EXTENDING RADEMACHER COMPLEXITY-BASED GENERALIZATION THEORY TO DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Under the independent and identically distributed (i.i.d.) assumption, Rademacher complexity-based statistical learning theory characterizes the model's generalization through a Rademacher complexity term and its high-probability confidence guarantee. However, this framework becomes inadequate in the domain generalization scenario where distribution shifts commonly occur, precisely because distribution shift itself gives rise to an additional, distribution-induced model complexity, whereas classical Rademacher complexity is conditioned solely on training samples and thus fails to capture the complexity variations induced by distribution shifts. To overcome this limitation, we propose to treat domain distributions as the fundamental objects for defining distribution-level complexity. Specifically, we introduce a risk functional class on the Wasserstein space and quantify the model complexity with respect to domain distributions via its Rademacher term. Furthermore, by exploiting the $\mathrm{CD}(K, \infty)$ curvature structure of the Wasserstein space, we modify the McDiarmid-type concentration inequality to obtain a Wasserstein-geometric confidence guarantee for the proposed complexity measure. Based on this complexity measure, we derive a domain generalization error bound that separates a distribution-level Rademacher complexity term, which captures the complexity caused by distribution shift, and that establishes a connection between the curvature of the sample space and domain generalization performance through its confidence term. This reveals a new principle of domain generalization: learning representation space with higher curvature facilitates better cross-domain generalization. Finally, we experimentally validate this conclusion.

1 INTRODUCTION

Generalization (Vapnik, 1968) is arguably the most fundamental problem in machine learning, as it quantifies a model's ability to perform well on unseen data after training on a finite set of examples. Among various approaches in statistical learning theory (Valiant, 1984; Bousquet & Elisseeff, 2002; Xu & Mannor, 2012; Bartlett et al., 1998; Haussler, 1995; Dudley, 1967), the Rademacher complexity-based generalization theory (Bartlett & Mendelson, 2002; Bartlett et al., 2002) has become one of the most influential frameworks for characterizing generalization performance. Rademacher complexity-based generalization theory characterizes generalization ability by quantifying the complexity of a hypothesis class with respect to training samples and deriving a high-probability confidence guarantee for this complexity measure via measure concentration inequalities (e.g., McDiarmid's inequality (McDiarmid, 1989)). Accordingly, the resulting generalization bound consists of a Rademacher complexity term and its associated confidence term. However, this theory is inherently built upon the i.i.d. assumption, under which training and test samples are drawn from an identical distribution. This assumption is violated in domain generalization (Zhou et al., 2022; Wang et al., 2022), where the model is trained across multiple diverse domains and is expected to generalize to unseen domains under distribution shift (Shimodaira, 2000). In this setting, distribution shift makes model complexity inherently distribution-induced: the complexity of a model is no longer determined solely by the training samples but is fundamentally governed by the domain distributions. The classical Rademacher complexity framework is therefore insufficient, as its

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.

Rademacher term is defined solely on the training samples and thus cannot capture the distribution-induced complexity that the domain distributions themselves impose.

Meanwhile, although existing domain generalization theories (Albuquerque et al., 2019; Sicilia et al., 2023) attempt to characterize the impact of distribution shifts through distribution discrepancy measures (Ben-David et al., 2006; Blitzer et al., 2007; Zhao et al., 2019), the resulting bounds usually rely on a specified target-domain distribution, which contradicts the target-invisible assumption of domain generalization. Furthermore, these approaches mainly quantify domain shifts through distribution discrepancy measure, while the effect of distribution variations on model complexity remains largely unexplored.

In this paper, we aim to extend Rademacher complexity-based statistical learning theory to the domain generalization setting. The key challenge is to construct a distribution-level complexity measure that captures how domain distribution shifts affect model complexity. To this end, we treat domain distributions as the fundamental objects to define a risk functional class over the Wasserstein space, and then leverage its Rademacher complexity to characterize model complexity with respect to domain distributions. Furthermore, to provide a confidence guarantee for the newly defined distribution-level complexity, we must recognize that its fundamental objects are probability distributions residing on the Wasserstein space. Consequently, the corresponding confidence bound ought to reflect the intrinsic geometry of this space. We approach this issue via the lens of measure concentration, which is known to be tightly linked to curvature properties. Specifically, we leverage the $\mathrm{CD}(K,\infty)$ curvature–dimension theory (Lott & Villani, 2009) to adapt McDiarmid's inequality to our setting. This adaptation modifies the standard bounded-difference condition so that the resulting inequality captures how the measure concentration of distribution-valued random variables depends on the Ricci curvature of the underlying sample space. We then derive a new domain generalization error bound that incorporates both the newly defined complexity term and its confidence term, with the latter depending on the Ricci curvature of the sample space. This bound not only quantifies the excess complexity required by distribution shifts, but also highlights the positive impact of high-curvature representation spaces on generalization across domains. We conclude with experimental results that validate these theoretical findings. The primary contributions of this paper are summarized as follows:

- We define a class of risk functionals on the Wasserstein space and leverage its Rademacher complexity to quantify the excess complexity required by distribution shifts in domain generalization scenario.
- By leveraging the $\mathrm{CD}(K,\infty)$ curvature theory to modify the McDiarmid-type concentration inequality, we adapt McDiarmid's inequality to reflect the Wasserstein geometric structure. This modification yields a confidence guarantee of the new complexity term and explicitly depends on the Ricci curvature of the sample space.
- We derive a new domain generalization error bound that incorporates both the newly defined complexity term and its confidence term. This bound not only quantifies the excess complexity required by distribution shifts, but also highlights the positive impact of high-curvature representation spaces on generalization across domains.

2 PRELIMINARIES

In this section, we introduce the main foundations of this paper, including key definitions, notations, formulas and some necessary assumptions.

2.1 RADEMACHER COMPLEXITY-BASED STATISTICAL LEARNING THEORY

We first introduce classical Rademacher complexity-based statistical learning theory, beginning with the generalization error—the key quantity that this theory aims to bound.

We consider a hypothesis class $\mathcal{H}$ defined over the input space $\mathbb{X} \subset \mathbb{R}^d$, where d denotes the dimension. Let $\hat{\mathbb{Y}} \subset [0,1]$ be the output space and $\mathbb{Y} \subset [0,1]$ the labeling space. Let P be a probability distribution on sample space $\mathbb{X} \times \mathbb{Y}$, and let $S = \{(x_i, y_i)\}_{i=1}^n$ be n i.i.d. samples drawn from P.

For any hypothesis $h \in \mathcal{H}$, the expected risk on distribution P is defined as:

$$\mathcal{E}_P(h) = E_{(x,y)\sim P}[\ell(h(x), y)], \tag{1}$$

while the empirical risk on samples $S = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{X}$ is:

$$\hat{\mathcal{E}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i), \tag{2}$$

where the loss function ℓ is q-Lipschitz continuous with respect to its first argument. The generalization error is defined as:

$$\mathrm{Err}(h) = |\mathcal{E}_P(h) - \hat{\mathcal{E}}_S(h)|. \tag{3}$$

Remark. In this paper, the following notations are considered equivalent: $E_{x\sim P}[y]$, $E_x[y]$ and $E_P[y]$.

Then, the empirical Rademacher complexity of $\mathcal{H}$ conditioned n samples is:

$$\hat{\mathcal{R}}_n(\mathcal{H}) = E_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i) \Big| x_1, \ldots, x_n\right], \tag{4}$$

where $\sigma_i \sim \mathrm{Unif}\{\pm 1\}$ are Rademacher variables. The (population) Rademacher complexity is defined as the expectation of $\hat{\mathcal{R}}_n(\mathcal{H})$ over the random draws of the samples:

$$\mathcal{R}_n(\mathcal{H}) = E_{x_1,\ldots,x_n}\left[\hat{\mathcal{R}}_n(\mathcal{H})\right]. \tag{5}$$

This measure characterizes the maximum extent to which a hypothesis class can fit purely random labels on the given samples. A larger Rademacher complexity indicates that the hypothesis class is rich enough to capture spurious patterns, thereby signaling a higher propensity for overfitting and worse generalization.

The following theorem provides an upper bound on the generalization error and reveals the relationship between model complexity and generalization.

Theorem 1 (Generalization Error Bound (Bartlett & Mendelson, 2002))**.** *For any $\delta \in [0, 1]$ and $h \in \mathcal{H}$, the following inequality holds with a confidence level of at least $1 - \delta$:*

$$E[h(x)] \le \frac{1}{n}\sum_{i=1}^{n} h(x_i) + 2\hat{\mathcal{R}}_n(\mathcal{H}) + \mathcal{O}\Big(\sqrt{\frac{\ln(1/\delta)}{n}}\Big), \tag{6}$$

More generally, for the generalization error:

$$\sup_{h\in\mathcal{H}} \mathrm{Err}(h) \le \underbrace{2q\hat{\mathcal{R}}_n(\mathcal{H})}_{\textit{Complexity term}} + \underbrace{\mathcal{O}\Big(\sqrt{\frac{\ln(1/\delta)}{n}}\Big)}_{\textit{Confidence term}}. \tag{7}$$

where the q is the Lipschitz coefficient of loss function ℓ with respect to its first argument.

The second term on the R.H.S. of Eqn. 7 is the high-probability confidence term for the Rademacher complexity, which is grounded in the following probabilistic foundation:

Measure Concentration Foundation: McDiarmid's inequality (McDiarmid, 1989) is a fundamental probabilistic tool in statistical learning theory, and is given as follows:
If a function $g : \mathbb{X}^n \to \mathbb{R}$ satisfies the bounded difference condition,

$$\sup_{x_1,\ldots,x_n,x_i'} |g(x_1, \ldots, x_i, \ldots, x_n) - g(x_1, \ldots, x_i', \ldots, x_n)| \le c_i, \tag{8}$$

then for any $t > 0$ it satisfies the following exponential concentration:

$$P\left(|g(x_1, \ldots, x_n) - E_{P^n}[g(x_1, \ldots, x_n)]| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right), \tag{9}$$

where $x_1, \cdots, x_n, x_i' \in \mathbb{X}$, $i \in [n]$.

The inequality in Theorem 1 gives a high-probability upper bound for the generalization error, comprising a Rademacher complexity term that quantifies the model complexity on the finite training sample, along with a confidence term that provides a probabilistic guarantee for the complexity measure. Here, the confidence term arises from statistical laws, while the essence of generalization lies in the model complexity. A model with high complexity tends to overfit the training samples and thus loses its generalization capability, resulting in a larger generalization bound.

2.2 Basic Notations and Definitions of Domain Generalization

Then we introduce the notations in domain generalization. Domain generalization (DG) considers the setting where models are trained on samples drawn from multiple source domains and are required to generalize to unseen target domains, with all domains following different distributions. Let $\mathbb{D}$ denote the set of all possible target domains, and let $\hat{\mathbb{D}}_m \subset \mathbb{D}$ be the set of m domains randomly drawn from $\mathbb{D}$, referred to as the source domains. Each domain is represented by its distribution, and the i-th source domain is denoted by D_i. We denote the distribution over $\mathbb{D}$ by $\mathcal{P}$ and call it the meta-distribution. Furthermore, we assume that each source domain contains n samples; the j-th sample from the i-th source domain is denoted by $(x_{i,j}, y_{i,j})$, the sample set of the i-th source domain is denoted by S_i, and S denotes the collection of all mn samples.

Under above concepts, the domain generalization error is defined as following:

$$\mathrm{Err}_D(h) = \left| E_{D\sim\mathcal{P}}\mathcal{E}_D(h) - \frac{1}{m}\sum_{i=1}^{m}\hat{\mathcal{E}}_{S_i}(h) \right|, \tag{10}$$

where $\mathcal{E}_D(h)$ denotes the risk over target domain D, $\hat{\mathcal{E}}_{D_i}(h)$ is the empirical risk on the i-th source domain's sample set S_i.

Based on the above definitions, there has been work that used Rademacher complexity to investigate domain generalization and provided a domain generalization error bound.

Theorem 2 (Domain Generalization Error Bound (Gouk et al., 2022))**.** *Assume the loss function ℓ takes values in $[0, 1]$. For any hypothesis class $\mathcal{H}$ and any domain $d \in \mathbb{D}$, with probability at least $1 - \delta$ over the random sampling of S, the following holds uniformly for all $h \in \mathcal{H}$:*

$$\mathrm{Err}_D(h) \le \frac{2}{m}\sum_{i=1}^{m} \underbrace{\hat{\mathcal{R}}_n^i(\ell \circ \mathcal{H})}_{\textit{Per-domain Complexity}} + 2\underbrace{\hat{\mathcal{R}}_{mn}(\ell \circ \mathcal{H})}_{\textit{Joint Complexity}} + 5\underbrace{\sqrt{\frac{\ln(3/\delta)}{2mn}}}_{\textit{Confidence term}}. \tag{11}$$

where $\hat{\mathcal{R}}_n^i(\ell \circ \mathcal{H})$ denotes the empirical Rademacher complexity with respect to samples draw from the i-th source domain, and $\hat{\mathcal{R}}_{mn}(\ell \circ \mathcal{H})$ denotes the empirical Rademacher complexity with respect to the training samples from all source domains.

Theorem 2 applies the classical Rademacher complexity generalization analysis both jointly over all training samples and separately over samples from each source domain, yielding a Joint Complexity term and an averaged Per-domain Complexity term, respectively. However, when deriving the high-probability bound for the joint complexity term, Theorem 2 implicitly treats all mn samples as i.i.d. data and applies McDiarmid's inequality accordingly. In the domain generalization setting, samples are generated hierarchically: m source domains are first drawn from the meta-distribution, and then n samples are drawn from each source domain. Consequently, the mn samples are not i.i.d., the joint complexity term and its associated confidence term are not directly justified under the hierarchical sampling model. In this paper, we adopt the following approach to address this issue: instead of measuring the complexity induced by all mn training samples (which are not i.i.d.), we define a distribution-level complexity in Wasserstein space to quantify the change in complexity caused by distribution shift.

3 Main Result

In this section, we extend Rademacher complexity-based generalization theory to domain generalization. We define a distribution-level complexity term on the Wasserstein space to characterize model complexity with respect to source domain distributions. Leveraging the geometry of this space, we provide a confidence guarantee for the new term. We then derive a domain generalization error bound that reflects model complexity on both training samples and source domain distributions, and separates the confidence biases from sample and domain sampling.

3.1 Distribution-Level Complexity on Wasserstein Space

We first introduce an assumption and some formal definitions of domain generalization on the Wasserstein space, covering both the sample space and the target domain space.

Definition 1 (Sample Space)**.** *We define the sample space as a metric-measure space $(\mathbb{Z}, d, \nu)$, where d and ν are the metric and reference measure on $\mathbb{Z}$, and $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$ is a product space.*

For the definition of metric-measure space, as well as the contents of the Wasserstein space and Wasserstein geometry used in the following, see Appendix B for details.

Assumption 1. *We assume that for any target domain, its distribution $D \in \mathbb{D}$ is absolutely continuous with respect to the reference measure ν on $\mathbb{Z}$, and the Radon–Nikodym derivative belongs to L^2 space, i.e. $\frac{\mathrm{dD}}{\mathrm{d}\nu} \in L^2(\mathbb{Z})$.*

Remark. Since unrestricted domain generalization is infeasible, we constrain the target domain by requiring all domain distributions to be sufficiently close to the reference distribution ν. This mirrors the reference-set idea in Albuquerque et al. (2019), Sicilia et al. (2023), and Ye et al. (2021), which also assume restricted target domains. In the following, we will further quantify the difficulty of domain generalization by defining the concept of the C-regular domain generalization problem.

Definition 2 (Domain Space)**.** *We define the space of all possible domains as the 2-Wasserstein space of $\mathbb{Z}$, denoted by $P_2(\mathbb{Z})$. Under Assumption 1, the target domain space is restricted to*

$$\mathbb{D} = \left\{ D \in P_2(\mathbb{Z}) \;\middle|\; \frac{\mathrm{d}D}{\mathrm{d}\nu} \in L^2(\mathbb{Z}) \right\}. \tag{12}$$

Let $\mathcal{P}$ be a meta-distribution over $\mathbb{D}$ satisfying

$$\nu = \mathbb{E}_{D\sim\mathcal{P}}[D]. \tag{13}$$

Definition 3 (C-regular domain generalization problem)**.** *Define*

$$\mathbb{D}^C = \left\{ D \in P_2(\mathbb{Z}) \;\middle|\; \left\| \frac{\mathrm{d}D}{\mathrm{d}\nu} \right\|^2_{L^2(\mathbb{Z})} \leq C \right\}, \tag{14}$$

*and we say that the domain generalization problem $(\hat{\mathbb{D}}^C_m, \mathbb{D}^C)$ is C-*regular*. The conditional probability measure of $\mathcal{P}$ restricted in $\mathbb{D}^C$ is denoted as $\mathcal{P}^C$, i.e.*

$$\mathrm{d}\mathcal{P}^C(D) = \frac{\mathrm{d}\mathcal{P}}{\mathcal{P}(\mathbb{D}^C)}(D), \qquad D \in \mathbb{D}^C. \tag{15}$$

Then, we define a functional class on domain space, and its Rademacher complexity term.

Definition 4 (Risk Functional Class)**.** *We define the risk functional induced by the loss function ℓ and a hypothesis $h \in \mathcal{H}$ as*

$$\phi_{\ell,h}(D) = E_D[\ell(h(x), y)], \tag{16}$$

the class of risk functional induced by the hypothesis class $\mathcal{H}$ and loss function ℓ is defined as

$$\Phi_{\ell,\mathcal{H}} = \{\phi_{\ell,h} : \mathbb{D} \to \mathbb{R} \mid \phi_{\ell,h}(D) = E_D[\ell(h(x), y)]; h \in \mathcal{H}\}. \tag{17}$$

Its restriction on $\mathbb{D}^C$ is defined as

$$\Phi^C_{\ell,\mathcal{H}} = \{\phi_{\ell,h} : \mathbb{D}^C \to \mathbb{R} \mid \phi_{\ell,h}(D) = E_D[\ell(h(x), y)]; h \in \mathcal{H}\}. \tag{18}$$

Definition 5 (Rademacher Complexity of Risk Functional Class)**.** *The empirical Rademacher complexity of the risk functional class $\Phi_{\ell,\mathcal{H}}$ with respect to the source domain distributions $D_1, \ldots, D_m$ is defined as*

$$\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi_{\ell,\mathcal{H}}) = E_\sigma[\sup_{\phi\in\Phi_{\ell,\mathcal{H}}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \phi(D_i)], \tag{19}$$

where the components of $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ are independent Rademacher variables (i.e., $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$ for any $i \in \{1, \cdots, m\}$).

Next, we leverage the distinctive geometric properties of the Wasserstein space to modify McDiarmid's inequality, thereby providing a Wasserstein-geometric confidence term for the distribution-level complexity, which directly characterizes the domain generalization performance in the error bound presented in the next subsection. We first introduce a new concept and assumption:

Definition 6 (Multivariate Risk Functional)**.** *For a C-regular domain generalization problem $(\hat{\mathbb{D}}_m^C, \mathbb{D}^C)$, we define a multivariate risk functional that takes its source domains as inputs:*

$$F(D_1, \cdots, D_m) = \sup_{\phi \in \Phi_{\ell,\mathcal{H}}^C} \left| E_{\mathcal{P}^C}[\phi(D)] - \frac{1}{m} \sum_{i=1}^{m} \phi(D_i) \right|, \tag{20}$$

where $D_1, \ldots, D_m \in \mathbb{D}^C$. The class of all multivariate risk functionals is denoted as $\mathcal{F}^C$.

Assumption 2. *The function $\ell \circ \mathcal{H}$ is q_z-Lipschitz continuous with respect to z, i.e.*

$$\left| \ell(h(x), y) - \ell(h(x'), y') \right| \leq q_z d((x, y), (x', y')), \qquad \forall (x, y), (x', y') \in \mathbb{Z}, \quad h \in \mathcal{H}, \tag{21}$$

where d is the metric of $\mathbb{Z}$. The cost function of the p-Wasserstein distance in Definition 7 is defined on $\mathbb{Z} \times \mathbb{Z}$ as

$$c(z, z') = d(z, z'), \qquad z, z' \in \mathbb{Z}. \tag{22}$$

Then we modify the bounded differences condition in the McDiarmid's inequality to make it reflect the distinctive properties induced by the Wasserstein geometry.

Theorem 3 (Bounded Difference Condition for $\mathcal{F}^C$)**.** *Under Assumptions 2, and assuming the sample space $(\mathbb{Z}, d, \nu)$ satisfies the $\mathrm{CD}(K, \infty)$ condition with $K > 0$ (Definition 8 in Appendix B), any multivariate risk functional $F \in \mathcal{F}^C$ satisfies the following bounded difference condition for any $k \in \{1, \ldots, m\}$:*

$$|F(D_1, \ldots, D_k, \ldots, D_m) - F(D_1, \ldots, D_k', \ldots, D_m)| \leq \frac{B}{m\sqrt{K}}, \tag{23}$$

where $B = 2q_z\sqrt{2\,(C-1)}$ is a constant.

The proof of Theorem 3 can be found in Appendix C, and we discuss the constant B in detail in Appendix E.

With the bounded difference property established, we can directly apply McDiarmid's inequality to obtain a concentration result for F around its expectation.

Corollary 1 (Measure Concentration with Curvature Dependence)**.** *For any $t > 0$, $F \in \mathcal{F}^C$ and event*

$$\begin{aligned} \Omega^m = \big\{ &(D_1, \cdots, D_k, \cdots, D_m) \in (\mathbb{D}^C)^m : \\ &F(D_1, \cdots, D_k, \cdots, D_m) - E_{(\mathcal{P}^c)^m}[F(D_1, \cdots, D_k, \cdots, D_m)] \geq t \big\}, \end{aligned} \tag{24}$$

the following inequality holds:

$$\int_{\Omega^m} \mathrm{d}(\mathcal{P}^C)^m \leq 2 \exp\left(-\frac{2mKt^2}{B^2} \right), \tag{25}$$

where $D_1, \cdots, D_m \in \mathbb{D}^C$.

It shows that the deviation of the worst-case empirical risk (over the hypothesis class) from its expectation decays exponentially in the number of source domains m, at a rate that depends on the lower curvature bound K of the underlying sample space. In the next subsection, we provide a domain generalization error bound to explain how the curvature of the sample space determines cross-domain generalization via the confidence term of the distribution-level complexity.

3.2 A Domain Generalization Error Bound Incorporating Distribution-Level Complexity and Wasserstein-Geometric Confidence Term

We now derive a high-probability upper bound on the domain generalization error $\mathrm{Err}_D(h)$, which incorporate a distribution-level complexity and its Wasserstein-geometric confidence term.

Theorem 4 (Domain Generalization Error Bound)**.** *For a C-regular domain generalization problem $(\hat{\mathbb{D}}_m^C, \mathbb{D}^C)$, under Assumptions 2, and assuming the sample space $(\mathbb{Z}, d, \nu)$ satisfies the $\mathrm{CD}(K, \infty)$ condition with $K > 0$, the following inequality holds uniformly for all $h \in \mathcal{H}$ with probability at*

least $1-3\delta$ *over the random sampling of the source domains* $D_1, \ldots, D_m \sim \mathcal{P}$ *and the intra-domain samples* $S_i = \{x_{i,j}\}_{j=1}^n$*:*

$$\sup_{h\in\mathcal{H}} Err_D(h) \leq 2\underbrace{\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})}_{\textit{Complexity term 1}} + \frac{2q}{m}\underbrace{\sum_{i=1}^{m}\hat{\mathcal{R}}_n^i(\mathcal{H})}_{\textit{Complexity term 2}} + \underbrace{\mathcal{O}\Big(\sqrt{\frac{\ln(2/\delta)}{n}}\Big)}_{\textit{Confidence term 1}} + \underbrace{\mathcal{O}\Big(\sqrt{\frac{\ln(2/\delta)}{mK}}\Big)}_{\textit{Confidence term 2}}. \tag{26}$$

Here $\hat{\mathcal{R}}_n^i(\mathcal{H})$ *is the Rademacher complexity term with respect to samples of* i*-th source domain.*

The proof of Theorem 4 can be found in Appendix D.

The derived bound consists of four terms: Complexity term 1 ($\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})$) is the Rademacher complexity of the risk functional class with respect to the source domain distributions, which shows that distribution shift introduces an additional complexity penalty; Complexity term 2 ($\sum_{i=1}^m \hat{\mathcal{R}}_n^i(\mathcal{H})$) is the Rademacher complexity computed on the samples from all source domains and reflects the degree to which the model fits the training samples; Confidence term 1 ($\mathcal{O}(\sqrt{\ln(2/\delta)/n})$) is a standard confidence term in generalization analysis that decays as $1/\sqrt{n}$, i.e., inversely proportional to the square root of the total number of samples; and Confidence term 2 ($\mathcal{O}(\sqrt{\ln(2/\delta)/(mK)})$) is an additional confidence term arising from the distribution-level randomness. It depends on the lower Ricci curvature bound K of the sample space; a larger K reduces the term, implying that learned representation spaces with a larger lower bound on Ricci curvature are more favorable for domain generalization.

It is interesting to compare our bound (Theorem 4) with existing domain generalization bounds. In contrast to Theorem 2, which treat all samples as i.i.d. and construct a joint complexity term, we define a distribution-level Rademacher complexity on the Wasserstein space to quantify the excess complexity induced by distribution shifts. This ensures that our bound is rigorously justified under the hierarchical sampling model inherent to domain generalization, and further reveals how the Ricci curvature of the representation space governs the confidence term, offering a more faithful and interpretable characterization of domain generalization. Moreover, unlike Albuquerque et al. (2019); Sicilia et al. (2023), our bound can be computed without specific target domain distribution, a benefit of extending Rademacher complexity bounds to this scenario.

4 Experiment

In this section, we conduct two sets of experiments to validate our conclusions: (1) whether latent space curvature affects domain generalization, and (2) whether the number of source domains influences generalization performance. We adopt an ERM encoder-classifier framework and evaluate on PACS, Office-Home, and Digits-DG, with dataset details provided in Appendix F. Following the leave-one-domain-out protocol in Li et al. (2017), Carlucci et al. (2019), and Li et al. (2019), one domain is used for testing and the remaining domains for training. We set the batch size to 64 and average results over three random seeds using the Dassl.pytorch toolbox. Due to space limitations, we present the PACS results here; the Office-Home and Digits-DG results are provided in Appendix F.1.

4.1 Impact of Latent Space Curvature for Domain Generalization

Our aim is to study the impact of latent space curvature on domain generalization. To this end, we require a manifold with controllable curvature and a probability distribution defined on its geometric support. A natural choice is the von Mises–Fisher (vMF) distribution, which is defined on a hyperspherical geometry and uses a single radius parameter R to control curvature. We further extend this framework to a mixture of multiple hyperspheres, each with equal curvature but learnable centers, which substantially increases representational capacity while retaining efficient curvature control.

In our framework, an encoder maps each input image x to the parameters $(\boldsymbol{\mu}, \kappa)$ of a vMF distribution, and a latent code $\mathbf{z}_{\text{unit}} \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$ is sampled on the unit hypersphere. To avoid the limited capacity of a single small-radius sphere, we construct a latent space as a mixture of N_{sph} hyperspheres sharing radius R but having distinct learnable centers $\{\mathbf{c}_i\}_{i=1}^{N_{\text{sph}}}$. A Gumbel–Softmax



Figure 1: Comparison of domain generalization errors in latent spaces with different curvatures on the PACS, Office-Home and Digits-DG datasets

assignment $\boldsymbol{\pi}$ selects the contribution of each sphere, and the final representation is

$$\mathbf{z} = R \cdot \mathbf{z}_{\text{unit}} + \sum_{i=1}^{N_{\text{sph}}} \pi_i \mathbf{c}_i. \tag{27}$$

The curvature of each individual sphere remains $\mathcal{C} = 1/R^2$, so controlling R directly governs the geometric regularization of the representation. The scaled latent code $\mathbf{z}$ is then passed to a classifier for label prediction, and the entire pipeline is trained end-to-end with a supervised classification loss. Detailed descriptions of the vMF distribution, network architecture, hyperparameter settings, and training procedure are provided in Appendix F.2.

Figure 1 present the experimental results on three benchmark datasets: PACS, Office-Home, and Digits-DG. Our results largely corroborate our theoretical findings: in most cases, increasing the latent space curvature (i.e., decreasing the hyperspherical radius) leads to a smaller domain generalization error. This trend is consistently observed across various target domains on PACS, with one notable exception: when Photo is used as the target domain. In this setting, the domain generalization error remains consistently low across all curvature values, and no clear relationship between curvature and generalization performance is evident. A detailed discussion of this anomaly is deferred to Appendix F.2.

4.2 Impact of Source Domain Number on Domain Generalization

We further investigate the relationship between the number of source domains and the domain generalization error. To this end, we design a controlled experiment that systematically varies the number and composition of source domains used during training, while keeping all other experimental settings fixed. For PACS, Office-Home, and Digits-DG, we consider four source configurations: a single source domain, two source domains, three source domains (the standard multi-source setting), and six source domains obtained by applying MixStyle (Zhou et al., 2024) to mix feature statistics of source-domain pairs. For each dataset and each target domain, we train a pretrained ResNet-18 backbone with a classification head, and measure the domain generalization error as the test error on the target domain minus the average training error on the source domains.

Figure 2 reports the results as bar charts. In most cases, the results are consistent with our theoretical bound: increasing the number of source domains generally reduces the domain generalization error.

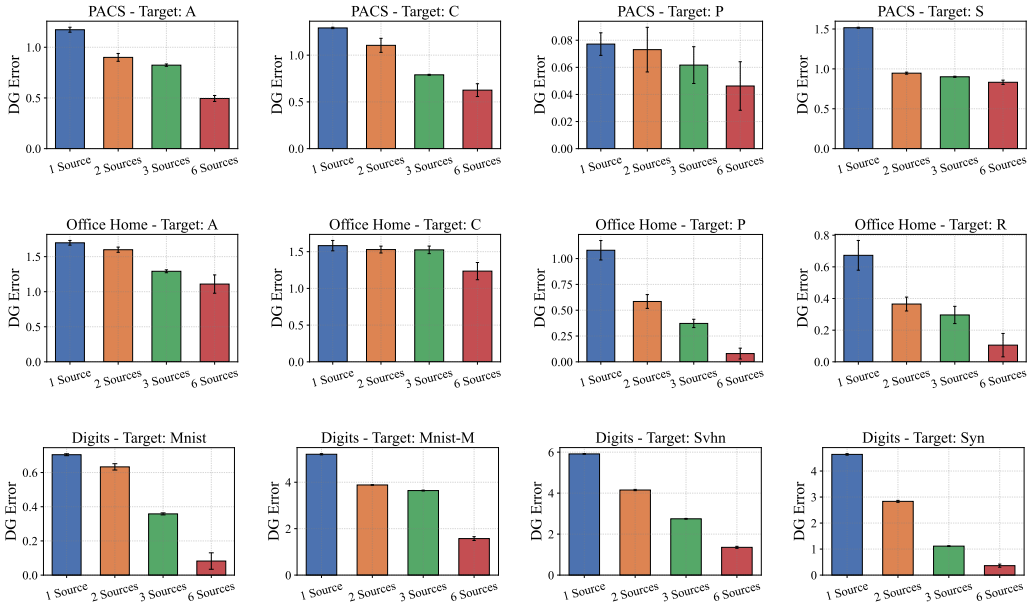

Figure 2: Comparison of domain generalization errors with different source domain number on the PACS, Office-Home and Digits-DG datasets

However, two exceptions are observed: when Photo is used as the target domain on PACS, the results exhibit notable random fluctuations, and when Clipart is used as the target domain on Office-Home, the influence of the number of source domains on the generalization error is less pronounced. We provide possible explanations for these exceptions in Appendix F.3.

5 Conclusion

This paper addresses the failure of classical Rademacher complexity-based statistical learning theory in domain generalization scenarios due to distribution shifts. We propose treating domain distributions as fundamental objects, defining a risk functional class on the Wasserstein space, and using its Rademacher term to characterize model complexity with respect to domain distributions. Furthermore, by leveraging the CD(K, ∞) curvature-dimension theory of the Wasserstein space, we modify the McDiarmid-type concentration inequality to obtain a confidence guarantee with Wasserstein-geometric meaning. From this, we derive a new domain generalization error bound that separates a distribution-level Rademacher complexity term, and establishes a connection between the curvature of the sample space and domain generalization performance through its confidence term. Finally, we conclude with experimental results that validate the theoretical findings.

AI Use Statement

In this work, we used generative AI tools to improve the readability and clarity of the manuscript and to assist with software code development. All AI-assisted edits and code were carefully reviewed, verified, and revised by the authors when necessary. We take responsibility for the final content of this work.

References

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pp. 254–263. PMLR, 2018.

Melih Barsbey, Milad Sefidgaran, Murat Erdogdu, Gael Richard, and Umut Simsekli. Heavy tails in sgd and compressibility of overparametrized neural networks. *Advances in neural information processing systems*, 34:29364–29378, 2021.

Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized rademacher complexities. In *International Conference on Computational Learning Theory*, pp. 44–58. Springer, 2002.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unseen task. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2178–2186, 2011.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007.

Olivier Bousquet and Andr'e Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 587–591. IEEE, 2019.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2229–2238, 2019.

Jiarui Ding and Aviv Regev. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *Nature communications*, 12(1):2554, 2021.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Henry Gouk, Ondrej Bohdal, Da Li, and Timothy Hospedales. On the limitations of general purpose domain generalisation methods. *arXiv preprint arXiv:2202.00563*, 2022.

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34:26370–26381, 2021.

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34:24670–24682, 2021.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1446–1455, 2019.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020.

John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, pp. 903–991, 2009.

Colin McDiarmid. *On the Method of Bounded Differences*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, Cambridge, 1989.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Milad Sefidgaran, Amin Gohari, Gaël Richard, and Umut Simsekli. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In *Conference on Learning Theory*, pp. 4416–4463. PMLR, 2022.

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

Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. Domain adversarial neural networks for domain generalization: When it works and how to improve. *Machine Learning*, 112(7):2685–2721, 2023.

Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir Vapnik. On the uniform convergence of relative frequencies of events to their probabilities. In *Doklady Akademii Nauk USSR*, volume 181, pp. 781–787, 1968.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 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.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*, 86:391–423, 2012.

Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4503–4513, 2018.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:23519–23531, 2021.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13025–13032, 2020.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3):822–836, 2024.

A Related Work

In this appendix, we introduce related work to provide a more comprehensive understanding of this research direction. The following topics are covered in order: classical generalization theory and domain generalization.

A.1 Generalization

To quantify the generalization of machine learning models, numerous foundational theoretical frameworks have been established over the past decades. In addition to the classical works in this field introduced in Section 1, we further review several studies that analyze generalization from different perspectives. Among them, the use of information-theoretic tools to analyze generalization error constitutes an important direction. Xu & Raginsky (2017) proposes an information-theoretic framework based on mutual information to quantify and control the bias induced by the selection process in adaptive data analysis, and demonstrates that this mutual-information bound provides tight and actionable upper bounds on bias in a variety of common analytical settings. Bu et al. (2019); Harutyunyan et al. (2021); Haghifam et al. (2021) extended it by using different conditional versions of mutual information. Moreover, some works such as Arora et al. (2018); Barsbey et al. (2021) introduce a compression-based generalization framework. These studies show that a higher level of compressibility indicates a lower generalization error, because the hypothesis class $\mathcal{W}$ can be approximated by a smaller "compressed" space, which intuitively leads to a lower worst-case error. Recently, Simsekli et al. (2020) proves that the Hausdorff dimension of SGD trajectories is governed by the heavy-tailedness of the gradient noise, and derives a generalization error bound based on this fractal dimension, thereby revealing that heavy-tailed noise enhances generalization by reducing the effective geometric complexity of the parameter trajectories as an implicit regularization mechanism. Finally, by unifying mutual information, compressibility, and fractal dimensions within a single rate-distortion framework, Sefidgaran et al. (2022) innovatively treats generalization error as a form of lossy distortion and requires compressibility to hold within any KL-divergence neighborhood of the true data distribution, thereby deriving tighter expected and high-probability generalization bounds.

A.2 Domain Generalization

The first formal characterization of the DG problem setting was provided by Blanchard et al. (2011). It worked within a reproducing kernel Hilbert space for the hypothesis class, learned from multiple related classification tasks, and established a distribution-free generalization error analysis. The subsequent work of Blanchard et al. (2021) further introduced a marginal transfer learning framework, which incorporates marginal distributions into the feature space to explicitly address domain shift. Li et al. (2020) posited a linear dependence in the latent feature space across domains, suggesting that any domain can be expressed as a linear combination of the others. Ye et al. (2021) introduced

a two-component quantification of feature distribution: a variation term capturing cross-domain stability and an informativeness term reflecting discriminative power. Using an expansion function driven by the variation term, they bounded the error on unseen domains, with the informativeness term indicating learnability. Although this approach relaxes prior assumptions, its resulting domain generalization error bound suffers from limited interpretability and offers little practical guidance. Wang et al. (2026) highlights the complementarity of the two different domain generalization strategies by decomposing the feature space into different subspaces and analyzing domain generalization methods based on domain-invariant representation learning and those based on domain augmentation within these subspaces.

B Wasserstein Geometry

In this appendix, we introduce the geometric tools from optimal transport that are needed to analyze domain distributions as fundamental objects. Unlike classical learning theory, which operates on a fixed sample space, our analysis treats each domain as a probability distribution and the collection of all domains as a subset of the 2-Wasserstein space. This space carries a rich metric structure, but its probabilistic behavior—such as the concentration of distribution-valued random variables—is governed by curvature.

Definition 7 (Metric-measure Space and Wasserstein Space)**.** *Let $\mathbb{X}$ be a compact length space and let ν be a Borel probability measure on $\mathbb{X}$, referred to as the reference measure. The triple $(\mathbb{X}, d, \nu)$ is called a metric-measure space, where d is the distance of $\mathbb{X}$.*

Let $P(\mathbb{X})$ be the set of all Borel probability measures on $\mathbb{X}$, For any two measures $\mu, \nu \in P(\mathbb{X})$, the p-Wasserstein distance is defined as

$$W_p(D_1, D_2) = \inf_{\gamma \in \Pi(D_1, D_2)} \Big(\int_{\mathbb{X}\times\mathbb{X}} c(x_1, x_2)^p d\gamma(x_1, x_2) \Big)^{\frac{1}{p}}, \tag{28}$$

where $\Pi(D_1, D_2)$ is a collection of all joint probability measures on $\mathbb{X} \times \mathbb{X}$ with marginals D_1 and D_2, $c(x_1, x_2)$ is a cost function for transporting one unit of mass x_1 to x_2. Equipped with the 2-Wasserstein distance $W_2(\cdot, \cdot)$, $P(\mathbb{X})$ becomes a metric space, denoted by $P_2(\mathbb{X})$ and called the 2-Wasserstein space.

Definition 8 (CD(K,∞) curvature condition (Lott & Villani, 2009))**.** *A metric measure space $(\mathbb{X}, d, \nu)$ is said to have Ricci curvature bounded below by $K \in \mathbb{R}$ (or to satisfy the $\mathrm{CD}(K, \infty)$ condition) if for every pair $\mu_0, \mu_1 \in P_2(\mathbb{X}, \nu)$, where $P_2(\mathbb{X}, \nu) = \{\mu \in P_2(\mathbb{X}) : \mathrm{supp}(\mu) \subset \mathrm{supp}(\nu)\}$, there exists a Wasserstein geodesic $\{\mu_t\}_{t\in[0,1]}$ connecting them such that the relative entropy $H_{\infty,\nu}(\mu) = \int_{\mathbb{X}} \rho \log \rho \mathrm{d}\nu$ (where $\rho = \mathrm{d}\mu/\mathrm{d}\nu$) satisfies the displacement K-convexity inequality along this geodesic:*

$$H_{\infty,\nu}(\mu_t) \le (1-t)H_{\infty,\nu}(\mu_0) + tH_{\infty,\nu}(\mu_1) - \frac{K}{2}t(1-t)W_2(\mu_0, \mu_1)^2, \quad \forall t \in [0, 1]. \tag{29}$$

C Proof of Theorem 3

Proof. For a C-regular domain generalization problem $(\hat{\mathbb{D}}_m^C, \mathbb{D}^C)$, we have $D_1, \dots D_m \in \mathbb{D}^C$. We consider the change in the functional $F \in \mathcal{F}^C$ when the k-th distribution D_k is replaced by $D_k' \in \mathbb{D}^C$. By definition,

$$\begin{aligned}
&F(D_1, \cdots, D_k, \cdots, D_m) - F(D_1, \cdots, D_k', \cdots, D_m) \\
&= \sup_{\phi \in \Phi_{\ell,\mathcal{H}}^C} \big| E_{\mathcal{P}^C}[\phi(D)] - \frac{1}{m}\sum_{i=1}^m \phi(D_i) \big| \\
&\quad - \sup_{\phi \in \Phi_{\ell,\mathcal{H}}^C} \big| E_{\mathcal{P}^C}[\phi(D)] - \frac{1}{m}\big(\sum_{i\neq k} \phi(D_i) + \phi(D_k') \big) \big| \\
&\le \sup_{\phi \in \Phi_{\ell,\mathcal{H}}^C} \frac{|\phi(D_k) - \phi(D_k')|}{m} \\
&= \frac{1}{m} \sup_{h\in\mathcal{H}} \big| E_{D_k}[\ell(h(x), y)] - E_{D_k'}[\ell(h(x), y)] \big|.
\end{aligned} \tag{30}$$

By Assumption 2, for any $h \in \mathcal{H}$, the function $z = (x, y) \mapsto \ell(h(x), y)$ is $q_{\mathbb{Z}}$-Lipschitz. Let $f_h(z) = \ell(h(x), y)/q_z$. Then f_h is 1-Lipschitz, and by the Kantorovich–Rubinstein duality,

$$\left|E_{D_k}[f_h] - E_{D'_k}[f_h]\right| \leq W_1(D_k, D'_k).$$

Multiplying by q_z yields

$$\left|E_{D_k}[\ell(h(x), y)] - E_{D'_k}[\ell(h(x), y)]\right| \leq q_z\, W_1(D_k, D'_k).$$

Taking the supremum over $h \in \mathcal{H}$, we obtain

$$\sup_{h\in\mathcal{H}} \left|E_{D_k}[\ell(h(x), y)] - E_{D'_k}[\ell(h(x), y)]\right| \leq q_z\, W_1(D_k, D'_k).$$

Therefore,

$$F(D_1, \ldots, D_k, \ldots, D_m) - F(D_1, \ldots, D'_k, \ldots, D_m) \leq \frac{q_z}{m} W_1(D_k, D'_k).$$

Then, due to the Hölder inequality, for any $\gamma \in \Pi(D_k, D'_k)$ and cost function $c(\cdot, \cdot)$, we have

$$\begin{aligned}\int_{\mathbb{Z}\times\mathbb{Z}} c(z_1, z_2) d\gamma(z_1, z_2) &\leq \Big(\int_{\mathbb{Z}\times\mathbb{Z}} c(z_1, z_2)^2 d\gamma(z_1, z_2)\Big)^{\frac{1}{2}} \cdot \Big(\int_{\mathbb{Z}\times\mathbb{Z}} 1^2 d\gamma(z_1, z_2)\Big)^{\frac{1}{2}} \\ &= \Big(\int_{\mathbb{Z}\times\mathbb{Z}} c(z_1, z_2)^2 d\gamma(z_1, z_2)\Big)^{\frac{1}{2}}\end{aligned} \tag{31}$$

Taking the infimum on both sides yields:

$$\begin{aligned}W_1(D_k, D'_k) &= \inf_{\gamma\in\Pi(D_k, D'_k)} \int_{\mathbb{Z}\times\mathbb{Z}} c(z_1, z_2) d\gamma(z_1, z_2) \\ &\leq \inf_{\gamma\in\Pi(D_k, D'_k)} \Big(\int_{\mathbb{Z}\times\mathbb{Z}} c(z_1, z_2)^2 d\gamma(z_1, z_2)\Big)^{\frac{1}{2}} \\ &= \Big(\inf_{\gamma\in\Pi(D_k, D'_k)} \int_{\mathbb{Z}\times\mathbb{Z}} c(z_1, z_2)^2 d\gamma(z_1, z_2)\Big)^{\frac{1}{2}} \\ &= W_2(D_k, D'_k)\end{aligned} \tag{32}$$

So we have

$$\begin{aligned}&F(D_1, \cdots, D_k, \cdots, D_m) - F(D_1, \cdots, D'_k, \cdots, D_m) \\ &\leq \frac{q_z W_1(D_k, D'_k)}{m} \leq \frac{q_z W_2(D_k, D'_k)}{m} \\ &\leq \frac{q_z(W_2(D_k, \nu) + W_2(D'_k, \nu))}{m} \\ &\leq \frac{2q_z}{m}\Big(\sqrt{\frac{H_{\infty,\nu}(D_k)}{K}} + \sqrt{\frac{H_{\infty,\nu}(D'_k)}{K}}\Big),\end{aligned} \tag{33}$$

The inequality in the last line is due to the Theorem 6.1 in Lott & Villani (2009).

Similarly we can obtain:

$$\begin{aligned}&F(D_1, \cdots, D'_k, \cdots, D_m) - F(D_1, \cdots, D_k, \cdots, D_m) \\ &\leq \frac{2q_z}{m}\Big(\sqrt{\frac{H_{\infty,\nu}(D_k)}{K}} + \sqrt{\frac{H_{\infty,\nu}(D'_k)}{K}}\Big).\end{aligned} \tag{34}$$

Let $\mathrm{dD} = \rho \mathrm{d}\nu$, we have:

$$\begin{aligned} H_{\infty,\nu}(D) &= \int_{\mathbb{X}} \rho \log \rho \mathrm{d}\nu \\ &= \int_{\mathbb{X}} \log \rho \mathrm{d}D \\ &\leq \log\Big(\int_{\mathbb{X}} \rho \mathrm{d}D\Big) \\ &= \log\Big(\int_{\mathbb{X}} \rho^2 \mathrm{d}\nu\Big) \\ &\leq \int_{\mathbb{X}} \rho^2 \mathrm{d}\nu - 1 \\ &= \|\rho\|_{L^2}^2 - 1. \end{aligned} \tag{35}$$

The first "$\leq$" is due to the $\log(\cdot)$ is a concave function and the Jensen's inequality. The second "$\leq$" is due to the $\log x \leq x - 1$.

Because the $D \in \mathbb{D}^C$, $\|\rho\|_{L^2}^2 \leq C$, so we have:

$$H_{\infty,\nu}(D) \leq C - 1. \tag{36}$$

According to Hölder's inequality, we have:

$$1 = \int \rho \mathrm{d}\nu \leq \left(\int \rho^2 \mathrm{d}\nu\right)^{1/2} = \|\rho\|_{L^2} \leq \sqrt{C}, \tag{37}$$

so

$$C - 1 \geq 0. \tag{38}$$

Letting $B = 2q_z\sqrt{2(C-1)}$, and combining Eqn. 33, Eqn. 34 and Eqn. 35, we obtain the bounded difference condition for the multivariate risk functional:

$$|F(D_1, \cdots, D_k, \cdots, D_m) - F(D_1, \cdots, D'_k, \cdots, D_m)| \leq \frac{B}{m\sqrt{K}} \tag{39}$$

□

D Proof of Theorem 4

Proof. We start by adding and subtracting the per-domain expected risks, then apply the triangle inequality:

$$\begin{aligned} \sup_{h\in\mathcal{H}} \mathrm{Err}_D(h) &= \sup_{h\in\mathcal{H}} \left| \mathcal{E}_{\mathcal{P}^C}(h) - \frac{1}{m}\sum_{i=1}^{m} \hat{\mathcal{E}}_{S_i}(h) \right| \\ &= \sup_{h\in\mathcal{H}} \left| E_{\mathcal{P}^C}\Big[E_D[\ell(h(x), y)]\Big] - \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n} \ell(h(x_{i,j}), y_{i,j}) \right| \\ &\leq \underbrace{\sup_{h\in\mathcal{H}} \left| E_{\mathcal{P}^C}\Big[E_D[\ell(h(x), y)]\Big] - \frac{1}{m}\sum_{i=1}^{m} E_{D_i}[\ell(h(x), y)] \right|}_{\mathrm{Term}_A} \\ &\quad + \underbrace{\sup_{h\in\mathcal{H}} \left| \frac{1}{m}\sum_{i=1}^{m} E_{D_i}[\ell(h(x), y)] - \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n} \ell(h(x_{i,j}), y_{i,j}) \right|}_{\mathrm{Term}_B} \end{aligned} \tag{40}$$

(1) We first handle Expression Term_B:

$$\begin{aligned}
\text{Term}_B &= \sup_{h\in\mathcal{H}} \Big| \frac{1}{m}\sum_{i=1}^{m} E_{D_i}[\ell(h(x),y)] - \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\ell(h(x_{i,j}),y_{i,j}) \Big| \\
&\le \frac{1}{m}\sum_{i=1}^{m} \sup_{h\in\mathcal{H}} \Big| E_{D_i}[\ell(h(x),y)] - \frac{1}{n}\sum_{j=1}^{n}\ell(h(x_{i,j}),y_{i,j}) \Big| \\
&= \frac{1}{m}\sum_{i=1}^{m} \sup_{h\in\mathcal{H}} \big| \mathcal{E}_{D_i}(h) - \hat{\mathcal{E}}_{S_i}(h) \big| \\
&\le \frac{2}{m}\sum_{i=1}^{m} \hat{\mathcal{R}}_n^i(\ell\circ\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}} \\
&\le \frac{2q}{m}\sum_{i=1}^{m} \hat{\mathcal{R}}_n^i(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}
\end{aligned} \tag{41}$$

The above inequality holds with probability at least $1-\delta$.

(2) We next handle Expression Term_A.

$$\begin{aligned}
\text{Term}_A &= \sup_{h\in\mathcal{H}} \Big| E_{\mathcal{P}^C}\Big[E_D[\ell(h(x),y]\Big] - \frac{1}{m}\sum_{i=1}^{m} E_{D_i}[\ell(h(x),y)] \Big| \\
&= \sup_{h\in\mathcal{H}} \Big| E_{\mathcal{P}^C}[\phi_{\ell,h}(D) - \frac{1}{m}\sum_{i=1}^{m}\phi_{\ell,h}(D_i)] \Big| \\
&= \sup_{\phi\in\Phi_{\ell,\mathcal{H}}^C} \Big| E_{\mathcal{P}^C}[\phi(D) - \frac{1}{m}\sum_{i=1}^{m}\phi(D_i)] \Big| \\
&= F(D_1,\cdots,D_k,\cdots,D_m)
\end{aligned} \tag{42}$$

According to Corollary 1, we have:

$$\int_{\Omega^m} \mathrm{d}(\mathcal{P}^C)^m \le 2\exp\left(-\frac{2mKt^2}{B^2}\right), \tag{43}$$

where

$$\begin{aligned}
\Omega^m = \big\{ &(D_1,\cdots,D_k,\cdots,D_m)\in(\mathbb{D}^C)^m : \\
&F(D_1,\cdots,D_k,\cdots,D_m) - E_{(\mathcal{P}^c)^m}[F(D_1,\cdots,D_k,\cdots,D_m)] \ge t \big\},
\end{aligned} \tag{44}$$

Let $\delta = 2\exp(-\frac{2mKt^2}{B^2})$, then we have:

$$\int_{\Omega_c^m} \mathrm{d}(\mathcal{P}^C)^m \ge 1-\delta, \tag{45}$$

where

$$\Omega_c^m = \left\{ (D_1,\cdots,D_k,\cdots,D_m)\in(\mathbb{D}^C)^m : F(D_1,\cdots,D_k,\cdots,D_m) - E_{(\mathcal{P}^C)^m}[F(D_1,\cdots,D_k,\cdots,D_m)] \le B\sqrt{\frac{\ln(2/\delta)}{2mK}} \right\}. \tag{46}$$

i.e. the following inequality holds with probability at least $1-\delta$:

$$\underbrace{\text{Term}_A - E_{(\mathcal{P}^C)^m}\Big[\sup_{h\in\mathcal{H}}\Big|E_{\mathcal{P}^C}\big[E_D[\ell(h(x),y]\big] - \frac{1}{m}\sum_{i=1}^{m}E_{D_i}[\ell(h(x),y)]\Big|\Big]}_{\text{Term}_C} \le B\sqrt{\frac{\ln(2/\delta)}{2mK}}. \tag{47}$$

Next, we handle Equation Term_C.

$$
\begin{aligned}
\text{Term}_C &= E_{D\sim(\mathcal{P}^C)^m}\Big[\sup_{h\in\mathcal{H}}\Big|E_{\mathcal{P}^C}\big[E_D[\ell(h(x),y]\big]-\frac{1}{m}\sum_{i=1}^{m}E_{D_i}[\ell(h(x),y)]\Big|\Big]\\
&= E_{D\sim(\mathcal{P}^C)^m}\Big[\sup_{h\in\mathcal{H}}\Big|E_{\mathcal{P}^C}[\phi_{\ell,h}(D)]-\frac{1}{m}\sum_{i=1}^{m}\phi_{\ell,h}(D_i)\Big|\Big]\\
&= E_{D\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}\Big|E_{\mathcal{P}^C}[\phi]-\hat{E}_D[\phi]\Big|\Big]\\
&= E_{D\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}E_{D'\sim(\mathcal{P}^C)^m}\big[\big|\hat{E}_{D'}[\phi]-\hat{E}_D[\phi]\big|\big]\Big]\\
&\le E_{D\sim(\mathcal{P}^C)^m}E_{D'\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}\big|\hat{E}_{D'}[\phi]-\hat{E}_D[\phi]\big|\Big]\\
&\le E_{D\sim(\mathcal{P}^C)^m}E_{D'\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}\frac{1}{m}\sum_{i=1}^{m}\big|\phi(D'_i)-\phi(D_i)\big|\Big]\\
&= E_\sigma E_{D\sim(\mathcal{P}^C)^m}E_{D'\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(\phi(D'_i)-\phi(D_i)\big)\Big]\\
&\le E_\sigma E_{D\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\phi(D_i)\Big]+E_\sigma E_{D'\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i\phi(D'_i)\Big]\\
&= 2E_\sigma E_{D\sim(\mathcal{P}^C)^m}\Big[\sup_{\phi\in\Phi^C_{\ell,\mathcal{H}}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\phi(D_i)\Big]\\
&= 2E_{D\sim(\mathcal{P}^C)^m}[\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi^C_{\ell,\mathcal{H}})]\\
&= 2E_{D\sim(\mathcal{P}^C)^m}[\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})]
\end{aligned}
\tag{48}
$$

Let $\hat{\mathcal{R}}_{D_1,\cdots,D'_k,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})$ be the Rademacher complexity after replacing the distribution of the k-th source domain. Analogously to the multivariate risk functionals F, we simplify the derivation in the proof of Theorem 3 to obtain:

$$
\begin{aligned}
\hat{\mathcal{R}}_{D_1,\cdots,D_k,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})-\hat{\mathcal{R}}_{D_1,\cdots,D'_k,\cdots,D_m}(\Phi_{\ell,\mathcal{H}}) &\le \sup_{h\in\mathcal{H}}\frac{|\phi_{\ell,h}(D_k)-\phi_{\ell,h}(D'_k)|}{m}\\
&\le \frac{1}{m}\Big(\sqrt{\frac{2H_{\infty,\nu}(D_k)}{K}}+\sqrt{\frac{2H_{\infty,\nu}(D'_k)}{K}}\Big)\\
&\le \frac{B}{m\sqrt{K}}
\end{aligned}
\tag{49}
$$

Similarly, we have:

$$
\hat{\mathcal{R}}_{D_1,\cdots,D'_k,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})-\hat{\mathcal{R}}_{D_1,\cdots,D_k,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})\le\frac{B}{m\sqrt{K}} \tag{50}
$$

so,

$$
\Big|\hat{\mathcal{R}}_{D_1,\cdots,D_k,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})-\hat{\mathcal{R}}_{D_1,\cdots,D'_k,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})\Big|\le\frac{B}{m\sqrt{K}} \tag{51}
$$

so $\hat{\mathcal{R}}_{D_1,\cdots,D_m}$ also satisfies Theorem 3, and applying Corollary 1, the following inequality holds with probability at least $1-\delta$:

$$
E_{D\sim(\mathcal{P}^C)^m}\big[\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})\big]\le\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi_{\ell,\mathcal{H}})+B\sqrt{\frac{\ln(2/\delta)}{2mK}} \tag{52}
$$

Combining Eqn. 41, Eqn. 47, Eqn. 48, and Eqn. 52, we obtain the final result:

$$\sup_{h\in\mathcal{H}} Err_D(h) \le 2\hat{\mathcal{R}}_{D_1,\cdots,D_m}(\Phi_{\ell,\mathcal{H}}) + \frac{2q}{m}\sum_{i=1}^{m}\hat{\mathcal{R}}_n^i(\mathcal{H}) + \mathcal{O}\Big(\sqrt{\frac{\ln(2/\delta)}{n}}\Big) + \mathcal{O}\Big(\sqrt{\frac{\ln(2/\delta)}{mK}}\Big). \quad (53)$$

The above inequality holds with probability at least $1-3\delta$ by a union bound over the three high-probability events used above. □

E Discussion on Constant B

The constant B appearing in Theorem 4 originates from the bounded-difference constant in the proof of Theorem 3. Following the derivation, B is defined as

$$B = 2q_z\sqrt{2\,(C-1)}, \quad (54)$$

where C is the upper bound of $\left\{\left\|\frac{\mathrm{d}D}{\mathrm{d}\nu}\right\|^2_{L^2(\nu)} : D\in\mathbb{D}^C\right\}$, with ν being the reference distribution of $\mathbb{Z}$.

Intuitively, B can be interpreted as the "degree of generalizability" for $\mathbb{D}$:

- If every domain coincides with ν (i.e., no distribution shift), then $C=1$, hence $B=0$.
- If there are huge differences between domains in the domain space, then C grows larger, making B larger, which reflects a more challenging cross-domain generalization task.

F Experimental Details

F.1 Dataset Introduction

We provide a brief overview of the datasets used in these experiments.

Digits-DG (Zhou et al., 2020) comprises four distinct digit datasets, namely MNIST (LeCun et al., 1998), MNIST-M (Ganin & Lempitsky, 2015), SVHN (Netzer et al., 2011), and SYN (Ganin & Lempitsky, 2015), which exhibit substantial variations in font style, stroke color, and background. Each domain contains ten categories.

PACS (Li et al., 2017) consists of four domains, namely Photo (1,670 images), Art Painting (2,048 images), Cartoon (2,344 images) and Sketch (3,929 images). Each domain contains seven categories.

Office-Home (Venkateswara et al., 2017) contains four domains,which are Artistic, Clipart, Product and Real-World. Each domain has 65 classes, which are related to office and home objects. There are around 15,500 images in total. The domain variations mainly take place in background, view point and image style.

F.2 Implementation Details in Section 4.1

von Mises–Fisher distribution. The von Mises–Fisher (vMF) distribution provides a natural framework for modeling data distributed on a hypersphere (Ding & Regev, 2021; Xu & Durrett, 2018). It is characterized by a mean direction $\boldsymbol{\mu}\in\mathcal{S}^{m-1}$ and a concentration parameter $\kappa\ge 0$, with support on the unit hypersphere $\mathcal{S}^{m-1}\subset\mathbb{R}^m$. Sampling from a vMF distribution yields latent codes that inherently reside on a hyperspherical manifold. The concentration parameter κ governs the tightness of samples around $\boldsymbol{\mu}$; in the limiting case where $\kappa=0$, the vMF distribution reduces to the uniform distribution over the sphere, thereby achieving maximum coverage.

Encoder and mixture-of-spheres construction. An encoder network maps each input image x to the parameters $(\boldsymbol{\mu},\kappa)$ of a vMF distribution. Specifically, the encoder is built upon a pretrained ResNet-18 convolutional backbone, followed by three independent linear heads that respectively predict the vMF parameters $(\boldsymbol{\mu},\kappa)$ and the mixture-assignment logits. We then draw a sample $\mathbf{z}_{\text{unit}}\sim\mathrm{vMF}(\boldsymbol{\mu},\kappa)$, which lies on the unit hypersphere.

Algorithm 1 Training procedure for the proposed model

Initialize all parameters randomly:
  encoder backbone,
  learnable sphere centers $\{\mathbf{c}_i\}_{i=1}^{N_{\text{sph}}}$,
  classifier f_{cls}.
repeat
    Sample a batch (x, y) from the union of source domains.
    Encode to obtain $(\boldsymbol{\mu}, \kappa) = \big(\text{Encoder}_1(x), \text{Encoder}_2(x)\big)$ with $\|\boldsymbol{\mu}\|_2 = 1$.
    Obtain mixture assignment logits $\boldsymbol{\pi}_{\text{logits}} = \text{Encoder}_3(x)$.
    Compute $\boldsymbol{\pi} = \text{GumbelSoftmax}(\boldsymbol{\pi}_{\text{logits}}, \tau)$.
    Sample $u \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$.
    Form the representation: $z = R \cdot u + \sum_i \pi_i \mathbf{c}_i$.
    Predict $\hat{y} = f_{\text{cls}}(z)$ and compute the cross-entropy loss.
    Update all learnable parameters.
until convergence

While the above approach is theoretically sound, we empirically observed that a single hypersphere with a small radius (high curvature) tends to cause training difficulties. This is due to a decrease in representational capacity, as the data points become densely packed together. To overcome this limitation while retaining the benefits of curvature control, we introduce a mixture of multiple hyperspheres with learnable centers. Concretely, instead of using a single spherical manifold, we construct a latent space consisting of a mixture of N_{sph} hyperspheres. These hyperspheres share an equal radius R, but have distinct centers. The center of the i-th sphere is represented by a learnable vector $\mathbf{c}_i \in \mathbb{R}^m$, where $i \in [1, N_{\text{sph}}]$. We stochastically sample a latent variable via standard mixture density sampling. Specifically, we first determine the sphere index using the mixture weights $\boldsymbol{\pi}$. These weights are derived by computing the logits $\boldsymbol{\pi}_{\text{logits}}$ of an N_{sph}-categorical distribution over the spheres, followed by a Gumbel–Softmax operation to obtain a differentiable assignment vector $\boldsymbol{\pi} \in \triangle^{N_{\text{sph}}}$, i.e., an N_{sph}-dimensional simplex, such that $\sum_{i=1}^{N_{\text{sph}}} \pi_i = 1$:

$$\boldsymbol{\pi} = \text{Gumbel-Softmax}(\boldsymbol{\pi}_{\text{logits}}, \tau), \tag{55}$$

where τ is a temperature parameter (annealed during training). The final latent representation is a convex combination of the sphere-specific points:

$$\mathbf{z} = R \cdot \mathbf{z}_{\text{unit}} + \sum_{i=1}^{N_{\text{sph}}} \pi_i \mathbf{c}_i. \tag{56}$$

Thus, $\mathbf{z}$ lies on the union of N_{sph} spheres of radius R, each centered at $\mathbf{c}_i$, and the assignment $\boldsymbol{\pi}$ controls which sphere contributes most to the representation. The learnable centers $\mathbf{c}_i$ are initialized on a sphere of radius 3 and are regularized to stay well-separated, ensuring that the composite geometry remains expressive while preserving curvature properties. Importantly, the curvature of each individual sphere is still $\mathcal{C} = 1/R^2$, so controlling R directly influences the geometric regularization of the representation.

Classifier architecture. The scaled latent code $\mathbf{z}$ is then passed to a classifier that predicts the class label. For PACS and Digits-DG, the classifier is a single linear layer, the latent space dimensions are set to $m = 5, 10$, with the number of hypersphere centers N_{sph} correspondingly chosen as $\{7, 15\}$. For Office-Home, which has 65 categories, we use a two-layer network with 64 hidden units, ReLU activation, and a dropout rate of 0.3. The latent space dimension is set to $m = 20$, and the number of hypersphere centers is set to $N_{\text{sph}} = 75$. The entire pipeline is trained end-to-end by minimizing a supervised classification loss; no reconstruction term or KL divergence is required, as our focus is solely on the discriminative power of the curved latent space. The radius R governs the curvature of the embedding space and is treated as a fixed hyperparameter.

Training procedure. We summarize the experimental procedure as Algorithm 1.

Radius and curvature settings. We configure the radius settings separately for different datasets in our experiments. Specifically, for PACS and Digits-DG, we choose radii $R \in \{5, 2, 1\}$, corresponding to curvatures $\mathcal{C} = 0.04,\ 0.25,\ 1$, respectively. For Office-Home, which contains 65

categories and is more challenging to train, we adopt larger radii $R \in \{9, 7, 5\}$, yielding curvatures $\mathcal{C} \approx 0.012,\ 0.020,\ 0.040$. This choice is motivated by the fact that Office-Home has a significantly larger number of classes than the other two datasets; a small radius in the latent space would lead to excessive representation collapse on the hypersphere, making it difficult to preserve sufficient discriminative information and thereby hindering training convergence and limiting accuracy improvement. Therefore, we employ larger radii on Office-Home to provide greater representational capacity, while still using the same mixture-of-spheres architecture across all datasets. This enables the model to adaptively assign different data patterns to different centers, effectively mitigating the training difficulty associated with a single sphere at small radii while continuing to benefit from the curvature-driven generalization guarantee.

Discussion of the Photo anomaly. In most cases, increasing the latent space curvature (i.e., decreasing the hyperspherical radius) leads to a smaller domain generalization error. A notable exception is the case where Photo is used as the target domain in PACS: the domain generalization error remains consistently low across all curvature values, and no clear relationship between curvature and generalization performance is evident. We hypothesize that this anomaly arises because the target Photo domain may be particularly similar to one of the source domains (e.g., Art Painting, Cartoon, or Sketch). In such a scenario, the distribution shift between the source domains and the target is minimal, allowing the model to achieve strong generalization even without curvature adjustments. Consequently, the beneficial effect of a more curved latent space becomes less pronounced, as the inherent simplicity of the transfer task overshadows any gains from curvature.

F.3 Implementation Details in Section 4.2

Source domain configurations. For the three benchmark datasets (PACS, Office-Home, and Digits-DG), we consider the following source domain configurations:

1. **Single source domain:** randomly select one domain as the source and evaluate its generalization performance on each of the remaining three domains.
2. **Two source domains:** randomly select a pair of domains as the source set and evaluate on the other two domains.
3. **Three source domains:** use all three domains as sources (standard multi-source setting) and evaluate on the remaining target domain.
4. **Six source domains via MixStyle:** apply the MixStyle (Zhou et al., 2024) strategy to mix the feature statistics of source domains, generating three mixed domains (each obtained by mixing a different pair of source domains), resulting in a total of six source domains.

MixStyle. MixStyle (Zhou et al., 2024) is a method that generates new domain samples by computing style statistics and mixing the style statistics of samples from different source domains.

Experimental setup. For each dataset and each target domain, we train models under the four source configurations described above, using a pretrained ResNet-18 as the backbone, where the hypothesis class corresponds to the classification head of the network. The domain generalization error is defined as the test error on the target domain minus the average training error on the source domains. The network is trained for 50 epochs, and the error is averaged over the final five epochs to reduce random fluctuations.

Visualization. We visualize the results as bar charts, where each bar corresponds to a specific triple of (dataset, target domain, source configuration). For each dataset, we show four subfigures corresponding to the four target domains, resulting in twelve bar charts in total. In each chart, the horizontal axis denotes the different source configurations (single source, two sources, three sources, six sources via MixStyle), and the vertical axis represents the domain generalization error. Error bars indicate the standard deviation across three runs.

Discussion of exceptions. As shown in Figure 2, the results are consistent with our theoretical bound in most cases: increasing the number of source domains generally reduces the domain generalization error. However, when Photo serves as the target domain on PACS, the results exhibit

notable random fluctuations. Moreover, on Office-Home, when Clipart is used as the target domain, the influence of the number of source domains on the generalization error is less pronounced.

We conjecture that this may be due to the inherent characteristics of these specific domains. For PACS with Photo as the target, the Photo domain may be relatively similar to the source domains (e.g., Art Painting, Cartoon, Sketch), making the model less sensitive to the number of source domains and leading to unstable estimates. In the case of Office-Home with Clipart as the target, Clipart may present a unique distribution that is less well-covered by the source domains, causing the model to rely more heavily on domain-invariant features and thereby diminishing the observable effect of varying source domain quantity.