Invariant Risk Minimization Is A Total Variation Model

Zhao-Rong Lai [1]   Weiwen Wang [1]

Abstract

Invariant risk minimization (IRM) is an arising approach to generalize invariant features to different environments in machine learning. While most related works focus on new IRM settings or new application scenarios, the mathematical essence of IRM remains to be properly explained. We verify that IRM is essentially a total variation based on L^2 norm (TV-ℓ_2) of the learning risk with respect to the classifier variable. Moreover, we propose a novel IRM framework based on the TV-ℓ_1 model. It not only expands the classes of functions that can be used as the learning risk and the feature extractor, but also has robust performance in denoising and invariant feature preservation based on the coarea formula. We also illustrate some requirements for IRM-TV-ℓ_1 to achieve out-of-distribution generalization. Experimental results show that the proposed framework achieves competitive performance in several benchmark machine learning scenarios.

1. Introduction

Many machine learning tasks can be reduced to minimizing the learning risk, which is properly designed according to the task. Specifically, the learning risk is empirically minimized on the training set, but the learned model should be used on the unseen data. If there are significant differences between the features of the training and unseen data, the performance of the learning system may deteriorate (Recht et al., 2019; Arjovsky et al., 2020; Lin et al., 2022). For example, if a convolutional neural network (CNN) is trained with pictures of cows in grass and camels in deserts, it may fail to classify easy samples of cows in deserts. This is because the CNN minimizes the training error by classifying grass and deserts as cows and camels, respectively.

The key to solving this misclassification problem is to distinguish between the invariant features (animal shapes) and the spurious features (landscapes) which lead to distributional shift. Invariant Risk Minimization (IRM, (Arjovsky et al., 2020)) emerges as a learning paradigm that estimates nonlinear, invariant, and causal predictors from multiple training environments for such out-of-distribution (OOD) generalization issue. It introduces a gradient norm penalty that measures the optimality of the dummy classifier at each environment. Different variants of IRM have been proposed since then, such as Risk Extrapolation (REx, Krueger et al. 2021), SparseIRM (Zhou et al., 2022), jointly-learning with auxiliary information (ZIN, Lin et al. 2022), and invariant feature learning through independent variables (TIVA, Tan et al. 2023). These works improve the robustness and extendability of IRM.

In this paper, we investigate the mathematical essence of IRM and find that it can be absorbed in a traditional and widely-used operator in mathematics: **total variation (TV)**. TV has long been used in various fields of mathematics and engineering, like optimal control, data transmission, sensing and denoising. For example, TV based on ℓ_1 norm (TV-ℓ_1, (Rudin et al., 1992)) is exploited to develop noise removal algorithms in signal processing. TV-ℓ_1 can also be used in image restoration and processing (Chan et al., 2006; Chen et al., 2006). In other aspects, TV-ℓ_1 is also a tractable approach to data-driven scale selection and function approximation (Chan & Esedoglu, 2005).

Based on this motivation, we reveal some useful mathematical properties of TV-based IRM models that can be exploited in machine learning scenarios. **1.** We formulate and verify that the existing IRM framework is essentially a TV-ℓ_2 model. **2.** We propose a novel IRM framework based on the TV-ℓ_1 model (IRM-TV-ℓ_1). It has two advantages: 1) The set of TV-ℓ_1 integrable functions is larger than that of TV-ℓ_2 integrable functions when the measure is finite (this is the case in IRM-TV-ℓ_2), hence more classes of functions are allowed as the learning risk and the feature extractor in IRM-TV-ℓ_1. 2) TV-ℓ_1 has robust performance in denoising and preserving sharp features based on the coarea formula. This property helps to shape a blocky (piecewise-constant) learning risk that is more robust to the environment change. **3.** We investigate some requirements for IRM-TV-ℓ_1 to achieve OOD generalization, such as a more flexible penalty

---

[1]Department of Mathematics, College of Information Science and Technology, Jinan University, Guangzhou, China. Correspondence to: Weiwen Wang <wangww29@jnu.edu.cn>.

parameter, extendability of the training environment set, and accuracy of the measure. These findings help to explain why IRM is effective theoretically.

2. Preliminaries and Related Works

Note that there are different formulations and interpretations for both IRM and TV models. For the purpose of understanding the proposed approach, we adopt the most convenient formulations without loss of generality.

2.1. Invariant Risk Minimization

Suppose we have a training data set of n samples $\mathcal{D} := \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^n$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the input space and the output space, respectively. These samples are collected from some training environments in the set $\mathcal{E}_{tr}$ (e.g., grass or deserts). Note that we take a general scenario where environment partition is absent, thus the true environment label e for a sample (x_i, y_i) is unknown. The machine learning task is to learn a predictor $\mathcal{X} \to \mathcal{Y}$ such that some predefined risk (or error) metric R can be minimized on the training data $\mathcal{D}$. In the IRM framework (Arjovsky et al., 2020), this predictor is composed of two operators $w \circ \Phi$, where $\Phi : \mathcal{X} \to \mathcal{H}$ is able to extract invariant features and $w : \mathcal{H} \to \mathcal{Y}$ is a classifier.

Suppose we use the loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \times \mathcal{E}_{tr} \to \mathbb{R}$ to compute the prediction error, where the environment variable becomes an argument of $\mathcal{L}$. The empirical risk of the training samples in the environment e is computed by

$$R(w \circ \Phi, e) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}(w \circ \Phi(x_i), y_i, e), \qquad (1)$$

which is the mean loss between the predicted value $w \circ \Phi(x_i)$ and the true value y_i in the environment e. The notations x_i and y_i can be omitted in R for concise expressions, as they are constants. The actual parameters that we should learn are w, Φ, and e.

The original IRM (Arjovsky et al., 2020) can be defined by

$$\min_{w, \Phi} \sum_{e \in \mathcal{E}_{tr}} R(w \circ \Phi, e)$$
$$\text{s. t. } w \in \arg\min_{\tilde{w}} R(\tilde{w} \circ \Phi, e), \quad \forall e \in \mathcal{E}_{tr}. \qquad (\text{IRM})$$

It aims to learn w and Φ simultaneously by minimizing the total risk and forcing w to be a uniform minimizer of the risk under each environment. However, since it is a challenging bi-level optimization problem, a practical surrogate can be used instead (Arjovsky et al., 2020):

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} \left\{ R(1 \circ \Phi, e) + \lambda \|\nabla_w|_{w=1} R(w \circ \Phi, e)\|_2^2 \right\}, \qquad (\text{IRMv1})$$

where Φ becomes the entire invariant predictor and the classifier w is fixed to be a scalar 1. The penalty parameter λ is non-negative, and the ℓ_2 norm is denoted by $\|\cdot\|_2$. The minimization in the constraint of (IRM) is converted into a gradient norm penalty in the objective of (IRMv1).

The REx (Krueger et al., 2021) approach extrapolates the training risks to their affine combinations. The Variance-REx (V-REx) is a simple, stable and effective form of REx that takes the variance of training risks in different environments as the regularizer:

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{tr}} R(1 \circ \Phi, e) + \lambda \mathbb{V}(\{R(1 \circ \Phi, e)\}_{e \in \mathcal{E}_{tr}}), \quad (\text{V-REx})$$

where $\mathbb{V}(\cdot)$ denotes the variance operator. The scalar $w = 1$ is retained to be consistent with (IRMv1).

As for other extensions of IRM, ZIN (Lin et al., 2022) further exploits additional auxiliary information $\{z_i \in \mathcal{Z}\}_{i=1}^n$ to simultaneously learn environment partition and invariant representation. It assumes that the training environment space $\mathcal{E}_{tr}$ is a convex hull of E linearly-independent fundamental environments (basis). Then each environment in $\mathcal{E}_{tr}$ is isomorphic to an element in the E-dimensional simplex

$$\Delta^E := \{v \in \mathbb{R}^E : v \geqslant 0_{(E)} \text{ and } v \cdot 1_{(E)} = 1\}, \quad (2)$$

where $0_{(E)}$ and $1_{(E)}$ denote E-dimensional vectors of 0 and 1, respectively. The inner product of two vectors is denoted by $\cdot$. ZIN aims to learn a mapping $\rho : \mathcal{Z} \to \Delta^E$ that directly serves as the weights for different environments:

$$R(w \circ \Phi, \rho) := \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}(w \circ \Phi(x_i), y_i) \rho(z_i) \in \mathbb{R}^E,$$

$$\min_{w, \Phi} \left\{ R(w \circ \Phi, \frac{1}{E} 1_{(E)}) \cdot 1_{(E)} + \lambda \max_\rho \|\nabla_w R(w \circ \Phi, \rho)\|_2^2 \right\}. \qquad (\text{ZIN})$$

Compared with (1), (ZIN) decomposes the environment argument into a convex combination of E components. The treatment of the environment lies in learning ρ. $\frac{1}{E} 1_{(E)}$ denotes the averaged environment, and $R(w \circ \Phi, \frac{1}{E} 1_{(E)}) \cdot 1_{(E)}$ measures the holistic empirical loss across all environments. The variation $\nabla_w R(w \circ \Phi, \rho) \in \mathbb{R}^{E \times d_{\mathcal{H}}}$ becomes a gradient matrix, and the matrix ℓ_2 norm is used as a penalty. The penalty is maximized w.r.t. ρ to identify the worst environment that causes the largest absolute variation. Then the whole risk and penalty are minimized w.r.t. (w, Φ).

TIVA (Tan et al., 2023) directly uses the auxiliary information as input variables $\{\tilde{x}_i := (x_i, z_i) \in \mathcal{X} \times \mathcal{Z}\}_{i=1}^n$ instead of sending it to the environment predictor ρ. It decomposes $\tilde{x} := (\tilde{x} \odot u^+) + (\tilde{x} \odot (1 - u^+))$ with a gate function $u^+ \in [0, 1]^{d_{\mathcal{X} \times \mathcal{Z}}}$ to identify environmental-related features.

The risk metric of TIVA is defined by

$$R(w \circ \Phi, \rho) := \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathcal{L}}(w \circ \Phi(\tilde{x} \odot u^+), y_i) \rho(\tilde{x} \odot (1 - u^+)). \tag{TIVA}$$

The optimization model is similar to (ZIN).

2.2. Total Variation

Total variation (TV, Rudin et al. 1992) is a useful operator that characterizes the varying property of a function. Let $\Omega \subseteq \mathbb{R}^d$ be an open set and $L^1(\Omega)$ be the space of functions defined on Ω with bounded Lebesgue integrals. Then for any function $f \in L^1(\Omega)$, its TV can be defined by (Chan et al., 2006)

$$\int_{\Omega} |\nabla f| := \sup\{ \int_{\Omega} f(x) \mathrm{div} g(x) \, \mathrm{d}x : \ g \in C_c^1(\Omega, \mathbb{R}^d),$$
$$\|g\|_{L^{\infty}(\Omega)} \leqslant 1\}, \tag{TV}$$

where g is a differentiable vector function with compact support contained in Ω and essential supremum no larger than 1. $\mathrm{div} g$ denotes the divergence of g. In brief, the TV operator measures the local change of f along all dimensions of x, then integrates all these local changes throughout the whole domain Ω. Note that f is not necessarily differentiable and the notation $|\nabla f|$ is symbolic. But if f is differentiable, $|\nabla f|$ truly becomes the modulus (i.e., ℓ_2 norm) of the gradient.

Only functions with bounded TV are meaningful in this realm, leading to the bounded variation (BV) space $\{f \in L^1(\Omega) : \int_{\Omega} |\nabla f| < \infty\}$. To simplify expressions, the term TV function actually means the BV function in this paper. In general, a TV model aims to minimize TV by default without explicitly stating it, because it makes little sense to maximize TV in practice.

The TV operator is known for preserving sharp discontinuities while removing noise and other unwanted fine scale detail of f. This property is based on the coarea formula (Chan et al., 2006)

$$\int_{\Omega} |\nabla f| := \int_{-\infty}^{\infty} \int_{f^{-1}(\gamma)} \mathrm{d}s \, \mathrm{d}\gamma, \tag{3}$$

where $f^{-1}(\gamma) := \{x \in \Omega : f(x) = \gamma\}$ denotes the level set (preimage) of f at γ. TV actually integrates along all contours of $f^{-1}(\gamma)$ for all γ where the differential $\mathrm{d}\gamma$ exists. Hence if f is more blocky (piecewise-constant), it will have a smaller TV. This property is exploited in many signal recovery applications via the following TV-ℓ_1 model (Rudin et al., 1992)

$$\inf_{f \in L^2(\Omega)} \left\{ \int_{\Omega} |\nabla f| + \lambda \int_{\Omega} (f - \tilde{f})^2 \, \mathrm{d}x \right\}, \tag{TV-ℓ_1}$$

where $\tilde{f} \in L^2(\Omega)$ is the ground-truth signal to be recovered and λ is the approximating accuracy parameter. The objective of this model is to preserve sharp discontinuities in the approximation f while leaving noise (especially Gaussian noise) and other unwanted fine scale detail in the residue $(f - \tilde{f})$.

If f satisfies the variation condition $\int_{\Omega} |\nabla f|^2 < \infty$, then the following TV-$\ell_2$ model (Mumford & Shah, 1985) is feasible:

$$\inf_{f \in L^2(\Omega)} \left\{ \int_{\Omega} |\nabla f|^2 + \lambda \int_{\Omega} (f - \tilde{f})^2 \, \mathrm{d}x \right\}. \tag{TV-ℓ_2}$$

In fact, TV-ℓ_2 appears earlier in history than TV-ℓ_1. However, the square-integral $\int_{\Omega} |\nabla f|^2$ does not have the coarea formula (3). Therefore, TV-ℓ_2 generally does not produce a blocky recovered signal f.

3. Methodology

Motivated by the properties of TV, we aim to develop a risk metric that is blocky (piecewise-constant) w.r.t. the environment, so that it can be better generalized to different environments. First, we establish the theoretical framework for IRM to be a TV-ℓ_2 model.

3.1. Conditions for IRM to be TV-ℓ_2

A necessary but not sufficient condition for the classifier w to achieve the minimums in the constraints of (IRM) is

$$\nabla_w R(w \circ \Phi, e) = 0, \quad \forall e \in \mathcal{E}_{tr}. \tag{4}$$

This leads to the surrogate model (IRMv1) which penalizes the gradient norms over all the environments. However, the risk function $R(w \circ \Phi, e)$ usually has a complex nonconvex geometric structure w.r.t. w in deep learning. Thus (4) may achieve a maximum instead of a minimum, and it is intractable to turn (IRMv1) back to (IRM).

From a different perspective, we interpret (IRMv1) as a TV-ℓ_2 model in two ways: 1. Consider the environment features as noise and try to remove it. 2. Recover the useful signal (e.g., animal shapes) from the background environment.

To begin with, we turn w back to an argument in the minimization of (IRMv1). This is reasonable because the learning of w and Φ is a dynamic process.

$$\min_{w, \Phi} \sum_{e \in \mathcal{E}_{tr}} \left\{ R(w \circ \Phi, e) + \lambda \|\nabla_w R(w \circ \Phi, e)\|_2^2 \right\}. \tag{5}$$

We also need the following successive conditions.

Condition 1. There exists a measure μ for e on $(\mathcal{E}_{tr}, \mathscr{F}_{tr})$ where $\mathscr{F}_{tr}$ denotes the σ-algebra on $\mathcal{E}_{tr}$ that includes all the environment combinations in the training set.

Condition 2. Under this measure μ, the feature extractor Φ is uncorrelated to the environment variable e. Moreover, the correlation of the risk metric R to e only lies on the classifier w, i.e., $R(w \circ \Phi, e) \, d\mu \equiv R(w(e) \circ \Phi) \, d\mu$.

Condition 3. w is a measurable function of e in the sense that w is parameterized (either fully or partly) by e. In other words, $w(\cdot) \circ \Phi : \mathcal{E}_{tr} \times \mathcal{H} \to \mathcal{Y}$ is measurable.

Condition 4. $R(w(e) \circ \Phi) \in L^1(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ and $|\nabla_w R(w(e) \circ \Phi)| \in L^2(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$. $L^1(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ and $L^2(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ denote the function spaces with integrable and square-integrable functions under measure μ defined on $(\mathcal{E}_{tr}, \mathscr{F}_{tr})$, respectively.

Condition 1 implies that we are able to measure the magnitude of the environments we encounter in the training set. For instance, (5) actually uses a counting measure on $\mathcal{E}_{tr}$ such that for any $e \in \mathcal{E}_{tr}$, $d\mu(\{e\}) = 1$. Condition 2 is weaker than $R(w \circ \Phi, e) \equiv R(w(e) \circ \Phi)$ and indicates that the environment influences only w but not Φ under an infinitesimal scale of environment change $d\mu$. Then w passes the environment representation to R. This is reasonable because Φ is the invariant feature extractor that we intend to learn. Condition 3 enables us to represent and measure w by e. Finally, Condition 4 is fundamental to ensure boundedness when we compute the total risk and its TV.

Theorem 3.1. *A risk metric $R(w \circ \Phi, e)$ that satisfies Conditions 1∼4 has the following well-defined finite integral and TV-ℓ_2 form w.r.t. w:*

$$\int_\Omega R(w \circ \Phi) \, d\nu \quad and \quad \int_\Omega |\nabla_w R(w \circ \Phi)|^2 \, d\nu, \quad (6)$$

where $\Omega = w(\mathcal{E}_{tr})$ is the image of $\mathcal{E}_{tr}$ under mapping w, and ν is the measure for w induced by μ. If μ is a probability measure, then the above integrals become mathematical expectations

$$\mathbb{E}_w[R(w \circ \Phi)] \quad and \quad \mathbb{E}_w[|\nabla_w R(w \circ \Phi)|^2]. \quad (7)$$

The following IRM-TV-ℓ_2 model is also well-defined:

$$\min_\Phi \mathbb{E}_w[R(w \circ \Phi) + \lambda |\nabla_w R(w \circ \Phi)|^2]. \quad \text{(IRM-TV-ℓ_2)}$$

The proof is presented in Appendix A.1. Note that only Φ is left in the minimizing argument because w has been taken expectation in the objective function. This reveals the mathematical essence of why w can be set as a dummy scalar classifier 1 in (IRMv1).

It follows from Theorem 3.1 and multiplying (5) by $1/|\mathcal{E}_{tr}|$ that the following corollary can be obtained.

Corollary 3.2. *Let $\mathcal{E}_{tr}$ be a finite discrete environment training set with a discrete uniform probability measure μ. Then given Conditions 1∼4, (5) can be generalized to (IRM-TV-ℓ_2).*

By slightly changing Condition 4, (V-REx) can also be generalized like Theorem 3.1.

Condition 4a. $R(\Phi, e) \in L^2(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ with μ being a probability measure.

Proposition 3.3. *Given Conditions 1 and 4a, (V-REx) can be generalized to the following V-REx-ℓ_2 model:*

$$\min_\Phi \mathbb{E}_e[R(\Phi, e)] + \lambda \mathbb{E}_e[(R(\Phi, e) - \mathbb{E}_e[R(\Phi, e)])^2]. \quad \text{(V-REx-ℓ_2)}$$

The proof is presented in Appendix A.2. In (V-REx-ℓ_2), $\mathcal{E}_{tr}$ is not necessarily a finite discrete set and μ is not necessarily a uniform probability measure. Besides, (V-REx-ℓ_2) does not require a classifier w to convey environment information (Conditions 2 and 3). This proposition reveals the mathematical essence of why V-REx can extrapolate risk to a larger region (see Appendix A.2). (V-REx-ℓ_2) can be considered as a variant of TV-ℓ_2 model, as variance is a similar tool to TV that assesses the deviation of a function.

To incorporate ZIN and TIVA into our TV-ℓ_2 framework, we present the following theorem.

Theorem 3.4. *The environment learner ρ is actually a probability measure satisfying Condition 1. Given Conditions 2∼4, ZIN and TIVA can be generalized to the following Minimax-TV-ℓ_2 model*

$$\min_\Phi \left\{ \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda \max_\rho \mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|^2] \right\}, \quad \text{(Minimax-TV-ℓ_2)}$$

where $\mathbb{E}_{w \leftarrow \rho}$ denotes the mathematical expectation w.r.t. w whose measure ν is induced by ρ, and $\frac{1_{(E)}}{E}$ denotes the uniform probability measure for e.

The proof is presented in Appendix A.3. This theorem indicates that this minimax scheme actually selects a probability measure ρ that can maximize the risk variation (the worst case), then minimizes this worst-case total objective by Φ.

3.2. IRM-TV-ℓ_1

In last section, we verify that some typical IRM models and related extensions are essentially TV-ℓ_2 models. However, IRM-TV-ℓ_2 does not have the coarea formula (Chan et al., 2006) that provides a geometric nature of sharp discontinuity preservation, despite having the denoising property. As a result, the processed learning risk R may still be environment-sensitive.

To address this issue, we establish a novel IRM-TV-ℓ_1 framework that enhances the robustness of learning risk to environment changes based on the coarea formula. Since $L^2(\Omega, \mathscr{F}_\Omega, \nu) \subsetneq L^1(\Omega, \mathscr{F}_\Omega, \nu)$ when $\nu(\Omega) < \infty$ (this is usually the case especially when ν is a probability measure,

including IRM-TV-ℓ_2 and Minimax-TV-ℓ_2), IRM-TV-ℓ_1 allows broader classes of functions to be the learning risk R and the feature extractor Φ, which enables us to address more complicated tasks and explore more feature extractors, respectively. To achieve this, we can relax Condition 4 as the following Condition 4b. It is a necessary but not sufficient condition for Condition 4 when $\mu(E_{tr}) < \infty$.

Condition 4b. $R(w(e) \circ \Phi), |\nabla_w R(w(e) \circ \Phi)| \in L^1(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$.

Theorem 3.5. Ω, ν, $\mathbb{E}_w$ and ρ are the same as those in Theorems 3.1 and 3.4. A risk metric $R(w \circ \Phi, e)$ satisfying Conditions 1~3 and 4b has the following well-defined integral, TV-ℓ_1 form and mathematical expectations w.r.t. w:

$$\int_\Omega R(w \circ \Phi)\, \mathrm{d}\nu \quad and \quad \int_\Omega |\nabla_w R(w \circ \Phi)|\, \mathrm{d}\nu, \quad (8)$$

$$\mathbb{E}_w[R(w \circ \Phi)] \quad and \quad \mathbb{E}_w[|\nabla_w R(w \circ \Phi)|]. \quad (9)$$

The following IRM-TV-ℓ_1 and Minimax-TV-ℓ_1 models are also well-defined:

$$\min_\Phi \left\{ \mathbb{E}_w[R(w \circ \Phi)] + \lambda (\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2 \right\},$$
$$\text{(IRM-TV-}\ell_1\text{)}$$

$$\min_\Phi \{ \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda \max_\rho (\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2 \}.$$
$$\text{(Minimax-TV-}\ell_1\text{)}$$

The proof is presented in Appendix A.4. Note that $(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2 = \|\nabla_w R(w \circ \Phi)\|_1^2$ if $\| \cdot \|_1$ is the L^1 norm for $L^1(\Omega, \mathscr{F}_\Omega, \nu)$, while $\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|^2] = \|\nabla_w R(w \circ \Phi)\|_2^2$ if $\| \cdot \|_2$ is the L^2 norm for $L^2(\Omega, \mathscr{F}_\Omega, \nu)$. Hence we retain a squared TV-ℓ_1 term in (IRM-TV-ℓ_1) to be consistent with the squared TV-ℓ_2 term in (IRM-TV-ℓ_2), so as (Minimax-TV-ℓ_1).

The term $|\nabla_w R(w \circ \Phi)|$ in (IRM-TV-ℓ_1) and (Minimax-TV-ℓ_1) is non-differentiable w.r.t. Φ. Therefore, conventional backpropagation or other gradient-type methods cannot be used in this model. We develop a closed-form subgradient computation to solve (IRM-TV-ℓ_1) and (Minimax-TV-ℓ_1), shown in Appendix B.1. It operates similarly to a gradient computation and will not increase computational complexity.

The following proposition ensures that IRM-TV-ℓ_1 has the coarea formula, which produces a learning risk that is blocky (piecewise-constant) w.r.t. the environment. It explains the mathematical essence of why IRM-TV-ℓ_1 is able to learn invariant features.

Proposition 3.6. *Assume the induced measure ν for w to be the Lebesgue measure on $\mathbb{R}^d$. Given Conditions 1~3 and*

4b, the following coarea formula holds:

$$\int_\Omega |\nabla_w R(w \circ \Phi)|\, \mathrm{d}\nu = \int_{-\infty}^\infty \int_{\{w \in \Omega: R(w \circ \Phi) = \gamma\}} \mathrm{d}s\, \mathrm{d}\gamma, \quad (10)$$

where s is the $(d-1)$-dimensional Hausdorff measure (i.e., area in $(d-1)$ dimensions).

The proof is given in Appendix A.5.

3.3. Out-of-distribution Generalization

OOD generalization (Ben-Tal et al., 2009) refers to the ability of a trained model to be generalized to an unseen domain. It can be formally defined as follows (Arjovsky et al., 2020; Sagawa et al., 2020; Duchi & Namkoong, 2021; Krueger et al., 2021; Lin et al., 2022):

$$\min_\Phi \max_{e \in \mathcal{E}_{all}} R(\Phi, e), \quad \text{(OOD)}$$

where $\mathcal{E}_{all}$ denotes the global environment set that contains all the possible environments that can occur in the test, and Φ denotes the model parameters for R. Two facts are implied in this framework: 1. $\max_{e \in \mathcal{E}_{all}} R(\Phi, e) < \infty$ for some Φ, or else it cannot be minimized by any Φ. 2. Both $\arg\max_{e \in \mathcal{E}_{all}} R(\Phi, e)$, $\forall \Phi$ and $\arg\min_\Phi \max_{e \in \mathcal{E}_{all}} R(\Phi, e)$ exist, or else the worst-case environment or the optimal model parameters cannot be identified. To absorb the environment variable e into the classifier w, Condition 2 should be strengthened:

Condition 2a. The representation of the risk metric R by e only lies on the classifier w, i.e., $R(\Phi, e) \equiv R(w(e) \circ \Phi)$.

We assume that the following Theorems 3.7, 3.8 and 3.9 satisfy Conditions 1,2a,3 and 4b on the probability space $(\mathcal{E}_{all}, \mathscr{F}_{all}, \mu_{all})$.

Theorem 3.7. (OOD) *is equivalent to*

$$\min_\Phi \max_{w \in \Omega_{all}} R(w \circ \Phi), \quad \text{(OOD-}w\text{)}$$

where $\Omega_{all} = w(\mathcal{E}_{all})$.

The proof is given in Appendix A.6. In fact, IRM, ZIN and TIVA implicitly exploit this theorem to set up their models.

On the other hand, (IRM-TV-ℓ_1) under the global environment set becomes

$$\min_\Phi \left\{ \mathbb{E}_w^{all}[R(w \circ \Phi)] + \lambda (\mathbb{E}_w^{all}[|\nabla_w R(w \circ \Phi)|])^2 \right\},$$
$$\text{(IRM-TV-}\ell_1\text{-global)}$$

where $\mathbb{E}_w^{all}$ denotes the expectation w.r.t. w on the induced probability space $(\Omega_{all}, \mathscr{F}_{\Omega_{all}}, \nu_{all})$. Our first step to establish OOD generalization is to investigate the conditions under which (IRM-TV-ℓ_1-global) is really equivalent to (OOD-w).

Theorem 3.8 (IRM-TV-ℓ_1-global Achieving OOD Generalization). *1) The penalty parameter λ should be allowed to vary with Φ. Otherwise, (IRM-TV-ℓ_1-global) cannot achieve (OOD-w).*

2) For each Φ, if $\mathbb{E}_w^{all}[|\nabla_w R(w \circ \Phi)|] > 0$, then there exists some $\lambda_\Phi \geqslant 0$ depending only on Φ such that

$$\max_{w \in \Omega_{all}} R(w \circ \Phi)$$
$$= \mathbb{E}_w^{all}[R(w \circ \Phi)] + \lambda_\Phi (\mathbb{E}_w^{all}[|\nabla_w R(w \circ \Phi)|])^2. \quad (11)$$

If $\mathbb{E}_w^{all}[|\nabla_w R(w \circ \Phi)|] = 0$, (11) still holds when $R(w \circ \Phi)$ is Lipschitz continuous w.r.t. w.

3) An optimal point $\Phi^\bullet$ of the following model

$$\min_\Phi \left\{ \mathbb{E}_w^{all}[R(w \circ \Phi)] + \lambda_\Phi (\mathbb{E}_w^{all}[|\nabla_w R(w \circ \Phi)|])^2 \right\}$$
(IRM-TV-ℓ_1-global-Φ)

is also an optimal point of (OOD-w), and vice versa.

The proof is given in Appendix A.7. Item 1) allows λ to vary with Φ. Without this variation, even a simple function fitting task cannot be generalized, as shown in Appendix A.7. Item 2) ensures the existence of λ_Φ that fills the gap between the maximum and the expectation of $R(w \circ \Phi)$. Item 3) establishes the equivalence between the optimal points of (IRM-TV-ℓ_1-global) and (OOD-w).

Theorem 3.9 (Minimax-TV-ℓ_1-global Achieving OOD Generalization). *1) The penalty parameter λ should be allowed to vary with Φ. Otherwise, the following model cannot achieve (OOD-w):*

$$\min_\Phi \left\{ \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}^{all}[R(w \circ \Phi)] + \lambda \max_\rho (\mathbb{E}_{w \leftarrow \rho}^{all}[|\nabla_w R(w \circ \Phi)|])^2 \right\}.$$
(Minimax-TV-ℓ_1-global)

2) For each Φ, if $\mathbb{E}_{w \leftarrow \rho}^{all}[|\nabla_w R(w \circ \Phi)|] > 0$ for some ρ, then there exists some $\lambda_\Phi \geqslant 0$ depending only on Φ such that

$$\max_{w \in \Omega_{all}} R(w \circ \Phi)$$
$$= \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}^{all}[R(w \circ \Phi)] + \lambda_\Phi \max_\rho (\mathbb{E}_{w \leftarrow \rho}^{all}[|\nabla_w R(w \circ \Phi)|])^2.$$
$$(12)$$

If $\mathbb{E}_{w \leftarrow \rho}^{all}[|\nabla_w R(w \circ \Phi)|] = 0$ for all ρ, (11) still holds when $R(w \circ \Phi)$ is Lipschitz continuous w.r.t. w.

3) An optimal point $\Phi^\bullet$ of the following model

$$\min_\Phi \left\{ \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}^{all}[R(w \circ \Phi)] + \lambda_\Phi \max_\rho (\mathbb{E}_{w \leftarrow \rho}^{all}[|\nabla_w R(w \circ \Phi)|])^2 \right\}$$
(Minimax-TV-ℓ_1-global-Φ)

is also an optimal point of (OOD-w), and vice versa.

The proof is given in Appendix A.8.

Theorems 3.7, 3.8, and 3.9 reveal the conditions under which TV-ℓ_1 models can achieve OOD generalization under the global environment set. We then investigate whether the training environment set can be generalized to the global environment set. In most cases, we encounter some fundamental environments in the training set that form a basis for environment representation, either explicitly (e.g., IRM, REx) or implicitly (e.g., ZIN, TIVA).

Definition 3.10 (Basis from Training Environments).

$$\mathcal{B}_{tr} := \{e_i\}_{i \in E}, \quad (13)$$

where E denotes the index set for this basis. E can be finite, countable or uncountable.

Normally, $\mathcal{B}_{tr} \subseteq \mathcal{E}_{all}$, and usually $\mathcal{B}_{tr} \subsetneq_{\not\supseteq} \mathcal{E}_{all}$ because there may be unseen environments outside the training set. This is why OOD generalization methods try to expand $\mathcal{B}_{tr}$ from different approaches. An intuitive and popular approach is to consider the linear space spanned by $\mathcal{B}_{tr}$:

$$\mathcal{E}_{tr}^{\mathscr{L}} := \mathscr{L}(\mathcal{B}_{tr}). \quad (14)$$

REx (Krueger et al., 2021) explicitly takes this form for the training environment set (see Appendix A.2). Although ZIN and TIVA aim to expand $\mathcal{B}_{tr}$ to its convex hull, they actually take $\mathcal{E}_{tr}^{\mathscr{L}}$ as well and adopt a probability measure ρ on it according to Theorem 3.4.

The following theorem reveals the minimum requirements for global environment set generalization.

Theorem 3.11. *Under the same conditions as in Theorem 3.5, $\mathcal{E}_{all} \in \mathscr{F}_{tr}$ is the minimum requirement for (IRM-TV-ℓ_1) and (Minimax-TV-ℓ_1) to be generalized to (IRM-TV-ℓ_1-global-Φ) and (Minimax-TV-ℓ_1-global-Φ), respectively.*

The proof is given in Appendix A.9. This theorem reveals two important facts:

1. $\mathcal{E}_{tr}$ should be abundant and diverse enough. In particular, $\mathcal{E}_{tr} \supseteq \mathcal{E}_{all}$ is necessary. It may seem impossible at first thought, but one can try to expand $\mathcal{B}_{tr}$ as demonstrated above, like $\mathcal{E}_{tr}^{\mathscr{L}}$. After all, a model can hardly foresee the environment of the cosmos in $\mathcal{E}_{all}$ with only grass and deserts in $\mathcal{B}_{tr}$. But it may be able to learn the environment with half deserts and half grass, or even the environment of a forest (see Figure 1). The maximum σ-algebra of $\mathcal{E}_{tr}$ is its power set $2^{\mathcal{E}_{tr}}$. If $\mathcal{E}_{all} \notin 2^{\mathcal{E}_{tr}}$, then distortion is inevitable in (IRM-TV-ℓ_1-global-Φ) and (Minimax-TV-ℓ_1-global-Φ). Recently, the spurious feature diversification strategy (Lin et al., 2023) has been proposed to expand $\mathcal{E}_{tr}$, which is consistent with the above analysis.

2. The measure μ for e on $(\mathcal{E}_{tr}, \mathscr{F}_{tr})$ should be accurate enough. Even if $\mathcal{E}_{tr} \supseteq \mathcal{E}_{all}$ cannot guarantee a measur-
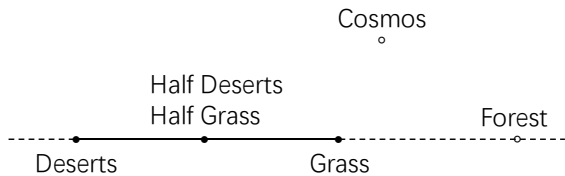
Cosmos

Half Deserts
Half Grass

Forest

Deserts   Grass

Figure 1. A simple example for the linear environment space $\mathcal{E}_{tr}^{\mathscr{L}}$ (the dashed line).

able $\mathcal{E}_{all}$. For example, let $\mathcal{E}_{tr} \supsetneq \mathcal{E}_{all} \supsetneq \emptyset$ and the measure μ be the coarsest one such that $\mathscr{F}_{tr} = \{\emptyset, \mathcal{E}_{tr}\}$, then $\mu(\mathcal{E}_{all})$ does not exist. For another example, let $\mathcal{E}_{tr} = \mathbb{R}^E$, then the most accurate measure that can be realized by modern computing technologies is the Lebesgue measure μ_L defined on the Lebesgue σ-algebra $\mathscr{F}_L$. However, $\mathscr{F}_L \subsetneq 2^{\mathbb{R}^E}$ and there are non-measurable sets such as the Vitali set. Hence if $\mathcal{E}_{all} \notin \mathscr{F}_L$, distortion is also inevitable in (IRM-TV-ℓ_1-global-Φ) and (Minimax-TV-ℓ_1-global-Φ). It makes little sense to only expand $\mathcal{E}_{tr}$ without equipping it with an accurate μ.

Theorem 3.11 actually deploys $\mathscr{F}_{tr}$ as the σ-algebra for $\mathcal{E}_{all}$. However, $\mathcal{E}_{all}$ may have its own information structure, represented by its own σ-algebra $\mathscr{F}_{all}$ and measure μ_{all}. Then it will be more complicated to explore (OOD-w), leading to the following sufficient requirements.

Corollary 3.12. *Under the same conditions as in Theorem 3.5, $\mathscr{F}_{tr} \supseteq \mathscr{F}_{all}$ is the sufficient requirement for* (IRM-TV-ℓ_1) *and* (Minimax-TV-ℓ_1) *to be generalized to* (IRM-TV-ℓ_1-global-Φ) *and* (Minimax-TV-ℓ_1-global-Φ), *respectively.*

To summarize this subsection, there are three ways to improve OOD generalization: 1. Allow λ_Φ to vary with Φ. 2. Expand the training environment basis $\mathcal{B}_{tr}$ to a larger training environment set $\mathcal{E}_{tr}$ by all means. 3. Equip $\mathcal{E}_{tr}$ with a more accurate measure μ.

4. Experiments

We assess the proposed IRM-TV-ℓ_1 framework by comparisons with existing OOD methods in simulation and real-world experiments, following (Lin et al., 2022; Tan et al., 2023). Seven state-of-the-art methods: IRM (Arjovsky et al., 2020), groupDRO (Sagawa et al., 2020), EIIL (Creager et al., 2021), LfF (Nam et al., 2020), HRM (Liu et al., 2021), ZIN (Lin et al., 2022), and TIVA (Tan et al., 2023), as well as the baseline Empirical risk minimization (ERM) are taken into comparisons. IRM, groupDRO, and IRM-TV-ℓ_1 require ground-truth environment partitions, while Minimax-TV-ℓ_1 is the counterpart of IRM-TV-ℓ_1 for non-environment-partition situations. Implementation details are presented in Appendix B. Each experiment is repeated for 10 times to compute the averaged result and the standard deviation (STD) for each method.

4.1. Simulation Study

Each synthetic data set is characterized by temporal heterogeneity with distributional shift w.r.t time and has been used for OOD generalization evaluation in (Lin et al., 2022; Tan et al., 2023). For time $t \in [0, 1]$, the interested binary outcome is denoted by $Y(t)$, whose cause and effect are the invariant features $X_v(t) \in \mathbb{R}$ and the spurious features $X_s(t) \in \mathbb{R}$, respectively (see Appendix B.2 for more details). The correlation between $X_v(t)$ and $Y(t)$ is stable and controlled by a parameter p_v, while the correlation between $X_s(t)$ and $Y(t)$ is changing over t and controlled by a parameter $p_s(t)$. In training sample generation, $p_s(t)$ is fixed as p_s^- for $t \in [0, 0.5)$ and as p_s^+ for $t \in [0.5, 1]$. Thus, the data generation process follows the settings of triplet (p_s^-, p_s^+, p_v). Evaluation on four test environments is performed with $p_s \in \{0.999, 0.8, 0.2, 0.001\}$ and p_v unchanged. Time t is used as the auxiliary variable in ZIN and Minimax-TV-ℓ_1 for environment inference.

We present the mean accuracy and the worst accuracy over the four test environments in Table 1. When environment partitions are unavailable, Minimax-TV-ℓ_1 outperforms the other competitors in all but one case where $(p_s^-, p_s^+) = (0.999, 0.7)$ and $p_v(t) = 0.8$. When environment partitions are available, IRM-TV-ℓ_1 is the best method. As the correlation between $X_s(t)$ and $Y(t)$ grows, the TV-ℓ_1-based models are more robust in two aspects: 1. The gap between the mean and the worst performance of the TV-ℓ_1-based models is generally smaller than that of the other competitors; 2. The performance of the TV-ℓ_1-based models is less affected by such change.

We also show how the TV-ℓ_1-based models improve identifiability of invariant features over the TV-ℓ_2-based models in Appendix C.

4.2. Real-world Experiments

House Price Prediction. We use the House Prices data set[1] to verify the TV-ℓ_1-based models in a regression task. In this experiment, 15 variables including the number of bathrooms, locations, etc., are used to predict the house price. Samples with built year in period $[1900, 1950]$ are used for training and those with built year in period $(1950, 2000]$ are used for test. The house price is normalized within the same built year. The built year is used as an auxiliary variable in both ZIN and Minimax-TV-ℓ_1 for environment inference. The training samples are divided into 5 segments with 10-year range in each segment. Then each segment is considered as having the same environment.

Table 2 reports the mean squared errors (MSE) of competing methods in this regression task. IRM-TV-ℓ_1 achieves the

---

[1]https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Table 1. Upper: accuracies (%) of competing methods on four test environments in simulation study averaged by 10 repetitions. Lower: STDs computed by 10 repetitions.

ENV PARTITION	(p_s^-, p_s^+)	(0.999, 0.7)				(0.999, 0.8)				(0.999, 0.9)			
	$p_v(t)$	0.9		0.8		0.9		0.8		0.9		0.8	
	METHOD	MEAN	WORST	MEAN	WORST	MEAN	WORST	MEAN	WORST	MEAN	WORST	MEAN	WORST
FALSE	ERM	76.22	58.81	59.80	25.95	69.34	43.06	55.96	15.60	60.62	23.30	53.10	8.04
	EIIL	39.43	18.22	64.95	48.45	50.26	47.02	68.86	54.91	61.33	52.70	69.82	58.58
	HRM	76.52	59.78	59.98	26.97	69.87	44.49	56.40	16.85	60.57	23.46	53.16	8.37
	TIVA	82.54	76.74	75.82	70.97	81.53	73.05	69.78	56.23	71.42	49.95	59.47	30.77
	ZIN	87.70	85.86	**78.33**	76.60	86.78	84.86	77.42	75.12	83.42	78.62	74.03	67.45
	MINMAX-TV-ℓ_1	**88.67**	**87.83**	78.14	**76.68**	**88.55**	**87.62**	**78.74**	**77.56**	**87.01**	**85.74**	**77.31**	**74.54**
TRUE	GROUPDRO	72.42	54.90	63.74	43.37	71.09	51.60	62.78	40.21	69.67	47.72	61.81	36.44
	IRM	87.84	86.20	78.33	76.58	86.84	84.42	77.48	74.80	84.16	77.89	74.53	68.72
	IRM-TV-ℓ_1	**88.03**	**86.40**	**78.49**	**76.88**	**87.10**	**84.90**	**77.95**	**75.65**	**84.84**	**80.06**	**75.55**	**70.77**
FALSE	ERM	1.17	2.06	1.04	2.06	1.23	2.47	0.76	1.42	1.10	2.01	0.62	0.95
	EIIL	1.52	3.18	1.46	1.72	1.70	3.09	1.43	2.26	2.46	1.99	1.58	2.04
	HRM	1.35	2.71	0.94	2.43	0.75	1.83	0.71	2.33	0.84	1.29	0.45	0.93
	TIVA	6.12	11.09	3.55	7.18	4.83	9.19	6.46	13.96	5.18	10.34	6.32	13.66
	ZIN	1.05	2.19	1	1.43	1.67	2.73	1.43	2.13	3.52	6.72	2.09	3.86
	MINMAX-TV-ℓ_1	0.57	0.60	0.84	1.03	0.45	0.50	0.67	0.74	1.28	1.66	0.65	1.13
TRUE	GROUPDRO	8.45	18.08	6.99	16.84	8.42	19.03	6.71	17.27	8.27	18.51	6.52	16.45
	IRM	0.82	2.01	0.91	1.49	1.16	2.34	1.82	3.01	1.98	4.11	3.14	4.52
	IRM-TV-ℓ_1	0.86	2.08	0.74	1.33	1.35	2.67	1.24	2.22	2.19	4.77	2.92	4.31

Table 2. Upper: mean squared errors of competing methods in house price prediction averaged by 10 repetitions. Lower: STDs computed by 10 repetitions.

ENV PARTITION	METHOD	TRAIN	TEST	WORST
FALSE	ERM	0.1057	0.4409	0.6206
	EIIL	0.1103	0.3939	0.5581
	HRM	0.5578	0.5949	0.7250
	TIVA	0.2575	0.4418	0.6145
	ZIN	0.2241	0.4293	0.6198
	MINMAX-TV-ℓ_1	0.2168	**0.3395**	**0.4983**
TRUE	GROUPDRO	0.1271	0.7358	1.0611
	IRM	0.5663	0.8168	1.1168
	IRM-TV-ℓ_1	0.3261	**0.4420**	**0.6096**
FALSE	ERM	0.0017	0.0435	0.0641
	EIIL	0.0020	0.0305	0.0460
	HRM	0.0593	0.0025	0.0052
	TIVA	0.0002	0.0019	0.0062
	ZIN	0.1137	0.1994	0.2869
	MINMAX-TV-ℓ_1	0.0652	0.0638	0.0958
TRUE	GROUPDRO	0.0029	0.0877	0.1287
	IRM	0.1389	0.3115	0.4511
	IRM-TV-ℓ_1	0.1279	0.2503	0.3342

Table 3. Upper: accuracies (%) of competing methods on CelebA averaged by 10 repetitions. Lower: STDs computed by 10 repetitions.

ENV PARTITION	METHOD	TRAIN	TEST	WORST
FALSE	ERM	63.76	63.99	62.05
	EIIL	59.12	58.15	54.22
	LFF	57.50	57.73	56.18
	TIVA	64.36	64.23	61.63
	ZIN	78.32	76.73	76.19
	MINMAX-TV-ℓ_1	85.12	**83.68**	**81.45**
TRUE	GROUPDRO	81.50	81.19	79.27
	IRM	85.59	82.54	80.75
	IRM-TV-ℓ_1	84.79	**83.47**	**81.21**
FALSE	ERM	14.45	14.16	14.16
	EIIL	8.74	8.48	10.23
	LFF	0.12	0.24	0.57
	TIVA	1.68	1.99	1.47
	ZIN	1.16	0.87	0.85
	MINMAX-TV-ℓ_1	0.92	0.33	0.43
TRUE	GROUPDRO	0.31	0.48	0.74
	IRM	1.49	1.35	0.99
	IRM-TV-ℓ_1	0.59	0.48	0.67

best results in the environment-partition case, and Minimax-TV-ℓ_1 achieves the best results in the non-environment-partition case. Again, the TV-ℓ_1-based models have smaller gaps between the mean and the worst performance than the other competitors.

CelebA. This data set contains face images of celebrities (Liu et al., 2015). The task is to identify the smiling faces, which is deliberately correlated with gender. 512-dimensional deep features of face images are extracted by a pre-trained ResNet18 (He et al., 2016), and the invariant features are learned by subsequent multilayer perceptrons. Seven additional descriptive variables including *Young*, *Blond Hair*, *Eyeglasses*, *High Cheekbones*, *Big Nose*, *Bags Under Eyes*, and *Chubby* are fed into ZIN and Minimax-TV-ℓ_1 for environment inference. The gender

variable is only used as the environment indicator for group-DRO, IRM, and IRM-TV-ℓ_1.

Table 3 presents the results. The TV-ℓ_1-based models achieve the best accuracies in the mean and the worst scenarios whether environment partitions are available or not. Moreover, Minimax-TV-ℓ_1 is closer to IRM and IRM-TV-ℓ_1 than other environment inference methods (EIIL, LfF, ZIN, and TIVA). It indicates that TV-ℓ_1 narrows the gap between absence and presence of environment information.

Landcover. The Landcover data set records time series and the corresponding land cover types from the satellite data (Gislason et al., 2006; Russwurm et al., 2020; Xie et al., 2021). Time series data with dimension 46×8 are used as input to identify one from six land cover types. The

Table 4. Upper: accuracies (%) of competing methods on Land-cover averaged by 10 repetitions. Lower: STDs computed by 10 repetitions.

METHOD	TRAIN	IID TEST	OOD TEST	WORST
ERM	66.61	66.44	61.54	60.80
EIIL	64.11	63.81	60.43	59.53
LfF	58.12	57.89	55.76	55.07
TIVA	67.49	64.79	52.02	51.46
ZIN	70.02	69.42	62.22	61.87
MINMAX-TV-ℓ_1	73.59	**71.95**	**63.77**	**63.25**
ERM	1.82	1.56	0.92	0.77
EIIL	1.66	1.72	0.88	1.21
LfF	2.73	2.45	1.96	1.93
TIVA	0.28	0.62	0.98	1.09
ZIN	1.09	1.14	1.09	1.21
MINMAX-TV-ℓ_1	0.69	0.63	1.17	1.37

invariant feature extractor Φ is instantiated as a 1D-CNN to handle the time series input, following (Xie et al., 2021; Lin et al., 2022). Ground-truth environment partitions are unavailable, thus latitude and longitude are taken as auxiliary information for environment inference. All methods are trained on non-African data, and then tested on both non-African (not overlapping the training data) and African data. Corresponding results are denoted by IID Test and OOD Test.

As shown in Table 4, Minimax-TV-ℓ_1 achieves the best performance in all of the IID, OOD and Worst Tests, with at least 1.4% higher than the second best competitor. Hence Minimax-TV-ℓ_1 can be better generalized to unseen environments.

Adult Income Prediction. In this task we use the Adult data set[2] to predict if the income of an individual exceeds \$50K/yr based on the census data. We split the data set into four subgroups regarded as separated environments according to $race \in \{\text{Black}, \text{Non-Black}\}$ and $sex \in \{\text{Male}, \text{Female}\}$. We randomly choose two thirds of data from the subgroups Black Male and Non-Black Female for training, and then verify models across all four subgroups with the rest data. Six integral variables: *Age, FNLWGT, Eduction-Number, Capital-Gain, Capital-Loss,* and *Hours-Per-Week* are fead into ZIN and Minimax-TV-ℓ_1 for environment inference. Ground-truth environment indicators are provided for groupDRO, IRM and IRM-TV-ℓ_1. Categorical variables except race and sex are encoded by one-hot coding, followed by the principal component analysis transform with over 99% cumulative explained variance ratio kept. The transformed features are combined with the integral variables, yielding 59-dimensional representations, which are subsequently normalized to have zero mean and unit variance for invariant feature learning.

Results are shown in Table 5. Minimax-TV-ℓ_1 and IRM-TV-

---

[2]https://archive.ics.uci.edu/dataset/2/adult

ℓ_1 achieve the best accuracies in the mean and the worst scenarios within their respective categories. Again, Minimax-TV-ℓ_1 is closer to IRM and IRM-TV-ℓ_1 than other environment inference methods. Hence Minimax-TV-ℓ_1 narrows the gap between absence and presence of environment information.

Table 5. Upper: accuracies (%) of competing methods in adult income prediction averaged by 10 repetitions. Lower: STDs computed by 10 repetitions.

ENV PARTITION	METHOD	TRAIN	TEST	WORST
FALSE	ERM	93.34	82.16	79.55
	EIIL	79.97	72.77	70.94
	LfF	82.03	75.04	73.00
	TIVA	91.45	81.95	79.28
	ZIN	93.16	82.26	79.67
	MINMAX-TV-ℓ_1	92.40	**83.33**	**80.95**
TRUE	GROUPDRO	87.51	76.42	73.07
	IRM	93.19	82.32	79.76
	IRM-TV-ℓ_1	92.42	**83.31**	**80.93**
FALSE	ERM	0.31	0.33	0.37
	EIIL	0.56	0.61	0.73
	LfF	5.54	3.01	2.45
	TIVA	0.12	0.39	0.46
	ZIN	0.17	0.27	0.29
	MINMAX-TV-ℓ_1	0.11	0.14	0.16
TRUE	GROUPDRO	0.59	1.29	1.54
	IRM	0.28	0.23	0.29
	IRM-TV-ℓ_1	0.19	0.18	0.19

5. Conclusion and Discussion

We theoretically show that IRM is essentially an IRM-TV-ℓ_2 model that simultaneously minimizes the total empirical risk and its total variation with respect to the classifier variable. Following this idea, we propose the IRM-TV-ℓ_1 and the Minimax-TV-ℓ_1 models for learning tasks with or without environment partitions, respectively. They allow broader classes of functions to be the learning risk and the feature extractor, and preserve invariant features based on the coarea formula. Moreover, we investigate the requirements for the TV-ℓ_1 framework to achieve out-of-distribution (OOD) generalization. Extensive experiments on both synthetic and real-world data sets show that the TV-ℓ_1 framework performs well in several OOD tests.

A direct impact of this work is to provide a new technical approach that identifies, analyzes and constructs different kinds of invariants with nonsmooth or even discontinuous modules. Although these properties are difficult to handle, they precisely show a reasonable nature of generalization ability and robustness. Further improvements on this work may lie in designing an adaptive penalty parameter to improve OOD generalization, constructing a diverse and representative training environment space, and developing new TV-ℓ_1 models for deep learning. These are nontrivial tasks and we shall put them in future works.

Acknowledgments

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are no potential societal consequences of our work that we feel must be specifically highlighted here.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2020.

Ben-Tal, A., Ghaoui, L. E., and Nemirovski, A. *Robust Optimization*, volume 28. Princeton University Press, 2009.

Chan, T. and Esedoglu, S. Aspects of total variation regularized l^1 function approximation. *SIAM Journal on Applied Mathematics*, (65):1817–1837, 2005.

Chan, T., Esedoglu, S., Park, F., and Yip, A. *Total Variation Image Restoration: Overview and Recent Developments*, pp. 17–31. Springer US, Boston, MA, 2006.

Chen, T., Yin, W., Zhou, X. S., Comaniciu, D., and Huang, T. S. Total variation models for variable lighting face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1519–1524, 2006.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2189–2200, 18–24 Jul 2021.

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

Federer, H. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5815–5826, 2021.

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

Lin, Y., Tan, L., Hao, Y., Wong, H., Dong, H., Zhang, W., Yang, Y., and Zhang, T. Spurious feature diversification improves out-of-distribution generalization, 2023.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Heterogeneous risk minimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6804–6814, 18–24 Jul 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Mumford, D. and Shah, J. Boundary detection by minimizing functionals, I. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22–26, 1985.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684, 2020.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–5400, 2019.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

Russwurm, M., Wang, S., Korner, M., and Lobell, D. Meta-learning for few-shot land cover classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR*, 2020.

Tan, X., Yong, L., Zhu, S., Qu, C., Qiu, X., Yinghui, X., Cui, P., and Qi, Y. Provably invariant learning without domain information. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 33563–33580, 2023.

Xie, S. M., Kumar, A., Jones, R., Khani, F., Ma, T., and Liang, P. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations*, 2021.

Zhou, X., Lin, Y., Zhang, W., and Zhang, T. Sparse invariant risk minimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 27222–27244, 2022.

A. Proofs

A.1. Proof of Theorem 3.1

A.1.1. INDUCED MEASURE ν

First, we verify that there is a measure ν for w induced by μ. From Condition 3, let $\Omega = w(\mathcal{E}_{tr})$ and $\mathscr{F}_\Omega := \{w(\mathcal{F}) : \mathcal{F} \in \mathscr{F}_{tr}\}$. By definition, $\mathscr{F}_\Omega$ is the σ-algebra for Ω under mapping w. Define a set function ν as follows:

$$\nu : \mathscr{F}_\Omega \to \mathbb{R}, \quad \nu(\mathcal{G}) := \mu(w^{-1}(\mathcal{G})). \tag{15}$$

It is well defined because $w^{-1}(\mathcal{G}) \in \mathscr{F}_{tr}$. Now we prove that ν is a measure. Since μ is a measure, the following non-negativity and zero measure are obvious:

$$\forall\, \mathcal{G} \in \mathscr{F}_\Omega, \quad \nu(\mathcal{G}) = \mu(w^{-1}(\mathcal{G})) \geqslant 0, \quad \nu(\emptyset) = \mu(w^{-1}(\emptyset)) = \mu(\emptyset) = 0. \tag{16}$$

To prove countable additivity, let $\{\mathcal{G}_i\}_{i=1}^\infty$ be any countable collection of mutually disjoint sets in $\mathscr{F}_\Omega$. From the property of measurable function, $\{w^{-1}(\mathcal{G}_i)\}_{i=1}^\infty$ is also a countable collection of mutually disjoint sets in $\mathscr{F}_{tr}$. Therefore,

$$\nu(\bigcup_{i=1}^\infty \mathcal{G}_i) = \mu(w^{-1}(\bigcup_{i=1}^\infty \mathcal{G}_i)) = \mu(\bigcup_{i=1}^\infty [w^{-1}(\mathcal{G}_i)]) = \sum_{i=1}^\infty \mu(w^{-1}(\mathcal{G}_i)) = \sum_{i=1}^\infty \nu(\mathcal{G}_i), \tag{17}$$

where the mutually disjoint $\{\mathcal{G}_i\}_{i=1}^\infty$ yields the second equality and the mutually disjoint $\{w^{-1}(\mathcal{G}_i)\}_{i=1}^\infty$ yields the third equality, respectively. Combining (16) and (17), we can conclude that ν is a measure.

A.1.2. CONSTRUCTING MEASURABLE FUNCTIONS $R(w \circ \Phi)$ AND $|\nabla_w R(w \circ \Phi)|$

From Conditions 2 and 4, $R(w(e) \circ \Phi) : \mathcal{E}_{tr} \to \mathbb{R}$ is a measurable function on $(\mathcal{E}_{tr}, \mathscr{F}_{tr})$ and

$$\int_{\mathcal{E}_{tr}} R(w \circ \Phi, e) \,\mathrm{d}\mu = \int_{\mathcal{E}_{tr}} R(w(e) \circ \Phi) \,\mathrm{d}\mu \in (-\infty, +\infty). \tag{18}$$

We can directly construct a function $R(w \circ \Phi) : \Omega \to \mathbb{R}$ on $(\Omega, \mathscr{F}_\Omega)$ and verify its measurability:

$$\forall \omega \in \Omega, \quad R(\omega \circ \Phi) := R(w[w^{-1}(\omega)] \circ \Phi), \tag{19}$$

where $w^{-1}(\omega) \subseteq \mathcal{E}_{tr}$. From Condition 3, we have $\mathcal{E} = w^{-1}(\mathcal{G}) \in \mathscr{F}_{tr}$ for any $\mathcal{G} \in \mathscr{F}_\Omega$. Hence

$$R(\mathcal{G} \circ \Phi) : \mathscr{F}_\Omega \to \mathbb{R}, \quad R(\mathcal{G} \circ \Phi) := R(w[w^{-1}(\mathcal{G})] \circ \Phi) = R(w(\mathcal{E}) \circ \Phi) \tag{20}$$

is well-defined. Next, let $\mathcal{R}_0 \in \mathscr{F}_\mathbb{R}$, where $\mathscr{F}_\mathbb{R}$ denotes the σ-algebra of all the Lebesgue measurable sets contained in $\mathbb{R}$. Then there exists some $\mathcal{E}_0 \in \mathscr{F}_{tr}$ such that

$$R(w(\mathcal{E}_0) \circ \Phi) = \mathcal{R}_0. \tag{21}$$

Let $\mathcal{G}_0 := w(\mathcal{E}_0) \in \mathscr{F}_\Omega$. Since $\mathcal{R}_0$ is arbitrary, we have

$$\mathring{\mathscr{F}}_\Omega := \{\mathcal{G}_0 : \ R(\mathcal{G}_0 \circ \Phi) = \mathcal{R}_0, \ \mathcal{R}_0 \in \mathscr{F}_\mathbb{R}\} \subseteq \mathscr{F}_\Omega. \tag{22}$$

It remains to verify that $\mathring{\mathscr{F}}_\Omega$ is a σ-algebra, forming a sub-σ-algebra of $\mathscr{F}_\Omega$. It is evident that $\emptyset, \Omega \in \mathring{\mathscr{F}}_\Omega$. For any $\mathcal{G}_0 \in \mathring{\mathscr{F}}_\Omega$,

$$\Omega \backslash \mathcal{G}_0 = w[w^{-1}(\Omega \backslash \mathcal{G}_0)] = w[w^{-1}(\Omega) \backslash w^{-1}(\mathcal{G}_0)] = w[\mathcal{E}_{tr} \backslash w^{-1}(\mathcal{G}_0)]. \tag{23}$$

Because $\mathcal{G}_0 \in \mathring{\mathscr{F}}_\Omega \subseteq \mathscr{F}_\Omega$, $w^{-1}(\mathcal{G}_0) \in \mathscr{F}_{tr}$. Since $\mathscr{F}_{tr}$ is closed under complement, $\mathcal{E}_{tr} \backslash w^{-1}(\mathcal{G}_0) \in \mathscr{F}_{tr}$. Thus $w[\mathcal{E}_{tr} \backslash w^{-1}(\mathcal{G}_0)] \in \mathscr{F}_\Omega$. Then

$$R((\Omega \backslash \mathcal{G}_0) \circ \Phi) = R(w[\mathcal{E}_{tr} \backslash w^{-1}(\mathcal{G}_0)] \circ \Phi) = \tilde{\mathcal{R}}_0 \text{ for some } \tilde{\mathcal{R}}_0 \in \mathscr{F}_\mathbb{R}. \tag{24}$$

The first equality follows from (23) and the second equality follows from the measurability of $R(w(e) \circ \Phi)$ w.r.t. e. (24) indicates that $\Omega \backslash \mathcal{G}_0 \in \mathring{\mathscr{F}}_\Omega$ according to (22). Hence $\mathring{\mathscr{F}}_\Omega$ is closed under complement. Finally, let any $\{\mathcal{G}_i \in \mathring{\mathscr{F}}_\Omega\}_{i=1}^\infty$. Then

$$\bigcup_{i=1}^\infty \mathcal{G}_i = w \left[w^{-1}(\bigcup_{i=1}^\infty \mathcal{G}_i) \right] = w \left[\bigcup_{i=1}^\infty w^{-1}(\mathcal{G}_i) \right]. \tag{25}$$

Because $\mathcal{G}_i \in \mathring{\mathscr{F}}_\Omega \subseteq \mathscr{F}_\Omega$, $w^{-1}(\mathcal{G}_i) \in \mathscr{F}_{tr}$. Since $\mathscr{F}_{tr}$ is closed under countable unions, $\bigcup_{i=1}^\infty w^{-1}(\mathcal{G}_i) \in \mathscr{F}_{tr}$. Thus $w\left[\bigcup_{i=1}^\infty w^{-1}(\mathcal{G}_i)\right] \in \mathscr{F}_\Omega$. Then

$$R\left(\left(\bigcup_{i=1}^\infty \mathcal{G}_i \right) \circ \Phi \right) = R \left(w \left[\bigcup_{i=1}^\infty w^{-1}(\mathcal{G}_i) \right] \circ \Phi \right) = \mathcal{R}_\infty \text{ for some } \mathcal{R}_\infty \in \mathscr{F}_\mathbb{R}. \tag{26}$$

The first equality follows from (25) and the second equality follows again from the measurability of $R(w(e) \circ \Phi)$ w.r.t. e. (26) indicates that $\bigcup_{i=1}^\infty \mathcal{G}_i \in \mathring{\mathscr{F}}_\Omega$ according to (22). Hence $\mathring{\mathscr{F}}_\Omega$ is closed under countable unions. To summarize, $\mathring{\mathscr{F}}_\Omega$ satisfies the definition of σ-algebra and forms a sub-σ-algebra of $\mathscr{F}_\Omega$.

On the other hand, the measurable function $|\nabla_w R(w \circ \Phi)|$ can be constructed via the same procedure. Condition 2 implies that $\nabla_w R(w \circ \Phi, e) \, \mathrm{d}\mu \equiv \nabla_w R(w(e) \circ \Phi) \, \mathrm{d}\mu$ since the partial differential is taken on w. Moreover, Condition 4 indicates that

$$\int_{\mathcal{E}_{tr}} |\nabla_w R(w \circ \Phi, e)|^2 \, \mathrm{d}\mu = \int_{\mathcal{E}_{tr}} |\nabla_w R(w(e) \circ \Phi)|^2 \, \mathrm{d}\mu \in (-\infty, +\infty). \tag{27}$$

We are now ready to establish the integrations w.r.t. w from (18) and (27).

A.1.3. INTEGRATIONS W.R.T. w

The integrations w.r.t. w can be established via the integrations w.r.t. e. First, we aim to establish the following:

$$\int_\Omega R(w \circ \Phi) \, \mathrm{d}\nu := \int_{\mathcal{E}_{tr}} R(w(e) \circ \Phi) \, \mathrm{d}\mu, \quad \forall R(w(e) \circ \Phi) \in L^1(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu). \tag{28}$$

A.1.1 and A.1.2 have already verified the eligible measure $\mathrm{d}\nu$ and the measurable function $R(w \circ \Phi)$. We start from the simplest case. Let $\mathcal{G}$ be any measurable set in $\mathscr{F}_\Omega$ and $\mathcal{E} = w^{-1}(\mathcal{G}) \in \mathscr{F}_{tr}$. Define the indicator function of $\mathcal{G}$ as follows:

$$\mathbb{I}_\mathcal{G}(w) = \begin{cases} 1 & \text{if} \quad w \in \mathcal{G}, \\ 0 & \text{if} \quad w \notin \mathcal{G}. \end{cases} \tag{29}$$

We have

$$\int_\Omega \mathbb{I}_\mathcal{G}(w) \, \mathrm{d}\nu = \nu(\mathcal{G}) = \mu(\mathcal{E}) = \int_{\mathcal{E}_{tr}} \mathbb{I}_\mathcal{E}(e) \, \mathrm{d}\mu. \tag{30}$$

This relationship also holds for $a\mathbb{I}_\mathcal{G}(w)$, $\forall a \geqslant 0$. Next, define the sets of non-negative simple functions as

$$\mathcal{S}_\Omega^+ := \left\{ \sum_{i=1}^m a_i \mathbb{I}_{\mathcal{G}_i}(w) : m \in \mathbb{N}^+, a_i \geqslant 0, \mathcal{G}_i \in \mathscr{F}_\Omega \right\} \quad \text{and} \quad \mathcal{S}_{tr}^+ := \left\{ \sum_{i=1}^m a_i \mathbb{I}_{\mathcal{E}_i}(e) : m \in \mathbb{N}^+, a_i \geqslant 0, \mathcal{E}_i \in \mathscr{F}_{tr} \right\}. \tag{31}$$

Then for any $\sum_{i=1}^m a_i \mathbb{I}_{\mathcal{G}_i}(w)$, denote $\mathcal{E}_i = w^{-1}(\mathcal{G}_i)$, $i = 1, \cdots, m$. We have

$$\int_\Omega \sum_{i=1}^m a_i \mathbb{I}_{\mathcal{G}_i}(w) \, \mathrm{d}\nu = \sum_{i=1}^m a_i \int_\Omega \mathbb{I}_{\mathcal{G}_i}(w) \, \mathrm{d}\nu = \sum_{i=1}^m a_i \nu(\mathcal{G}_i) = \sum_{i=1}^m a_i \mu(\mathcal{E}_i) = \sum_{i=1}^m a_i \int_{\mathcal{E}_{tr}} \mathbb{I}_{\mathcal{E}_i}(e) \, \mathrm{d}\mu = \int_{\mathcal{E}_{tr}} \sum_{i=1}^m a_i \mathbb{I}_{\mathcal{E}_i}(e) \, \mathrm{d}\mu. \tag{32}$$

That is, for any $\sum_{i=1}^m a_i \mathbb{I}_{\mathcal{G}_i}(w) \in \mathcal{S}_\Omega^+$, there exists some $\sum_{i=1}^m a_i \mathbb{I}_{\mathcal{E}_i}(e) \in \mathcal{S}_{tr}^+$ satisfying (32).

Next, suppose $R^+(w \circ \Phi)$ is a **non-negative measurable function** w.r.t. w. Then it can be approached by a sequence of functions in $\mathcal{S}_\Omega^+$. Specifically, let

$$a_i := \frac{i}{2^n}, \mathcal{G}_i := \{w : a_i \leqslant R(w \circ \Phi) < a_{i+1}\}, \ i = 0, 1, \cdots, n2^n - 1; \ \mathcal{G}_{R \geqslant n} := \{w : R(w \circ \Phi) \geqslant n\}. \quad (33)$$

Define the sequence of approaching functions as follows:

$$R_n^+(w \circ \Phi) = \sum_{i=0}^{n2^n - 1} a_i \mathbb{I}_{\mathcal{G}_i}(w) + n \mathbb{I}_{\mathcal{G}_{R \geqslant n}}(w), \quad n = 1, 2, 3, \cdots. \quad (34)$$

It can be easily observed that $R_n^+(w \circ \Phi) \in \mathcal{S}_\Omega^+$ and $R_n^+(w \circ \Phi) \uparrow R^+(w \circ \Phi)$ (point-wise monotonically non-decreasing and convergent). Let $\mathcal{E}_i = w^{-1}(\mathcal{G}_i)$, $i = 0, 1, \cdots, n2^n - 1$, and $\mathcal{E}_{R \geqslant n} = w^{-1}(\mathcal{G}_{R \geqslant n})$. Then

$$R_n^+(w(e) \circ \Phi) = \sum_{i=0}^{n2^n - 1} a_i \mathbb{I}_{\mathcal{E}_i}(e) + n \mathbb{I}_{\mathcal{E}_{R \geqslant n}}(e), \quad n = 1, 2, 3, \cdots. \quad (35)$$

We also have $R_n^+(w(e) \circ \Phi) \in \mathcal{S}_{tr}^+$ and $R_n^+(w(e) \circ \Phi) \uparrow R^+(w(e) \circ \Phi)$ for some non-negative measurable function $R^+(w(e) \circ \Phi)$ w.r.t. e. From (32),

$$\int_\Omega R_n^+(w \circ \Phi) \, \mathrm{d}\nu = \int_{\mathcal{E}_{tr}} R_n^+(w(e) \circ \Phi) \, \mathrm{d}\mu, \quad n = 1, 2, 3, \cdots. \quad (36)$$

Moreover, this sequence of integrals is monotonically increasing and thus converges to $\int_{\mathcal{E}_{tr}} R^+(w(e) \circ \Phi) \, \mathrm{d}\mu \in \bar{\mathbb{R}}$ (the extended real space including $+\infty$ and $-\infty$). We can directly define

$$\int_\Omega R^+(w \circ \Phi) \, \mathrm{d}\nu := \int_{\mathcal{E}_{tr}} R^+(w(e) \circ \Phi) \, \mathrm{d}\mu. \quad (37)$$

Since we only consider integrable functions $R^+(w(e) \circ \Phi) \in L^1(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ by Condition 4, (37) indicates that $\int_\Omega R^+(w \circ \Phi) \, \mathrm{d}\nu < +\infty$.

Last, suppose $R(w \circ \Phi)$ is a **measurable function** w.r.t. w. Following conventional methodology, it can be decomposed into a positive part and a negative part:

$$R(w \circ \Phi) = R^+(w \circ \Phi) - R^-(w \circ \Phi), \quad R^+(w \circ \Phi) := \max\{R(w \circ \Phi), 0\}, \ R^-(w \circ \Phi) := \max\{-R(w \circ \Phi), 0\}. \quad (38)$$

Both $R^+(w \circ \Phi)$ and $R^-(w \circ \Phi)$ are non-negative measurable functions. By (37), there exist non-negative measurable functions $R^+(w(e) \circ \Phi)$ and $R^-(w(e) \circ \Phi)$ w.r.t. e such that

$$\int_\Omega R^+(w \circ \Phi) \, \mathrm{d}\nu := \int_{\mathcal{E}_{tr}} R^+(w(e) \circ \Phi) \, \mathrm{d}\mu, \quad \int_\Omega R^-(w \circ \Phi) \, \mathrm{d}\nu := \int_{\mathcal{E}_{tr}} R^-(w(e) \circ \Phi) \, \mathrm{d}\mu. \quad (39)$$

Let $R(w(e) \circ \Phi) = R^+(w(e) \circ \Phi) - R^-(w(e) \circ \Phi)$. Again, we only consider $R(w(e) \circ \Phi) \in L^1(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ by Condition 4, thus both integrals in (39) are finite. Define

$$\int_\Omega R(w \circ \Phi) \, \mathrm{d}\nu := \int_\Omega R^+(w \circ \Phi) \, \mathrm{d}\nu - \int_\Omega R^-(w \circ \Phi) \, \mathrm{d}\nu, \quad (40)$$

then $\int_\Omega R(w \circ \Phi) \, \mathrm{d}\nu \in (-\infty, +\infty)$.

One can apply the above procedure to $|\nabla_w R(w \circ \Phi)|^2$ and exploit $|\nabla_w R(w \circ \Phi)| \in L^2(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ from Condition 4 to define

$$\int_\Omega |\nabla_w R(w \circ \Phi)|^2 \, \mathrm{d}\nu \in [0, +\infty). \quad (41)$$

Note that $|\nabla_w R(w \circ \Phi)|$ is a non-negative measurable function.

A.1.4. EXPECTATIONS W.R.T. w AND WELL-DEFINED (IRM-TV-ℓ_2)

To ensure that $\int_\Omega R(w \circ \Phi) \, \mathrm{d}\nu$ and $\int_\Omega |\nabla_w R(w \circ \Phi)|^2 \, \mathrm{d}\nu$ are mathematical expectations, we need to verify that ν is also a probability measure induced by μ:

$$\nu(\Omega) = \mu(\mathcal{E}_{tr}) = 1, \tag{42}$$

where $\mu(\mathcal{E}_{tr}) = 1$ because μ is a probability measure. $\mathbb{E}_w[R(w \circ \Phi)]$ and $\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|^2]$ denote the two integrals when ν is a probability measure. Since both integrals are finite, $\mathbb{E}_w[R(w \circ \Phi) + \lambda|\nabla_w R(w \circ \Phi)|^2]$ is also finite and can be minimized w.r.t. Φ. Hence (IRM-TV-ℓ_2) is well-defined.

For an extension, one can establish a probability measure as long as $\nu(\Omega) \in (0, +\infty)$:

$$\tilde{\nu}(\cdot) := \nu(\cdot)/\nu(\Omega). \tag{43}$$

It is very useful when Ω is a bounded subset of $\mathbb{R}^d$ and ν is the Lebesgue measure.

A.2. Proof of Proposition 3.3

Given Conditions 1 and 4a, $\mathbb{E}_e[R(\Phi, e)]$ and $\mathbb{E}_e[R^2(\Phi, e)]$ are well-defined and finite. We can use the integral with probability measure μ to represent the first term in (V-REx):

$$\sum_{e \in \mathcal{E}_{tr}} R(\Phi, e) := \mathbb{E}_e[R(\Phi, e)] \in (-\infty, +\infty). \tag{44}$$

Additionally, the variance of $R(\Phi, e)$ in (V-REx) can be generalized to

$$\mathbb{V}(\{R(\Phi, e)\}_{e \in \mathcal{E}_{tr}}) := \mathbb{E}_e[(R(\Phi, e) - \mathbb{E}_e[R(\Phi, e)])^2] = \mathbb{E}_e[R^2(\Phi, e)] - (\mathbb{E}_e[R(\Phi, e)])^2 \in (-\infty, +\infty). \tag{45}$$

Therefore, (V-REx-ℓ_2) is well-defined and finite, and can be minimized.

To illustrate how V-REx-ℓ_2 extrapolates risk to a larger region, we need to examine the graph space of the learning risk: $\mathcal{E}_{tr} \times \mathbb{R} := \{(e, R(\Phi, e)) : e \in \mathcal{E}_{tr}\}$. $\mathbb{E}_e[R(\Phi, e)]$ can be seen as the mean risk of all the environments, which is located in the center. Under some mild assumptions (such as the continuity of $R(\Phi, e)$ w.r.t. e), $\mathbb{E}_e[R(\Phi, e)]$ falls into the range of $R(\Phi, e)$. Then there exists some $\tilde{e}$ such that $R(\Phi, \tilde{e}) = \mathbb{E}_e[R(\Phi, e)]$. This $\tilde{e}$ can be seen as some mean-risk environment, and $(\tilde{e}, \mathbb{E}_e[R(\Phi, e)])$ is in the midst of the graph $\{(e, R(\Phi, e)) : e \in \mathcal{E}_{tr}\}$. The variance in (45) actually involves a set of vectors in the graph space $\{(e, R(\Phi, e)) - (\tilde{e}, \mathbb{E}_e[R(\Phi, e)]) : e \in \mathcal{E}_{tr}\}$ which can span a linear space to extrapolate risk, as shown in Figure A1. The performance of such risk extrapolation depends on the diversity of the training environment set $\mathcal{E}_{tr}$ and the measure μ.
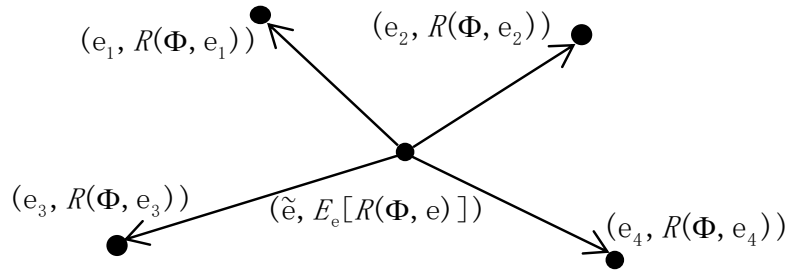

Figure A1. Risk extrapolation of V-REx-ℓ_2 to a larger region.

A.3. Proof of Theorem 3.4

To prove this theorem, we first absorb (ZIN) into the framework of this paper. Suppose the learning model uses a finite set of E fundamental environments (not necessarily known by the model) as the basis defined in Definition 3.10: $\mathcal{B}_{tr} := \{e_i\}_{i=1}^E$. Then the first term in (ZIN) can be rewritten as follows:

$$\tilde{R}(w \circ \Phi, \frac{1}{E} 1_{(E)}) \cdot 1_{(E)} = \sum_{i=1}^E [\tilde{R}(w \circ \Phi, \frac{1}{E} 1_{(E)})]_i = \frac{1}{E} \sum_{i=1}^E R(w \circ \Phi, e_i), \tag{46}$$

where $[\cdot]_i$ denotes the i-th dimension of a vector. Here we use $\tilde{R}$ to denote the original form of learning risk in (ZIN), in order to distinguish it from R in our framework. In fact, $\tilde{R}$ is an E-dimensional vector that represents the learning risk in the E fundamental environments. This representation is equivalent to interpreting R as a function of e_i with e_i varying in the E fundamental environments. Therefore, (46) is true.

Similarly, the second term of (ZIN) can be rewritten as follows:

$$\|\nabla_w \tilde{R}(w \circ \Phi, \rho)\|_2^2 = \sum_{i=1}^{E} [|\nabla_w \tilde{R}(w \circ \Phi, \rho)|]_i^2 = \sum_{i=1}^{E} |\nabla_w R(w \circ \Phi, e_i)|^2 \rho_i, \quad \rho \in \Delta^E. \tag{47}$$

It further assigns a probability $\rho \in \Delta^E$ (see Eq. 2 for the definition) to the fundamental environments, while (46) uses the uniform probability $\frac{1}{E} 1_{(E)}$.

Both (46) and (47) are essentially mathematical expectations:

$$\mathbb{E}_{\frac{1_{(E)}}{E}}[R(w \circ \Phi, e)] := \int_{\mathcal{E}_{tr}} R(w \circ \Phi, e) \, d(\frac{1_{(E)}}{E}) = \frac{1}{E} \sum_{i=1}^{E} R(w \circ \Phi, e_i), \tag{48}$$

$$\mathbb{E}_{\rho}[|\nabla_w R(w \circ \Phi, e)|^2] := \int_{\mathcal{E}_{tr}} |\nabla_w R(w \circ \Phi, e)|^2 \, d\rho = \sum_{i=1}^{E} |\nabla_w R(w \circ \Phi, e_i)|^2 \rho_i. \tag{49}$$

From Theorem 3.1, there exist induced probability measures for w such that

$$\mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] = \mathbb{E}_{\frac{1_{(E)}}{E}}[R(w \circ \Phi, e)], \quad \mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|^2] = \mathbb{E}_{\rho}[|\nabla_w R(w \circ \Phi, e)|^2]. \tag{50}$$

Both ZIN and TIVA aim to learn the probability ρ from the training samples with environment information. They differ mainly in their learning approach. Therefore, both ZIN and TIVA can be generalized to (Minimax-TV-ℓ_2).

A.4. Proof of Theorem 3.5

The proof is almost identical to that of A.1 and A.3, except that $|\nabla_w R(w \circ \Phi)|^2$ is replaced by $|\nabla_w R(w \circ \Phi)|$. Note that the condition $|\nabla_w R(w(e) \circ \Phi)| \in L^1(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ in Theorem 3.5 is weaker than $|\nabla_w R(w(e) \circ \Phi)| \in L^2(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu)$ in Theorem 3.1 when $\mu(\mathcal{E}_{tr}) < \infty$. Therefore, (IRM-TV-$\ell_1$) and (Minimax-TV-$\ell_1$) allow broader classes of $R(w \circ \Phi)$ to be used than (IRM-TV-ℓ_2) and (Minimax-TV-ℓ_2).

A.5. Proof of Proposition 3.6

The key to prove this proposition is to partition the real space into countable continuous intervals such that the level sets of $R(w \circ \Phi)$ w.r.t. these intervals have positive volumes. Specifically, we establish a set of $K \in \mathbb{N}^+ \cup \{+\infty\}$ intervals

$$\{B_k := [a_k, b_k)\}_{k=1}^{K} \quad \text{s. t.} \quad a_k < b_k \leqslant a_{k+1}, \quad k = 1, 2, \cdots, K,$$
$$R(w \circ \Phi) \text{ is Lipschitz continuous on } \mathcal{G}_k := \{w \in \Omega : a_k < R(w \circ \Phi) < b_k\}. \tag{51}$$

Since $a_k < b_k$, each B_k contains at least one rational number. Besides, since $B_k \cap B_j = \emptyset$ for any $k \neq j$, the rational numbers in B_k cannot overlap those in B_j. Since all the rational numbers are dense in $\mathbb{R}$ and the volume $|B_k| := (b_k - a_k) > 0$ for all k, there are at most countable intervals $\{B_k\}_{k=1}^{K}$. On the contrary, if none of such interval B_k exists, it is obvious that the integral in (10) equals 0.

Given the conditions, $\int_\Omega |\nabla_w R(w \circ \Phi)| \, d\nu < \infty$. Since $B_k \subseteq \Omega$ and $|\nabla_w R(w \circ \Phi)| \geqslant 0$, $\int_{B_k} |\nabla_w R(w \circ \Phi)| \, d\nu \leqslant \int_\Omega |\nabla_w R(w \circ \Phi)| \, d\nu < \infty$ holds for all k. Thus, we can decompose the integral as follows:

$$\int_\Omega |\nabla_w R(w \circ \Phi)| \, d\nu = \int_{\Omega \backslash (\bigcup_{k=1}^{K} \mathcal{G}_k)} |\nabla_w R(w \circ \Phi)| \, d\nu + \int_{\bigcup_{k=1}^{K} \mathcal{G}_k} |\nabla_w R(w \circ \Phi)| \, d\nu. \tag{52}$$

Since $\Omega \backslash (\bigcup_{k=1}^{K} \mathcal{G}_k)$ corresponds to $\mathbb{R} \backslash (\bigcup_{k=1}^{K} B_k)$ where the function value $R(w \circ \Phi)$ is discontinuous almost everywhere (a.e.), the Lebesgue measure (i.e., the d-dimensional volume) $d\nu = 0$ on $\Omega \backslash (\bigcup_{k=1}^{K} \mathcal{G}_k)$. Hence, the first term in (52)

$\int_{\Omega \setminus (\bigcup_{k=1}^{K} \mathcal{G}_k)} |\nabla_w R(w \circ \Phi)| \, d\nu = 0$. From (Federer, 1959), the coarea formula holds for all the Lipschitz continuous parts:

$$\int_{\mathcal{G}_k} |\nabla_w R(w \circ \Phi)| \, d\nu = \int_{a_k}^{b_k} \int_{\{w \in \Omega : R(w \circ \Phi) = \gamma\}} ds \, d\gamma, \quad k = 1, 2, \cdots, K. \tag{53}$$

Exploiting the σ-additivity of Lebesgue measure yields

$$\int_{\bigcup_{k=1}^{K} \mathcal{G}_k} |\nabla_w R(w \circ \Phi)| \, d\nu = \sum_{k=1}^{K} \int_{\mathcal{G}_k} |\nabla_w R(w \circ \Phi)| \, d\nu = \sum_{k=1}^{K} \int_{a_k}^{b_k} \int_{\{w \in \Omega : R(w \circ \Phi) = \gamma\}} ds \, d\gamma$$

$$= \int_{\bigcup_{k=1}^{K} B_k} \int_{\{w \in \Omega : R(w \circ \Phi) = \gamma\}} ds \, d\gamma. \tag{54}$$

Besides, it is also obvious to see that $d\gamma = 0$ on $\mathbb{R} \setminus (\bigcup_{k=1}^{K} B_k)$ where $\gamma = R(w \circ \Phi)$ is discontinuous a.e. Hence $\int_{\mathbb{R} \setminus (\bigcup_{k=1}^{K} B_k)} \int_{\{w \in \Omega : R(w \circ \Phi) = \gamma\}} ds \, d\gamma = 0$. Together with (52) and (54), we have

$$\int_{-\infty}^{\infty} \int_{\{w \in \Omega : R(w \circ \Phi) = \gamma\}} ds \, d\gamma = \int_{\bigcup_{k=1}^{K} B_k} \int_{\{w \in \Omega : R(w \circ \Phi) = \gamma\}} ds \, d\gamma = \int_{\bigcup_{k=1}^{K} \mathcal{G}_k} |\nabla_w R(w \circ \Phi)| \, d\nu$$

$$= \int_{\Omega} |\nabla_w R(w \circ \Phi)| \, d\nu. \tag{55}$$

It proves the coarea formula in Proposition 3.6.

This proof also explains the mathematical essence of why TV-ℓ_1 models can learn invariant features. To achieve a smaller TV for $R(w \circ \Phi)$, the continuous interval B_k should take a smaller length, especially on the level set $\{w \in \Omega : R(w \circ \Phi) = \gamma\}$ with a large area. To do this, the optimizing procedure may squash the graph of $R(w \circ \Phi)$ w.r.t. w in every continuous level set $\mathcal{G}_k$, making the graph more blocky (piece-wise constant). This makes $R(w \circ \Phi)$ more robust w.r.t. w in each $\mathcal{G}_k$.

A.6. Proof of Theorem 3.7

From the implied facts of (OOD), we only need to prove that for any Φ such that $\max_{e \in \mathcal{E}_{all}} R(\Phi, e) < \infty$, $\max_{e \in \mathcal{E}_{all}} R(\Phi, e) = \max_{w \in \Omega_{all}} R(w \circ \Phi)$. From Conditions 2a and 3, it suffices to prove that $\forall \Phi$, $\max_{e \in \mathcal{E}_{all}} R(w(e) \circ \Phi) = \max_{w \in \Omega_{all}} R(w \circ \Phi)$. First, we examine how the left side yields the right side. Denote $e^* = \arg \max_{e \in \mathcal{E}_{all}} R(w(e) \circ \Phi)$. Then $R(w(e^*) \circ \Phi) \geqslant \max_{w \in \Omega_{all}} R(w \circ \Phi)$. If not, then there exists $\omega^\bullet \in \Omega_{all}$ such that $R(w(e^*) \circ \Phi) < R(\omega^\bullet \circ \Phi)$. Since w is a surjective mapping, we have $w^{-1}(\omega^\bullet) \neq \emptyset$ and $e^* \notin w^{-1}(\omega^\bullet)$. Picking up any $e^\bullet \in w^{-1}(\omega^\bullet)$, we have $R(w(e^\bullet) \circ \Phi) > R(w(e^*) \circ \Phi)$, which violates the optimality of e^*.

Second, we examine how the right side yields the left side. Denote $w^* = \arg \max_{w \in \Omega_{all}} R(w \circ \Phi)$. Then $R(w^* \circ \Phi) \geqslant \max_{e \in \mathcal{E}_{all}} R(w(e) \circ \Phi)$. If not, then there exists $e^\bullet \in \mathcal{E}_{all}$ such that $R(w^* \circ \Phi) < R(w(e^\bullet) \circ \Phi)$. Hence $R(w^* \circ \Phi) < R(w^\bullet \circ \Phi)$ with $w^\bullet := w(e^\bullet) \in \Omega_{all}$, which violates the optimality of w^*.

Summarizing the inequalities of both directions, we have $\max_{e \in \mathcal{E}_{all}} R(w(e) \circ \Phi) = \max_{w \in \Omega_{all}} R(w \circ \Phi)$, $\forall \Phi$. Thus minimizing either side of this equality w.r.t. Φ leads to the same optimization model.

A.7. Proof of Theorem 3.8

A.7.1. ALLOWING λ TO VARY WITH Φ

To prove that λ should be allowed to vary with Φ, we only need to provide a counterexample with a fixed λ. Consider a simple function fitting problem that uses $w \cdot \Phi$ to fit the constant function -1. We use the absolute fitting error as the learning risk $R(w \circ \Phi) := |w \cdot \Phi - (-1)| = |w \cdot \Phi + 1|$. Suppose we aim to learn the best feature parameter $\Phi \in [-1, 1]$. The classifier w lies uniformly in $[-0.9, 0.1]$, but its center deviates from 0 after conveying the environment information e. Equip w with the uniform probability measure on $[-0.9, 0.1]$ as ν. Now $R(w \circ \Phi) = 1 + w \cdot \Phi$ according to the domains of w and Φ. It can be easily calculated that

$$\max_{w \in \Omega_{all}} R(w \circ \Phi) = \max_{w \in [-0.9, 0.1]} \{1 + w \cdot \Phi\} = \begin{cases} 1 + 0.1\Phi & \text{if} \quad \Phi \geqslant 0 \\ 1 - 0.9\Phi & \text{if} \quad \Phi < 0 \end{cases}, \tag{56}$$

$$\min_{\Phi\in[-1,1]} \max_{w\in\Omega_{all}} R(w\circ\Phi) = 1, \quad \arg\min_{\Phi\in[-1,1]} \max_{w\in\Omega_{all}} R(w\circ\Phi) = 0. \tag{57}$$

However,

$$\mathbb{E}_w^{all}[R(w\circ\Phi)] + \lambda(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|])^2 = \int_{-0.9}^{0.1}(1+w\cdot\Phi)\,\mathrm{d}\nu + \lambda(|\Phi|\int_{-0.9}^{0.1}\mathrm{d}\nu)^2 = 1 - \frac{2}{5}\Phi + \lambda\Phi^2, \tag{58}$$

$$\min_{\Phi\in[-1,1]}\{\mathbb{E}_w^{all}[R(w\circ\Phi)] + \lambda(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|])^2\} = \begin{cases} 1 - \frac{1}{25\lambda} & \text{if} \quad \lambda > \frac{1}{5} \\ \frac{3}{5}+\lambda & \text{if} \quad 0 < \lambda \leqslant \frac{1}{5} \\ \frac{3}{5} & \text{if} \quad \lambda = 0 \end{cases},$$

$$\arg\min_{\Phi\in[-1,1]}\{\mathbb{E}_w^{all}[R(w\circ\Phi)] + \lambda(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|])^2\} = \begin{cases} \frac{1}{5\lambda} & \text{if} \quad \lambda > \frac{1}{5} \\ 1 & \text{if} \quad 0 \leqslant \lambda \leqslant \frac{1}{5} \end{cases}. \tag{59}$$

Comparing (59) with (57), we have

$$\min_{\Phi\in[-1,1]}\{\mathbb{E}_w^{all}[R(w\circ\Phi)] + \lambda(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|])^2\} \neq \min_{\Phi\in[-1,1]} \max_{w\in\Omega_{all}} R(w\circ\Phi),$$

$$\arg\min_{\Phi\in[-1,1]}\{\mathbb{E}_w^{all}[R(w\circ\Phi)] + \lambda(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|])^2\} \neq \arg\min_{\Phi\in[-1,1]} \max_{w\in\Omega_{all}} R(w\circ\Phi), \tag{60}$$

for any fixed $\lambda \geqslant 0$. Hence (IRM-TV-ℓ_1-global) cannot achieve (OOD-w) in terms of the objective value and the argument.

This counterexample can be explained as follows. The learning risk R increases with $|\Phi|$ for the worst-case w. Therefore, $\Phi = 0$ is the optimal point in (OOD-w). However, the deviation of w from 0 also deviates the expectation of R w.r.t. w, leading to the deviation of optimal Φ from 0 in (IRM-TV-ℓ_1-global). This deviation of Φ cannot be offset by the TV term in (IRM-TV-ℓ_1-global), no matter how we set the fixed λ.

A.7.2. EXISTENCE OF λ_Φ

First, we consider the trivial case $\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|] = 0$, which indicates a zero TV term. When $R(w\circ\Phi)$ is Lipschitz continuous w.r.t. w, $\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|] = 0$ implies that $R(w\circ\Phi)$ is constant w.r.t. w. In this case, $\mathbb{E}_w^{all}[R(w\circ\Phi)] = \max_{w\in\Omega_{all}} R(w\circ\Phi)$ and λ can be set arbitrarily.

Next, we investigate the nontrivial case $\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|] > 0$. We have

$$\mathbb{E}_w^{all}[R(w\circ\Phi)] \leqslant \mathbb{E}_w^{all}[\max_{\omega\in\Omega_{all}} R(\omega\circ\Phi)] \leqslant (\max_{\omega\in\Omega_{all}} R(\omega\circ\Phi))\cdot\mathbb{E}_w^{all}[1] = \max_{\omega\in\Omega_{all}} R(\omega\circ\Phi). \tag{61}$$

Note that $\max_{\omega\in\Omega_{all}} R(\omega\circ\Phi)$ is constant w.r.t. w and thus can be moved outside the expectation. Then λ_Φ exists by the form:

$$\lambda_\Phi := \frac{(\max_{w\in\Omega_{all}} R(w\circ\Phi)) - \mathbb{E}_w^{all}[R(w\circ\Phi)]}{(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|])^2} \geqslant 0. \tag{62}$$

This quotient depends only on Φ, but not w.

A.7.3. ACHIEVING OPTIMALITY IN (OOD-w)

$\Phi^\bullet$ being the optimal point of (IRM-TV-ℓ_1-global-Φ) indicates

$$\mathbb{E}_w^{all}[R(w\circ\Phi^\bullet)] + \lambda_\Phi^\bullet(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi^\bullet)|])^2 \leqslant \mathbb{E}_w^{all}[R(w\circ\Phi)] + \lambda_\Phi(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi)|])^2, \forall\Phi. \tag{63}$$

In particular, Φ can be any optimal point Φ^* of (OOD-w). Together with (11), we have

$$\max_{w\in\Omega_{all}} R(w\circ\Phi^\bullet) = \mathbb{E}_w^{all}[R(w\circ\Phi^\bullet)] + \lambda_\Phi^\bullet(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi^\bullet)|])^2$$

$$\leqslant \mathbb{E}_w^{all}[R(w\circ\Phi^*)] + \lambda_\Phi^*(\mathbb{E}_w^{all}[|\nabla_w R(w\circ\Phi^*)|])^2 = \max_{w\in\Omega_{all}} R(w\circ\Phi^*). \tag{64}$$

Hence $\Phi^\bullet$ is also an optimal point of (OOD-w).

Conversely, Φ^* being the optimal point of (OOD-w) indicates

$$\mathbb{E}_w^{all}[R(w \circ \Phi^*)] + \lambda_\Phi^*(\mathbb{E}_w^{all}[|\nabla_w R(w \circ \Phi^*)|])^2 = \max_{w \in \Omega_{all}} R(w \circ \Phi^*)$$

$$\leqslant \max_{w \in \Omega_{all}} R(w \circ \Phi^\bullet) = \mathbb{E}_w^{all}[R(w \circ \Phi^\bullet)] + \lambda_\Phi^\bullet(\mathbb{E}_w^{all}[|\nabla_w R(w \circ \Phi^\bullet)|])^2. \tag{65}$$

Hence Φ^* is also an optimal point of (IRM-TV-ℓ_1-global-Φ).

A.8. Proof of Theorem 3.9

A.8.1. ALLOWING λ TO VARY WITH Φ

The counterexample is similar to that in A.7.1 except for some changes. First, the measure $\nu_{\frac{1_{(E)}}{E}}$ for w induced by $\frac{1_{(E)}}{E}$ can be set as the uniform probability measure on $[-0.9, 0.1]$. The measure ν_ρ for w induced by ρ can be any arbitrary probability measure. Then

$$\mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda \max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2$$

$$= \int_{-0.9}^{0.1} (1 + w \cdot \Phi) \, d\nu_{\frac{1_{(E)}}{E}} + \lambda \max_\rho (|\Phi| \int_{-0.9}^{0.1} d\nu_\rho)^2 = 1 - \frac{2}{5}\Phi + \lambda \Phi^2, \tag{66}$$

which is the same as that in (58). Similar results to (59) and (60) also hold for any fixed $\lambda \geqslant 0$. Thus (OOD-w) cannot be achieved by (Minimax-TV-ℓ_1-global).

A.8.2. EXISTENCE OF λ_Φ

If $\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|] = 0$ for all ρ and $R(w \circ \Phi)$ is Lipschitz continuous w.r.t. w, then $R(w \circ \Phi)$ is constant w.r.t. w. In this case, $\mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] = \max_{w \in \Omega_{all}} R(w \circ \Phi)$ and λ can be set arbitrarily.

If $\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|] > 0$ for some ρ, then $\max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2 > 0$. Similar to (61),

$$\mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] \leqslant \max_{\omega \in \Omega_{all}} R(\omega \circ \Phi). \tag{67}$$

Then λ_Φ exists by the form:

$$\lambda_\Phi := \frac{(\max_{w \in \Omega_{all}} R(w \circ \Phi)) - \mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)]}{\max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2} \geqslant 0. \tag{68}$$

This quotient depends only on Φ, but not w.

A.8.3. ACHIEVING OPTIMALITY IN (OOD-w)

An optimal point $\Phi^\bullet$ of (Minimax-TV-ℓ_1-global-Φ) satisfies

$$\mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi^\bullet)] + \lambda_\Phi^\bullet \max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi^\bullet)|])^2 \leqslant \mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda_\Phi \max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2, \ \forall \Phi. \tag{69}$$

Φ can be any optimal point Φ^* of (OOD-w). By combining (12), we can deduce that

$$\max_{w \in \Omega_{all}} R(w \circ \Phi^\bullet) = \mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi^\bullet)] + \lambda_\Phi^\bullet \max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi^\bullet)|])^2$$

$$\leqslant \mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi^*)] + \lambda_\Phi^* \max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi^*)|])^2 = \max_{w \in \Omega_{all}} R(w \circ \Phi^*). \tag{70}$$

Hence $\Phi^\bullet$ is also an optimal point of (OOD-w).

Conversely, an optimal point Φ^* of (OOD-w) satisfies

$$\mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi^*)] + \lambda_\Phi^* \max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi^*)|])^2 = \max_{w \in \Omega_{all}} R(w \circ \Phi^*)$$

$$\leqslant \max_{w \in \Omega_{all}} R(w \circ \Phi^\bullet) = \mathbb{E}^{all}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi^\bullet)] + \lambda_\Phi^\bullet \max_\rho (\mathbb{E}^{all}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi^\bullet)|])^2. \tag{71}$$

Hence Φ^* is also an optimal point of (Minimax-TV-ℓ_1-global-Φ).

A.9. Proof of Theorem 3.11

We denote the probability space for the training environment set as $(\mathcal{E}_{tr}, \mathscr{F}_{tr}, \mu_{tr})$ and its induced probability space for w as $(\Omega_{tr}, \mathscr{F}_{\Omega_{tr}}, \nu_{tr})$. Since $\mathcal{E}_{all} \in \mathscr{F}_{tr}$, we have $\mathcal{E}_{all} \subseteq \mathcal{E}_{tr}$. Define $\mathscr{F}_{tr|all}$ as the restricted σ-algebra on $\mathcal{E}_{all}$:

$$\mathscr{F}_{tr|all} := \{A \cap \mathcal{E}_{all} : A \in \mathscr{F}_{tr}\}. \tag{72}$$

Since $\mathcal{E}_{all} \in \mathscr{F}_{tr}$, we have $A \cap \mathcal{E}_{all} \in \mathscr{F}_{tr}$. Hence $\mathscr{F}_{tr|all}$ is a sub-σ-algebra of $\mathscr{F}_{tr}$, and it can still use the probability measure μ_{tr}. Then $(\mathcal{E}_{all}, \mathscr{F}_{tr|all}, \mu_{tr})$ can be used as a coarse probability space for $\mathcal{E}_{all}$, and its induced probability space for w is $(\Omega_{all}, \mathscr{F}_{\Omega_{tr|all}}, \nu_{tr})$, where $\Omega_{all} \subseteq \Omega_{tr}$ and $\mathscr{F}_{\Omega_{tr|all}} \subseteq \mathscr{F}_{\Omega_{tr}}$.

Since the learning risk $R(w \circ \Phi)$ and $|\nabla_w R(w \circ \Phi)|$ are defined on $(\Omega_{tr}, \mathscr{F}_{\Omega_{tr}}, \nu_{tr})$, the following objective function and optimization model are well-defined:

$$\int_{\Omega_{all}} R(w \circ \Phi)\, d\nu_{tr} + \lambda_\Phi \Big(\int_{\Omega_{all}} |\nabla_w R(w \circ \Phi)|\, d\nu_{tr}\Big)^2 =: \mathbb{E}_w^{tr}[R(w \circ \Phi)] + \lambda_\Phi (\mathbb{E}_w^{tr}[|\nabla_w R(w \circ \Phi)|])^2, \tag{73}$$

$$\min_\Phi \left\{ \mathbb{E}_w^{tr}[R(w \circ \Phi)] + \lambda_\Phi (\mathbb{E}_w^{tr}[|\nabla_w R(w \circ \Phi)|])^2 \right\}. \tag{IRM-TV-ℓ_1-training-Φ}$$

The only difference of (IRM-TV-ℓ_1-training-Φ) from (IRM-TV-ℓ_1) is the variable λ_Φ, which does not depend on w and its probability space. It can be easily verified that (IRM-TV-ℓ_1-training-Φ) has the same properties and capabilities as (IRM-TV-ℓ_1-global-Φ) in Theorem 3.8.

As for (Minimax-TV-ℓ_1), we denote the probability space for w induced by ρ as $(\Omega_{tr}, \mathscr{F}_{\Omega_{tr}}, \nu_{tr}^\rho)$. Following the above procedure, we obtain $(\Omega_{all}, \mathscr{F}_{\Omega_{tr|all}}, \nu_{tr}^\rho)$ as a coarse probability space for all ρ. Then the following objective function and optimization model are also well-defined:

$$\int_{\Omega_{all}} R(w \circ \Phi)\, d\nu_{tr}^{\frac{1_{(E)}}{E}} + \lambda_\Phi \max_\rho \Big(\int_{\Omega_{all}} |\nabla_w R(w \circ \Phi)|\, d\nu_{tr}^\rho\Big)^2 =: \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}^{tr}[R(w \circ \Phi)] + \lambda_\Phi \max_\rho (\mathbb{E}_{w \leftarrow \rho}^{tr}[|\nabla_w R(w \circ \Phi)|])^2, \tag{74}$$

$$\min_\Phi \left\{ \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}^{tr}[R(w \circ \Phi)] + \lambda_\Phi \max_\rho (\mathbb{E}_{w \leftarrow \rho}^{tr}[|\nabla_w R(w \circ \Phi)|])^2 \right\}. \tag{Minimax-TV-ℓ_1-training-Φ}$$

Similarly, (Minimax-TV-ℓ_1-training-Φ) has the same properties and capabilities as (Minimax-TV-ℓ_1-global-Φ) in Theorem 3.9.

Although (IRM-TV-ℓ_1-training-Φ) and (Minimax-TV-ℓ_1-training-Φ) have the same capabilities as (IRM-TV-ℓ_1-global-Φ) and (Minimax-TV-ℓ_1-global-Φ) to achieve OOD generalization, the former two are not necessarily equivalent to the latter two. Denote the probability space for the global environment set as $(\mathcal{E}_{all}, \mathscr{F}_{all}, \mu_{all})$ and its induced probability space for w as $(\Omega_{all}, \mathscr{F}_{\Omega_{all}}, \nu_{all})$. Let $\mathscr{F}_{\Omega_{tr \cap all}} := \mathscr{F}_{\Omega_{tr|all}} \cap \mathscr{F}_{\Omega_{all}}$ be the intersection of the two σ-algebras, which is also a σ-algebra. It can be anticipated that $\nu_{tr}(A) = \nu_{all}(A), \forall A \in \mathscr{F}_{\Omega_{tr \cap all}}$. That is, the measure for both training and test environments should be the same. However, calculating the integrals of $R(w \circ \Phi)$ and $|\nabla_w R(w \circ \Phi)|$ on $\mathscr{F}_{\Omega_{tr \cap all}}$ may result in distortion when $\mathscr{F}_{\Omega_{tr \cap all}} \subsetneq \mathscr{F}_{\Omega_{tr|all}}$. Because for some $A \in \mathscr{F}_\mathbb{R}$, $\{w : R(w \circ \Phi) \in A\} \in \mathscr{F}_{\Omega_{tr|all}} \backslash \mathscr{F}_{\Omega_{tr \cap all}}$ and thus $\mathbb{I}_{\{w:R(w \circ \Phi) \in A\}}$ cannot be used in A.1.3 to construct $\int_{\Omega_{all}} R(w \circ \Phi)\, d\nu_{tr}$.

Another similar case is $\Omega_{tr} = \Omega_{all}$ but $\mathscr{F}_{\Omega_{tr}} \subsetneq \mathscr{F}_{\Omega_{all}}$. In this scenario, calculating the integrals of $R(w \circ \Phi)$ and $|\nabla_w R(w \circ \Phi)|$ on $\mathscr{F}_{\Omega_{tr}}$ may result in distortion of (IRM-TV-ℓ_1-global-Φ) and (Minimax-TV-ℓ_1-global-Φ). This is because for some $A \in \mathscr{F}_\mathbb{R}$, $\{w : R(w \circ \Phi) \in A\} \in \mathscr{F}_{\Omega_{all}} \backslash \mathscr{F}_{\Omega_{tr}}$ and thus $\mathbb{I}_{\{w:R(w \circ \Phi) \in A\}}$ cannot be used in A.1.3 to construct $\int_{\Omega_{all}} R(w \circ \Phi)\, d\nu_{all}$.

Last, if $\mathscr{F}_{\Omega_{tr}} \supseteq \mathscr{F}_{\Omega_{all}}$, then (IRM-TV-$\ell_1$-training-$\Phi$) and (Minimax-TV-$\ell_1$-training-$\Phi$) imply (IRM-TV-$\ell_1$-global-$\Phi$) and (Minimax-TV-$\ell_1$-global-$\Phi$), as concluded in Corollary 3.12.

B. Implementation Details

B.1. Subgradient Approach to Solve IRM-TV-ℓ_1 and Minimax-TV-ℓ_1

First, we give the definition of the Fréchet subdifferential of $f : \mathbb{R}^{d_\mathcal{Y}} \to \mathbb{R}$ at w denoted by $\partial f(w)$:

Definition B.1 (The Fréchet Subdifferential)**.**

$$\partial f(w) := \left\{ v \in \mathbb{R}^{d_{\mathcal{Y}}} : \liminf_{\substack{u \to w \\ u \neq w}} \frac{f(u) - f(w) - v \cdot (u - w)}{\|u - w\|_2} \geqslant 0 \right\}. \tag{75}$$

The subgradient of the modulus function (i.e., ℓ_2 norm) can be easily calculated:

$$\partial |w| = \begin{cases} \frac{w}{|w|} & \text{if} \quad w \neq 0, \\ \{v \in \mathbb{R}^{d_{\mathcal{Y}}} : |v| \leqslant 1\} & \text{if} \quad w = 0. \end{cases} \tag{76}$$

In this paper, we directly take 0 as the convenient subgradient for $\partial |w|$ at $w = 0$, then $\partial |w|$ can be used in a similar way to a gradient. In particular, the chain rule yields:

$$\partial_{\Phi} |\nabla_w R(w \circ \Phi)| = \begin{cases} \frac{J_{\Phi}^{\top} [\nabla_w R(w \circ \Phi)] * \nabla_w R(w \circ \Phi)}{|\nabla_w R(w \circ \Phi)|} & \text{if} \quad \nabla_w R(w \circ \Phi) \neq 0, \\ 0 & \text{if} \quad \nabla_w R(w \circ \Phi) = 0, \end{cases} \tag{77}$$

where $J_{\Phi}^{\top}[\cdot]$ denotes the transpose of the Jacobian matrix w.r.t. Φ, and $*$ denotes the matrix multiplication. For a simpler form with a one-dimensional w,

$$\partial_{\Phi} |R'_w(w \circ \Phi)| = \text{sign}(R'_w(w \circ \Phi)) \nabla_{\Phi} [R'_w(w \circ \Phi)]. \tag{78}$$

The objective function in (IRM-TV-ℓ_1) and its subgradient are

$$g(\Phi) = \mathbb{E}_w[R(w \circ \Phi)] + \lambda(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2, \tag{79}$$

$$\partial_{\Phi} g(\Phi) = \mathbb{E}_w[\nabla_{\Phi} R(w \circ \Phi)] + 2\lambda \mathbb{E}_w[|\nabla_w R(w \circ \Phi)|] \mathbb{E}_w[\partial_{\Phi} |\nabla_w R(w \circ \Phi)|]. \tag{80}$$

Then the update formula for Φ at the $(k+1)$-th iteration is given by the following equation, where $\eta > 0$ is the learning rate:

$$\Phi^{(k+1)} = \Phi^{(k)} - \eta \partial_{\Phi} g(\Phi^{(k)}). \tag{81}$$

We adopt the adversarial training architecture of (Lin et al., 2022) to solve (Minimax-TV-ℓ_1). Its objective function, gradient and subgradient are

$$h(\Phi, \rho) = \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda(\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2, \tag{82}$$

$$\nabla_{\rho} h(\Phi, \rho) = 2\lambda \mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|] \nabla_{\rho}(\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|]), \tag{83}$$

$$\partial_{\Phi} h(\Phi, \rho) = \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[\nabla_{\Phi} R(w \circ \Phi)] + 2\lambda \mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|] \mathbb{E}_{w \leftarrow \rho}[\partial_{\Phi} |\nabla_w R(w \circ \Phi)|]. \tag{84}$$

Then the update formulae for ρ and Φ at the $(k+1)$-th iteration are given by the following equations, where $\eta_1, \eta_2 > 0$ are the learning rates:

$$\rho^{(k+1)} = \rho^{(k)} + \eta_1 \nabla_{\rho} h(\Phi^{(k)}, \rho^{(k)}), \tag{85}$$

$$\Phi^{(k+1)} = \Phi^{(k)} - \eta_2 \partial_{\Phi} h(\Phi^{(k)}, \rho^{(k+1)}). \tag{86}$$

We use the Adam scheme (Kingma & Ba, 2015) for Pytorch[3] as the optimizer. We apply min-batch subgradients with batch size 1024 in Landcover, and full-batch subgradients in the other data sets. More implementing details can be found in the code link, such as the learning rate, the number of training epochs, etc.

To verify that IRM-TV-ℓ_1 (or Minimax-TV-ℓ_1) will not induce higher computational complexity than IRM-TV-ℓ_2 (or ZIN/Minimax-TV-ℓ_2), we run IRM-TV-ℓ_1, Minimax-TV-ℓ_1, IRM-TV-ℓ_2 and ZIN for 5300 epochs and 10 repetitions, then report the average running times (in seconds) in the following Table B1. The STDs are all less than 1.5 seconds. It indicates that IRM-TV-ℓ_1 (or Minimax-TV-ℓ_1) has almost the same average running times as those of IRM-TV-ℓ_2 (or ZIN).

---

[3]https://pytorch.org

Table B1. Average running times (in seconds) of IRM-TV models by 10 repetitions.

METHOD	SIMULATION	HOUSE PRICES	CELEBA	LANDCOVER	ADULT
ZIN	13.5424	18.7798	23.5480	68.0905	32.6906
MINIMAX-TV-ℓ_1	12.8399	18.7685	23.2884	68.3897	33.1776
IRM-TV-ℓ_2	8.9668	8.9001	22.9603	N/A	11.1240
IRM-TV-ℓ_1	8.4198	9.4552	24.0967	N/A	11.7834

B.2. Synthetic Data Set Generation

We follow (Lin et al., 2022; Tan et al., 2023) to generate synthetic data sets. Let t be the time in $[0, 1]$, and $X_v(t) \in \mathbb{R}$ and $X_s(t) \in \mathbb{R}$ be the invariant and spurious features, respectively. We have

$$X_v(t) \sim \begin{cases} \mathcal{N}(1,1), \text{ w.p. } 0.5; \\ \mathcal{N}(-1,1), \text{ w.p. } 0.5. \end{cases} \quad Y(t) \sim \begin{cases} \text{sign}(X_v(t)), & \text{w.p. } p_v; \\ -\text{sign}(X_v(t)), & \text{w.p. } 1-p_v. \end{cases} \quad X_s(t) \sim \begin{cases} \mathcal{N}(Y(t),1), & \text{w.p. } p_s(t); \\ \mathcal{N}(-Y(t),1), & \text{w.p. } 1-p_s(t). \end{cases}$$

Moreover, $X_v(t)$ and $X_s(t)$ are extended to 5 and 10 dimensional sequences by adding up a standard Gaussian noise, respectively.

B.3. Architectures of Φ and ρ

We follow (Lin et al., 2022; Tan et al., 2023) to instantiate the Pytorch-style architectures of the invariant feature extractor Φ and the environment inferring measure ρ on each data set, shown in Table B2. Φ takes primary variables to learn invariant features, while ρ takes auxiliary variables to learn environment partition. All the compared methods except HRM and TIVA share the same Φ, while HRM and TIVA adopt a linear model for Φ. Minimax-TV-ℓ_1 and ZIN share the same ρ, while TIVA directly uses auxiliary variables as primary variables in Φ. The cross entropy and the mean squared error are used as prediction loss functions for classification tasks and regression tasks, respectively.

Table B2. Pytorch-style architectures of the invariant feature extractor Φ and the environment inferring measure ρ.

DATA SET	Φ	ρ
SIMULATION	LINEAR(15, 1)	LINEAR(1, 16)→RELU()→LINEAR(16, 1) →SIGMOID()
HOUSE PRICES	LINEAR(15, 32)→RELU()→LINEAR(32, 1)	LINEAR(1, 64)→RELU()→LINEAR(64, 4)→SOFTMAX()
CELEBA	LINEAR(512, 16)→RELU()→LINEAR(16, 1)	LINEAR(7, 16)→RELU()→LINEAR(16, 1)→SIGMOID()
LANDCOVER	CONV1D(8, 32, 5)→RELU()→CONV1D(32, 32, 3) →RELU()→MAXPOOL1D(2, 2)→CONV1D(32, 64, 3) →RELU()→MAXPOOL1D(2, 2)→CONV1D(64, 6, 3) →RELU()→AVEPOOL1D(1)	LINEAR(2, 16)→RELU()→LINEAR(16, 2)→SOFTMAX()
ADULT	LINEAR(59, 16)→RELU()→LINEAR(16, 1)	LINEAR(6, 16)→RELU()→LINEAR(16, 4)→SOFTMAX()

C. Invariant Feature Identifiability

We take the simulation study with $(p_s^-, p_s^+, p_v(t)) = (0.999, 0.9, 0.8)$ as an example. There are 15 features in this experiment, where **No. 1~5** are **invariant** features and **No. 6~15** are **spurious** features. We run 10 times of this experiment with different simulated samples and compute the normalized average absolute values of the feature weights, shown in Table C1 and Figure C1. The invariant features of IRM-TV-ℓ_1 and Minimax-TV-ℓ_1 take up 64.88% and 65.08% of the total feature weights, respectively. In contrast, the invariant features of IRM-TV-ℓ_2 and ZIN take up only 58.88% and 55.90% of the total feature weights, respectively. Hence the TV-ℓ_1 models extract more invariant features than the TV-ℓ_2 models. Moreover, the gap between the invariant and spurious feature weights in Figures C1(b) or C1(d) is larger than that in Figures C1(a) or C1(c), respectively. For example, the invariant feature weight w2= 0.0884 and the spurious feature weights w6= 0.0679

and w8= 0.0586 for ZIN, while w2= 0.1210, w6= 0.0403, and w8= 0.0446 for Minimax-TV-ℓ_1, respectively. It indicates that Minimax-TV-ℓ_1 enhances the invariant feature w2 while suppresses the spurious features w6 and w8.

Table C1. Normalized absolute values of feature weights for IRM-TV models in simulation study with $(p_s^-, p_s^+, p_v(t)) = (0.999, 0.9, 0.8)$. **w1~w5** correspond to **invariant** features and **w6~w15** correspond to **spurious** features.

METHOD	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	w15
IRM-TV-ℓ_2	0.1091	0.1143	0.1452	0.0912	0.1290	0.0412	0.0439	0.0454	0.0323	0.0483	0.0361	0.0438	0.0282	0.0526	0.0395
IRM-TV-ℓ_1	0.1232	0.1292	0.1548	0.1090	0.1326	0.0329	0.0329	0.0446	0.0249	0.0371	0.0317	0.0386	0.0262	0.0482	0.0341
ZIN	0.1253	0.0884	0.1118	0.1336	0.0999	0.0679	0.0454	0.0586	0.0353	0.0261	0.0267	0.0376	0.0510	0.0505	0.0419
MINIMAX-TV-ℓ_1	0.1320	0.1210	0.1410	0.1353	0.1215	0.0403	0.0356	0.0446	0.0221	0.0279	0.0285	0.0341	0.0352	0.0395	0.0412

(a) IRM-TV-ℓ_2

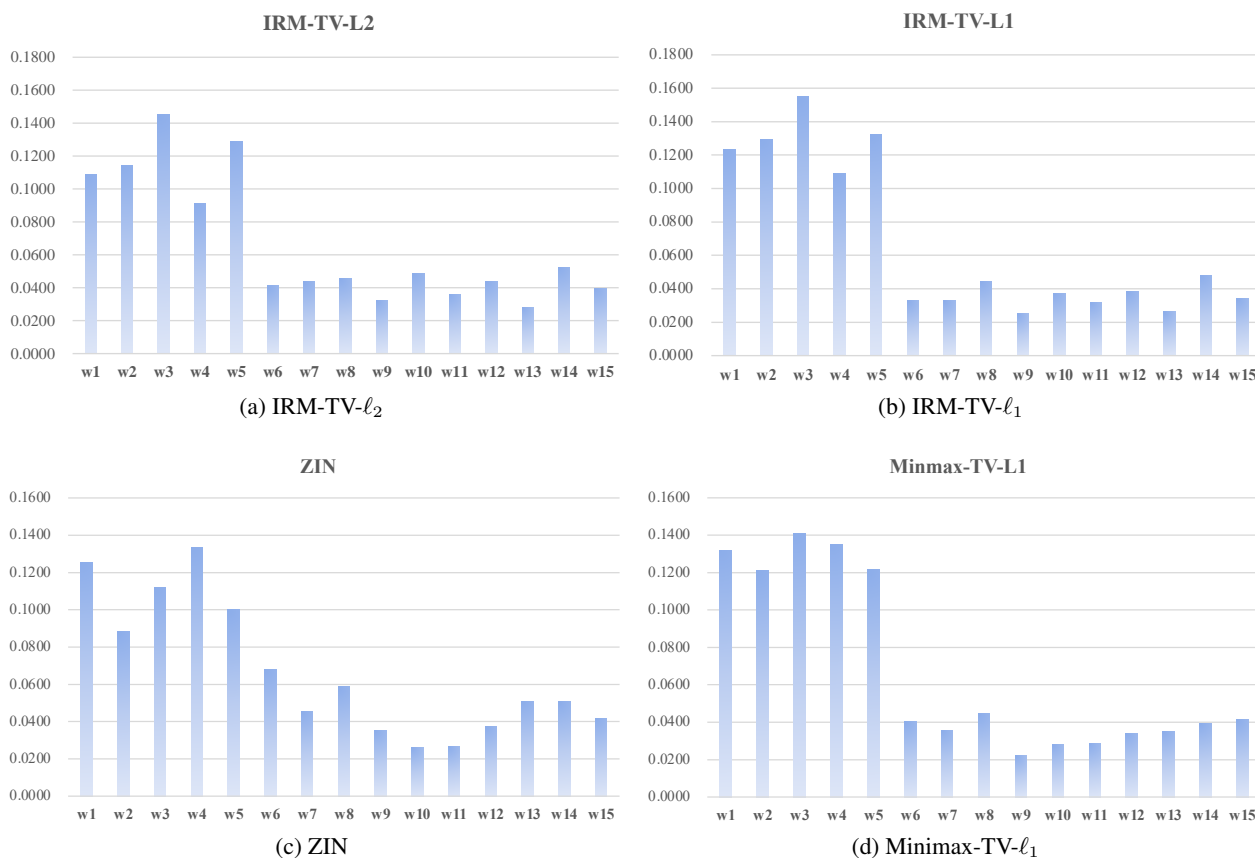(b) IRM-TV-ℓ_1

(c) ZIN

(d) Minimax-TV-ℓ_1

Figure C1. Normalized absolute values of feature weights for IRM-TV models in simulation study with $(p_s^-, p_s^+, p_v(t)) = (0.999, 0.9, 0.8)$. **w1~w5** correspond to **invariant** features and **w6~w15** correspond to **spurious** features.