# ON THE NATURE OF DOMAIN GENERALIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Domain generalization essentially faces the challenge where distribution shift leads to a generalization gap between the source and target domains, while any information on the target domains is unavailable during training. In this work, by introducing representation, we reveal that matching representation marginals and learning invariant features are equally important and both necessary for narrowing the generalization gap. In particular, we show that marginal matching can be regarded as a prerequisite for invariance learning. We then uncover a "domino effect" in marginal matching under a mild assumption, which guarantees that, without requiring knowledge of any target marginals, matching source marginals alone is sufficient to address the prerequisite. Additionally, we analyze the generalization discrepancy between the minimizer of the empirical risk subject to marginal matching and invariance constraints and the optimal minimizers of the expected risk, showing that domain generalization methods have no theoretical advantage in reducing this generalization discrepancy compared with Empirical Risk Minimization. This suggests that research focusing merely on algorithmic-level improvement cannot lead to real breakthroughs.

## 1 INTRODUCTION

Representation learning is widely regarded as an effective approach to mitigating the effects of distribution shift in Domain Generalization (DG), as it facilitates the learning of domain-invariant features (Muandet et al., 2013). Representation distribution matching (Ganin et al., 2016; Zhao et al., 2018) is among the representative approaches for learning invariant features. Recently, Dong et al. (2025) pointed out that distribution matching alone may not be sufficient to fully learn invariant representations so as to solve the DG problem. Based on mutual information theory, they argued that gradient matching (Shi et al., 2022; Nguyen et al., 2024) should also be considered in conjunction with distribution matching. Despite the success of the above methods in practical applications, several limitations remain. **(A)** Mutual information-based methods typically "blind" a discriminator to prevent it from distinguishing the domain of a learned representation, thereby reducing the corresponding mutual information, i.e., the distribution discrepancy. However, even a totally "blind" discriminator does not necessarily imply a small distribution discrepancy. **(B)** Distribution matching in DG typically focuses on source domains, as target-domain information is unavailable during training. Although empirically common, the theoretical mechanism underlying this practice remains unclear. While **(A)** can be addressed by classical distribution matching methods, such as (Bonneel et al., 2015; Sun et al., 2016), **(B)** remains unresolved.

In this work, we address the above limitation by adopting the most fundamental divergence-based information-theoretic perspective. We establish that the generalization gap between the source and target domains is upper-bounded by their joint distribution discrepancy, which can be decomposed, via representation mappings, into marginal and conditional discrepancies. We then recognize that most domain-invariant feature learning methods (Arjovsky et al., 2019; Pham et al., 2023) essentially minimize only the conditional discrepancy, leaving the marginal discrepancy unaddressed. Influenced by the marginal discrepancy, the generalization gap could still be large. This implies that marginal matching and invariance learning are both critical. Although akin to the main argument in (Dong et al., 2025), our analysis reveals a deeper relationship between the two methods, which enables invariance learning under a weaker assumption. Now, regarding **(B)**, conditional matching depends on theorems from previous invariance-based learning studies, while for marginal matching, we present a theorem under measurability and approximability assumptions on the representation



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

function space, revealing a "domino effect": minimizing marginal discrepancy across source domains triggers a cascade that reduces marginal discrepancies across all domains.

Additionally, by analyzing the generalization discrepancy between the empirical-risk minimizer under marginal matching and invariance constraints and the optimal expected-risk minimizer, we reveal a limitation of DG methods: without further justification for some arguments related to the optimal hypothesis, DG methods have no theoretical advantage over Empirical Risk Minimization (ERM) in reducing the generalization discrepancy. This suggests that under standard, general-purpose DG settings, the room for algorithmic improvement is far smaller than we imagined.

In conclusion, we contribute by (1) demonstrating that marginal matching and invariance learning are equally important and both necessary for DG, as well as their deep relationship, (2) theoretically explaining why matching source marginals leads to a reduction in the discrepancy between source and target marginals, and (3) revealing that that real breakthroughs may be achieved by simultaneously focusing on algorithmic-level improvement and in-depth analysis of the given data.

## 2 Theoretical Results

### 2.1 Preliminaries

Considering $\mathcal{X} \subset \mathbb{R}$ and $\mathcal{Y} \subset \mathbb{R}$ being respectively an input and output space, we have a training set

$$D = \{\xi_1 = (X, Y)_1, \ldots, \xi_E = (X, Y)_E\}$$

of $E$ domains in $\mathcal{X} \times \mathcal{Y}$ drawn from unknown distributions in $\mathcal{P}$. Assume that each $\xi \in D$ contains $N$ data pairs $(x_l, y_l)$, i.e., $(X, Y) := \{(x_l, y_l)\}_{l=1}^{N}$. Under DG, $(X, Y)_i \sim^{i.i.d.} P_i^{XY}$ for all $i \in [E]$, and $P_i^{XY} \neq P_j^{XY}$ for all $i \neq j \in [E]$. Thus, the overall samples follow an independent non-identically distributed (i.n.d.) assumption. The training set $D$ is typically regarded as the source set, while the testing set $D_t$ is regarded as the target set. Similarly, $P_i^{XY} \neq P_t^{XY}$ for all $i \in [E]$ and all $t \in [E_t]$, where $E_t$ denotes the number of target domains in $D_t$. Furthermore, no information from $D_t$ is available during training. In an idealized setting, we consider there is no label noise.

If considering representation learning, by assuming a *latent representation space* $\mathcal{Z} \subset \mathbb{R}$, we can use the capital letter $\hat{Z}$, corresponding to $X$ and $Y$, to denote representation variables, and the lowercase letter $\hat{z}$, corresponding to $x$ and $y$, to denote their values. We use the hat notation for representations, as they are typically learned during training.

We consider a hypothesis $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{H}$ is a hypothesis space. When considering representation learning, a hypothesis $h$ can be decomposed as $h = f \circ g$, where $g : \mathcal{X} \to \mathcal{Z}$ is a representation function and $f : \mathcal{Z} \to \mathcal{Y}$ is a prediction function. Accordingly, $\mathcal{H} = \mathcal{F} \circ \mathcal{G} = \{f \circ g : f \in \mathcal{F}, g \in \mathcal{G}\}$, where $\mathcal{G}$ and $\mathcal{F}$ denote the representation and prediction function spaces, respectively. Then, $\hat{Y} = h(X)$, which can be reformulated as $\hat{Z} = g(X)$ and $\hat{Y} = f(\hat{Z})$. We assume that all functions are measurable and all datasets are countable, which does not restrict the generality of the results presented hereafter.

To measure the accuracy of predictions from $h$, we consider the cost function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and the loss of $h$ with respect to data pairs $\xi$ is then defined as

$$\ell(h, \xi) = \ell(h, (X, Y)) = c(h(X), Y).$$

Correspondingly, given the training (source) set $D$, the *expected* risk of $h$ is defined by

$$R(h) = \frac{1}{E} \sum_{i=1}^{E} \mathbb{E}_{\xi \sim P_i^{XY}}[\ell(h, \xi)] = \frac{1}{E} \sum_{i=1}^{E} \mathbb{E}_{\xi_i}[\ell(h, \xi_i)],$$

and the *empirical* risk of $h$ is defined by

$$\hat{R}(h) = \frac{1}{EN} \sum_{i=1}^{E} \sum_{l=1}^{N} \ell(h, (x_l, y_l)_i) = \frac{1}{EN} \sum_{i=1}^{E} \ell(h, \xi_i).$$

For the target set $D_t$, we define only the *expected* risk of $h$ as follows,

$$R_T(h) = \frac{1}{E_t} \sum_{t=1}^{E_t} \mathbb{E}_{\xi_t}[\ell(h, \xi_t)].$$

Note that $R(h)$ is typically defined as $R(h) = \mathbb{E}_{P^{XY} \in \mathcal{P}_E} \mathbb{E}_{\xi \in P^{XY}}[\ell(h, \xi)]$, where $\mathcal{P}_E$ denotes the set of source distributions. Due to distribution shift, the outer expectation reduces to the empirical average over domains, i.e., $\mathbb{E}_{P \in \mathcal{P}_E}[\cdot] = E^{-1} \sum_{i=1}^{E} [\cdot]$, and the same argument also applies to $R_T(h)$.

### 2.2 Analysis

In this subsection, we provide a series of analyses based on the simplest theorems in statistical learning theory and information theory, as the most significant results often stem from the simplest proofs. We begin with the following theorem. *The proofs of all theoretical results are provided in Appendix A.3, and we do not repeat this statement hereafter.*

**Theorem 1.** *Given the definitions of $R_T(h)$ and $R(h)$, and assuming that $0 \leq \ell \leq M$, the following inequality holds for any $h \in \mathcal{H}$:*

$$R_T(h) \leq R(h) + C \sum_{t=1}^{E_t} \sum_{i=1}^{E} d_{JS}(P_t^{Y,h(X)}, P_i^{Y,h(X)}).$$

*Here, $C = (2\sqrt{2}M)/EE_t$ and $d_{JS}(P, Q) = \sqrt{JS(P||Q)}$, where $JS(\cdot||\cdot)$ denotes the Jensen-Shannon divergence.*

**Remarks**. (a) In essence, the above inequality can be expressed as $|R_T(h) - R(h)| \leq C \sum d_{JS}(P_t^{Y\hat{Y}}, P_i^{Y\hat{Y}})$, indicating that the generalization gap between the expected risks in source and target domains is bounded by the discrepancy between their joint distributions. Apparently, the smaller $d_{JS}(\cdot, \cdot)$, the tighter the bound. (b) Since $d_{JS}(P_t^{Y\hat{Y}}, P_i^{Y\hat{Y}})$ is upper-bounded by $d_{JS}(P_t^{YX}, P_i^{YX})$, without considering representation, most studies adopt the latter as an inherently uncontrollable factor. (c) $d_{JS}(\cdot, \cdot)$ is a metric that satisfies symmetry and the triangle inequality. When the JS divergence is defined using the natural logarithm, $0 \leq d_{JS}(\cdot, \cdot) \leq \sqrt{\ln 2}$. (d) For this upper bound, the lower bound $\sup_{\|\ell\|_\infty \leq M} |R_T(h) - R(h)| \geq c \sum_t \sum_i JS(P_t^{Y,h(X)}||P_i^{Y,h(X)})$, where $c = 2M/(\ln 2EE_t)$, proves its non-improvable convergence.

The term $\sum_t \sum_i d_{JS}(P_t^{Y,h(X)}, P_i^{Y,h(X)})$ stands for the total JS distance between the target and source domains in the joint distribution of the label $Y$ and the predictions by hypothesis $h$. It seems that we can control the learned $h$ during training to minimize this joint distribution discrepancy, i.e., we need to ensure that $P_t^{Y,h(X)}$ and $P_i^{Y,h(X)}$ are similar, thereby bringing the analysis to an end. However, **(i)** estimating the joint distribution of $Y$ and $h(X)$ can be challenging, particularly when the number of possible predictions is large; and **(ii)** most importantly, by optimization theory, imposing increasingly stringent constraints on $h$ progressively restricts the feasible hypothesis space $\mathcal{H}$, potentially excluding the optimal hypothesis of the original learning objective. Representation learning provides an effective solution to problem **(ii)** by introducing an additional representation function space $\mathcal{G}$, thereby shifting the constraints from $\mathcal{H}$ to $\mathcal{G}$, i.e., transforming the global constraints into local ones. By this, the following result naturally follows.

**Corollary 1.** *Under the setting of Theorem 1, let $h = f \circ g$. It follows that:*

$$R_T(h) \leq R(h) + C \sum_{t=1}^{E_t} \sum_{i=1}^{E} d_{JS}(P_t^{Y,g(X)}, P_i^{Y,g(X)}).$$

**Remarks**. The above inequality follows directly from the data processing inequality in information theory under the data processing sequence $X \to \hat{Z} \to \hat{Y}$. Namely, $d_{JS}(P_t^{Y\hat{Y}}, P_i^{Y\hat{Y}}) \leq d_{JS}(P_t^{Y\hat{Z}}, P_i^{Y\hat{Z}})$.

Now, we turn the distribution discrepancy from the prediction-related level to the representation-related level. Before proceeding further, we first address the term $R(h)$, the expected risk in the source domains, i.e., how this term can be made small. Based on classical learning theory, we obtain the following result.

**Theorem 2.** *Let $\epsilon_S(h) = R(h) - \hat{R}(h)$ for any $h \in \mathcal{H}$, and assume that $0 \leq \ell \leq M$. Then, with probability at least $1 - \delta$, the following inequality holds:*

$$\epsilon_S(h) \leq M \sqrt{\frac{\ln(1/\delta)}{2EN}}.$$

**Remarks**. (a) This characterizes a classical learning consistency bound, which follows from the Hoeffding concentration inequality. (b) Without any further information, the convergence rate cannot be further improved solely under the assumption of a bounded loss function, and is $\mathcal{O}(1/\sqrt{EN})$. Thus, there is no need to discuss the relative lower bound.

According to the above theorem, when $EN \to \infty$, we obtain the tightest upper bound for $\epsilon_S(h)$, such that a small $\hat{R}(h)$ leads to a small $R(h)$. This is a standard result in statistical learning theory, and we do not provide further explanation or analysis. Let $\epsilon_T(h) = R_T(h) - \hat{R}(h)$. We rewrite the inequality in Corollary 1 as follows:

$$\epsilon_T(h) \leq M\sqrt{\frac{\ln(1/\delta)}{2EN}} + C\sum_{t=1}^{E_t}\sum_{i=1}^{E} d_{JS}(P_t^{Y,g(X)}, P_i^{Y,g(X)}).$$

Namely, we hereafter consider the *generalization gap* $\epsilon_T(h)$ between the *expected* risk in the target domains and the *empirical* risk in the source domains for $h$. Then, by the above inequality, when $EN \to \infty$ and $d_{JS}(\cdot,\cdot) \to 0$, we obtain the tightest upper bound on $\epsilon_T(h)$, such that a small $\hat{R}(h)$ leads to a small $R_T(h)$.

We now proceed to further address the distribution discrepancy. For convenience, we ONLY consider one $t \in [E_t]$ and one $i \in [E]$ below; $\sum$ over all $t$ and $i$ is understood. Introducing $g$ addresses problem **(ii)**, while problem **(i)** can be addressed by the following result.

**Theorem 3.** *For any $g \in \mathcal{G}$, any $t \in [E_t]$, and any $i \in [E]$, $d_{JS}(P_t^{Y,g(X)}, P_i^{Y,g(X)})$ admits the following decomposition:*

$$d_{JS}(P_t^{Y,g(X)}, P_i^{Y,g(X)}) \leq d_{JS}(P_t^{g(X)}, P_i^{g(X)}) + \frac{1}{\sqrt{2}}\rho(g)$$

$$\rho(g) := \sqrt{\mathbb{E}_{\hat{z}\sim P_t^{g(X)}}[JS(P_t^{Y|\hat{z}}||P_i^{Y|\hat{z}})] + \mathbb{E}_{\hat{z}\sim P_i^{g(X)}}[JS(P_t^{Y|\hat{z}}||P_i^{Y|\hat{z}})]}.$$

**Remarks**. In essence, the above decomposition expresses the joint distribution $P^{Y\hat{Z}}$ in terms of the conditional distribution $P^{Y|\hat{Z}}$ and the marginal distribution $P^{\hat{Z}}$. Namely, the JS discrepancy between two joint distributions can be decomposed into the JS discrepancies between their corresponding marginals and conditionals.

Note that we first address **(ii)** and then **(i)**, with the reason for this ordering further discussed in Section 2.3. By the above decomposition, we have

$$\epsilon_T(h) \leq d_{JS}(P_t^g, P_i^g) + \rho(g) + \mathcal{O}(1/\sqrt{EN}),$$

where the terms are abbreviated for brevity. Here, if one is familiar with invariance learning methods, the meaning of $\rho(g)$ is readily apparent. If not, we explain it subsequently. Then, we argue that **both representation marginal matching ($\min d_{JS}(P_t^g, P_i^g)$) and invariance learning ($\min \rho(g)$) are necessary for narrowing** $\epsilon_T(h)$.

For $d_{JS}(P_t^{g(X)}, P_i^{g(X)})$, directly minimizing it is unrealistic, as information about the target domains is unavailable during training in DG. Thus, studies typically only match the representation marginals across the source domains, as in (Pham et al., 2023; Dong et al., 2025), and representatively adopt a discriminator for matching based on mutual information theory (this will be further discussed in Section 2.3). However, there is no theoretical explanation for that matching mechanism. We address this limitation with the following results. First, we give an assumption on the representation function space $\mathcal{G}$.

**Assumption 1.** *There exists a sequence of measurable representation functions $(g_k) \subset \mathcal{G}$ such that, for any $g \in \mathcal{G}$ and any $X \in \mathcal{X}$, $\lim_{k\to\infty} P^{g_k(X)} = P^{g(X)}$.*

**Remarks**. (a) In simple terms, for any representation marginal induced by a representator $g$, the function class $\mathcal{G}$ is rich enough to approximate it. Here, we emphasize that $g$ is arbitrary and does not necessarily correspond to an optimal or true one. (b) A detailed analysis of this assumption is provided in Section 2.3.

Under this assumption, we now present the following result, which constitutes the most important finding of our work.

**Theorem 4.** *Assuming that $\mathcal{G}$ satisfies Assumption 1, for any $t \in [E_t]$, any $i \in [E]$, and any distribution $p$ with $p \neq P_i^{g(X)}$ on the representation space $\mathcal{Z}$,*

$$d_{JS}(P_i^{g(X)}, p) \to 0 \implies d_{JS}(P_i^{g(X)}, P_t^{g(X)}) \to 0.$$

**Remarks**. (a) This $p$ may be a standard Gaussian, or any other fixed distribution. (b) This theorem implies a **"domino effect"** for matching representational marginals: "pushing down" $d_{JS}(P_i^g, p)$ causes $d_{JS}(P_i^g, P_t^g)$ to fall in turn. (c) The converse of this theorem does not hold unless $p$ is induced by $g$. (d) Expanding to all $t$ and $i$ naturally holds.

By the above theorem, to achieve a minimal $d_{JS}(P_i^g, P_t^g)$, we can safely match the source marginals to any representation marginal $p$ in the space $\mathcal{Z}$ without worrying about the target marginals. Note that $d_{JS}(P_i^{g(X)}, P_t^{g(X)}) \to 0$ does not imply that invariance is learned. Assume (temporarily) that $0 \leq f \circ g(X) \leq W$ for any $f$, $g$, and $X$. By Theorem 1, we have

$$|\mathbb{E}_{P_t}[f \circ g(X)] - \mathbb{E}_{P_i}[f \circ g(X)]| \leq 2\sqrt{2}W d_{JS}(P_t^g, P_i^g).$$

In theory, as $d_{JS}(P_i^g, P_t^g) \to 0$, we only obtain $\mathbb{E}_{P_t}[f \circ g(X)] \to \mathbb{E}_{P_i}[f \circ g(X)]$. Here, assuming that $f \circ g$ is bounded is only for explanation and does not affect the subsequent results.

Nevertheless, when $d_{JS}(P_t^g, P_i^g) \to 0$, the following result holds:

$$\epsilon_T(h) \leq \rho'(g) + \mathcal{O}(1/\sqrt{EN}),\ \rho'(g) = \sqrt{\mathbb{E}_{\hat{z} \in g(X)}[JS(P_t^{Y|\hat{z}} || P_i^{Y|\hat{z}})]}.$$

Now, $\rho'(g)$ is very clear for referring to learning invariance, and we can further give the following informal assumption, which is commonly adopted in studies of learning invariance, such as (Muandet et al., 2013; Li et al., 2018; Arjovsky et al., 2019; Choraria et al., 2023), to deepen the explanation.

**(iii)**: For any $i, j \in [E + E_t]$, assume that $P_i(Y|Z = z) = P_j(Y|Z = z)$, $\forall z \in \text{supp}(P_i^Z) \cap \text{supp}(P_j^Z)$.

This assumption essentially characterizes the underlying data generation mechanism. Under this assumption, learning invariance corresponds to minimizing $\rho(g)$ or $\rho'(g)$. By the change from $\rho(g)$ to $\rho'(g)$, we see that $d_{JS}(P_t^g, P_i^g) \to 0$ implies that the distribution of $\hat{Z}$ is no longer important. This implies that **matching marginals can serve as a prerequisite for learning invariance**, thereby relaxing the support set assumption in Assumption **(iii)**. How to specifically make the term $\rho'(g)$ small is beyond the scope of our analysis here, and many studies have proposed methods for this purpose, such as causal representation learning (Schölkopf et al., 2021; Zhang et al., 2024) and invariance learning (Wang et al., 2022; Janetzky et al., 2026). All in all, we conclude that, for DG, minimizing the following risk

$$\hat{R}_\lambda(h) = \hat{R}(h) + \lambda N(g),$$

where $\lambda$ is a penalty factor, and $N(g)$ represents the constraints of BOTH marginal matching and invariance learning, is "in an idealized sense" sufficient to achieve a tighter bound for $\epsilon_T(h)$, i.e., to learn an $h$ that performs well on $D_t$, when $EN \to \infty$ and Assumptions 1 and **(iii)** hold. Here, we say the idealized sense, as we exclude the scenarios where *label noise* exists. In practice, we should analyze how the generalization performance of the minimizer of $\hat{R}_\lambda(h)$ compares to that of the theoretically optimal minimizer of $R(h)$. We first make an informal assumption about the optimal minimizers of $R(h)$ under standard, general-purpose DG settings.

**(iv)**: Assume that for any $e \in [E + E_t]$, there exists $s_e = f_e' \circ g_e' = \arg\min_{h \in \mathcal{H}} R_e(h)$, where $R_e(h) = \mathbb{E}_{\xi_e}[\ell(h, \xi_e)]$. Here, $s_e(X_e)$ is not necessarily equal to $Y_e$.

This indicates that each domain has its own optimal hypothesis, which is a milder assumption than the existence of a single optimal hypothesis shared across all domains. One important reason is that the latter requires $\cap_{e=1}^{E_{all}} \arg\min \mathbb{E}_{\xi_e}[\ell(h)] \neq \emptyset$, whereas the former only requires $\arg\min \mathbb{E}_{\xi_e}[\ell(h)] \neq \emptyset, \forall e \in [E_{all}]$, where $E_{all} = E + E_t$. Note that here we did NOT assume that $g_e'$ is an invariant representator or that it leads to a diminishing discrepancy between the representation marginals. To evaluate the prediction discrepancy between any two hypotheses $h_1$ and $h_2$, we give the relative loss:

$$\gamma(h_1, h_2) := \ell(h_1, \cdot) - \ell(h_2, \cdot)$$

Then, we present the following theorem on the generalization discrepancy between any two hypotheses $h_1$ and $h_2$.

**Theorem 5.** *Given the definition of $\gamma$, and assuming that $0 \leq \ell \leq M$ and $\hat{R}_\lambda(h_1) \leq \hat{R}_\lambda(h_2)$ for any $h_1 = f_1 \circ g_1, h_2 = f_2 \circ g_2 \in \mathcal{H}$, the following inequality holds with probability at least $1-\delta$:*

$$\frac{1}{E_t}\sum_{t=1}^{E_t}\mathbb{E}_t[\gamma(h_1,h_2)] \leq \lambda\,|N(g_2)-N(g_1)| + M\sqrt{\frac{2\ln(1/\delta)}{EN}} + C\sum_{t=1}^{E_t}\sum_{i=1}^{E}\rho(g_1,g_2)$$

$$\rho(g_1,g_2) := d_{JS}(P_t^{Y,g_1(X)}, P_i^{Y,g_1(X)}) + d_{JS}(P_t^{Y,g_2(X)}, P_i^{Y,g_2(X)}).$$

*Here, $C = (2\sqrt{2}M)/EE_t$.*

**Remarks**. There are three terms on the right-hand side of the inequality. The first term represents the regularization error induced by $g_1$ and $g_2$, the second term represents the main order of the convergence rate, and the last term represents the distribution discrepancy between the sums of the joint distributions induced by $g_1$ and $g_2$. Namely, besides the parameters $E$ and $N$ of the source set, $\gamma(h_1,h_2)$ is primarily governed by the discrepancy between $g_1$ and $g_2$ in terms of their regularization and the corresponding distributions. The smaller the discrepancy, the tighter the bound.

By the above theorem, we then consider the generalization discrepancy between the minimizer of $\hat{R}_\lambda(h)$ and $s_e,\ \forall e \in [E+E_t]$, which leads to the following result.

**Corollary 2.** *Under the settings of Theorem 4, assume that $\hat{h} = \hat{f}\circ\hat{g} = \arg\min_{h\in\mathcal{H}}\hat{R}_\lambda(h)$. When $N(\hat{g}) \to 0$ such that $d_{JS}(P_t^{Y,\hat{g}(X)}, P_i^{Y,\hat{g}(X)}) \to 0$, for $s_e = f'_e \circ g'_e,\ \forall e \in [E+E_t]$, the following inequality holds with probability at least $1-\delta$:*

$$\frac{1}{E_t}\sum_{t=1}^{E_t}\mathbb{E}_t[\gamma(\hat{h},s_t)] \leq \frac{\lambda}{E}\sum_{i=1}^{E}N(g'_i) + C\sum_{t=1}^{E_t}\sum_{i=1}^{E}d_{JS}(P_t^{Y,g'_t(X)}, P_i^{Y,g'_i(X)}) + M\sqrt{\frac{2\ln(1/\delta)}{EN}}.$$

**Remarks**. (a) Note that $\hat{R}_\lambda(\hat{h}) \leq \hat{R}_\lambda(s_i)$ for any $i \in [E]$, which satisfies the assumption in Theorem 4. (b) This result essentially indicates that unless $N(g') \to 0$, there is still a risk of obtaining a loose or even vacuous bound for $\gamma(\hat{h},s)$. (c) Here, the assumption $N(\hat{g}) \to 0$ is strong.

We observe that even if $s_e(X_e) = Y_e,\ \forall e \in [E+E_t]$, i.e., $\mathbb{E}_T[\gamma(\hat{h},s)] \to \epsilon_T(h)$, the terms $N(g')$ and $d_{JS}(P_t^{Y,g'(X)}, P_i^{Y,g'(X)})$ remain. This is because we did not assume that $g'$ necessarily needs to satisfy $N(\cdot)$, i.e., it remains unclear how the properties of $g'$ itself affect the bound, even if the given data satisfy Assumption **(iii)**.

We then consider the minimizer of $\hat{R}(h)$ (ERM) and let $\hat{h}_{erm} = \hat{f}\circ\hat{g}_{erm} = \arg\min_{h\in\mathcal{H}}\hat{R}(h)$, by the above theorems, we obtain

$$\sum_t \mathbb{E}_t[\gamma(\hat{h}_{erm},s)] \leq \sum_t\sum_i \rho(\hat{g}_{erm},g') + \mathcal{O}(\frac{1}{\sqrt{EN}}).$$

Here, we omit some symbols without loss of generality. Comparing the above bounds for $\gamma(\hat{h}_{erm},s)$ and $\gamma(\hat{h},s)$, in theory, we cannot determine which bound is tighter, i.e., which of the minimizers $\hat{h}$ and $\hat{h}_{erm}$ performs closer to $s$. This suggests that we CANNOT ascertain **whether** $\min\hat{R}_\lambda(h)$ **(which can be extended to encompass various DG methods) has a theoretical advantage over ERM in reducing the generalization discrepancy** $\gamma(h,s)$. Even if the given data satisfy Assumption **(iii)**, it is unclear under which conditions the assumed optimal hypothesis $s$ satisfies $N(\cdot)$, i.e., lacking related justification. By this, we recognize that, under the general DG setting, merely pursuing algorithmic-level improvements, i.e., designing $N(\cdot)$, cannot lead to a genuine breakthrough, i.e., $\mathbb{E}_T[\gamma(\cdot,s)] \leq \kappa + \mathcal{O}(1/\sqrt{EN})$, where $\kappa \to 0$ may represent a residual bias. If we aim to achieve such a breakthrough, it is necessary to analyze the properties of $s$, i.e., conduct an in-depth analysis of the given data, while simultaneously designing $N(\cdot)$. We then complete our analysis.

### 2.3 Discussion

**(1) Difference in the Processing Order of Problems (i) and (ii)**. By first addressing **(i)** and then **(ii)**, i.e., first performing the decomposition and then introducing $g$, we obtain the following result:

$$d_{JS}(P_t^{Y,h(X)}, P_i^{Y,h(X)}) \leq d_{JS}(P_t^{g(X)}, P_i^{g(X)}) + \frac{1}{\sqrt{2}}\rho(h).$$

This leads to the problem **(ii)** not being sufficiently addressed.

**(2) Mutual Information-based Methods**. For a representation $\hat{Z}$ and a domain indicator (i.e., a random variable) $V$ that is assumed to be uniformly distributed over $\{1, 2, \ldots, E\}$, forcing the discriminator $\phi$ to increase the probability of incorrect predictions will decrease $I(\hat{Z}; V)$. When $I(\hat{Z}; V) = 0$, i.e., $\hat{Z} \perp\!\!\!\perp V$, the distribution discrepancy vanishes among $\{\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_E\}$. Let $\phi$ take values in $V$, and define $P_E(\phi) = E^{-1} \sum_{i \in [E]} P_i(\phi(\hat{Z}) \neq i)$. By Lemma 5, we have

$$I(\hat{Z}; V) \geq \log \left( \frac{E^{1-P_E(\phi)}}{2} \right).$$

The above result implies that increasing the probability of incorrect predictions from $\phi$ such that $I(\hat{Z}; V)$ decreases is theoretically sound. However, in fact, even if $P_E(\phi) = 1.0$, $I(\hat{Z}; V)$ can still be larger than 0. In simple terms, the discriminator may already be nearly guessing blindly, while the distributions are still not properly aligned.

**(3) Analysis of Assumption 1**. In theoretical machine learning, causal representation learning, and domain generalization, this type of assumption is common and mild. We provide the following further explanations: (a) The limit is taken with respect to the induced distributions $P^g$, rather than pointwise or uniform convergence of $g_k \to g$, without imposing unnecessary constraints on the pointwise outputs of $g$. (b) The assumption means that the set of induced distributions $P^{g(X)} : g \in \mathcal{G}$ is separable, with measurable representations forming a dense subset. Many natural function classes (e.g., suitably regularized neural networks) satisfy similar separability properties.

We then provide the following analysis of the necessity of the conclusion in Theorem 4. First, we give the standard formulation of Theorem 4 as follows:

$$A_1 \wedge A_2 \implies C_o$$

Here, $A_1$ is Assumption 1 (the implied separability assumption is excluded from the analysis), $A_2$ denotes $d_{JS}(P_i^g, p) \to 0$, and $C_o$ represents the conclusion $d_{JS}(P_i^g, P_t^g) \to 0$.

In fact, $A_2$ is a condition rather than an assumption; however, it is harmless to treat it as an assumption here. We fix $A_1$ and violate $A_2$ by defining $A_2' := d_{JS}(P_i^g, p) \to \tau \gg 0$. By $A_1$, we have $d_{JS}(P_i^g, p) = d_{JS}(P_t^g, p)$. Whereas, by $A_2'$, we cannot obtain $C_o$. Then, $A_2'$ serves as a counterexample, showing that $A_2$ is necessary. Next, we fix $A_2$ and violate $A_1$ by defining $A_1'$, where we consider $\lim_{k \to \infty} P^{g_k(X)} \neq P^{g(X)}$. More directly, we discard $A_1$. Then, by $A_2$, we only have $d_{JS}(P_t^g, P_i^g) \to d(P_t^g, p)$, thereby preventing us from obtaining $C_o$. This shows that $A_1'$ serves as a counterexample, demonstrating that $A_1$ is necessary. The above demonstrates that both $A_1$ and $A_2$ are important for $C_o$.

Then, we verify whether $A_1$ can be weakened. The sequence $g_k$ is kept unchanged. We first consider restricting $X \in \mathcal{X}$, i.e., we consider a specific or fixed $X$, such that $g_k(X)$ is replaced by $g_k^{(X)}(X)$. This consideration is unrealistic in DG, as we need to consider $X$ from both $D$ and $D_t$; therefore, $X \in \mathcal{X}$ cannot be weakened. Then, we consider replacing the weak convergence $\lim P^{g_k} = P^g$ with $\mathbb{E}[f(g_k(X))] \to \mathbb{E}[f(g(X))],\ \forall f \in \mathcal{F}$. In fact, this weakening is equivalent to $\lim P^{g_k} = P^g$, as we can bound the difference between the expectations by the JS divergence. In summary, the main role of $A_1$ is technical regularization (avoiding non-measurable pathologies and obtaining countable approximation). At this level, the above "weakenings" are usually just equivalent or nearly equivalent reformulations, rather than genuine relaxations. Therefore, it is reasonable to keep the original $A_1$.

## 3 Experiments ∘

*This experimental part is intended solely to provide partial empirical support for our main theoretical contribution and is mainly based on synthetic data, which does not constitute an independent contribution. Readers who are not interested in experimental results may skip this part.*

In this section, we conduct TWO validations: **(A)** the validation of Theorem 4, i.e., the domino effect of marginal matching, and **(B)** the validation of our main argument that marginal matching and invariance learning are necessary for tightening the generalization gap.

Table 1: Marginal matching results ($d_S^{\mathcal{M}}/d_{S,T}^{\mathcal{M}}$). According to these results, the marginal discrepancy is consistently reduced by applying any matching method across both the source and target domains, while LDM achieving the minimum discrepancy.

| | $\mathcal{M}_{COR}$ | $\mathcal{M}_{SLW}$ | $\mathcal{M}_{LDM}$ |
|---|---|---|---|
| ERM | $1.66 \times 10^5$ / $2.45 \times 10^5$ | $1.61 \times 10^3$ / $1.65 \times 10^3$ | 28.14 / 29.70 |
| COR | $2.60 \times 10^3$ / $4.28 \times 10^3$ | $6.71 \times 10^1$ / $8.04 \times 10^1$ | 5.70 / 6.55 |
| SLW | $4.83 \times 10^3$ / $7.93 \times 10^3$ | $7.66 \times 10^1$ / $7.74 \times 10^1$ | 6.25 / 7.22 |
| LDM | $1.07 \times 10^3$ / $1.76 \times 10^3$ | $3.64 \times 10^1$ / $4.77 \times 10^1$ | 4.29 / 4.96 |

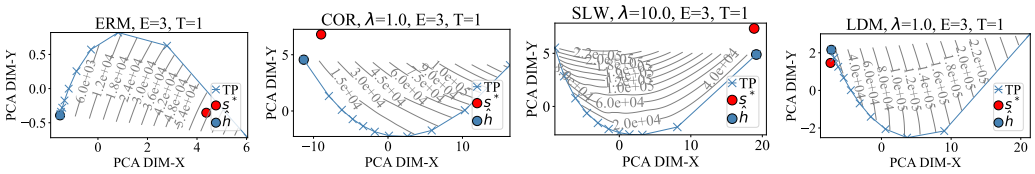


Figure 1: Learning convergence analysis of the learned $\hat{h}$ under different methods. The two-dimensional PCA projections of the optimal $s^*$ and the learned $\hat{h}$ are shown as red and blue points, respectively. TP denotes the evolution of $\hat{h}$ across epochs, while the gray lines represent the corresponding loss values. As observed, matching marginals leads to better convergence.

For **fairness**, all matching and invariance learning methods adopt existing and well-established approaches. Specifically, for marginal matching, we use **SL**iced and radon **W**asserstein (Bonneel et al., 2015), **COR**al (Sun et al., 2016), and **L**atent **D**istribution **M**atching (Mikulasch & Zenke, 2026). For invariance learning, we adopt the well-known IRM (Arjovsky et al., 2019). *Data details, additional experimental results, experimental details, and hyperparameter studies are provided in Section A.4. The main code is provided in the technical supplementary material.*

### 3.1 Synthetic Data Experiments

**Evaluations & Learning Methods**. Since we adopt COR, SLW, and LDM as marginal matching methods, we also use these methods as evaluation metrics, denoted by $\mathcal{M}_{COR}$, $\mathcal{M}_{SLW}$, and $\mathcal{M}_{LDM}$, respectively. Specifically, we aim to evaluate the marginal discrepancies within the source domains and between the source and target domains. We therefore provide the following specific formulation:

$$d_S^{\mathcal{M}} = \frac{2}{E(E-1)} \sum_{i=1}^{E} \sum_{j=i+1}^{E} \mathcal{M}(P_i^{g(X)}, P_j^{g(X)}),\ d_{S,T}^{\mathcal{M}} = \frac{1}{EE_t} \sum_{i=1}^{E} \sum_{t=1}^{E_t} \mathcal{M}(P_i^{g(X)}, P_t^{g(X)}).$$

Here, $\mathcal{M}$ denotes the matching method, $d_S$ denotes the discrepancy among the source domains, and $d_{S,T}$ denotes the discrepancy between the source and target domains.

For the learning methods, we use ERM as the baseline without any matching method, and compare it with COR: ERM+$\lambda\mathcal{M}_{COR}$, SLW: ERM+$\lambda\mathcal{M}_{SLW}$, and LDM: ERM+$\lambda\mathcal{M}_{LDM}$. *Note that $d_S$ is a metric defined only over the source domains. For learning, we match $P_i^{g(X)}$ to a standard Gaussian distribution $p$, i.e.,* $\min_g \mathcal{M}(P_i^{g(X)}, p),\ \forall i \in [E]$.

**Results**. Here, we only provide the results for the dataset with $E = 3$ and $E_t = 1$. First, we present the direct marginal matching evaluations in Table 1. *The results demonstrate that both $d_S$ and $d_{S,T}$ decrease when applying $\mathcal{M}$, which is consistent with the conclusion of Theorem 4. This consistency further demonstrates the mildness and soundness of Assumption 1.* Additionally, we provide a learning convergence analysis, as shown in Figure 1. To explain these results, we give the following bound:

$$\frac{1}{N} \sum_{i=1}^{E} \mathbb{E}_i[\gamma(h, s^*)] \leq \frac{2\sqrt{2}M}{E} \sum_{i=1}^{E} d_{JS}(P_i^{Y,g(X)}, P_i^{Y,\Phi(X)}).$$

Table 2: Prediction accuracy on the testing data of CMNIST obtained by different learning methods. Here, $s_1$ and $s_2$ denote the two source domains, and $t_1$ denotes one target domain. "Avg." denotes the mean accuracy across all domains, while "Oracle" represents the results obtained by ERM trained on gray MNIST. The optimal accuracy is $75.0\%$ due to the $25.0\%$ label-shifted noise.

| | $s_1$ | $s_2$ | $t_1$ | Avg. |
|---|---|---|---|---|
| ERM | 87.1±1.0 | 83.4±1.5 | 21.9±6.0 | 64.1±2.8 |
| ELDM | 69.5±2.6 | 67.5±3.5 | 42.3±3.4 | 59.7±3.2 |
| IRM | 66.4±6.5 | 65.7±5.9 | 39.6±6.3 | 57.2±6.2 |
| ILDM | 72.0±3.0 | 68.8±3.1 | 43.3±1.0 | 61.4±2.4 |
| Oracle | 74.1±0.9 | 74.8±1.2 | 70.7±0.6 | 73.2±0.9 |
| Optimal | 75.0 | 75.0 | 75.0 | 75.0 |

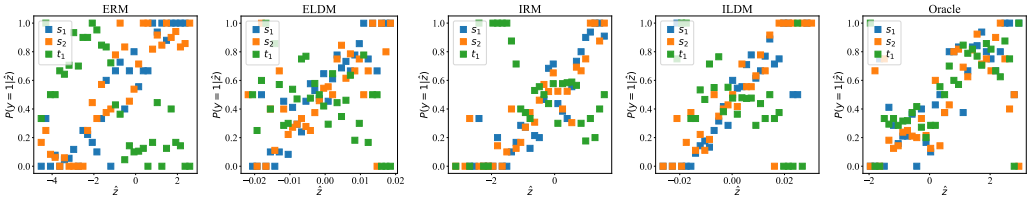


Figure 2: $P(y=1|\hat{g})$ as function of $\hat{g}$ for different $\hat{h}$ trained on CMNIST. As observed, invariance is gradually learned from ERM to ILDM.

For ERM, $\hat{h}_{erm}$, the right-hand side remains unchanged. For the other methods, $\hat{h}$, the right-hand side becomes smaller as $d_{JS}(P_i^{\hat{g}}, p) \to 0$. Since the given data satisfy Assumption **(iii)** and we know $\Phi$ is an invariant representator, using the above inequality for explanation is suitable. Then, validation **(A)** is complete.

### 3.2 Semi-Synthetic Data Experiments

**Learning Methods**. We take ERM and IRM: ERM + $\lambda_2 Inv$ as the baselines. Since the previous experiments showed that LDM achieved good matching, we construct two methods: ELDM: ERM+$\lambda_1 \mathcal{M}_{LDM}$, and ILDM: IRM + $\lambda_1 \mathcal{M}_{LDM}$.

**Results**. We exhibit the prediction accuracy in Table 2 and $P(y=1|g)$ in Figure 2. From Table 2, it seems that $\hat{h}$ learned by $\hat{R}_\lambda(h)$ (DG methods) performs well compared to ERM. However, the improvement in $t_1$ is not significant when comparing the results from DG methods, which aligns with our final analysis. Nevertheless, from the results in Figure 2, we find that, for $t_1$, the learned invariance gradually concentrates toward the middle from the surroundings, in the order from ERM to ILDM. More interestingly, the results of ILDM seem like the sum of those of ELDM and IRM, and the "Avg" distance to the oracle is approached. This demonstrates that marginal matching and invariance learning complement each other and are both necessary, completing the validation of **(B)**.

## 4 Conclusion

In this work, we theoretically analyze the nature of DG, i.e., how to mitigate the generalization gap and other issues. Although we can theoretically obtain a tighter upper bound for $\epsilon_T(h)$ in the idealized setting, obtaining a tighter upper bound for $\gamma(h, s)$ remains challenging in practical scenarios. This limitation is closely related to the in-depth analysis of the given data, which, as we argue, is important for DG. The deeper the analysis, the tighter the bound for $\gamma(h, s)$ will be. Our analysis relies on the assumptions that all functions are measurable and that the loss function is bounded. These assumptions constitute a limitation of our work when they do not hold. A natural direction for future work is to relax these assumptions while preserving the main theoretical guarantees. Specifically, we aim to extend our analysis to unbounded loss functions and more general function classes, thereby broadening the applicability of our theoretical framework.

## AI USE STATEMENT

In this work, we used generative AI tools solely to edit and polish the language of the abstract, introduction, and related works in order to improve readability. We have not used generative AI tools for any other tasks with required disclosure (including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting with proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, interpreting results, or any other research ideation or technical content generation). These tasks are not applicable to this work, as all theorems, proofs, and experiments were developed and conducted entirely by the authors without any assistance from generative AI tools. We have carefully reviewed all AI-assisted text. We take full responsibility for the final content of this work.

## ETHICS STATEMENT

This work does not involve any human subjects, sensitive data, potentially harmful insights or applications, fairness/bias concerns, privacy or security issues, or other ethical risks. All authors have read and adhere to the ICLR Code of Ethics. We confirm that this research complies fully with the ICLR Code of Ethics.

## REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, all theorem derivations and complete proofs are provided in Section A.3. The key source code for the experiments is included in the supplementary materials. Additional implementation details and experimental settings are described in the main text and appendix.

## REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020. doi: 10.1214/19-STS721.

Ziliang Chen, Yongsen Zheng, Zhao-Rong Lai, Quanlong Guan, and Liang Lin. Diagnosing and rectifying fake ood invariance: A restructured causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11471–11479, 2024.

Moulik Choraria, Ibtihal Ferwana, Ankur Mani, and Lav R. Varshney. Learning optimal features via partial invariance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7175–7183, Jun. 2023.

Yuxin Dong, Tieliang Gong, Hong Chen, Shuangyong Song, Weizhan Zhang, and Chen Li. How does distribution matching help domain generalization: An information-theoretic analysis. *IEEE Transactions on Information Theory*, 71(3):2028–2053, 2025. doi: 10.1109/TIT.2025.3531136.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Pascal Janetzky, Tobias Schlagenhauf, and Stefan Feuerriegel. Continual learning of domain-invariant representations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Fabian A. Mikulasch and Friedemann Zenke. Understanding self-supervised learning via latent distribution matching. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Hiroshi Morioka and Aapo Hyvarinen. Causal representation learning made identifiable by grouping of observational variables. In *Forty-first International Conference on Machine Learning*, 2024.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pp. 10–18. PMLR, 2013.

Toan Nguyen, Kien Do, Bao Duong, and Thin Nguyen. Domain generalisation via risk distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2790–2799, 2024.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. In *International Conference on Learning Representations (ICLR)*, 2023.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2020.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 2058–2065, 2016.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *International Conference on Machine Learning*, pp. 23018–23033. PMLR, 2022.

Xi Wang, Liang Bai, Xian Yang, Richard Yi Da Xu, and Jiye Liang. Bridging domain invariance and diversity: A fine-grained risk bound for domain generalization. *Journal of Machine Learning Research*, 27(137):1–54, 2026.

Yimu Wang, Yihan Wu, and Hongyang Zhang. Lost domain generalization is a natural consequence of lack of training domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15689–15697, 2024.

Danru Xu, Dingling Yao, Sebastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *Forty-first International Conference on Machine Learning*, 2024.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *Forty-first International Conference on Machine Learning*, 2024.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Han Zhao, Chen Dan, Bryon Aragam, Tommi S Jaakkola, Geoffrey J Gordon, and Pradeep Ravikumar. Fundamental limits and tradeoffs in invariant representation learning. *The Journal of Machine Learning Research*, 23(1):15356–15404, 2022.

Hong Zheng and Fei Teng. Distribution shift is key to learning invariant prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):28812–28820, Mar. 2026.

# A Appendix

## A.1 Related Works

In this section, we briefly introduce related research on invariance learning-based methods, representation distribution matching-based methods, and theoretical works. For further details, we refer the reader to the corresponding citations.

**Invariance Learning Methods**. Since the early development of domain generalization (DG) (Muandet et al., 2013), learning invariant representations has been regarded as a promising approach to addressing distribution shift. The key intuition is that, despite varying data distributions, certain properties of the data-generating process remain invariant, enabling models to retain predictive information while discarding domain-specific variations. Causality provides a closely related perspective on invariance in DG. Peters et al. (2016) argued that causal relationships constitute a fundamental source of invariance. However, causality and invariance should not be treated as synonymous, as invariance may arise from sources beyond causality (Bühlmann, 2020). More explicitly, Wang et al. (2026) characterized invariance through three distinct components, with causality representing only one of them. Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) is a representative approach that learns a representation and classifier whose optimal predictor remains invariant across training environments. This formulation has inspired various extensions, including Partial IRM (Chorariа et al., 2023) and IRM-games (Ahuja et al., 2020), as well as studies on alternative notions of invariant risk (Rosenfeld et al., 2020). In parallel, causal representation learning approaches aim to disentangle causal factors from domain-specific or spurious factors to improve robustness under distribution shifts (Schölkopf et al., 2021; Zhang et al., 2024; Morioka & Hyvarinen, 2024; Xu et al., 2024). Information-theoretic approaches based on the Information Bottleneck (IB) principle seek to retain predictive information while discarding irrelevant or domain-specific information, with invariant IB methods incorporating invariance into this compression process (Li et al., 2022; Ahuja et al., 2021).

**Distribution Matching Methods**. In fact, representation distribution matching can be viewed as one of a family of approaches for learning invariance from observed data. We discuss these methods separately because they are particularly relevant to and have directly inspired our work. Representative methods explicitly match representation distributions across source domains, often using generative

models such as GANs (Sun et al., 2016), with some approaches focusing on conditional distributions to reduce domain-specific discrepancies while preserving task-relevant information (Zhao et al., 2018; Pham et al., 2023). These methods are typically motivated by information-theoretic principles, particularly mutual information. Beyond distribution matching, gradient matching provides another perspective for learning cross-domain invariance Shi et al. (2022), seeking to reduce discrepancies between optimization signals across source domains rather than directly aligning representation distributions (Nguyen et al., 2024; Eastwood et al., 2022). More recently, Dong et al. (2025) argued that distribution matching and gradient matching capture complementary aspects of domain invariance and proposed a mutual-information-based framework that jointly considers both distribution and gradient matching to characterize cross-domain consistency more comprehensively.

**Theoretical Works**. Turning to the theoretical foundations of learning across domains, Ben-David et al. (2010) established a seminal theoretical framework for domain adaptation (DA) by relating target-domain error to source-domain error, the discrepancy between source and target distributions, and the error of the ideal joint hypothesis. This result provides a fundamental basis for understanding the role of distribution discrepancy in cross-domain generalization. Compared with DA, domain generalization (DG) must infer transferable structures solely from observed source domains, raising fundamental questions about the feasibility and limitations of invariant learning from a finite set of domains. In this regard, Chen et al. (2024) and Zhao et al. (2022) showed that enforcing invariance across observed source domains does not necessarily guarantee generalization to unseen domains, since multiple predictors or representations may satisfy the same invariance constraints while behaving differently on unseen domains. Further, Wang et al. (2024) established a lower bound on the number of training domains required for reliable DG, showing that insufficient domain diversity cannot generally be compensated for by increasing the number of samples within each domain. Conversely, Zheng & Teng (2026) demonstrated that distribution shift can itself facilitate invariant prediction under certain conditions, as greater heterogeneity across source domains can help ERM distinguish stable predictive relationships from domain-specific correlations.

### A.2 LEMMAS

**Lemma 1** (Expectation Upper Bound via Jansen-Shannon divergence)**.** *Let $P$ and $Q$ be two probability distributions. Let $F : \mathcal{V} \to \mathbb{R}$ be a measurable function satisfying*

$$\|F\|_\infty := \sup_{V \in \mathcal{V}} |F(V)| \leq M \leq \infty.$$

*Then,*

$$|\mathbb{E}_P[F] - \mathbb{E}_Q[F]| \leq 2M\sqrt{2JS(P^V||Q^V)},$$

*or*

$$|\mathbb{E}_P[F] - \mathbb{E}_Q[F]| \leq 2\sqrt{2}Md_{JS}(P^V, Q^V).$$

*Proof.* The result follows from standard inequalities in information theory. First, by a standard total variation bound,

$$|\mathbb{E}_P[F] - \mathbb{E}_Q[F]| \leq 2\,\|F\|_\infty\, d_{TV}(P, Q).$$

Applying the Pinsker's inequality, we have

$$\left(\frac{1}{2}d_{TV}(P, Q)\right)^2 \leq \frac{1}{2}KL(P||M)$$
$$\left(\frac{1}{2}d_{TV}(P, Q)\right)^2 \leq \frac{1}{2}KL(Q||M).$$

Here, $M = (P + Q)/2$, $d_{TV}(P, M) = (1/2)d_{TV}(P, Q)$, and $d_{TV}(Q, M) = (1/2)d_{TV}(P, Q)$. Then, by the definition of the JS divergence, i.e., $JS(P; Q) = 1/2KL(P; M) + 1/2KL(Q; M)$, we obtain

$$2\left(\frac{1}{2}d_{TV}(P, Q)\right)^2 \leq \frac{1}{2}JS(P||Q).$$

In this work, $V$ and $F$ are regarded as $(h, \xi)$ and $\ell$, respectively. □

**Lemma 2** (Hoeffding Inequality under Non-i.i.d. Settings)**.** *Let $V_1, \ldots, V_n$ be n independent random variables with $V_i \in [a_i, b_i]$ for each $i \in [n]$. Define $S_n = \sum_{i=1}^{n} V_i$. Then for any $\varepsilon > 0$,*

$$P\left\{S_n - \mathbb{E}[S_n] \geq \varepsilon\right\} \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

*Proof.* This is a classical inequality, and thus we omit the proof here. For more details, please refer to (Boucheron et al., 2003).

For clarity on how this lemma is applied in our case, we provide the following simple example. Consider $V_i = \ell(h, \xi_i)$ as the $i$-th random variable and $S_E = \sum_{i=1}^{E} V_i$, and assume that $V_i \in [0, K_i]$. Then, based on this inequality, we have

$$P\left\{S_E - \mathbb{E}[S_E] \geq \varepsilon\right\} \leq \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^{E} K_i^2}\right\}.$$

□

**Lemma 3** (Properties of JS Divergences)**.** *For arbitrary random variables $A$ and $B$ with distributions $P$ and $Q$, respectively, we have*

- *Conditional Divergence:*

  $$JS(P^{B|A}||Q^{B|A}|P^A) = \mathbb{E}_{a\sim P}\left[JS(P^{B|A=a}||Q^{B|A=a})\right]$$

  *and*

  $$JS(P^{B|A}||Q^{B|A}|Q^A) = \mathbb{E}_{a\sim Q}\left[JS(P^{B|A=a}||Q^{B|A=a})\right].$$

- *Chain Rule:*

  $$JS(P^{BA}||Q^{BA}) \leq JS(P^A||Q^A) + \frac{1}{2}\left(JS(P^{B|A}||Q^{B|A}|P^A) + JS(P^{B|A}||Q^{B|A}|Q^A)\right).$$

- *Data Processing:*

  $$B = \phi(A) \Rightarrow JS(P^B||Q^B) \leq JS(P^A||Q^A),$$

  *with equality if $\phi$ is* 1-1.

*Proof.* Assuming $M = \frac{1}{2}(P + Q)$, we use the properties of KL divergence to derive our results. The proofs of the KL divergence properties used in our derivations can be found in (Polyanskiy & Wu, 2014), specifically Theorem 2.2 and its proof.

- By the definition of conditional KL divergence, we have

  $$KL(P^{B|A}||M^{B|A}|P^A) = \mathbb{E}_{a\sim P}[KL(P^{B|A=a}||M^{B|A=a})]$$

  and

  $$KL(Q^{B|A}||M^{B|A}|P^A) = \mathbb{E}_{a\sim P}[KL(Q^{B|A=a}||M^{B|A=a})].$$

  Similarly, we can also obtain

  $$KL(P^{B|A}||M^{B|A}|Q^A) = \mathbb{E}_{a\sim Q}[KL(P^{B|A=a}||M^{B|A=a})]$$

  and

  $$KL(Q^{B|A}||M^{B|A}|Q^A) = \mathbb{E}_{a\sim Q}[KL(Q^{B|A=a}||M^{B|A=a})].$$

  Then, by the definition of JS divergence, the proof of this part is completed.

- Applying the chain rule of KL divergence, we have

$$\begin{aligned}
JS(P^{BA}||Q^{BA}) =& \frac{1}{2}KL(P^{B|A}||M^{B|A}|P^A) + \frac{1}{2}KL(P^A||M^A) \\
&+ \frac{1}{2}KL(Q^{B|A}||M^{B|A}|Q^A) + \frac{1}{2}KL(Q^A||M^A) \\
=& JS(P^A||Q^A) + \frac{1}{2}KL(P^{B|A}||M^{B|A}|P^A) \\
&+ \frac{1}{2}KL(Q^{B|A}||M^{B|A}|Q^A) \\
\leq& JS(P^A||Q^A) + \frac{1}{2}KL(P^{B|A}||M^{B|A}|P^A) \\
&+ \frac{1}{2}KL(Q^{B|A}||M^{B|A}|P^A) \\
&+ \frac{1}{2}KL(Q^{B|A}||M^{B|A}|Q^A) \\
&+ \frac{1}{2}KL(P^{B|A}||M^{B|A}|Q^A) \\
=& JS(P^A||Q^A) + \frac{1}{2}JS(P^{B|A}||Q^{B|A}|P^A) \\
&+ \frac{1}{2}JS(P^{B|A}||Q^{B|A}|Q^A)
\end{aligned}$$

- Applying the data processing inequality of KL divergence, we have

$$KL(P^B||M^B) \leq KL(P^A||M^A)$$

and

$$KL(Q^B||M^B) \leq KL(Q^A||M^A).$$

Then, according to the definition of JS divergence, the proof of this part is completed.

In this work, $P$ and $Q$ are considered as $P_i$ and $P_t$ for any $i, t \in [E + E_t]$, respectively, and $A$ and $B$ can be considered as arbitrary variables in $\{X, \hat{Z}, \hat{Y}, Y\}$. □

**Lemma 4** (Expectation Lower Bound via JS divergence)**.** *Let $P$ and $Q$ be two probability distributions. Let $F : \mathcal{V} \to \mathbb{R}$ be a measurable function satisfying*

$$\|F\|_\infty := \sup_{V \in \mathcal{V}} |F(V)| \leq M \leq \infty.$$

*Then,*

$$\sup_F |\mathbb{E}_P[F] - \mathbb{E}_Q[F]| \geq \frac{2M}{\ln 2} JS(P^V||Q^V).$$

*Proof.* Let the total variation

$$\delta(P, Q) = \frac{1}{2} \int |dP - dQ|,$$

then we have

$$\sup_F |\mathbb{E}_P[F] - \mathbb{E}_Q[F]| \geq 2M\delta(P, Q).$$

By the classical inequality between the Jensen–Shannon divergence and the total variation

$$JS(P||Q) \leq \ln 2 \cdot \delta(P, Q),$$

the proof is complete. □

**Lemma 5** (Fano's Lemma). *Let $S = \{P_1, \ldots, P_m\}$ be a finite collection of probability measures on $\mathcal{X}$. Let the random variable $V$ be uniformly distributed on $\{1, \ldots, m\}$, and conditional on $V = i$, let $X \sim P_i$. Then for any estimator $\phi = \phi(X)$ taking values in $\{1, \ldots, m\}$,*

$$m^{-1} \sum_{i=1}^{m} P_i \left(\phi \neq i\right) \geq 1 - \frac{I(X;V) + \log(2)}{\log(m)}.$$

*Proof.* This is a classical form of Fano's inequality in information theory (see the proof on page 39 in (Polyanskiy & Wu, 2014)). It characterizes a fundamental difficulty in distribution estimation: as the discrepancy between distributions decreases, reflected by a decrease in the mutual information $I(X;V)$, the estimation error necessarily increases. □

### A.3 PROOFS

*Proof of Theorem 1 and Corollary 1.* According to Lemma 1, for any $h$, and any $P_i$ and $Q_j$, we have

$$\left|\mathbb{E}_{P_i}[\ell] - \mathbb{E}_{Q_j}[\ell]\right| \leq 2\sqrt{2} M d_{JS}(P_i^{Y,\hat{Y}}, Q_j^{Y,\hat{Y}}).$$

Extending this inequality to all source and target domains, we obtain

$$\left| \frac{1}{E_t} \sum_t \mathbb{E}_{P_t}[\ell] - \frac{1}{E} \sum_i \mathbb{E}_{P_i}[\ell] \right| \leq C \sum_t \sum_i d_{JS}(P_t^{Y,h(X)}, P_i^{Y,h(X)}),$$

where $C = (2\sqrt{2}M)/EE_t$. By Lemma 3, the data processing inequality, considering $B = (Y, \hat{Y})$ and $A = (Y, \hat{Z})$, and by $X \to \hat{Z} \to \hat{Y}$, we obtain

$$d_{JS}(P_t^{Y,h(X)}, P_i^{Y,h(X)}) \leq d_{JS}(P_t^{Y,g(X)}, P_i^{Y,g(X)}).$$

Then, the proof is completed by applying the inequality $|a| - |b| \leq |a - b|$, and the proof of the lower bound follows directly from Lemma 4. □

*Proof of Theorem 2.* By defining $V_i = \ell(h, (x_l, y_l)_i)$, it follows that

$$S_{EN} = \sum_{i=1}^{E} \sum_{l=1}^{N} V_i.$$

Then, under the assumption $0 \leq \ell \leq M$, by Lemma 2, we have

$$P\left\{\hat{R}(h) - R(h) \geq \varepsilon\right\} \leq \exp\left\{-\frac{2(EN\varepsilon)^2}{ENM^2}\right\},$$

completing the proof. □

*Proof of Theorem 3.* By Lemma 3, the chain rule, for any domains $t$ and $i$, we have

$$JS(P_t^{Y,\hat{Z}} || P_i^{Y,\hat{Z}}) \leq A + \frac{1}{2}\left(B + C\right).$$

Here, $A := JS(P_t^{\hat{Z}} || P_i^{\hat{Z}})$, $B := JS(P_t^{Y|\hat{Z}} || P_i^{Y|\hat{Z}} | P_t^{\hat{Z}})$, and $C := JS(P_t^{Y|\hat{Z}} || P_i^{Y|\hat{Z}} | P_i^{\hat{Z}})$.

Since $d_{JS}(\cdot,\cdot) = \sqrt{JS(\cdot||\cdot)}$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, it follows that

$$\sqrt{JS(P_t^{Y,\hat{Z}} || P_i^{Y,\hat{Z}})} \leq \sqrt{A} + \frac{1}{\sqrt{2}}\sqrt{B + C}.$$

By the conditional divergence, the proof is complete. □

*Proof of Theorem 4.* For simplicity, let $P_t$ and $P_i$ be denoted by $P_t^{g(X)}$ and $P_i^{g(X)}$, respectively, and let $d$ denote $d_{JS}$. By the triangle inequality of the JS distance, we have

$$d(P_t, P_i) \leq d(P_i, p) + d(p, P_t)$$
$$|d(P_t, P_i) - d(P_t, p)| \leq d(P_i, p).$$

This implies that $d(P_i, p) \to 0$ leads to $|d(P_t, P_i) - d(P_t, p)| \to 0$, i.e., $d(P_t, P_i) \to d(P_t, p)$. Similarly, by the triangle inequality for the JS distance, for an intermediate distribution $q$, we also have

$$d(P_t, p) \leq d(P_t, q) + d(q, P_i) + d(P_i, p)$$
$$d(P_i, p) \leq d(P_i, q) + d(q, P_t) + d(P_t, p).$$

Then, let $q$ be $P^{g_k}$ and consider the limit as $k \to \infty$. According to Assumption 1, for any $X \in D, D_t$, it follows that $P^{g_k(X)} = P^{g(X)}$. Consequently, we obtain

$$d(P_t, p) = d(P_i, p).$$

Unless $p = P^g$, i.e., $p$ is induced by $g$, the converse theorem does not hold. Since we need to satisfy Assumption 1, i.e., $d(P^{g_k}, p) \to 0$. □

*Proof of Theorem 5 and Corollary 2.* We begin with the following

$$\begin{aligned}|\mathbb{E}_T[\gamma(h_1, h_2)] - \mathbb{E}_S[\gamma(h_1, h_2)]| &= |\mathbb{E}_T[\ell(h_1)] - \mathbb{E}_T[\ell(h_2)] - \mathbb{E}_S[\ell(h_1)] + \mathbb{E}_S[\ell(h_2)]| \\ &= |(\mathbb{E}_T[\ell(h_1)] - \mathbb{E}_S[\ell(h_1)]) + (\mathbb{E}_S[\ell(h_2)] - \mathbb{E}_T[\ell(h_2)])| \\ &\leq |\mathbb{E}_T[\ell(h_1)] - \mathbb{E}_S[\ell(h_1)]| + |\mathbb{E}_S[\ell(h_2)] - \mathbb{E}_T[\ell(h_2)]|,\end{aligned}$$

where the last inequality follows from the triangle inequality $|a + b| \leq |a| + |b|$.

By $|a| - |b| \leq |a - b|$, we obtain

$$|\mathbb{E}_T[\gamma(h_1, h_2)]| \leq |\mathbb{E}_S[\gamma(h_1, h_2)]| + |\mathbb{E}_T[\ell(h_1)] - \mathbb{E}_S[\ell(h_1)]| + |\mathbb{E}_S[\ell(h_2)] - \mathbb{E}_T[\ell(h_2)]|.$$

Now, we consider these terms on the RHS one by one. For $\mathbb{E}_S[\gamma(h_1, h_2)]$, by Lemma 2, the following inequality holds with probability at least $1 - \delta$:

$$\mathbb{E}_S[\gamma(h_1, h_2)] \leq \sum_S \gamma(h_1, h_2) + M\sqrt{\frac{2\ln(1/\delta)}{EN}},$$

where $\sum_S$ denotes the empirical risk for brevity, and the coefficient 2 for $\ln$ is due to $-M \leq \gamma \leq M$.

By Corollary 1, we have

$$|\mathbb{E}_T[\ell(h)] - \mathbb{E}_S[\ell(h)]| \leq C \sum_S \sum_T d_{JS}(P_T^{Y,g(X)}, P_S^{Y,g(X)}).$$

Consequently, we obtain

$$\begin{aligned}|\mathbb{E}_T[\gamma(h_1, h_2)]| \leq &\left|\sum_S \gamma(h_1, h_2)\right| + C \sum_S \sum_T d_{JS}(P_T^{Y,g_1(X)}, P_S^{Y,g_1(X)}) \\ &+ C \sum_S \sum_T d_{JS}(P_T^{Y,g_2(X)}, P_S^{Y,g_2(X)}) + M\sqrt{\frac{2\ln(1/\delta)}{EN}},\end{aligned}$$

where $h_1 = f_1 \circ g_1$ and $h_2 = f_2 \circ g_2$. Since $\sum_S \gamma(h_1, h_2) = \hat{R}(h_1) - \hat{R}(h_2)$, we have

$$\begin{aligned}\left|\hat{R}(h_1) - \hat{R}(h_2)\right| &= \left|\hat{R}(h_1) + \hat{R}_\lambda(h_1) - \hat{R}_\lambda(h_1) - \hat{R}(h_2)\right| \\ &\leq \left|\hat{R}(h_1) + \hat{R}_\lambda(h_2) - \hat{R}_\lambda(h_1) - \hat{R}(h_2)\right| \\ &= \left|\hat{R}(h_1) - \hat{R}(h_1) - \lambda N(g_1) + \hat{R}(h_2) + \lambda N(g_2) - \hat{R}(h_2)\right| \\ &= \lambda\,|N(g_2) - N(g_1)|\end{aligned}$$

where the first $\leq$ follows from the assumption that $\hat{R}_\lambda(h_1) \leq \hat{R}_\lambda(h_2)$. Then, let $h_1 = \hat{h}$ and let $h_2$ be $s_e,\ \forall e \in [E + E_t]$, we obtain

$$\begin{aligned}\left|\mathbb{E}_T[\gamma(\hat{h}, s_T)]\right| \leq & \lambda \sum_S |N(g'_S) - N(\hat{g})| + C \sum_S \sum_T d_{JS}(P_T^{Y,\hat{g}(X)}, P_S^{Y,\hat{g}(X)}) \\ & + C \sum_S \sum_T d_{JS}(P_T^{Y,g'_T(X)}, P_S^{Y,g'_S(X)}) + M\sqrt{\frac{2\ln(1/\delta)}{EN}},\end{aligned}$$

and the proof is complete. □

### A.4 Experiments

**Synthetic Data**. The data pair $(X, Y)$ is generated as follows, consistent with the ideal in (Peters et al., 2016). Specifically, we consider invariant features to generate the response $Y$ and environmental interventions on $X$:

$$\begin{aligned}&X \sim \mathcal{N}(0, e),\ Z_{inv} = X\Phi, \\ &Y = \mathbf{1}^{\mathbf{T}} Z_{inv} + \varepsilon_w,\ \varepsilon_w \perp Z_{inv}.\end{aligned}$$

Here, $X \in \mathbb{R}^{N\times 2d}$, $Z_{inv} \in \mathbb{R}^{N\times d}$, $\Phi \in \mathbb{R}^{2d\times d}$, $\mathbf{1}^{\mathbf{T}} \in \mathbb{R}^{1\times d}$, $\varepsilon_w$ is Gaussian white noise that is orthogonal to $Z_{inv}$, and $e$ represents the domain intervention on the variance of the Gaussian distributions. In these experiments, we set $N = 1000$ and $d = 20$, and let $e$ take values in $\{1.0, 2.0, 3.0, \ldots, E+E_t\}$. For example, when $E = 3$ and $E_t = 1$, we have $e \in \{1.0, 2.0, 3.0, 4.0\}$, where $e \in \{4.0\}$ corresponds to $D_t$, and $e \in \{1.0, 2.0, 3.0\}$ corresponds to $D$. Since we aim for $Z_{inv}$ to represent the invariant features, we keep $\Phi$ fixed across domains, i.e., $\Phi$ serves as an invariant representator. In fact, $\mathbf{1}$ is a predictor, which we let take this simple form without loss of generality, and we denote the optimal model by $s^* = \mathbf{1} \circ \Phi$. Note that this $\Phi$ cannot match the representation marginals, i.e., $P^{\Phi(X_e)} \neq P^{\Phi(X_{e'})}$ for any $e \neq e' \in [E + E_t]$. We adopt this setting to align with the settings of most DG methods, i.e., only Assumption **(iii)** is satisfied. Specifically, we aim to conduct experiments on general synthetic data, thereby providing experimental evidence that is more representative.

**Semi-Synthetic Data**. Colored MNIST (Arjovsky et al., 2019) is often used to check whether a model has learned invariant representations by plotting $P(y = 1|g)$; thus, we adopt this dataset and follow the same validation procedure. This dataset applies color interventions and label shifts to the original MNIST data. The color intervention variable $v$ takes values in $\{0.1, 0.2, 0.9\}$, where a smaller value indicates a stronger correlation between labels and colors. Then, the labels are shifted by $25\%$ to introduce label noise. There are two source domains with $v = 0.1$ and $v = 0.2$, respectively, and one target domain with $v = 0.9$. In this experiment, due to our limited computational resources, we use only $10\%$ of the original data, which may lead to accuracy evaluations that are not SOTA; however, this does not affect our main results.

**Experimental Conditions**: The entire set of experiments was performed on a single Nvidia RTX 4090 24 GB GPU, utilizing the Tensorflow-gpu 2.6.0 deep learning library.

**Learning Objective on Synthetic Data**. The learning objective is

$$\min_{\mathbf{1}\circ g \in \mathcal{G}} \hat{R}(\mathbf{1} \circ g) + \lambda \mathcal{M}(P^g, p).$$

Here, $p$ is a standard Gaussian distribution, i.e., $p \sim \mathcal{N}(0, 1)$. COR, SLW, and LDM denote different choices of $\mathcal{M}$.

**Additional Synthetic Experimental Results** . Here, we provide additional results on different datasets with $E \in \{3, 5, 7\}$ and $E_t \in \{1, 3, 5, 7\}$. The results are presented in Tables 3, 4, and 5. The purpose of this experiment is to examine whether the decrease in marginal discrepancy in the target domains is affected by the magnitude of $E$, i.e., whether the decrease occurs when $E \gg E_t$ but does not occur when $E \ll E_t$. In other words, we aim to determine whether our theory is affected by the numbers of source and target domains. Through this experiment, we observe that the decrease occurs regardless of the numbers of source and target domains. Therefore, we believe that our theory can be adopted as a fundamental theory for machine learning.

Table 3: Marginal matching results when $E = 3$ and $E_t \in \{1, 3, 5, 7\}$. As observed, regardless of the number of target domains, matching source marginals reduces the discrepancy between target marginals.

| | | $d^{\mathcal{M}}_{S(E=3)}$ | $d^{\mathcal{M}}_{S,T(E_t=1)}$ | $d^{\mathcal{M}}_{S,T(E_t=3)}$ | $d^{\mathcal{M}}_{S,T(E_t=5)}$ | $d^{\mathcal{M}}_{S,T(E_t=7)}$ |
|---|---|---|---|---|---|---|
| $\mathcal{M}_{COR}$ | ERM | $1.66 \times 10^5$ | $2.45 \times 10^5$ | $3.95 \times 10^5$ | $5.63 \times 10^5$ | $9.04 \times 10^5$ |
| | ELDM | $1.07 \times 10^3$ | $1.76 \times 10^3$ | $2.46 \times 10^3$ | $4.74 \times 10^3$ | $7.06 \times 10^3$ |
| $\mathcal{M}_{SLW}$ | ERM | $1.61 \times 10^3$ | $1.65 \times 10^3$ | $2.19 \times 10^3$ | $2.53 \times 10^3$ | $3.21 \times 10^3$ |
| | ELDM | $3.64 \times 10^1$ | $4.77 \times 10^1$ | $4.82 \times 10^1$ | $7.00 \times 10^1$ | $8.00 \times 10^1$ |
| $\mathcal{M}_{LDM}$ | ERM | 28.14 | 29.70 | 36.92 | 42.03 | 48.35 |
| | ELDM | 4.29 | 4.96 | 5.92 | 7.25 | 8.26 |

Table 4: Marginal matching results when $E = 5$ and $E_t \in \{1, 3, 5, 7\}$. The conclusion drawn from this table is consistent with those of Tables 3.

| | | $d^{\mathcal{M}}_{S(E=5)}$ | $d^{\mathcal{M}}_{S,T(E_t=1)}$ | $d^{\mathcal{M}}_{S,T(E_t=3)}$ | $d^{\mathcal{M}}_{S,T(E_t=5)}$ | $d^{\mathcal{M}}_{S,T(E_t=7)}$ |
|---|---|---|---|---|---|---|
| $\mathcal{M}_{COR}$ | ERM | $1.97 \times 10^5$ | $3.58 \times 10^5$ | $6.53 \times 10^5$ | $9.04 \times 10^5$ | $9.43 \times 10^5$ |
| | ELDM | $1.12 \times 10^3$ | $1.71 \times 10^3$ | $3.79 \times 10^3$ | $5.50 \times 10^3$ | $5.95 \times 10^3$ |
| $\mathcal{M}_{SLW}$ | ERM | $1.44 \times 10^3$ | $1.61 \times 10^3$ | $2.06 \times 10^3$ | $2.54 \times 10^3$ | $2.79 \times 10^3$ |
| | ELDM | $2.91 \times 10^1$ | $3.65 \times 10^1$ | $4.41 \times 10^1$ | $6.01 \times 10^1$ | $6.27 \times 10^1$ |
| $\mathcal{M}_{LDM}$ | ERM | 26.39 | 30.45 | 36.56 | 42.60 | 45.61 |
| | ELDM | 4.09 | 4.66 | 6.05 | 6.97 | 7.35 |

Moreover, we provide a convergence analysis of the learning of $\hat{h}$ by ERM and LDM on different datasets, as shown in Figures 3 and 4. For the same reason, we aim to exclude the effects of the number of domains from the learning convergence analysis. In simple terms, as the number of source domains increases, we expect ERM to achieve better convergence. However, the convergence behavior remains the same as before, while LDM also maintains the same convergence behavior. Therefore, we can conclude that the number of source domains is irrelevant to the results.

**Hyperparameter Exp. on Synthetic Data**. We exhibit the selection of the hyperparameters for COR, SLW, and LDM in Table 6. Still, LDM matches better than the others even when setting $\lambda = 100.0$. Moreover, we provide the learning convergence analysis of the learned $\hat{h}$ by these methods under the corresponding hyperparameters, as shown in Figure 5. Basically, the gaps are the same, except for SLW when $\lambda = 1.0$. Thus, based on this result, we select $\lambda = 10.0$ as the hyperparameter for SLW, which differs from those used for the other methods. However, we project the original $s^*$ and $\hat{h}$ onto the PCA space, and thus the related results can only serve as a reference. Namely, we need further validation to check whether $s^*$ and $\hat{h}$ approach each other in the space $\mathcal{G}$ if we aim to make this point completely certain. Despite $\lambda = 100.0$ leading to smaller values of the distribution discrepancy and exhibiting similar learning convergence, we did not select it, as the training loss is much higher; see TP in Figure 5.

**ELDM & ILDM**. ELDM is

$$\hat{R}_{\lambda,LDM}(h) = \hat{R}(h) + \lambda_1 LDM(P^g, p),$$

and ILDM is

$$\hat{R}_{\lambda,LDM,Inv}(h) = \hat{R}(h) + \lambda_1 LDM(P^g, p) + \lambda_2 Inv(h),$$

where $Inv(\cdot)$ represents the constraint of IRMv1 in (Arjovsky et al., 2019) and $p$ is a standard Gaussian distribution.

**Hyperparameter Exp. on CMNIST**. We here present the experiments for selecting the hyperparameters for ELDM and IRM, as shown in Tables 7 and 8. We can observe that some accuracies for $t_1$ are much higher than those we selected, but we do not take them into consideration. For binary classification, random guessing can lead to an accuracy of $50\%$; therefore, we consider these results to be meaningless. For IRM, $\lambda = 100.0$ is the original choice. As we only use $10\%$ of the data, we therefore provide new experiments for its hyperparameters.

Table 5: Marginal matching results when $E = 7$ and $E_t \in \{1, 3, 5, 7\}$. The conclusion drawn from this table is consistent with those of Tables 3 and 4.

| | | $d^{\mathcal{M}}_{S(E=7)}$ | $d^{\mathcal{M}}_{S,T(E_t=1)}$ | $d^{\mathcal{M}}_{S,T(E_t=3)}$ | $d^{\mathcal{M}}_{S,T(E_t=5)}$ | $d^{\mathcal{M}}_{S,T(E_t=7)}$ |
|---|---|---|---|---|---|---|
| $\mathcal{M}_{COR}$ | ERM | $2.99 \times 10^5$ | $6.56 \times 10^5$ | $9.75 \times 10^5$ | $1.01 \times 10^6$ | $1.56 \times 10^6$ |
| | ELDM | $1.36 \times 10^3$ | $3.85 \times 10^3$ | $4.75 \times 10^3$ | $5.05 \times 10^3$ | $7.10 \times 10^3$ |
| $\mathcal{M}_{SLW}$ | ERM | $1.49 \times 10^3$ | $1.88 \times 10^3$ | $2.40 \times 10^3$ | $2.87 \times 10^3$ | $3.47 \times 10^3$ |
| | ELDM | $2.89 \times 10^1$ | $4.68 \times 10^1$ | $4.64 \times 10^1$ | $5.60 \times 10^1$ | $6.17 \times 10^1$ |
| $\mathcal{M}_{LDM}$ | ERM | 27.38 | 33.91 | 40.77 | 46.34 | 51.45 |
| | ELDM | 4.20 | 5.75 | 6.36 | 7.00 | 7.54 |

Table 6: Hyperparameter results for the matching penalty $\lambda$, where $*$ denotes the final selection.

| | | $\mathcal{M}_{COR}$ | $\mathcal{M}_{SLW}$ | $\mathcal{M}_{LDM}$ |
|---|---|---|---|---|
| COR | 100.0 | $4.93 \times 10^1$ / $8.19 \times 10^1$ | $1.87 \times 10^1$ / $2.20 \times 10^1$ | 3.15 / 3.49 |
| | 10.0 | $6.03 \times 10^2$ / $9.96 \times 10^2$ | $6.80 \times 10^1$ / $8.09 \times 10^1$ | 5.90 / 6.59 |
| | *1.0 | $2.60 \times 10^3$ / $4.28 \times 10^3$ | $6.71 \times 10^1$ / $8.04 \times 10^1$ | 5.70 / 6.55 |
| SLW | 100.0 | $2.35 \times 10^3$ / $3.86 \times 10^3$ | $5.73 \times 10^1$ / $6.05 \times 10^1$ | 5.22 / 6.03 |
| | *10.0 | $4.83 \times 10^3$ / $7.93 \times 10^3$ | $7.66 \times 10^1$ / $7.74 \times 10^1$ | 6.25 / 7.22 |
| | 1.0 | $5.22 \times 10^3$ / $8.59 \times 10^3$ | $9.00 \times 10^1$ / $8.93 \times 10^1$ | 6.57 / 7.55 |
| LDM | 100.0 | 3.11 / 5.16 | 5.65 / 6.54 | 1.64 / 1.85 |
| | 10.0 | $9.45 \times 10^1$ / $1.55 \times 10^2$ | $1.21 \times 10^1$ / $1.50 \times 10^1$ | 2.53 / 2.91 |
| | *1.0 | $1.07 \times 10^3$ / $1.76 \times 10^3$ | $3.64 \times 10^1$ / $4.77 \times 10^1$ | 4.29 / 4.96 |

Table 7: Hyperparameter results for the ELDM penalty $\lambda_1$, where $*$ denotes the final selection.

| $\lambda_1$ | $s_1$ | $s_2$ | $t_1$ | Avg. |
|---|---|---|---|---|
| 2.0 | 53.2±2.5 | 54.2±2.7 | 49.1±4.4 | 52.2±3.2 |
| 1.0 | 61.4±4.1 | 58.0±2.7 | 44.6±3.7 | 54.7±3.5 |
| *0.5 | 69.5±2.6 | 67.5±3.5 | 42.3±3.4 | 59.7±3.2 |
| 0.25 | 79.8±3.5 | 76.4±2.6 | 32.7±2.9 | 63.0±3.0 |
| 0.1 | 84.4±1.4 | 81.0±1.4 | 27.4±1.6 | 64.3±1.5 |
| 0.05 | 84.8±1.3 | 81.3±1.0 | 27.5±1.3 | 64.5±1.2 |

Table 8: Hyperparameter results for the IRM penalty $\lambda_2$, where $*$ denotes the final selection.

| $\lambda_2$ | $s_1$ | $s_2$ | $t_1$ | Avg. |
|---|---|---|---|---|
| 1000.0 | 51.3±5.4 | 51.2±3.7 | 52.2±5.7 | 51.6±4.9 |
| 100.0 | 50.4±3.8 | 49.8±2.9 | 48.8±6.1 | 49.7±4.3 |
| 10.0 | 49.1±3.7 | 50.3±2.8 | 47.3±4.4 | 48.9±3.6 |
| 1.0 | 52.5±3.3 | 53.9±3.0 | 48.9±5.9 | 51.7±4.1 |
| *0.1 | 66.4±6.5 | 65.7±5.9 | 39.6±6.3 | 57.2±6.2 |
| 0.01 | 88.1±0.5 | 85.2±0.5 | 15.8±1.4 | 63.0±0.8 |

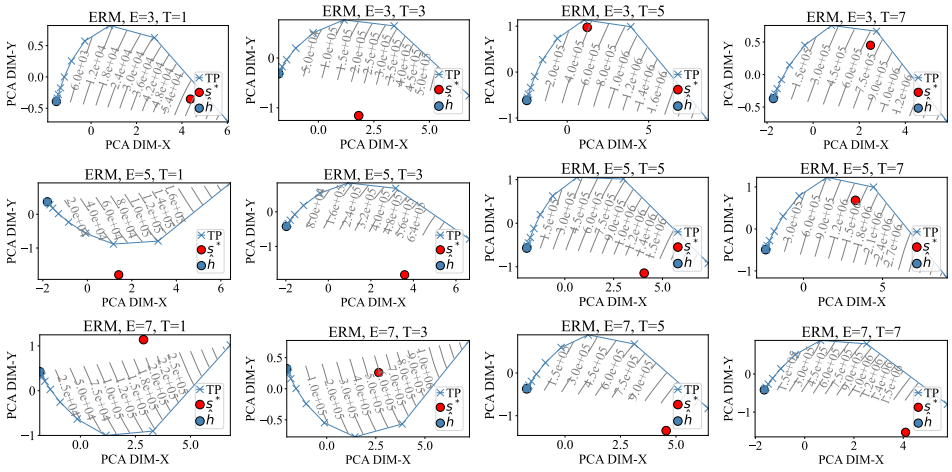


Figure 3: Learning convergence analysis of the learned $\hat{h}$ by ERM across different datasets. As observed, $\hat{h}$ basically does not approach $s^*$ for any of the datasets.

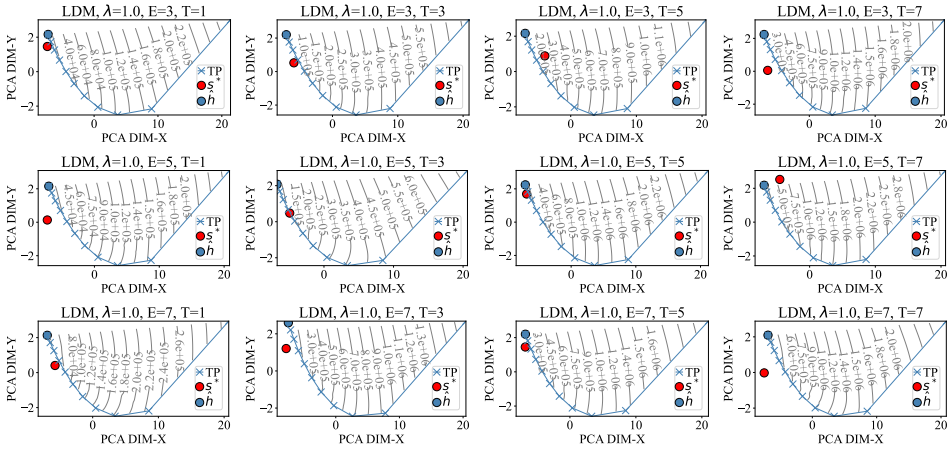


Figure 4: Learning convergence analysis of the learned $\hat{h}$ by LDM across different datasets. As observed, the gap between $\hat{h}$ and $s^*$ is smaller than that observed in Figure 3.

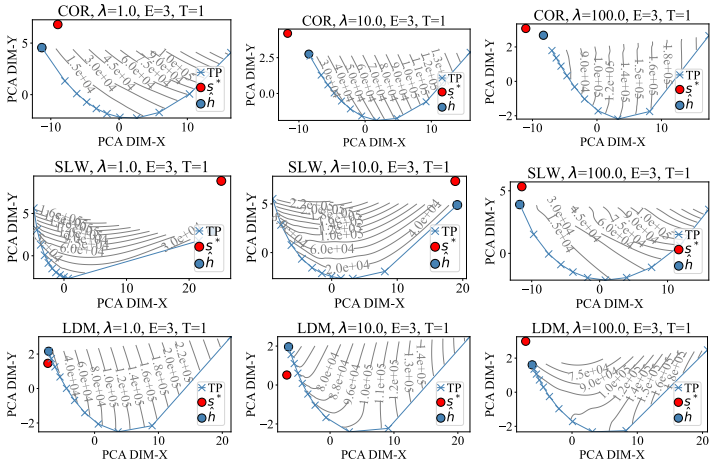


Figure 5: Learning convergence analysis of the learned $\hat{h}$ across different matching methods and hyperparameter settings. As observed, except for SLW, the other methods yield similar gaps.