Distribution Shift Is Key to Learning Invariant Prediction

Hong Zheng[1], Fei Teng[1,2]*

[1]School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, 611756, China
[2]Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 611756, China
fteng@swjtu.edu.cn

Abstract

An interesting phenomenon arises: Empirical Risk Minimization (ERM) sometimes outperforms methods specifically designed for out-of-distribution tasks. This motivates an investigation into the reasons behind such behavior beyond algorithmic design. In this study, we find that one such reason lies in the distribution shift across training domains. A large degree of distribution shift can lead to better performance even under ERM. Specifically, we derive several theoretical and empirical findings demonstrating that distribution shift plays a crucial role in model learning and benefits learning invariant prediction. Firstly, the proposed upper bounds indicate that the degree of distribution shift directly affects the prediction ability of the learned models. If it is large, the models' ability can increase, approximating invariant prediction models that make stable predictions under arbitrary known or unseen domains; and vice versa. We also prove that, under certain data conditions, ERM solutions can achieve performance comparable to that of invariant prediction models. Secondly, the empirical validation results demonstrated that the predictions of learned models approximate those of Oracle or Optimal models, provided that the degree of distribution shift in the training data increases.

Introduction

Due to distribution shift within the data, Empirical Risk Minimization (ERM) exhibits poor performance (Muandet, Balduzzi, and Schölkopf 2013). To address this, several methods have been proposed for learning invariant prediction in domain generalization (DG), such as Invariant Risk Minimization (IRM) (Arjovsky et al. 2019), IRM-games (Ahuja et al. 2020), among others. The invariant prediction implies stable prediction across environments (domains), which essentially amounts to discovering the labeling mechanism from data (Peters, Bühlmann, and Meinshausen 2016). Since the causality between covariates X and a response Y reflects the strongest stabilizing associations, discovering such causal relationships appears to be the best choice for learning invariant prediction (Bühlmann 2020). However, an interesting phenomenon has arisen (Gulrajani and Lopez-Paz 2020): when carefully implemented, ERM achieves state-of-the-art performance, despite the existence of numerous algorithms specifically designed for out-of-distribution tasks. So, what is the reason causing this? Despite being attributed as a reason for model selection in their work, we argue that there should be a deeper reason behind it. Addressing it could lead to a clearer understanding of the learning process in machine learning.

Several studies provided explanations for the above phenomenon. One study attributed the phenomenon to the limited number of training domains (Wang, Wu, and Zhang 2024), i.e., poor generalization arises due to insufficient training domains. However, while this explanation accounts for why performance can be poor, it does not explain why ERM can perform well in certain cases with a relatively small number of domains. Since this study made the explanation via their lower bound, we then recall the bounds provided by (Ben-David et al. 2010). They provided bounds that consider the complexity of the data condition by using a constant λ. For a large λ, it means no classifier performs well on both the testing and training domains, i.e., the data are too complex for generalization; vice versa. Consequently, the good performance of ERM indicates that λ is relatively small for certain datasets. However, under DG, we are only allowed to access training domains, indicating that the explanation by λ is limited. But it enlightens us that analyzing the training data condition may be a good way.

In this study, we find that the performance of learning algorithms is driven by distribution shift across training data. A large degree of distribution shift leads to better performance, which is counterintuitive. Through an analysis of a regression case, we observed that the shift in the distributions of the training data (i.e., the joint distribution of (X, Y)) determines the learning of regression models. When the shift is large, the learned regression models tend to approach the preset ground-truth model. Inspired by this finding and building on studies (Birgé 1986; Massart and Nédélec 2006; Peters, Bühlmann, and Meinshausen 2016), we derive several theoretical results, which constitute our contributions as follows:

- ERM solutions exhibit similar performance to invariant prediction models. We prove that the optimal solution of ERM corresponds to an invariant prediction model, provided that the causality-related data assumption

*Corresponding author.

holds.

- Distribution shift plays a critical role in model learning. The proposed theorems tell us that if one wants to learn predictors from multi-domain data and expects them to perform as well as the realizable or Bayes optimal predictor under each domain, then distribution shift is necessary for such data and should be sufficiently large as the number of data domains increases.
- Distribution shift benefits learning invariant prediction. Analytically, a sufficiently large degree of shift among the training domains implies that the domains are well-separated, which in turn ensures that any estimator has minimal generalization error and performs stably, i.e., it approximates invariant prediction models.

According to the above, we can now explain why ERM performs well even with a relatively small number of domains. This is because the introduced data either satisfy a causality-related data assumption or exhibit a large degree of distribution shift in the training domains, making the carefully implemented prerequisite unnecessary under such data conditions. Additionally, we conduct experiments to empirically validate this. The results of a classification task on CMNIST (Arjovsky et al. 2019), for example, show that the performance of trained models is linearly correlated with the degree of distribution shift, aligning with our main argument.

Analytical Findings

Preliminaries

To better understand our work, we first outline several key concepts that will appear in the following sections.

Distribution shift: We use the superscript e to denote that a variable is associated with environment e, such as X^e and Y^e. Following the conventions presented in (Peters, Bühlmann, and Meinshausen 2016), (Arjovsky et al. 2019), and (Choraria et al. 2023), we consider the multiple environment data consist of several data environments (domains) indexed by the set $\mathcal{E} = \{1, 2, \ldots\}$, each associated with several data samples that follow a certain distribution. Correspondingly, the *distribution shift* denotes that $P^e \neq P^{e'}, \forall e \neq e' \in \mathcal{E}$.

Measurement of Distribution shift: We use the Kullback–Leibler (KL) divergence to measure the difference (i.e., the shift degree) between two given distributions. Other metrics are omitted here, as the KL is only related to our theorem findings. The KL divergence is defined as follows: Let P and Q be two probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. Assume that $P \ll Q$, and denote by $p = dP/d\mu$ and $q = dQ/d\mu$ the Radon-Nikodym derivatives of P and Q with respect to a common dominating measure μ. Then, the KL divergence is defined as

$$KL(P;Q) = \int \log \frac{p}{q} p d\mu,$$

whenever the integral is well-defined.

Deductive Analysis

Beginning with a simple case, we present a deductive analysis to explore the factors that determine the performance of ERM solutions. This case aligns with the examples discussed in studies (Arjovsky et al. 2019; Rosenfeld, Ravikumar, and Risteski 2020), and is as follows:

Example 1. *Assume that we have several latent variables following Gaussian distributions as follows:*

$$\begin{aligned} &z_1 \leftarrow \mathcal{N}(0, a), \varepsilon_1 \leftarrow \mathcal{N}(0, b), \\ &z_1 \perp \varepsilon_1, \varepsilon_2 \leftarrow \mathcal{N}(0, c), \\ &y = \gamma z_1 + \varepsilon_1, \gamma \leftarrow \mathcal{N}(0, 1), \\ &z_2 = y + \varepsilon_2, x = [z_1, z_2]. \end{aligned}$$

We aim to estimate a regression model ω based on (x, y) by solving $\min_{\omega \in \mathbb{R}} \mathbb{E}[(\omega^T x - y)^2]$, while the optimal regression model has already been defined above, i.e, $\omega^* = (\gamma \quad \mathbf{0})^T$. According to the normal equation, we have

$$\omega = \begin{pmatrix} \frac{\gamma a(c+q) - bp - rp^2 - qp}{a(b+c+2q) - p^2} \\ \frac{1a(b+q)}{a(b+c+2q) - p^2} \end{pmatrix},$$

where $p = Cov(z_1, \varepsilon_2)$ and $q = Cov(\varepsilon_1, \varepsilon_2)$, and the computational details are provided in the Appendix. Assume $p \to 0$ and $q \to 0$, then we have

$$\omega \approx \begin{pmatrix} \gamma \frac{c}{b+c} & \mathbf{1} \frac{b}{b+c} \end{pmatrix}^T. \tag{1}$$

Based on the result (1), we arrive at *a first finding*: the learning of ω is governed by the distribution P of x and y, as ω is determined solely by the variances b and c. Specifically, we find that if $c \gg b$, $\omega \to \omega^*$, as the weight of γ approaches 1 and the weight of $\mathbf{1}$ approaches 0. In this example, z_1, z_2, and ε_1 can be regarded as the causal variable, the spurious correlation variable, and the generative noise for y, respectively, according to the study (Bühlmann 2020). For ease of computation, we simply set them to follow Gaussian distributions with zero mean and variances a, b, and c.

Next, we consider the scenario of multi-domain data, i.e., using (x^e, y^e) for all $e \in \mathcal{E}$ to estimate a regression model ω. ω^* can then be considered as the *invariant prediction model* across domains. Following the settings in Example 1, i.e., letting a_e, b_e, c_e vary with e, we have

$$\omega \approx \begin{pmatrix} \gamma \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \frac{c_e}{b_e + c_e} & \mathbf{1} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \frac{b_e}{b_e + c_e} \end{pmatrix}^T$$

Base on the above result, we obtain *a second finding*: the data distributions still determine the learning of regression models, as the weights of γ and $\mathbf{1}$ are essentially similar to the result (1), but represent an average over domains. Specifically, if $\sum c_e \gg \sum b_e$, we have $\omega \to \omega^*$. Based on studies (Peters, Bühlmann, and Meinshausen 2016; Yong et al. 2024), the generative noise ε_1 for y is forbidden to change across all $e \in \mathcal{E}$, allowing the causal relationship $z_1 \to y$ to remain invariant. This implies that $\sum b_e$ can be treated as a constant. Therefore, we are aware that if the condition $\sum c_e \gg \sum b_e$ holds, the variance c_e must increase as e grows. This implies that the $KL(P^e; P^{e'}) \neq 0, \forall e \neq e' \in \mathcal{E}$, as the KL divergence of Gaussian-type variables depends on their mean and variance, and this value should be large. Then, a factor, *distribution shift*, influences learning, completing the deductive analysis.

Theoretical Findings

The analysis results from the previous section interestingly indicate that *distribution shift facilitates the learning of invariant prediction models*. In this section, we provide a theoretical justification for this argument. For each assumption and theorem, we include remarks to help better understanding.

We begin by recalling an assumption outlined in (Peters, Bühlmann, and Meinshausen 2016) to explain what invariant prediction is, as follows:

Assumption 1 (Invariant Prediction (Peters, Bühlmann, and Meinshausen 2016))**.** *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies* $\forall e \in \mathcal{E}$*:* X^e *has an arbitrary distribution and*

$$Y^e = \mu + \gamma^* X^e + \varepsilon^e,$$

$\varepsilon^e \sim F_\varepsilon$ *and* $\varepsilon^e \perp X^e_{S^*}$*, where* $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance, and the same distribution* F_ε *across all* $e \in \mathcal{E}$*.*

Remarks: (a) The nature of an invariant prediction model is that of a *labeling function*, ensuring stable predictions across domains. (b) Since the error ε^e follows Gaussian white noise, estimating $\hat{\gamma}$ is feasible for both log-likelihood and least-squares problems. (c) The variable $X^e_{S^*} = \gamma^* X^e$ differs from the causal variable, as this assumption does not imply that $PA(Y^e) = X^e_{S^*}$, where $PA(Y^e)$ denotes the parent of Y^e in a directed acyclic causal graph. Consequently, considering invariant prediction as causal prediction is theoretically problematic. Instead, (Peters, Bühlmann, and Meinshausen 2016) suggests that causal prediction is a special case under this assumption.

Then, we consider the data scenario involving causality between X^e and Y^e, which is similar to the data structure in Example 1, and present the following data assumption.

Assumption 2. *Assume a collection of datasets from multiple environments indexed by the set* $\mathcal{E}$*, where each environment* $e \in \mathcal{E}$ *is associated with an arbitrary distribution on* $\mathcal{X} \times \mathcal{Y}$*. We then assume that the causality between* X^e *and* Y^e *remains linearly invariant to changes in* e *and other latent variables in* X^e*.*

Remarks: (a) The causal relationship between X^e and Y^e is required to remain unaffected not only by e but also by any latent variables. This excludes situations involving confounders, mediators, or other variables that can affect the graph from $PA(Y^e) \to Y^e$. (b) This assumption is, in fact, limited, as real-world data often surprise us with their complex structures. Meanwhile, the invariance implies that $PA(Y^e) \subseteq X^e, \forall e \in \mathcal{E}$ holds.

Based on this causality-related data assumption, we have the following proposition of interest.

Proposition 1. *Assume* $(X^e, Y^e), \forall e \in \mathcal{E}$*, with* $|\mathcal{E}| \geq 1$*, satisfies Assumption 2. Then, the optimal solution* ω^* *of the learning objective*

$$\min_{\omega \in \mathbb{R}^n} \sum\nolimits_{e \in \mathcal{E}} \left\| \omega^T X^e - Y^e \right\|_2^2$$

satisfies Assumption 1, namely $\gamma^* = \omega^*$*.*

Remarks: (a) This proposition reveals that ERM solutions can also be considered invariant prediction models, such that they perform well in DG, provided that the data conditions satisfy Assumption 2. (b) There are several examples of such data, including those used for simple causal analyses in medical diagnosis, air quality assessment, etc.

The proof of this proposition, along with the proofs of the subsequent theorems, is provided in the Appendix. The above proposition, however, does not take into account the main argument concerning data distributions. To address this, we then consider the following assumptions for noise-free and Massart-noisy distribution scenarios.

Assumption 3 (Clean Distributions)**.** *Assume a hypothesis space* $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$*, and let* $\mathcal{P} = \{P \in \mathcal{P}(\mathcal{H})\}$ *denote a family data distributions, where* $\mathcal{P}(\mathcal{H}) \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{Y})$*.*

Remarks: (a) This assumption implies that the distribution family $\mathcal{P}$ is induced by certain functions in $\mathcal{H}$, thereby transforming the problem of probability estimation into a function estimation problem. Correspondingly, given the realizable functions, the associated distributions can be determined. (b) Note that for the quantity of data samples in any assumed distribution, we consider it is sufficient and do not mention it thereafter.

Assumption 4 (Massart-noisy Distributions)**.** *Assume a hypothesis space* $\mathcal{H} \subseteq \{h : \mathcal{X} \to \{0, 1\}\}$*. For* $m \in [0, 1]$*, let* $\mathcal{P}^m = \{P \in \mathcal{P}(\mathcal{H}) : |2\eta(x) - 1| \geq m \text{ for all } x \in \mathcal{X}\}$ *denote the family of data distributions satisfying the Massart noise condition (Massart and Nédélec 2006), where* $\mathcal{P}(\mathcal{H}) \subseteq \mathcal{P}(\mathcal{X} \times \{0, 1\})$*, and* $\eta(x) = \mathbb{E}[Y|X = x] = P(Y = 1|X = x)$*.*

Remarks: (a) This assumption aligns with the setting considered in prior studies (Massart and Nédélec 2006; Wang, Wu, and Zhang 2024) for binary classification. (b) Under this assumption, the Bayes-optimal classifiers remain valid for all distributions in $\mathcal{P}^m$ (Wang, Wu, and Zhang 2024). (c) The Massart noise condition constrains labeling randomness by requiring that, for any $x \in \mathcal{X}$, the label distribution must not be perfectly ambiguous. Specifically, when $m = 0$, the condition imposes no constraint, permitting arbitrary noise; when $m = 1$, it corresponds to a noise-free scenario. Hence, the Massart noise regime interpolates between the realizable and agnostic settings.

Assumption 4 is specifically related to the binary classification problem, whereas we do not mention any problem type in Assumption 3. To measure the difference between any $h \in \mathcal{H}$, we present the following measurement assumption.

Assumption 5. *Let* $\mathcal{H}$ *be a hypothesis space consisting of measurable functions* $h : \mathcal{X} \to \mathcal{Y}$*, where* $\mathcal{X}$ *and* $\mathcal{Y}$ *are measurable spaces, and* μ *be a probability measure on* $\mathcal{X}$*. Define a distance between any two hypotheses* $h, h' \in \mathcal{H}$ *by the* $\mathbb{L}_1(\mu)$*-norm:*

$$d(h, h') := \|h(X) - h'(X)\|_1. \tag{2}$$

We assume that this distance is uniformly bounded, i.e.,

$$0 \leq d(h, h') \leq \beta, \quad \forall h, h' \in \mathcal{H},$$

for some finite constant $\beta > 0$*.*

Remarks: (a) By definition of the distance, $d(h, h') = 0$ if and only if $h(X) = h'(X)$ μ-almost everywhere on $\mathcal{X}$. (b) Note that we assume the distance is upper bounded only to exclude the case of infinity.

Based on the above assumptions, we prove the following two theorems.

Theorem 1. *Let $\mathcal{P}_E = \{P_1, P_2, \ldots, P_E\} \subseteq \mathcal{P}$, where $\mathcal{P}$ is defined in Assumption 3. Assume that*

$$\sup_{P,P' \in \mathcal{P}_E} [KL(P; P')] \leq \alpha,$$

where $\alpha > 0$ is a finite constant. For each $P \in \mathcal{P}_E$, let $s^* \in \mathcal{H}$ denote a labeling (realizable) function such that $s^*(X) = Y$ almost surely under P. Then, we define $S_E = \{s_i^* | i \in \{1, \ldots, E\}\}$ as the set of such functions.

Given a metric d satisfying the distance definition (2) *in Assumption 5, for any estimate $\hat{s} \in \mathcal{H}$ under $\mathcal{P}_E$, one has*

$$\frac{1}{E} \sum_{s \in S_E} \mathbf{1}_{\{d(\hat{s}, s) \geq \varepsilon\}} \leq (1 - \sigma) + \sqrt{\frac{\log(1/\delta)}{2E}} \tag{3}$$

with probability at least $1 - \delta$, where

$$\sigma = \frac{\alpha + \log 2}{\log(E - 1)}$$

and $\delta \in (0, 1)$, if

$$\alpha \leq \log\left(\frac{E - 1}{2}\right) \tag{4}$$

for any small constant $\varepsilon \geq 0$.

Remarks. (a) The assumption on distributions implies that $\mathcal{P}_E$ constitutes a relatively compact family of distributions and is KL-bounded. (b) The differences between the assumed realizable functions are not important due to the clean data distribution assumption. (c) Inequality (3) is governed by α and E, indicating that the probability of the prediction error being larger than ε is bounded by the RHS. (d) Vacuous bound: When $E \to \infty$ and $\alpha = 0$, LHS ≤ 1.0 always holds. Worst-case: When $\alpha = 0$ and $E = 3$, LHS $\leq C$, where C denotes a large value for the second term on the RHS. According to these results, the significance of α is emphasized, as when its value is zero there is a risk of obtaining both a vacuous bound and the worst case. (e) Tightness: Inequality (3), a Hoeffding-type bound, is tight and converges at the rate of $\mathcal{O}(1/\sqrt{E})$, provided that $\alpha > 0$. Here, $\alpha > 0$ serves as a prerequisite condition for convergence, and when $\alpha \to \log$ and $E \to \infty$, the LHS $\to 0$. This implies that α should be as large as possible within the given bound.

Theorem 2. *Let $\mathcal{P}_E^m = \{P_1, P_2, \ldots, P_E\} \subseteq \mathcal{P}^m$, where $\mathcal{P}^m$ is defined in Assumption 4, such that*

$$\sup_{P,P' \in \mathcal{P}_E^m} [KL(P; P')] \leq \frac{2\beta m^2}{1 - m^2},$$

$m \in [0, 1)$, under Assumption 5. Let H_E denote the set of Bayes optimal classifiers corresponding to distributions in $\mathcal{P}_E^m$, i.e.,

$$H_E = \{h_P^* = \mathbf{1}_{\eta(X) \geq 1/2} : P \in \mathcal{P}_E^m\}.$$

Given a metric d satisfying the distance definition (2) *in Assumption 5, for any estimate $\hat{h} \in \mathcal{H}$ under $\mathcal{P}_E^m$, one has*

$$\frac{1}{E} \sum_{h \in H_E} \mathbf{1}_{\{d(\hat{h}, h) \geq \varepsilon\}} \leq (1 - \sigma) + \sqrt{\frac{\log(1/\delta)}{2E}} \tag{5}$$

with probability at least $1 - \delta$, where

$$\sigma = \frac{2\beta m^2 + (1 - m^2)\log 2}{(1 - m^2)\log(E - 1)}$$

and $\delta \in (0, 1)$, if

$$\beta \leq \frac{1 - m^2}{2m^2} \log\left(\frac{E - 1}{2}\right) \tag{6}$$

for any small constant $\varepsilon \geq 0$ and $m \neq 0$.

Remarks. (a) Due to the Massart-noisy condition, the differences between Bayes optimal classifiers affect the distribution shifts (see Proof). (b) The conditions (4) and (6) indicate that $\sigma \leq 1$. (c) Others, such as vacuous bound, worst-case, and tightness, are similar to Inequality (3). (d) For estimating $\hat{s}$ or $\hat{h}$, we do not specify particular learning methods, such as supervised or semi-supervised learning, nor learning algorithms, such as ERM or IRM. As long as our assumptions are satisfied, the above results hold. (e) *The theorems tell us that if one wants to learn predictors from multi-domain data and expects them to perform as well as the realizable or Bayes optimal predictor under each domain, then distribution shift is necessary for such data and should be sufficiently large under the given constraints as the number of data domains increases*.

Limitations of Theorems: (a) Inequalities (3) and (5) do not indicate the effect of the number of data samples in each domain on the estimation. (b) The complexity of the hypothesis space, such as the VC-dimension, is not considered. (c) Require $E \geq 2e^{\alpha} + 1$, where the logarithm is taken to the natural base and $\alpha > 0$.

Despite limitations, the theorems above indicate that distribution shift indeed affects the learning of models in a positive way, provided that some conditions are satisfied. Based on these results, we state the following corollary without proof to conclude our theoretical justification.

Corollary 1. *Let $\hat{s}$ be an estimator trained on data sampled from $\mathcal{P}_E$ (or from $\mathcal{P}_E^m$ in the binary classification setting). If the distribution shift measure α associated with $\mathcal{P}_E$ satisfies condition* (4) *(or the shift measure β associated with $\mathcal{P}_E^m$ satisfies condition* (6)*), then $\hat{s} \approx \gamma^*$, provided that α (or β) is sufficiently large under the given constraints. Here, γ^* serves as an invariant prediction model.*

Remarks: (a) Distribution shift is key to learning invariant prediction. (b) In the linear case, Assumption 1 holds for γ^*, under the distribution condition $\mathcal{P}_E$, we have $d(\hat{s}, \gamma^*) \leq \varepsilon$. Under $\mathcal{P}_E^m$, the difference between γ^* and the Bayes classifier s^* is given by $\|\gamma^* - s^*\|_1 = KL(P_{\gamma^*}; P_{s^*})/m \log((1 + m)/(1 - m)) = K$. Then, we have $d(\hat{s}, \gamma^*) \leq \varepsilon + K$ (see Lemma 3 in the Appendix). (c) For the nonlinear case, recalling that the invariant prediction models naturally correspond to the labeling model, the above also holds.

Empirical Validations

The previous section provides theoretical justifications for a key argument: *distribution shift benefits learning invariant prediction*. In this section, we conduct experiments to validate this key argument, not theorems; therefore, the constraint on E is not strictly enforced. Note that, due to the intervenable nature of real-world data, we only experimented with synthetic data.

Settings

We first illustrate the data information (two synthetic datasets: one for regression and the other for classification) and then list learning algorithms used for validation.

Synthetic Data for Regression: Following the data setting in Example 1, we build the synthetic data. To control the distributions, we let variances a and c be the variance e and set $b = 1.0$ as a constant. Then, by setting different values for e, we construct two datasets: $D1$ and $D2$. The $D1$ dataset has three trainable domains, i.e., $e \in \mathcal{E} = \{1.0, 2.0, 3.0\}$, and the $D2$ dataset has thirty training domains, i.e., $e \in \mathcal{E} = \{1.0, 2.0, 3.0, \cdots, 30.0\}$. In each domain of both datasets, we have ten thousand samples. Comparing the two datasets, we observe that the degree of distribution shift in $D2$ is greater than in $D1$, not only due to the variance settings but also because of the number of domains. The code used to generate them is provided in the Appendix.

CMNIST for Classification: The environment setting of CMNIST (Arjovsky et al. 2019) is originally defined as $\mathcal{E}_{all} = \{0.1, 0.2, 0.9\}$, where the decimal value represents the degree of distribution shift (i.e., the degree of correlation between colors and estimated labels), following the Bernoulli distribution. For instance, if $e = 0.0$, all color interventions are entirely correlated with the labels. Conversely, if $e = 1.0$, the color interventions are completely independent of the labels. In our experiments, we compare two datasets: $C1$ and $C2$. The dataset $C1$ has training domains $\mathcal{E} = \{0.1, 0.5\}$, while $C2$ has $\mathcal{E} = \{0.1, 0.2\}$. The domain $e = 0.9$ remains unchanged as the unseen domain. As observed, the degree of shift in $C1$ is larger than that in $C2$.

Learning Algorithms: We incorporate the baseline ERM, VREx (Krueger et al. 2021), and ERM++ (Teterwak et al. 2023), as well as the invariance-based IRM (Arjovsky et al. 2019) and P-IRM (Choraria et al. 2023), and the non-invariance-based FISH (Shi et al. 2022), EQRM (Eastwood et al. 2022), and RDM (Nguyen et al. 2024). The rationale for this choice is that these algorithms are open-source, allowing us to observe the learning outcomes of both invariance-based and non-invariance-based methods. Other details refer to the Appendix.

Experiments on Synthetic Data

Conclusion: The learned regression model approaches the true model when the training data exhibits a large degree of distribution shift and includes a large number of domains, supporting our main argument.

Experiments: Since the data is synthetic, we can access the preset GT model. Thus, we observe only the difference between the model ω, trained solely using ERM, and the GT model ω^*, as shown in Figure 1. Note that ω^* is an invariant prediction model that satisfies Assumption 1, as it is the generative model for the response Y. As observed from (a) to (b) in this figure, the learned model approximates the true one when the training data has sufficiently large distributional variation. Moreover, the result indicates that increasing the number of domains fundamentally expands the degree of distribution shift. This explains why the trained Machine Learning models perform better when the datasets contain a large number of domains, as presented by study (Wang, Wu, and Zhang 2024).

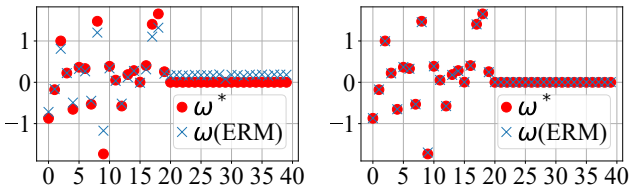

Figure 1: Models trained using ERM on the datasets (a) $D1$ and (b) $D2$. As observed, since the shift degree of distributions in $D2$ is larger than that in $D1$, the ω in (b) approaches ω^* more closely than the model in (a). Here, ω^* denotes the preset GT model.

Algorithms	$e_{0.1}$	$e_{0.5}$	$e_{0.9}$	Mean
ERM	**76.3±2.1**	66.3±1.6	49.5±0.9	64.0±1.5
IRM_Ω	58.3±5.6	55.6±2.7	50.1±4.9	54.7±4.4
IRM	63.8±5.2	65.0±1.5	57.4±4.8	62.1±3.8
$PIRM_\Omega$	59.6±4.6	53.3±2.8	44.9±2.8	52.6±3.4
$PIRM$	65.3±4.3	66.1±2.5	58.5±4.2	63.3±3.6
$ERM++$	**75.8±1.3**	67.2±1.6	49.3±1.7	64.1±1.5
$FISH$	**74.6±0.5**	65.2±1.8	47.9±0.4	62.6±0.9
RDM	70.6±4.1	66.5±6.1	54.6±3.5	63.9±4.6
$VREx$	**74.7±1.4**	67.0±1.3	49.4±2.2	63.7±1.7
$EQRM$	**74.3±1.0**	66.6±1.2	48.3±0.7	63.1±1.0
Oracle	64.6±1.6	66.2±1.4	64.5±1.3	65.1±1.4
Optimal	75.0±0.0	75.0±0.0	75.0±0.0	75.0±0.0

Table 1: Accuracy metrics obtained by different learning algorithms on the testing set for the $C1$ dataset. Here, the values represent the mean and standard deviation. *Oracle* refers to the results obtained by ERM grayscale models, and *Optimal* represents the results for hypothetically optimal invariant models. The **bold** values indicate abnormal values that exceed the *Optimal* results.

Experiments on CMNIST

Conclusion: The improvement in evaluation metrics for classification problems benefits from a greater degree of distribution shift in the data, i.e., the performance of trained models is linearly correlated with the degree of distribution shift. This also supports our main argument.

Experiments: We first validate the classification ability of

Algorithms	$e_{0.1}$	$e_{0.2}$	$e_{0.9}$	Mean
ERM	**83.6±0.8**	**79.2±1.2**	28.3±1.6	63.7±1.2
IRM_Ω	58.2±7.0	57.0±6.6	42.8±8.5	52.7±7.4
IRM	**85.2±1.1**	**82.1±1.2**	15.2±2.6	60.8±1.6
$PIRM_\Omega$	64.9±3.8	64.0±3.9	32.2±5.0	53.7±4.2
$PIRM$	**83.7±3.3**	**79.7±3.6**	18.4±4.9	60.6±3.9
$ERM++$	**84.9±1.4**	**78.6±1.1**	27.3±1.3	63.6±1.3
$FISH$	**84.6±0.8**	**79.9±1.3**	26.3±1.0	63.6±1.1
RDM	**81.5±2.7**	**78.2±2.0**	33.1±6.3	64.2±4.0
$VREx$	**83.1±1.1**	**80.8±0.5**	28.0±2.5	64.0±1.4
$EQRM$	**84.1±1.3**	**78.9±2.5**	27.8±2.7	63.6±2.2
Oracle	64.6±1.6	66.2±1.4	64.5±1.3	65.1±1.4
Optimal	75.0±0.0	75.0±0.0	75.0±0.0	75.0±0.0

Table 2: Accuracy metrics obtained by different learning algorithms on the testing set for the $C2$ dataset.

models trained by different algorithms, which can be seen as factual predictions. Then, we perform hypothesis testing on the factual predictions. Lastly, we validate the classification ability of the models when changing the colors in the original testing data, which can be seen as counterfactual predictions. *Note* that we only use accuracy to evaluate the results, as it sufficiently reflects model performance in both factual and counterfactual prediction scenarios. The Area Under the Curve (AUC) is not informative in this setting, as there is 25% label noise in CMNIST.

Factual predictions: The comparison between Tables 1 and 2 represents the main result of this validation. As observed, the evaluation metrics on domain $e_{0.9}$ in Table 1 are almost twice as high as those in Table 1. However, it is NOT important. The *decrease* in evaluations on domain $e_{0.1}$ is the most valuable for this validation experiment. It indicates that the performance of models trained using $C1$ is closer to the oracle model, even the optimal model, under any domain.

Notably, to validate that the improvements observed in the domain $e_{0.9}$ and $e_{0.1}$ are directly related to the distribution changes, we rely on the following hypothesis testing results for validation.

Hypothesis Testing for Results on $e_{0.9}$: First, we state the null hypothesis

$H_{e_{0.9},0}$: *The improved results in domain* $e_{0.9}$ *are unrelated to changes in environmental variables*.

The alternative hypothesis aligns with our argument and is omitted here. Second, to validate $H_{e_{0.9},0}$, we collected data by varying one environmental variable v_{e_2} in a training domain while keeping the v_{e_1} stable, as shown in Figure 2. The descriptor v_{e_1} represents the values in the first training domain in CMNIST, which are kept unchanged (i.e., $v_{e_1} = 0.1$). The descriptor v_{e_2} represents the values in the second training domain, which are changed (i.e., $v_{e_2} = \{0.2, 0.25, 0.3, 0.35, \cdots, 0.55\}$). As observed, the shits $\|v_{e_2} - v_{e_1}\|$ gradually increase. The descriptor y represents the classification accuracy in domain $e_{0.9}$. Then, we can use the parameter β to estimate a regression model for these data, specifically $y = \beta_0 + \beta_1 e_1 + \beta_2 e_2 + \varepsilon_{gaussian}$. Using the Python library statsmodels, we obtained the

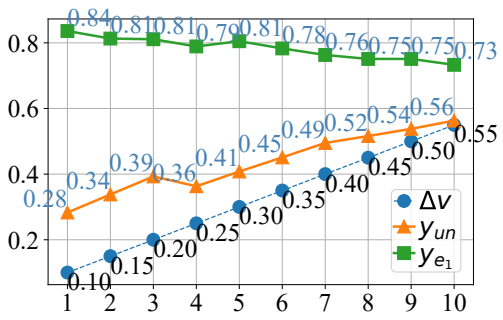

Figure 2: Accuracies (y_{un} and y_{e1}) as the shits $\Delta V = \|v_{e_2} - v_{e_1}\|$ increase. All results are obtained from ERM using CMNIST data, where v_{e_2} and v_{e_1} denote the distribution shift settings for spurious correlation colors in domains e_2 and e_1, respectively. In 10 sampling trials, we fixed v_{e_1} and varied v_{e_2}; thus, the shifts ΔV increased as sampling progressed. Correspondingly, we obtained 10 prediction results y_{un} for the unseen domain and 10 results y_{e1} for domain e_1. One observed phenomenon is that as the shifts ΔV increase, all prediction accuracies tend to converge to values between the oracle and optimal levels.

	coef	std err	t	$P > \|t\|$	[0.025	0.975]
β_1	1.7856	0.179	9.974	0.000	1.373	2.198
β_2	0.6029	0.040	15.108	0.000	0.511	0.695

Table 3: Results of t-testing for $H_{e_{0.9},0}$.

results shown in Table 3. The p-values are clearly significant, leading us to reject $H_{e_{0.9},0}$. Additionally, we observed that the t-value of β_2 is larger than β_1, suggesting that $e2$ has a greater impact on the results in the domain $e_{0.9}$.

Hypothesis Testing for Results on $e_{0.1}$: Similarly, define the null hypothesis

$H_{e_{0.1},0}$: *The decreased accuracy in domain* $e_{0.1}$ *is unrelated to changes in environmental variables*.

The alternative hypothesis, being the opposite, is omitted here. The data for changed and unchanged domains are consistent with the previous hypothesis testing. However, the difference is that we use $y1$ (see Figure 2) to denote the classification accuracy in domain $e_{0.1}$. We then use the parameter β to estimate a regression model for these data, specifically $y1 = \beta_0 + \beta_1 e_1 + \beta_2 e_2 + \varepsilon_{gaussian}$. Using the Python library statsmodels, we obtained the results shown in Table 4. The p-values are clearly significant, leading us to reject $H_{e_{0.1},0}$. Moreover, we observed that the t-value of β_2 is negative, indicating that $e2$ is the sole reason for the reduced accuracy in domain $e_{0.1}$.

The validation is complete, as supported by the above.

Counterfactual Predictions: For counterfactual predictions, we altered the colors in the testing data. For example, an original image matrix with a color layer of $[1, 0, 0]$, representing red, was changed to $[0, 1, 0]$, which represents green. This allows us to assess whether the model's prediction is invariant to non-causal features

	coef	std err	t	$P > \|t\|$	[0.025	0.975]
β_1	8.7422	0.083	105.669	0.000	8.551	8.933
β_2	-0.2135	0.018	-11.575	0.000	-0.256	-0.171

Table 4: Results of t-testing for $H_{e_{0.1},0}$.

Algorithms	Type	$e_{0.1}$	$e_{0.5}$	$e_{0.9}$
ERM	CF	43.6±2.5	62.6±1.2	76.7±1.8
	F	76.3±2.1	66.3±1.6	49.5±0.9
	$\|CF-F\|$	**32.7±0.4**	3.7±0.4	**27.2±0.9**
IRM	CF	53.9±5.6	65.5±1.4	67.0±4.7
	F	63.8±5.2	65.0±1.5	57.4±4.8
	$\|CF-F\|$	9.9±0.4	0.5±0.1	9.6±0.1
$PIRM$	CF	54.7±2.8	66.5±2.0	67.8±3.9
	F	65.3±4.3	66.1±2.5	58.5±4.2
	$\|CF-F\|$	10.6±1.5	0.4±0.5	9.3±0.3
RDM	CF	50.5±3.0	64.2±7.9	73.9±3.8
	F	70.6±4.1	66.5±6.1	54.6±3.5
	$\|CF-F\|$	20.1±1.1	2.3±1.8	19.3±0.3

Table 5: Counterfactual (CF) vs. Factual (F) predictions by different methods on the testing set for the $C1$ dataset. The **bold** values indicate errors exceeding 25%.

(color), and hence test its robustness under distributional shift. If predictions vary significantly, this indicates that the model has learned to rely on color as a shortcut, rather than capturing the shape-dependent causal structure.

The results are exhibited in Tables 5 and 6, which are obtained by four algorithms and are sufficient for illustration purposes. The results of $\|CF–F\|$ exhibit that models trained using $C1$ are less sensible for colors, approach the oracle models, and perform better than those trained using $C2$. We know that the shift degree of distributions in $C1$ is large than $C2$, and the counterfactual prediction results support our main argument. We consider a value of $\|CF–F\|$ greater than 25% to be strongly associated with color, since there is 25% labeling noise in the original test data, and we should allow for up to 25% tolerance under the counterfactual prediction scenario.

Related Works

We briefly illustrate several studies in Domain Generalization on learning invariant prediction and introduce theoretical bounds that guide generalization.

Learning Methods: An invariance-based learning revolution is brewing to address the out-of-distribution problems. For example, there are methods for learning invariant representations (Wang et al. 2022; Chuang, Torralba, and Jegelka 2020; Li et al. 2018) and causal representations (Wang et al. 2023; Bagi et al. 2023; Wang, Zhang, and Zhang 2025). However, several limitations of learning invariance have been discussed by researchers, such as fundamental limits and trade-offs (Zhao et al. 2022), a fundamental design flaw in IRM (Huh and Baidya 2022), and fake invariance (Chen et al. 2024). Moreover, several studies have questioned learning invariance, such as whether the invariances learned by deep neural networks align with human perception (Nanda et al. 2023), and how to learn invariances in the absence of distinct environments (Lin et al. 2022). Furthermore, some researchers have raised concerns that causal representation learning is inherently ill-posed (Morioka and Hyvarinen 2024). In conclusion, regardless of the advantages and open questions surrounding learning invariance, it has pushed this field of study a huge step forward.

Algorithms	Type	$e_{0.1}$	$e_{0.2}$	$e_{0.9}$
ERM	CF	28.9±1.6	36.7±1.8	84.8±0.8
	F	83.6±0.8	79.2±1.2	28.3±1.6
	$\|CF-F\|$	**54.7±0.8**	**42.5±0.6**	**56.5±0.8**
IRM	CF	18.0±2.1	21.5±2.1	87.6±1.4
	F	85.2±1.1	82.1±1.2	15.2±2.6
	$\|CF-F\|$	**67.2±1.0**	**60.6±0.9**	**72.4±1.2**
$PIRM$	CF	20.0±3.6	23.4±4.0	84.7±3.9
	F	83.7±3.3	79.7±3.6	18.4±4.9
	$\|CF-F\|$	**63.7±0.3**	**56.3±0.4**	**66.3±1.0**
RDM	CF	33.1±6.5	37.0±4.9	84.9±3.1
	F	81.5±2.7	78.2±2.0	33.1±6.3
	$\|CF-F\|$	**48.4±3.8**	**41.2±2.9**	**51.8±3.2**

Table 6: Counterfactual (CF) vs. Factual (F) predictions by different methods on the testing set for the $C2$ dataset.

Theoretical Bounds: The theoretical bounds in learning theory provide a guide for how much better the estimator is. From the VC bounds (Vapnik 2013) to the Massart-noisy bounds (Massart and Nédélec 2006), the mystery of the learning pattern in machine learning has been gradually revealed. These methods provide several useful fundamental inequalities to analyze different problems, inspiring future studies, including ours. Ben-David et al. (2010) were the first to focus on learning from different domains, providing generalization bounds based on the VC-class hypothesis and the $\mathcal{H}$-divergence. However, labeling noise is not considered in this work. A recent study (Wang, Wu, and Zhang 2024) addresses this by presenting a lower bound under the VC-class and Massart-noisy condition. All of these works are highly significant for machine learning, as we consider the advancement of fundamental theories to be true progress.

Conclusion

Our findings, in fact, reveal a pessimistic reality: model learning is overly dependent on data conditions. For example, ERM can outperform methods specifically designed for out-of-distribution tasks, provided that certain data conditions are met. When can we develop a learning method that depends less on data conditions? Finally, there are limitations in our work. The real-world data conditions are complex, such that only a few scenarios satisfy Assumption 2. In most cases, the determining factor is still distribution shift. However, the degree of distribution shift serves as a prior condition, which is typically unknown for most data conditions.

Ethical Statement

This study explores the theoretical mechanisms underlying domain generalization and, incidentally, explains why Empirical Risk Minimization (ERM) sometimes can outperform methods specifically designed for out-of-distribution (OOD) tasks. The insights provided by this study aim to promote a deeper understanding of learning under domain generalization, such as the role of distribution shift in the data, thereby supporting the ethical development and deployment of machine learning technologies for societal benefit. We stress that the outcomes of our research should be utilized with caution, upholding standards of transparency and accountability.

The positive societal implication of this research lies in understanding how much better or worse the performance of trained machine learning models can be when using complex and variable environmental data, which may originate from autonomous driving, medical diagnosis, financial analysis, or applications in the natural and social sciences. This understanding can promote the development of more robust AI systems for real-world applications. Nonetheless, we acknowledge that improving model robustness across diverse domains may give rise to certain ethical concerns. For example, models with strong generalization capabilities could be deployed in contexts lacking sufficient attention to data privacy, informed consent, or fairness, thereby exacerbating pre-existing social inequalities.

Acknowledgments

This research is supported by the grants from the National Natural Science Foundation of China (No.62272398) and Sichuan Science and Technology Program (No.2024NSFJQ0019).

References

Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*, 145–155. PMLR.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bagi, S. S. G.; Gharaee, Z.; Schulte, O.; and Crowley, M. 2023. Generative causal representation learning for out-of-distribution motion forecasting. In *International Conference on Machine Learning*, 31596–31612. PMLR.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79(1): 151–175.

Birgé, L. 1986. On estimating a density using Hellinger distance and some other strange facts. *Probability theory and related fields*, 71(2): 271–291.

Bühlmann, P. 2020. Invariance, causality and robustness. *Statistical Science*, 35(3): 404–426.

Chen, Z.; Zheng, Y.; Lai, Z.-R.; Guan, Q.; and Lin, L. 2024. Diagnosing and Rectifying Fake OOD Invariance: A Restructured Causal Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11471–11479.

Choraria, M.; Ferwana, I.; Mani, A.; and Varshney, L. R. 2023. Learning Optimal Features via Partial Invariance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 7175–7183.

Chuang, C.-Y.; Torralba, A.; and Jegelka, S. 2020. Estimating Generalization under Distribution Shifts via Domain-Invariant Representations. *Proceedings of Machine Learning Research*, 119.

Eastwood, C.; Robey, A.; Singh, S.; Von Kügelgen, J.; Hassani, H.; Pappas, G. J.; and Schölkopf, B. 2022. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35: 17340–17358.

Gulrajani, I.; and Lopez-Paz, D. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Huh, D.; and Baidya, A. 2022. The Missing Invariance Principle found–the Reciprocal Twin of Invariant Risk Minimization. *Advances in Neural Information Processing Systems*, 35: 23023–23035.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, 5815–5826. PMLR.

Li, Y.; Gong, M.; Tian, X.; Liu, T.; and Tao, D. 2018. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Lin, Y.; Zhu, S.; Tan, L.; and Cui, P. 2022. ZIN: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542.

Massart, P.; and Nédélec, É. 2006. Risk Bounds for Statistical Learning. *The Annals of Statistics*, 34(5): 2326–2366.

Morioka, H.; and Hyvarinen, A. 2024. Causal Representation Learning Made Identifiable by Grouping of Observational Variables. In *Forty-first International Conference on Machine Learning*.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International conference on machine learning*, 10–18. PMLR.

Nanda, V.; Majumdar, A.; Kolling, C.; Dickerson, J. P.; Gummadi, K. P.; Love, B. C.; and Weller, A. 2023. Do invariances in deep neural networks align with human perception? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9277–9285.

Nguyen, T.; Do, K.; Duong, B.; and Nguyen, T. 2024. Domain Generalisation via Risk Distribution Matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2790–2799.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and

confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5): 947–1012.

Rosenfeld, E.; Ravikumar, P. K.; and Risteski, A. 2020. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*.

Shi, Y.; Seely, J.; Torr, P.; N, S.; Hannun, A.; Usunier, N.; and Synnaeve, G. 2022. Gradient Matching for Domain Generalization. In *International Conference on Learning Representations*.

Teterwak, P.; Saito, K.; Tsiligkaridis, T.; Saenko, K.; and Plummer, B. A. 2023. Erm++: An improved baseline for domain generalization. *arXiv preprint arXiv:2304.01973*.

Vapnik, V. 2013. *The nature of statistical learning theory*. Springer science & business media.

Wang, H.; Si, H.; Li, B.; and Zhao, H. 2022. Provable domain generalization via invariant-feature subspace recovery. In *International Conference on Machine Learning*, 23018–23033. PMLR.

Wang, X.; Saxon, M.; Li, J.; Zhang, H.; Zhang, K.; and Wang, W. Y. 2023. Causal Balancing for Domain Generalization. Eleventh International Conference on Learning Representations (ICLR 2023).

Wang, Y.; Wu, Y.; and Zhang, H. 2024. Lost Domain Generalization Is a Natural Consequence of Lack of Training Domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15689–15697.

Wang, Y.; Zhang, W.; and Zhang, M.-L. 2025. Partial Label Causal Representation Learning for Instance-Dependent Supervision and Domain Generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21366–21374.

Yong, L.; Zhou, F.; Tan, L.; Ma, L.; Liu, J.; HE, Y.; Yuan, Y.; Liu, Y.; Zhang, J. Y.; Yang, Y.; and Wang, H. 2024. Continuous Invariance Learning. In *The Twelfth International Conference on Learning Representations*.

Zhao, H.; Dan, C.; Aragam, B.; Jaakkola, T. S.; Gordon, G. J.; and Ravikumar, P. 2022. Fundamental limits and tradeoffs in invariant representation learning. *The Journal of Machine Learning Research*, 23(1): 15356–15404.