasonably good hypothesis that generalizes across all domains. However, in practice, one may have access to a collection of candidate classes. How should we select the best class—one that contains a reasonably good hypothesis but is not overly complex—for generalization? This question relates to structured ERM in the learning theory literature and presents an interesting open direction.

Gap between upper and lower bounds. Although we show that the domain shattering dimension provides both upper and lower bounds on the domain sample complexity, a small gap remains. Our upper bound guarantees that $O(\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha))$ sampled domains are sufficient to learn a predictor with error at most τ on most domains, while the lower bound shows that $\Omega(\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha))$ domains are necessary to achieve an error threshold $\tau' \in (\tau - \alpha, \tau)$. Additionally, the lower bound does not apply to all domain families $\mathcal{G}$; in some cases, the construction requires augmenting $\mathcal{G}$ with a few extra domains. It would be of moderate interest to close this gap.

Unlabeled data from some unseen groups. In this work, we assume no information is available from unobserved domains. However, in real-world applications such as healthcare, labeled data can be expensive—particularly in rural regions where diagnostic tools may be limited—while unlabeled data are often cheaper and more accessible. This brings us full circle to the motivating scenario in the original paper on domain generalization Blanchard et al. [2011], what is there termed automatic gating of flow cytometry data. In this problem, each patient yields a patient-specific distribution (i.e. domain) of d -vectors of attributes of cells. The label might capture whether or not the given cell is a lymphocyte. Our work strengthens the results of Blanchard et al. [2011] by obtaining generalization to all unseen patients, and not just in expectation over unseen patients. This inspiring scenario, in which unlabeled data are plentiful, raises the following question: If we permit the (fixed) learned hypothesis to be modified based on unlabeled samples from the test domain, can we beat our lower bound and get away with training on *fewer* than $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha)$ domains, while maintaining the learning objective of performing well on *every* unseen domain?

References

- Maryam Aliakbarpour, Konstantina Bairaktari, Gavin Brown, Adam Smith, Nathan Srebro, and Jonathan Ullman. Metalearning with very few samples per task. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 46–93. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/aliakbarpour24a.html>.
- Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671. IEEE, 2022.
- Yannay Alon, Steve Hanneke, Shay Moran, and Uri Shalit. On the ERM principle in meta-learning. *arXiv preprint arXiv:2411.17898*, 2024.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- Pranjal Awasthi, Nika Haghtalab, and Eric Zhao. Open problem: The sample complexity of multi-distribution learning for VC classes. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5943–5949. PMLR, 12–15 Jul 2023. URL <https://proceedings.mlr.press/v195/awasthi23a.html>.

- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.
- Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.
- Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative PAC learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/186a157b2992e7daed3677ce8e9fe40f-Paper.pdf.
- Jiecao Chen, Qin Zhang, and Yuan Zhou. Tight bounds for collaborative PAC learning via multiplicative weights. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/ed519dacc89b2bead3f453b0b05a4a8b-Paper.pdf.
- Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2089–2099. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/collins21a.html>.
- Zhun Deng, Frances Ding, Cynthia Dwork, Rachel Hong, Giovanni Parmigiani, Prasad Patil, and Pragma Sur. Representation via representations: Domain generalization via adversarially learned invariant representations. *arXiv preprint arXiv:2006.11478*, 2020.
- Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=pW2Q2xLwIMD>.
- John C Duchi, Vitaly Feldman, Lunjia Hu, and Kunal Talwar. Subspace recovery from heterogeneous data with non-isotropic noise. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5854–5866. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/26ded5c8ee8ec1bc4caced4e1c9b1584-Paper-Conference.pdf.
- Vikas Garg, Adam Tauman Kalai, Katrina Ligett, and Steven Wu. Learn to expect the unexpected: Probably approximately correct domain generalization. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3574–3582. PMLR, 13–15 Apr 2021. URL <https://proceedings.mlr.press/v130/garg21a.html>.
- Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.
- Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 406–419. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/02917acec264a52a729b99d9bc857909-Paper-Conference.pdf.
- Lunjia Hu and Charlotte Peale. Comparative Learning: A Sample Complexity Theory for Two Hypothesis Classes. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 72:1–72:30, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.72. URL <https://drops.dagstuhl.de/opus/volltexte/2023/17575>.

- Lunjia Hu, Charlotte Peale, and Omer Reingold. Metric entropy duality and the sample complexity of outcome indistinguishability. In Sanjoy Dasgupta and Nika Haghtalab, editors, *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 515–552. PMLR, 29 Mar–01 Apr 2022. URL <https://proceedings.mlr.press/v167/hu22a.html>.
- Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.
- Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.
- Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5394–5404. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/kong20a.html>.
- Kasper Green Larsen, Omar Montasser, and Nikita Zhivotovskiy. Derandomizing multi-distribution learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 94246–94264. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ab63d1eb181e920273504411fe0942dc-Paper-Conference.pdf.
- David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/madras18a.html>.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016. URL <http://jmlr.org/papers/v17/15-242.html>.
- Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4615–4625. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/mohri19a.html>.
- Omar Montasser, Han Shao, and Emmanuel Abbe. Transformation-invariant learning and theoretical guarantees for OOD generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=u2gzfXRLaN>.
- Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.
- Huy Nguyen and Lydia Zakynthinou. Improved algorithms for collaborative PAC learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3569df159ec477451530c4455b2a9e86-Paper.pdf.
- Binghui Peng. The sample complexity of multi-distribution learning. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4185–4204. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/peng24b.html>.
- N Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972. ISSN 0097-3165. doi: [https://doi.org/10.1016/0097-3165\(72\)90019-2](https://doi.org/10.1016/0097-3165(72)90019-2). URL <https://www.sciencedirect.com/science/article/pii/0097316572900192>.

- Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.
- Han Shao, Omar Montasser, and Avrim Blum. A theory of PAC learnability under transformation invariances. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13989–14001. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5a829e299ebc1c1615ddb09e98fb6ce8-Paper-Conference.pdf.
- Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.
- Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Sample efficient linear meta-learning by alternating minimization. *arXiv preprint arXiv:2105.08306*, 2021.
- Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7852–7862. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/59587bffec1c7846f3e34230141556ae-Paper.pdf.
- Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10434–10443. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/tripuraneni21a.html>.
- L. G. Valiant. A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC ’84, page 436–445, New York, NY, USA, 1984. Association for Computing Machinery. ISBN 0897911334. doi: 10.1145/800057.808710. URL <https://doi.org/10.1145/800057.808710>.
- V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL <https://doi.org/10.1137/1116025>.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.
- Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- Zihan Zhang, Wenhao Zhan, Yuxin Chen, Simon S Du, and Jason D Lee. Optimal multi-distribution learning. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 5220–5223. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/zhang24b.html>.
- Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We have included references to all main theorems and lemmas in Section 1.1.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: See discussion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We provided all the assumptions and complete proofs (some are in the appendix).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper does not include experiments. All results are carefully proven and do not require experiments to establish correctness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This is a theoretical paper. All results are carefully proven and do not require experiments to establish correctness.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This is a theoretical paper. All results are carefully proven and do not require experiments to establish correctness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This is a theoretical paper. All results are carefully proven and do not require experiments to establish correctness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: This is a theoretical paper. All results are carefully proven and do not require experiments to establish correctness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have carefully read and comply with the NeurIPS code of ethics. This work is theoretical and we do not foresee harmful consequences.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: We note that our work is theoretical. We do not foresee any societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The work is entirely theoretical; we do not release new data or models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We do not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only used LLMs for editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Deferred Related Work

Many previous theoretical works on domain generalization make explicit assumptions on how domains are related. Blanchard et al. [2011] formulated the problem and approached it using kernel methods. They and Muandet et al. [2013] differ from our work in assuming similarity among the various domains (possibly after a learned transformation); moreover, they achieve low error only in expectation over the choice of test domain. More recently, Mohri et al. [2019] assumes that the learner has access to a family of training domains, and the test domains are a family of mixtures (i.e., convex combinations) of the training domains. The coefficients of the convex combinations are assumed to be in a restricted range. The works of Shao et al. [2022], Montasser et al. [2024] assume that the domains are related by transformations. These works [Mohri et al., 2019, Shao et al., 2022, Montasser et al., 2024] all use a min-max objective similar to our work. Ben-David et al. [2010] studies learning with a source (training) domain and a target (test) domain with the assumption that the each hypothesis $h \in \mathcal{H}$ has similar overall average on the two domains (they define $\mathcal{H}$ -divergence as the largest gap between the two averages). Kim et al. [2022] assume that the ratios of the probability density between domains (i.e. propensity scores) are bounded and come from a restricted family. Schölkopf et al. [2012], Zhang et al. [2015], Gong et al. [2016] formulate the problem from a causal perspective and make corresponding causal assumptions.

Inspired by causal reasoning, and seeking to eliminate spurious correlations, Arjovsky et al. [2019] introduced *Invariant Risk Minimization*, in which the different domains correspond to training data collected in different environments which “could represent different measuring circumstances locations, times, experimental conditions” and so on. Invariant Risk Minimization seeks a representation mapping for instances, under which the optimal classifier is identical for all environments. Deng et al. [2020], whose setting and assumptions exactly match ours, also seeks a representation approach, specifically, via adversarial learning using techniques developed in Madras et al. [2018]; their algorithmic results have provable guarantees for unseen domains only when the observed domains form a cover with respect to $\mathcal{H}$ -divergence for the set of possible domains.

The work of Garg et al. [2021] studies domain generalization with the goal of minimizing the average error across domains, which is equivalent to the error on the single large domain defined as the mixture of all the domains in the problem. This makes their learning task similar to the standard PAC learning with the additional advantage that the training data is grouped by domain. This advantage allows them to show computationally efficient learning algorithms for specific learning problems that do not have known efficient algorithms in the standard PAC learning setting.

Alon et al. [2024] study a related meta-learning problem, where instead of outputting a single hypothesis h , their learner outputs a hypothesis class which can be used later to learn a good hypothesis on a new task (analogous to a new domain) given additional data from that task. They focus on the realizable setting with performance measured using the average accuracy over random new tasks. A large body of work studies meta-learning for linear regression and related problems under the assumption that the underlying parameters of different tasks can be represented in a shared low-dimensional subspace [Maurer et al., 2016, Kong et al., 2020, Tripuraneni et al., 2020, 2021, Du et al., 2021, Collins et al., 2021, Thekumparampil et al., 2021, Duchi et al., 2022, Aliakbarpour et al., 2024].

There is a large body of recent and earlier work on multi-distribution learning [Larsen et al., 2024, Zhang et al., 2024, Peng, 2024, Awasthi et al., 2023, Haghtalab et al., 2022, Chen et al., 2018, Nguyen and Zakyntinou, 2018, Blum et al., 2017]. This problem is closely related to ours in that they also do not impose structural assumptions on the domains, and they also consider the min-max objective. However, their goal is to learn a hypothesis that performs well on the *observed* domains, so their focus lies in the total number of data points required for learning (i.e., the sample or query complexity). In contrast, we aim to generalize to *unobserved* domains and focus on the number of *domains* the learner needs to observe—the domain sample complexity.

B Deferred Proofs

B.1 Proof of Lemma 4.2

Lemma 4.2 (Uniform convergence for partial concepts). *Let $\mathcal{F}$ be a class of partial concepts $f : Z \rightarrow \{0, 1, \perp\}$ on an arbitrary input space Z . Assume that $\mathcal{F}$ has VC dimension d . For every*

$n \in \mathbf{Z}_{>0}$ and $\delta \in (0, 1/2)$, there exists

$$\gamma = O\left(\frac{d \log^2 n + \log(1/\delta)}{n}\right) \quad (6)$$

such that for every distribution $\mathcal{P}$ over Z , for n i.i.d. data points $z_1, \dots, z_n$ drawn from $\mathcal{P}$,

$$\Pr_{z_1, \dots, z_n} [\exists f \in \mathcal{F} \text{ s.t. } \Pr_{z \sim \mathcal{P}}[f(z) = 1] > \gamma \text{ and } \forall i \in [n], f(z_i) = 0] \leq \delta. \quad (7)$$

We will use the following result.

Theorem B.1 (Quasipolynomial Sauer-Shelah-Perles Lemma for Disambiguations of Partial Concepts [Alon et al., 2022]). *Let $\mathcal{F}$ be a class of partial concepts $f : Z \rightarrow \{0, 1, \perp\}$ defined on an arbitrary domain Z with VC dimension d . Let S be a subset of Z with size $|S| = n > 1$. Then there exists a class $\overline{\mathcal{F}}$ of total concepts $\overline{f} : S \rightarrow \{0, 1\}$ with size*

$$|\overline{\mathcal{F}}| = n^{O(d \log n)}$$

that satisfies the following. For every $f \in \mathcal{F}$, there exists $\overline{f} \in \overline{\mathcal{F}}$ such that for every $s \in S$ with $f(s) \in \{0, 1\}$, it holds that $\overline{f}(s) = f(s)$.

Proof of Lemma 4.2. We apply the symmetrization trick. Specifically, we independently draw another n data points $z'_1, \dots, z'_n$ i.i.d. from $\mathcal{P}$. By the multiplicative Chernoff bound, assuming $n > C/\gamma$ for a sufficiently large absolute constant $C > 0$ (which can be guaranteed by the choice of γ in (6)), for a fixed $f \in \mathcal{F}$ satisfying $\Pr_{z \sim \mathcal{P}}[f(z) = 1] > \gamma$, with probability at least $1/2$ we have

$$|\{i \in [n] : f(z'_i) = 1\}| > \gamma n/2.$$

Therefore, to show (7), it suffices to show that

$$\Pr_{z_1, \dots, z_n; z'_1, \dots, z'_n} [\exists f \in \mathcal{F} \text{ s.t. } |\{i \in [n] : f(z'_i) = 1\}| > \gamma n/2 \text{ and } \forall i \in [n], f(z_i) = 0] \leq \delta/2. \quad (12)$$

The $2n$ data points $z_1, \dots, z_n, z'_1, \dots, z'_n$ can be sampled equivalently as follows. We first draw $2n$ i.i.d. data points $r_1, \dots, r_{2n}$ from $\mathcal{P}$, choose n data points from them uniformly at random without replacement as $z_1, \dots, z_n$, and define the remaining n data points as $z'_1, \dots, z'_n$. We define $\mathbf{r} := (r_1, \dots, r_{2n})$, $\mathbf{z} := (z_1, \dots, z_n)$, $\mathbf{z}' := (z'_1, \dots, z'_n)$.

For a fixed $\mathbf{r}$, by Theorem B.1, there exists a class $\overline{\mathcal{F}}$ of total concepts $\overline{f} : \{r_1, \dots, r_{2n}\} \rightarrow \{0, 1\}$ with

$$|\overline{\mathcal{F}}| \leq (2n)^{O(d \log(2n))} \quad (13)$$

satisfying the following property: for every $f \in \mathcal{F}$, there exists $\overline{f} \in \overline{\mathcal{F}}$ such that for every $i \in [2n]$ with $f(r_i) \in \{0, 1\}$, it holds that $\overline{f}(r_i) = f(r_i)$. Therefore, to show (12), it suffices to show that for every fixed choice of $\mathbf{r}$, we can bound the following conditional probability given $\mathbf{r}$:

$$\Pr_{\mathbf{z}, \mathbf{z}'} [\exists \overline{f} \in \overline{\mathcal{F}} \text{ s.t. } |\{i \in [n] : \overline{f}(z'_i) = 1\}| > \gamma n/2 \text{ and } \forall i \in [n], \overline{f}(z_i) = 0 \mid \mathbf{r}] \leq \delta/2,$$

where the probability is over the random partitioning of $\mathbf{r}$ into $\mathbf{z}$ and $\mathbf{z}'$. By the union bound, it suffices to show that for every $\overline{f} \in \overline{\mathcal{F}}$,

$$\Pr_{\mathbf{z}, \mathbf{z}'} [|\{i \in [n] : \overline{f}(z'_i) = 1\}| > \gamma n/2 \text{ and } \forall i \in [n], \overline{f}(z_i) = 0 \mid \mathbf{r}] \leq \delta/(2|\overline{\mathcal{F}}|).$$

Note that $|\{i \in [n] : \overline{f}(z'_i) = 1\}| > \gamma n/2$ implies that among the $2n$ data points $r_1, \dots, r_{2n}$, more than $\gamma/4$ fraction of the data points r_i satisfy $\overline{f}(r_i) = 1$. Conditioned on $\mathbf{r}$ satisfying this property, since $z_1, \dots, z_n$ are chosen randomly without replacement from $r_1, \dots, r_{2n}$, the probability that all the n data points $z_1, \dots, z_n$ satisfy $\overline{f}(z_i) = 0$ is at most $(1 - \gamma/4)^n$. Therefore, it suffices to prove that

$$(1 - \gamma/4)^n \leq \delta/(2|\overline{\mathcal{F}}|).$$

This holds when

$$\gamma \geq \frac{C}{n} \log(|\overline{\mathcal{F}}|/\delta)$$

for a sufficiently large absolute constant $C > 0$. By (13), the above inequality can be achieved by a choice of γ satisfying (6). $\square$

B.2 Proof of Theorem 4.3

Theorem 4.3 (Lower bound). Consider an arbitrary class $\mathcal{H}$ of hypotheses $h : X \rightarrow \{0, 1\}$ and an arbitrary family $\mathcal{G}$ of domains $\mathcal{D}$ (i.e. distributions over $X \times \{0, 1\}$). Assume that $\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha) = d$ for some $\tau, \alpha \in \mathbb{R}$ and $d \in \mathbb{Z}_{>0}$, where $0 \leq \alpha < \tau \leq 1/2$. Let $\mathcal{D}_0$ be a distribution of $(x, y) \in X \times \{0, 1\}$ satisfying $\text{err}_{\mathcal{D}_0}(h) = 0$ for every $h \in \mathcal{H}$. Let $\mathcal{D}_1, \dots, \mathcal{D}_d \in \mathcal{G}$ be d domains shattered by $\mathcal{H}$, and define $\mathcal{G}' = \mathcal{G} \cup \{\mathcal{D}_0, \mathcal{D}'_1, \dots, \mathcal{D}'_d\}$ as in (8) for

$$\lambda = \frac{\tau - \alpha}{1 - \tau} \in (0, 1]. \quad (9)$$

Then $\text{Gdim}(\mathcal{H}, \mathcal{G}', \tau, \alpha) = d$. Moreover, for some $\gamma > 0, \delta \in (0, 1/4), n \in \mathbb{Z}_{>0}$, and an error threshold $\tau' < \tau - \frac{1-\tau}{1-\alpha} \cdot \alpha \in (\tau - \alpha, \tau]$, suppose there is an algorithm A with the following property on every distribution $\mathcal{P}$ over $\mathcal{G}'$ satisfying $\tau_{\mathcal{P}, \mathcal{H}}^* < \tau - \alpha$: it takes n domains drawn i.i.d. from $\mathcal{P}$ as input, and with probability at least $1 - \delta$, outputs a hypothesis $\hat{h}$ such that

$$\text{Er}_{\mathcal{P}, \tau'}(\hat{h}) \leq \gamma.$$

Then

$$\gamma = \Omega \left(\min \left\{ 1, \frac{d + \log(1/\delta)}{n} \right\} \right). \quad (10)$$

Claim B.2. In the setting of Theorem 4.3, for every $i = 1, \dots, d$, and every hypothesis $h : X \rightarrow \{0, 1\}$, we have

$$\text{err}_{\mathcal{D}'_i}(h) \geq \lambda(1 - \text{err}_{\mathcal{D}_i}(h)), \quad (14)$$

$$\max\{\text{err}_{\mathcal{D}'_i}(h), \text{err}_{\mathcal{D}_i}(h)\} \geq \frac{\lambda}{1 + \lambda} = \tau - \frac{1 - \tau}{1 - \alpha} \cdot \alpha > \tau'. \quad (15)$$

Moreover, the inequality (14) becomes an equality when $h \in \mathcal{H}$.

Proof. By (8), for any hypothesis h , we have

$$\text{err}_{\mathcal{D}'_i}(h) = (1 - \lambda)\text{err}_{\mathcal{D}_0}(h) + \lambda(1 - \text{err}_{\mathcal{D}_i}(h)).$$

Inequality (14) follows immediately from the trivial fact that $\text{err}_{\mathcal{D}_0}(h) \geq 0$. When $h \in \mathcal{H}$, we have $\text{err}_{\mathcal{D}_0}(h) = 0$, so (14) becomes an equality.

It remains to prove (15). Suppose $\text{err}_{\mathcal{D}_i}(h) < \frac{\lambda}{1 + \lambda}$. By (14),

$$\text{err}_{\mathcal{D}'_i}(h) \geq \lambda \left(1 - \frac{\lambda}{1 + \lambda} \right) = \frac{\lambda}{1 + \lambda}.$$

This proves the first inequality in (15). The equality in (15) follows from the definition of λ in (9). The last inequality in (15) is our assumption about τ' in Theorem 4.3. $\square$

Proof of Theorem 4.3. We first prove that $\text{Gdim}(\mathcal{H}, \mathcal{G}', \tau, \alpha) = d$. By Claim B.2, for any $\mathcal{D}_i$ and the corresponding $\mathcal{D}'_i$ in (8), the errors of any hypothesis $h \in \mathcal{H}$ on $\mathcal{D}_i$ and $\mathcal{D}'_i$ are related as follows:

$$\text{err}_{\mathcal{D}_i}(h) = 1 - \text{err}_{\mathcal{D}'_i}(h)/\lambda.$$

Therefore,

$$\begin{aligned} \text{err}_{\mathcal{D}'_i}(h) < \tau - \alpha &\implies \text{err}_{\mathcal{D}_i}(h) > 1 - (\tau - \alpha)/\lambda = \tau, \\ \text{err}_{\mathcal{D}'_i}(h) > \tau &\implies \text{err}_{\mathcal{D}_i}(h) < 1 - \tau/\lambda \leq 1 - (\tau - \alpha)/\lambda - \alpha = \tau - \alpha, \end{aligned} \quad (16)$$

where we used the definition of λ in (9). Consider any set $S \subseteq \mathcal{G}'$ shattered by $\mathcal{H}$. Clearly $\mathcal{D}_0$ cannot appear in S because $\text{err}_{\mathcal{D}_0}(h) = 0$ for every $h \in \mathcal{H}$. By (16), $\mathcal{D}_i$ and $\mathcal{D}'_i$ cannot both belong to S , and replacing $\mathcal{D}'_i$ with $\mathcal{D}_i$ in S still leads to a shattered set. We can thus change S into a shattered subset of $\mathcal{G}$ without changing its size. This implies that the domain shattering dimension of $\mathcal{H}$ on $\mathcal{G}'$ does not exceed its domain shattering dimension on $\mathcal{G}$. The reverse direction is trivial, so the two domain shattering dimensions are both equal to d .

Now we prove the lower bound (10) on γ . We can assume without loss of generality that $\gamma < 1/8$, because otherwise we already have $\gamma = \Omega(1)$ as needed to establish the theorem. For every bit string $b \in \{0, 1\}^d$, we define domains $\overline{\mathcal{D}}_0, \dots, \overline{\mathcal{D}}_d$ as follows:

$$\begin{aligned}\overline{\mathcal{D}}_0 &= \mathcal{D}_0, \\ \overline{\mathcal{D}}_i &= \begin{cases} \mathcal{D}_i, & \text{if } b_i = 0, \\ \mathcal{D}'_i, & \text{if } b_i = 1, \end{cases} \quad \text{for } i = 1, \dots, d. \end{aligned} \quad (17)$$

We define $\mathcal{P}_b$ to be the distribution that puts $1 - 4\gamma$ probability mass on $\overline{\mathcal{D}}_0$, and distributes the remaining 4γ probability mass uniformly on $\overline{\mathcal{D}}_1, \dots, \overline{\mathcal{D}}_d$.

We first show that $\tau_{\mathcal{P}_b, \mathcal{H}}^* < \tau - \alpha$. Since $\mathcal{D}_1, \dots, \mathcal{D}_d$ are shattered by $\mathcal{H}$, there exists a hypothesis $h_b \in \mathcal{H}$ such that for every $i = 1, \dots, d$,

$$\begin{aligned}\text{err}_{\mathcal{D}_i}(h_b) &< \tau - \alpha, & \text{if } b_i = 0, \\ \text{err}_{\mathcal{D}_i}(h_b) &> \tau, & \text{if } b_i = 1.\end{aligned}$$

When $\text{err}_{\mathcal{D}_i}(h_b) > \tau$, by Claim B.2,

$$\text{err}_{\mathcal{D}'_i}(h_b) < \lambda(1 - \tau) = \frac{\tau - \alpha}{1 - \tau} \cdot (1 - \tau) = \tau - \alpha.$$

Therefore, $\text{err}_{\overline{\mathcal{D}}_i}(h_b) < \tau - \alpha$ for every $i = 0, 1, \dots, d$. This completes the proof of our claim that $\tau_{\mathcal{P}_b, \mathcal{H}}^* < \tau - \alpha$.

Now let us consider the following process. We first draw b uniformly at random from $\{0, 1\}^d$, and then draw n i.i.d. domains from $\mathcal{P}_b$ to form a training set G . Let $\hat{h}$ be the output of algorithm A given G as input. With probability at least $1 - \delta$, the output $\hat{h}$ satisfies

$$\Pr_{\mathcal{D} \sim \mathcal{P}_b} [\text{err}_{\mathcal{D}}(\hat{h}) > \tau'] \leq \gamma. \quad (18)$$

Let m be the number of domains among $\overline{\mathcal{D}}_1, \dots, \overline{\mathcal{D}}_d$ that do not appear in the training set S . By Claim B.2, conditioned on m , with probability at least $1/2$, at least $m/2$ of these domains $\overline{\mathcal{D}}_i$ satisfy

$$\text{err}_{\overline{\mathcal{D}}_i}(\hat{h}) > \tau'.$$

That is, with probability at least $1/2$,

$$\Pr_{\mathcal{D} \sim \mathcal{P}_b} [\text{err}_{\mathcal{D}}(\hat{h}) > \tau'] \geq (m/2) \cdot (4\gamma/d) = (2m/d) \cdot \gamma.$$

Therefore, conditioned on $m > d/2$, with probability at least $1/2$, (18) does not hold. Since (18) holds with probability at least $1 - \delta$, we have

$$\Pr[m > d/2] \leq 2\delta,$$

or equivalently,

$$\Pr[d - m < d/2] \leq 2\delta. \quad (19)$$

The nonnegative random variable $d - m$ is the number of domains among $\overline{\mathcal{D}}_1, \dots, \overline{\mathcal{D}}_d$ that *do* appear in the training set S . Since the training set S is formed by n i.i.d. examples drawn from $\mathcal{P}_b$ which puts 4γ total probability mass on $\overline{\mathcal{D}}_1, \dots, \overline{\mathcal{D}}_d$, we have $\mathbb{E}[d - m] \leq 4\gamma n$. By Markov's inequality,

$$\Pr[d - m \leq 8\gamma n] \geq 1/2. \quad (20)$$

Combining (19) and (20) with our assumption that $\delta < 1/4$, we have

$$8\gamma n \geq d/2,$$

which implies that $\gamma \geq d/(16n)$. Moreover, with probability $(1 - 4\gamma)^n$, the training set S contains no domains among $\overline{\mathcal{D}}_1, \dots, \overline{\mathcal{D}}_d$, giving $d - m = 0$. Therefore, by (19),

$$2\delta \geq \Pr[d - m < d/2] \geq \Pr[d - m = 0] = (1 - 4\gamma)^n.$$

This implies that $\gamma \geq \Omega(\log(1/\delta)/n)$. In summary,

$$\gamma \geq \max \left\{ \frac{d}{16n}, \Omega \left(\frac{\log(1/\delta)}{n} \right) \right\} = \Omega \left(\frac{d + \log(1/\delta)}{n} \right). \quad \square$$

B.3 Proof of Theorem 5.1

Theorem 5.1. *For every positive integer d , there exists a hypothesis class $\mathcal{H}$ with $\text{VCdim}(\mathcal{H}) = d$ satisfying the following property. For any $\alpha \in (0, 1/12)$, there exist $k = \Omega(d \log(1/\alpha))$ domains $\mathcal{D}_1, \dots, \mathcal{D}_k$ such that*

$$\text{Gdim}(\mathcal{H}, \{\mathcal{D}_1, \dots, \mathcal{D}_k\}, 0.3, \alpha) = k.$$

We need the following helper lemma.

Lemma B.3. *For an odd positive integer m , on domain $X = \{0, 1, \dots, m\}$, consider the hypothesis class $\mathcal{H} = \{h_1, \dots, h_m\}$ where $h_i(x) = \mathbb{I}[x \geq i]$ for every $i = 1, \dots, m$ and $x \in X$. There exists a distribution $\mathcal{D}$ of $(x, y) \in X \times \{0, 1\}$ such that*

$$\text{err}_{\mathcal{D}}(h_i) = \begin{cases} 0.3 - \frac{1}{4m}, & \text{if } i \text{ is odd;} \\ 0.3 + \frac{1}{4m}, & \text{if } i \text{ is even.} \end{cases}$$

Proof. We construct the distribution $\mathcal{D}$ of (x, y) as follows. With probability $1/2$, we choose $x = 0$. With the remaining probability $1/2$, we choose x uniformly at random from $1, \dots, m$. That is, $\Pr[x = i] = 1/(2m)$ for every $i = 1, \dots, m$. Conditioned on $x = 0$, we draw $y \in \{0, 1\}$ such that $\Pr[y = 1 | x = 0] = 0.1$. Conditioned on $x \neq 0$, we choose $y = 1$ (deterministically) if x is odd, and $y = 0$ if x is even.

For odd $i \in \{1, \dots, m\}$, we have

$$\begin{aligned} \text{err}_{\mathcal{D}}(h_i) &= \frac{1}{2} \cdot 0.1 + \frac{1}{2m} \sum_{x=1}^m \mathbb{I}[h_i(x) \neq \mathbb{I}[x \text{ is odd}]] \\ &= 0.05 + \frac{1}{2m} \sum_{x=1}^m (\mathbb{I}[x < i \text{ and } x \text{ is odd}] + \mathbb{I}[x \geq i \text{ and } x \text{ is even}]) \\ &= 0.05 + \frac{1}{2m} \left(\frac{i-1}{2} + \frac{m-i}{2} \right) \\ &= 0.3 - \frac{1}{4m}. \end{aligned}$$

Similarly, for even $i \in \{1, \dots, m\}$, we have

$$\begin{aligned} \text{err}_{\mathcal{D}}(h_i) &= 0.05 + \frac{1}{2m} \sum_{x=1}^m (\mathbb{I}[x < i \text{ and } x \text{ is odd}] + \mathbb{I}[x \geq i \text{ and } x \text{ is even}]) \\ &= 0.05 + \frac{1}{2m} \left(\frac{i}{2} + \frac{m-i+1}{2} \right) \\ &= 0.3 + \frac{1}{4m}. \end{aligned} \quad \square$$

Proof of Theorem 5.1. We first prove the theorem in the special case where $d = 1$. On domain $X := \mathbf{Z}_{\geq 0} = \{0, 1, \dots\}$, consider the hypothesis class $\mathcal{H} = \{h_1, h_2, \dots\}$, where $h_i(x) = \mathbb{I}[x \geq i]$ for every $i = 1, 2, \dots$ and $x \in X$. Clearly, $\text{VCdim}(\mathcal{H}) = 1$.

Let k be the largest integer satisfying $2^{k+2} + 4 < 1/\alpha$. By our assumption that $\alpha \in (0, 1/12)$, we have $k = \Omega(\log(1/\alpha))$. Our goal is to construct k distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$ that are shattered by $h_1, \dots, h_K$ for $K = 2^k$. Let $E_1, \dots, E_K$ be all the K subsets of $\{1, \dots, k\}$. For every $j = 1, \dots, k$, we will construct $\mathcal{D}_j$ such that for every $i = 1, \dots, K$,

$$\begin{aligned} \text{err}_{\mathcal{D}_j}(h_i) &< 0.3 - \alpha, & \text{if } j \in E_i; \\ \text{err}_{\mathcal{D}_j}(h_i) &> 0.3, & \text{if } j \notin E_i. \end{aligned} \quad (21)$$

This ensures that $\mathcal{D}_1, \dots, \mathcal{D}_k$ are shattered by $h_1, \dots, h_K$, as needed to prove the lemma.

It remains to show the construction of $\mathcal{D}_j$ for every $j = 1, \dots, k$. It will become convenient to choose E_1 to be the whole set $E_1 = \{1, \dots, k\}$. This ensures that $j \in E_1$ for every $j = 1, \dots, k$. For a fixed j , we partition $\{1, \dots, K\}$ into consecutive non-empty blocks $S_1, \dots, S_m$ for some $m \leq K$. Each

block S_ℓ has the form $S_\ell = \{x_{\ell-1} + 1, x_{\ell-1} + 2, \dots, x_\ell\}$, where $0 = x_0 < x_1 < \dots < x_m = K$. We define these blocks $S_1, \dots, S_m$ so that

$$\begin{aligned} j &\in E_i && \text{for every odd } \ell \in \{1, \dots, m\} \text{ and every } i \in S_\ell; \\ j &\notin E_i && \text{for every even } \ell \in \{1, \dots, m\} \text{ and every } i \in S_\ell. \end{aligned} \quad (22)$$

If m is odd, we define $X' = \{x_0, x_1, \dots, x_m\}$. If m is even, we define $X' = \{x_0, x_1, \dots, x_m, K + 1\}$. By Lemma B.3, there exists a distribution $\mathcal{D}_j$ over $X' \times \{0, 1\}$ such that

$$\begin{aligned} \text{err}_{\mathcal{D}_j}(h_{x_\ell}) &\leq 0.3 - \frac{1}{4(m+1)} < 0.3 - \alpha && \text{for odd } \ell \in \{1, \dots, m\}, \\ \text{err}_{\mathcal{D}_j}(h_{x_\ell}) &\geq 0.3 + \frac{1}{4(m+1)} > 0.3 && \text{for even } \ell \in \{1, \dots, m\}, \end{aligned} \quad (23)$$

where we used $4(m+1) \leq 4(K+1) = 2^{k+2} + 4 < 1/\alpha$. Recall that for each $\ell = 1, \dots, m$, the largest element in block S_ℓ is x_ℓ , so for every $i \in S_\ell$,

$$h_i(x) = h_{x_\ell}(x) \quad \text{for every } x \in X'.$$

Therefore, (23) implies

$$\begin{aligned} \text{err}_{\mathcal{D}_j}(h_i) &< 0.3 - \alpha && \text{for every odd } \ell \in \{1, \dots, m\} \text{ and every } i \in S_\ell; \\ \text{err}_{\mathcal{D}_j}(h_i) &> 0.3 && \text{for every even } \ell \in \{1, \dots, m\} \text{ and every } i \in S_\ell. \end{aligned} \quad (24)$$

Combining (22) and (24) proves (21).

Finally, we prove the theorem for a general positive integer d . On input space $X' := [d] \times X$, we construct a hypothesis class $\mathcal{H}'$ using our hypothesis class $\mathcal{H}$ above as follows:

$$\mathcal{H}' := \{h : X' \rightarrow \{0, 1\} \mid \exists h_1, \dots, h_d \in \mathcal{H} \text{ s.t. } h'(i, x) = h_i(x) \text{ for every } (i, x) \in X'\}.$$

This construction ensures that $\text{VCdim}(\mathcal{H}') = d \cdot \text{VCdim}(\mathcal{H}) = d$. By our analysis above, there exist $k = \Omega(\log(1/\alpha))$ domains $\mathcal{D}_1, \dots, \mathcal{D}_k$ on X that are shattered by $\mathcal{H}$. Now for every $i \in [d]$ and $j \in [k]$, we define $\mathcal{D}_{i,j}$ to be the distribution of $(i, x) \in X'$ with $x \sim \mathcal{D}_j$. These $kd = \Omega(d \log(1/\alpha))$ domains $(\mathcal{D}_{i,j})_{i \in [d], j \in [k]}$ are shattered by $\mathcal{H}'$, completing the proof. $\square$

B.4 Proof of Theorem 5.2

Theorem 5.2. *Let $\mathcal{H}$ be an arbitrary hypothesis class with VC dimension d . For any set $\mathcal{G}$ of domains, any threshold $\tau \in \mathbb{R}$, and any margin $\alpha \in (0, 1/2)$,*

$$\text{Gdim}(\mathcal{H}, \mathcal{G}, \tau, \alpha) = O(d \log(1/\alpha)). \quad (11)$$

We will use the following classic result:

Theorem B.4 (Sauer-Shelah-Perles Lemma [Sauer, 1972, Shelah, 1972]). *Let $\mathcal{F}$ be a class of concepts $f : Z \rightarrow \{0, 1\}$ defined on an arbitrary domain Z with VC dimension d . Let S be a subset of Z with size $|S| = n \geq d$. Then*

$$|\mathcal{F}_S| \leq (en/d)^d,$$

where $\mathcal{F}_S$ is the restriction of $\mathcal{F}$ to the subset $S \subseteq Z$.

Proof of Theorem 5.2. Let $\mathcal{D}_1, \dots, \mathcal{D}_k \in \mathcal{G}$ be k domains shattered by $\mathcal{H}$. By the definition of shattering, there exist $K = 2^k$ hypotheses $h_1, \dots, h_K \in \mathcal{H}$ with the following property: for every pair of hypotheses h_{j_1}, h_{j_2} with $1 \leq j_1 < j_2 \leq K$, there exists $i \in \{1, \dots, k\}$ such that

$$|\text{err}_{\mathcal{D}_i}(h_{j_1}) - \text{err}_{\mathcal{D}_i}(h_{j_2})| > \alpha. \quad (25)$$

Let us consider a fixed domain $\mathcal{D}_i$ for an arbitrary $i = 1, \dots, k$. We show that for some $m = O(\alpha^{-2}k)$, there exists a dataset S_i consisting of m points $(x_1, y_1), \dots, (x_m, y_m)$ such that

$$|\text{err}_{\mathcal{D}_i}(h_j) - \text{err}_{S_i}(h_j)| < \alpha/2 \quad \text{for every } j = 1, \dots, K, \quad (26)$$

where $\text{err}_{S_i}(h_j)$ is the empirical error on S_i :

$$\text{err}_{S_i}(h_j) = \frac{1}{m} \sum_{\ell=1}^m \mathbb{I}[y_\ell \neq h_j(x_\ell)].$$

That is, S_i is a representative dataset for distribution $\mathcal{D}_i$ in terms of measuring the error of the K shattering hypotheses. We prove the existence of S_i by the probabilistic method. For i.i.d. points $(x_1, y_1), \dots, (x_m, y_m)$ drawn from $\mathcal{D}_i$, by the Chernoff bound, for a fixed j , (26) holds with probability at least $1 - 1/(2K)$ as long as $m \geq C\alpha^{-2}k$ for a sufficiently large absolute constant $C > 0$. By the union bound over $j = 1, \dots, K$, with probability at least $1/2$, the m data points satisfy (26), proving the existence of S_i .

We have now proved that for every $i = 1, \dots, k$, there exists a size- m dataset S_i satisfying (26), where $m = O(\alpha^{-2}k)$. Combining this with (25), we know that for any pair of hypotheses h_{j_1}, h_{j_2} with $1 \leq j_1 < j_2 \leq K$, there exists $i \in \{1, \dots, k\}$ such that

$$\text{err}_{S_i}(h_{j_1}) \neq \text{err}_{S_i}(h_{j_2}).$$

This implies that $h_{j_1}(x) \neq h_{j_2}(x)$ for some $(x, y) \in S_i \subseteq (S_1 \cup \dots \cup S_k)$. By Theorem B.4, for

$$n := |S_1 \cup \dots \cup S_k| \leq mk = O(\alpha^{-2}k^2), \quad (27)$$

it holds that

$$2^k \leq (2 + en/d)^d. \quad (28)$$

Plugging (27) into (28) and taking the logarithm of both sides, we get

$$k/d \leq O(\log(1/\alpha) + \log(2 + k/d)).$$

This implies $k/d = O(\log(1/\alpha))$, proving (11). $\square$