OUT-OF-DISTRIBUTION GENERALIZATION FOR TOTAL VARIATION BASED INVARIANT RISK MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Invariant risk minimization is an important general machine learning framework that has recently been interpreted as a total variation model (IRM-TV). However, how to improve out-of-distribution (OOD) generalization in the IRM-TV setting remains unsolved. In this paper, we extend IRM-TV to a Lagrangian multiplier model, named OOD-TV-IRM. We find that the autonomous TV penalty hyperparameter is exactly the Lagrangian multiplier. Thus OOD-TV-IRM is essentially a primal-dual optimization model, where the primal optimization reduces the entire invariant risk and the dual optimization strengthens the TV penalty. The objective is to reach a semi-Nash-equilibrium where the balance between the training loss and OOD generalization is kept. We also develop a convergent primal-dual solving algorithm that facilitates an adversarial learning scheme. Experimental results show that OOD-TV-IRM outperforms IRM-TV in most situations.

1 INTRODUCTION

Traditional risk minimization methods such as Empirical Risk Minimization (ERM) are widely used in machine learning. ERM generally assumes that both training and test data come from the same distribution. Based on this assumption, ERM learns model parameters by minimizing the average loss on the training data. However, the distributions between training and test data or even within the training or test data are often different in practical situations, which affects the model's performance in new environments. Besides, ERM tends to exploit correlations in the training data, even if these correlations do not hold in different environments. This may cause the model to learn some spurious features that are irrelevant to the target, so as to perform poorly in new environments. Because ERM only focuses on the average performance on a given data distribution, the model may overfit to specific training data and thus perform unstably when facing distribution changes or outliers. These problems lead to the poor generalization ability of ERM across different environments (Recht et al., 2019; Arjovsky et al., 2019; Lin et al., 2022).

The key to solving the above problems is to distinguish between invariant features and spurious features that cause distribution shifts, so that a trained model can be generalized to an unseen domain. This leads to the concept of out-of-distribution (OOD) generalization (Ben-Tal et al., 2009; Huang et al., 2023). One important methodology is the Invariant Risk Minimization (IRM, Arjovsky et al. 2019) criterion. It aims to extract invariant features across different environments to improve the generalization and robustness of the model (Yang et al., 2023; Xin et al., 2023). Specifically, it introduces a gradient norm penalty that measures the optimality of the virtual classifier in each environment. Then it minimizes the risk in all potential environments to ensure that the model can still work effectively when facing distribution drift. In summary, IRM copes with distribution changes by introducing environmental invariance constraints in order to make the trained model more robust in a wider range of application scenarios (Ahuja et al., 2020). There are various variants and improvements of IRM, such as Heterogeneous Risk Minimization (HRM, Liu et al. 2021), 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). On the other hand, diversifying spurious features (Lin et al., 2023) is also an effective approach to improve IRM. These works not only provide new IRM settings or new application scenarios, but also improve the robustness and extensibility of IRM. In addition, they enrich related theoretical results in OOD generalization.

A recent work reveals that the mathematical essence of IRM is a total variation (TV) model (Lai & Wang, 2024). TV measures the locally varying nature of a function, which is widely applied to different areas of mathematics and engineering, such as signal processing and image restoration. Its core idea is to eliminate noise by minimizing the TV of a function while preserving sharp discontinuities in the function (e.g., edges in an image (Dey et al., 2006)). Interpreting IRM as a TV model not only provides a unified mathematical framework, but also reveals why the IRM approach can work effectively across different environments. In the case of the original IRM, it actually contains an ℓ_2 norm based TV (TV-ℓ_2) term. Compared with TV-ℓ_2, TV-ℓ_1 further has the coarea formula that provides a geometric nature of sharp discontinuity preservation. This property is well-fitted to the invariant feature extraction of IRM. Hence an IRM-TV-ℓ_1 framework is proposed and it shows better performance than the IRM-TV-ℓ_2 framework. By considering the learning risk as part of the full-variance model, IRM-TV-ℓ_1 performs better in OOD generalization. This finding helps to deal with distributional drift and improve model robustness and generalization in machine learning.

However, IRM-TV-ℓ_1 may not achieve complete OOD generalization, due to insufficient diversity in the training environments and inflexible TV penalty. Lai & Wang (2024) investigate some additional requirements for IRM-TV-ℓ_1 to achieve OOD generalization. A key point is to let the penalty parameter vary according to the invariant feature extractor. However, the authors do not specify a tractable implementation. Thus, how to construct and optimize this framework for broader practical applications remains unresolved, resulting in an urgent need for effective solutions.

To fill this gap, we find that the TV penalty hyperparameter can be used as a Lagrangian multiplier. Then we extend the IRM-TV framework to a **Lagrangian multiplier framework**, named **OOD-TV-IRM**. It is essentially a primal-dual optimization framework. The primal optimization reduces the entire invariant risk, in order to learn invariant features , while the dual optimization strengthens the TV penalty, in order to provide an adversarial interference with spurious features. The objective is to reach a **semi-Nash-equilibrium** where the balance between the training loss and OOD generalization is kept. We also develop a **convergent primal-dual solving algorithm** that facilitates an **adversarial learning** scheme. This work not only extends the theoretical foundation of previous works, but also provides a concrete implementation scheme to improve OOD generalization of machine learning methods across unknown environments.

2 Preliminaries and Related Works

2.1 Invariant Risk Minimization

In machine learning tasks, we often work with a data set consisting of multiple samples, where each sample includes input and output variables. These samples are typically drawn from various environments, but in many scenarios, the specific environment labels are not explicitly available. As a result, the challenge is to train a model that can generalize well across these different environments, even when the environment information is unknown.

IRM (Arjovsky et al., 2019) addresses this challenge by structuring the model into two components: a feature extractor and a classifier. The feature extractor, denoted as Φ, is responsible for identifying invariant features from the input data, while the classifier w makes predictions based on these features. IRM aims to minimize the invariant risk across different training environments, thereby finding a representation that performs consistently well across all environments.

For a given training data set of n samples $\mathcal{D} := \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{n}$, the empirical risk in a given environment e is computed using a loss function $\mathcal{L}$, which measures the discrepancy between the predicted value and the true output. This risk is expressed as:

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

which represents the mean loss over n samples in environment e. IRM simultaneously learns an invariant feature extractor Φ and a classifier w across multiple training environments, such that the classifier minimizes the risk in all environments:

$$\min_{w,\Phi} \sum_{e \in \mathcal{E}_{\text{tr}}} R(w \circ \Phi, e), \quad \text{s. t.} \quad w \in \arg\min_{w'} R(w' \circ \Phi, e), \forall e \in \mathcal{E}_{\text{tr}}, \tag{2}$$

where the classifier w is required to be the optimal solution for minimizing the risk in each environment e. However, it is a challenging bi-level optimization problem. Hence Arjovsky et al. (2019) further propose a practical variant of IRM as follows:

$$\min_{\Phi} \sum_{e \in \mathcal{E}_{\text{tr}}} \left\{ R(1 \circ \Phi, e) + \lambda \|\nabla_w|_{w=1} R(w \circ \Phi, e)\|_2^2 \right\}. \tag{3}$$

In this version, the classifier w is fixed to a scalar value of 1, and a gradient norm penalty term is introduced to convert the objective into a single-level optimization problem.

ZIN (Lin et al., 2022) further extends the IRM framework by introducing auxiliary information $z_i \in \mathcal{Z}$ to simultaneously learn environment partitioning and invariant feature representation. It assumes that the training environment space $\mathcal{E}_{\text{tr}}$ is the convex hull of E linearly independent fundamental environments. By learning a mapping $\rho : \mathcal{Z} \to \Delta_E$ that provides weights for different environments, ZIN optimizes the following objective:

$$\min_{w,\Phi} \left\{ R(w \circ \Phi, \frac{1}{E} 1_{(E)}) \cdot 1_{(E)} + \lambda \max_{\rho} \|\nabla_w R(w \circ \Phi, \rho)\|_2^2 \right\}. \tag{4}$$

In this way, ZIN can effectively learn invariant features even without explicit environment partition information.

2.2 Total Variation

TV is a mathematical operator widely recognized for its ability to measure the global variability of a function. It has been extensively applied in various fields, such as optimal control, data transmission, and signal processing. The central concept behind TV is to quantify the overall amount of change in a function across its domain. It is able to preserve sharp discontinuities (such as edges in images) while removing noise, making it a valuable tool in image restoration and signal denoising. This property is further elucidated through the coarea formula (Chen et al., 2006). For a function $f \in L^1(\Omega)$, where Ω is an open subset of $\mathbb{R}^d$,

$$\int_{\Omega} |\nabla f| := \int_{-\infty}^{\infty} \int_{f^{-1}(\gamma)} ds \, d\gamma, \tag{5}$$

where $f^{-1}(\gamma)$ represents the level set of f at the value γ. This formulation shows that TV integrates over all the contours of the function, reinforcing its capability in capturing piecewise-constant features. Mumford & Shah (1985) and Rudin et al. (1992) propose the following TV-ℓ_2 and TV-ℓ_1 models, respectively.

$$\inf_{f \in L^2(\Omega)} \left\{ \int_{\Omega} |\nabla f|^2 + \lambda \int_{\Omega} (f - \tilde{f})^2 \, dx \right\}, \quad \inf_{f \in L^2(\Omega)} \left\{ \int_{\Omega} |\nabla f| + \lambda \int_{\Omega} (f - \tilde{f})^2 \, \mathrm{d}x \right\}, \tag{6}$$

where $\tilde{f} \in L^2(\Omega)$ is the ground-truth signal and λ is a hyperparameter that affects accuracy. These models aim to preserve target features in the approximation f and leave noise in the residual $(f - \tilde{f})$. In general, TV-ℓ_1 can better preserve useful features and sharp discontinuities in the approximation f.

2.3 Invariant Risk Minimization based on Total Variation

It can be seen that the penalties of (3) and (4) are similar to a TV-ℓ_2 term, which has been verified by Lai & Wang (2024) under some mild conditions, such as measurability of the involved functions, non-correlation between the feature extractor and the environment variable, representability of w by e, etc. In this sense, (3) and (4) are actually IRM-TV-ℓ_2 and Minimax-TV-ℓ_2 models. Considering the superiority of TV-ℓ_1 over TV-ℓ_2 in sharp feature preservation, Lai & Wang (2024) further propose the following IRM-TV-ℓ_1 (7) and Minimax-TV-ℓ_1 (8) models:

$$\min_{\Phi} \left\{ \mathbb{E}_w[R(w \circ \Phi)] + \lambda (\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2 \right\}, \tag{7}$$

$$\min_{\Phi} \left\{ \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda \max_{\rho} (\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2 \right\}, \tag{8}$$

where $\mathbb{E}_{w \leftarrow \rho}$ denotes the mathematical expectation with respect to (w.r.t.) w whose measure is induced by ρ, and $\frac{1_{(E)}}{E}$ denotes the uniform probability measure for the environment e. The maximization with respect to (w.r.t.) ρ identifies the worst-case inferred environment causing the greatest variation in the risk function. Hence, Minimax-TV-ℓ_1 can be seen as the TV-ℓ_1 version of ZIN. The corresponding coarea formula for IRM-TV-ℓ_1 or Minimax-TV-ℓ_1 is

$$\int_{\Omega} |\nabla_w R(w \circ \Phi)| \, d\nu = \int_{-\infty}^{\infty} \int_{\{w \in \Omega : R(w \circ \Phi) = \gamma\}} ds \, d\gamma, \tag{9}$$

where $\{w \in \Omega : R(w \circ \Phi) = \gamma\}$ denotes the level set, and s is the Hausdorff measure in the corresponding dimensions. This formula ensures that IRM-TV-ℓ_1 or Minimax-TV-ℓ_1 preserves essential structural features while maintaining robustness across different environments. In the rest of this paper, we abbreviate IRM-TV-ℓ_1 as IRM-TV if not specified otherwise.

3 Methodology

3.1 Autonomous Total Variation Penalty and Lagrangian Multiplier

OOD generalization represents the generalization ability of a trained model to an unseen domain. From the perspective of IRM-TV, it can be represented by (Ben-Tal et al., 2009; Arjovsky et al., 2019; Lin et al., 2022; Lai & Wang, 2024)

$$\min_{\Phi} \max_{w} R(w \circ \Phi). \tag{10}$$

In general, IRM-TV cannot achieve OOD generalization without additional conditions, due to insufficient diversity in the training environments and inflexible TV penalty. However, Lai & Wang (2024) indicate that the penalty hyperparameter λ_Φ of IRM-TV should be variable according to the invariant feature extractor Φ:

$$\min_{\Phi} \left\{ \mathbb{E}_w[R(w \circ \Phi)] + \lambda_\Phi (\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2 \right\}, \tag{11}$$

$$\min_{\Phi} \left\{ \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda_\Phi \max_{\rho} (\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2 \right\}. \tag{12}$$

The main reason is that a variable λ_Φ can fill the gap between the objective value of (11) or (12) and the objective value of (10). Lai & Wang (2024) further indicate that this variable λ_Φ exists in general situations. However, they do not specify any tractable and concrete implementation of λ_Φ.

To fill this gap, we find that λ_Φ actually serves as a Lagrangian multiplier in (11) and (12). Hence we use it as an autonomous parameter, which is also taken into the training scheme. Specifically, the parameters for the invariant feature extractor, also denoted by Φ, are directly taken as inputs for λ. Next, λ can be parameterized by another set of parameters Ψ. In this way, λ can be represented as a function of both Ψ and Φ: $\lambda(\Psi, \Phi)$. Then λ not only depends on Φ, but also adjusts its strength through Ψ. Now it can be deployed in a primal-dual optimization model, as illustrated in the next section.

3.2 Primal-dual Optimization and Semi-Nash-equilibrium

Replacing λ_Φ by $\lambda(\Psi, \Phi)$ in (11) and (12), we obtain the corresponding Lagrangian functions:

$$g(\Psi, \Phi) := \mathbb{E}_w[R(w \circ \Phi)] + \lambda(\Psi, \Phi)(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2, \tag{13}$$

$$h(\rho, \Psi, \Phi) := \mathbb{E}_{w \leftarrow \frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \lambda(\Psi, \Phi)(\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2. \tag{14}$$

Applying the OOD generalization criterion (10) in the literature of IRM, we adopt a minimax scheme to optimize Φ and Ψ.

- **Outer Minimization (Primal):** the primal variable Φ is trained to minimize the entire invariant risk, in order to captures invariant features across different environments.
- **Inner Maximization (Dual):** the dual variable Ψ (or (ρ, Ψ)) is trained to maximize the autonomous TV penalty $\lambda(\Psi, \Phi)(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2$ (or $\lambda(\Psi, \Phi)(\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2$), which confronts the most adverse scenarios with spurious features.

They lead to the following proposed OOD-TV-IRM (15) and OOD-TV-Minimax (16) models:

$$\min_{\Phi}\max_{\Psi} g(\Psi,\Phi) := \min_{\Phi}\ \left\{\mathbb{E}_w[R(w\circ\Phi)] + \max_{\Psi}\left[\lambda(\Psi,\Phi)(\mathbb{E}_w[|\nabla_w R(w\circ\Phi)|])^2\right]\right\}, \tag{15}$$

$$\min_{\Phi}\max_{\rho,\Psi} h(\rho,\Psi,\Phi) := \min_{\Phi}\ \left\{\mathbb{E}_{w\leftarrow\frac{1(E)}{E}}[R(w\circ\Phi)] + \max_{\rho,\Psi}\left[\lambda(\Psi,\Phi)(\mathbb{E}_{w\leftarrow\rho}[|\nabla_w R(w\circ\Phi)|])^2\right]\right\}. \tag{16}$$

They are essentially primal-dual optimization models. To further understand their functions, we introduce the concept of semi-Nash-equilibrium in the context of this paper, without loss of generality.

Definition 1. A semi-Nash-equilibrium (Ψ^*,Φ^*) of a Lagrangian function $g(\Psi,\Phi)$ satisfies the following conditions:

1. $g(\Psi^*,\Phi^*) \geqslant g(\Psi,\Phi^*)$ for any Ψ in the parameter space.
2. $g(\Psi^*,\Phi^*) = g(\Psi_{max}(\Phi^*),\Phi^*) \leqslant g(\Psi_{max}(\Phi),\Phi)$ for any Φ in the parameter space, where $\Psi_{max}(\Phi) \in \arg\max_{\Psi}\left[\lambda(\Psi,\Phi)(\mathbb{E}_w[|\nabla_w R(w\circ\Phi)|])^2\right]$. In other words,$\Psi_{max}(\Phi)$ is a solution to the dual optimization with Φ fixed.

A semi-Nash-equilibrium (ρ^*,Ψ^*,Φ^*) of a Lagrangian function $h(\rho,\Psi,\Phi)$ follows the same definition with the unified dual variable (ρ,Ψ).

Remark: in the context of game theory, OOD-TV-IRM (15) (or OOD-TV-Minimax Eq. 16) can be seen as a two-player zero-sum game, where $g(\Psi,\Phi)$ and $-g(\Psi,\Phi)$ are the gains for players Ψ and Φ, respectively. Item 1 in Definition 1 indicates that Ψ^* is the best strategy for dual optimization given Φ^* (which is independent of Ψ). On the other hand, Item 2 indicates that Φ^* is the conditional best strategy for primal optimization only when $\Psi_{max}(\Phi)$ is optimized over each Φ. In this sense, (Ψ^*,Φ^*) is only a semi-Nash-equilibrium instead of a full Nash equilibrium. It accords with our task where Φ is the feature extractor and primary variable that should confront the most adverse scenarios, while Ψ is the TV penalty parameter and secondary variable that provides OOD generalization depending on Φ.

Theorem 2. *Assume that $R(w\circ\Phi)$ is continuous w.r.t. (w,Φ) and differentiable w.r.t. w, $\lambda(\Psi,\Phi)$ is continuous w.r.t. (Ψ,Φ), and the feasible set of parameters (ρ,Ψ,Φ) is bounded and closed. Then any solution to OOD-TV-IRM (15) or OOD-TV-Minimax (16) is a semi-Nash-equilibrium.*

Proof. The proof is given in Appendix A.1. □

Most of the widely-used loss functions and deep neural networks satisfy the conditions of Theorem 2, such as the squared loss, cross-entropy loss, multilayer perceptron, convolutional neural networks, etc. In practical computation, a bounded and closed feasible parameter set is also satisfied due to the maximum float point value. Hence Theorem 2 is applicable to general situations. It indicates that OOD-TV-IRM and OOD-TV-Minimax balance between the training loss and OOD generalization at a semi-Nash-equilibrium.

3.3 PRIMAL-DUAL ALGORITHM AND ADVERSARIAL LEARNING

We develop a primal-dual algorithm to solve OOD-TV-IRM (15) or OOD-TV-Minimax (16). To do this, we further assume that $R(w\circ\Phi)$ and $\lambda(\Psi,\Phi)$ are differentiable w.r.t. their corresponding arguments w, Φ, and Ψ. As indicated in (Lai & Wang, 2024), $|\nabla_w R(w\circ\Phi)|$ in (15) or (16) is non-differentiable w.r.t. Φ. Hence we adopt the subgradient descent method to update these parameters, as illustrated in Appendix A.2. Then the primal and dual updates for (15) or (16) are:

$$\begin{cases}\Phi^{(k+1)} = \Phi^{(k)} - \eta_1^{(k)}\partial_\Phi g(\Psi^{(k)},\Phi^{(k)}),\\ \Psi^{(k+1)} = \Psi^{(k)} + \eta_2^{(k)}\nabla_\Psi g(\Psi^{(k)},\Phi^{(k+1)});\end{cases} \tag{17}$$

$$\begin{cases}\Phi^{(k+1)} = \Phi^{(k)} - \eta_1^{(k)}\partial_\Phi h(\rho^{(k)},\Psi^{(k)},\Phi^{(k)}),\\ (\rho^{(k+1)},\Psi^{(k+1)}) = (\rho^{(k)},\Psi^{(k)}) + \eta_2^{(k)}\nabla_{(\rho,\Psi)} h(\rho^{(k)},\Psi^{(k)},\Phi^{(k+1)}),\end{cases} \tag{18}$$

where $\eta_1^{(k)}$,$\eta_2^{(k)} > 0$ are the learning rates for the primal and dual updates, respectively. The computation of the above (sub)gradients is illustrated in Appendix A.2. In general machine learning scenarios, the smoothness and convexity of the Lagrangian functions g and h are usually unknown, thus $\eta_1^{(k)}$ and $\eta_2^{(k)}$ should be set according to different specific tasks. In addition, we provide a convergent scheme as follows.

Theorem 3. *The same assumptions in Theorem 2 are used. By setting*

$$\begin{aligned}
\eta_1^{(k)} &:= \begin{cases} \frac{1}{k^p \|\partial_\Phi g(\Psi^{(k)}, \Phi^{(k)})\|_2}, & \textit{if} \quad \partial_\Phi g(\Psi^{(k)}, \Phi^{(k)}) \neq 0; \\ 0, & \textit{if} \quad \partial_\Phi g(\Psi^{(k)}, \Phi^{(k)}) = 0. \end{cases} \\
\eta_2^{(k)} &:= \begin{cases} \frac{1}{k^p \|\nabla_\Psi g(\Psi^{(k)}, \Phi^{(k+1)})\|_2}, & \textit{if} \quad \nabla_\Psi g(\Psi^{(k)}, \Phi^{(k+1)}) \neq 0; \\ 0, & \textit{if} \quad \nabla_\Psi g(\Psi^{(k)}, \Phi^{(k+1)}) = 0. \end{cases}
\end{aligned} \tag{19}$$

$$\begin{aligned}
\textit{or} \quad \eta_1^{(k)} &:= \begin{cases} \frac{1}{k^p \|\partial_\Phi h(\rho^{(k)}, \Psi^{(k)}, \Phi^{(k)})\|_2}, & \textit{if} \quad \partial_\Phi h(\rho^{(k)}, \Psi^{(k)}, \Phi^{(k)}) \neq 0; \\ 0, & \textit{if} \quad \partial_\Phi h(\rho^{(k)}, \Psi^{(k)}, \Phi^{(k)}) = 0. \end{cases} \\
\eta_2^{(k)} &:= \begin{cases} \frac{1}{k^p \|\nabla_{(\rho,\Psi)} h(\rho^{(k)}, \Psi^{(k)}, \Phi^{(k+1)})\|_2}, & \textit{if} \quad \nabla_{(\rho,\Psi)} h(\rho^{(k)}, \Psi^{(k)}, \Phi^{(k+1)}) \neq 0; \\ 0, & \textit{if} \quad \nabla_{(\rho,\Psi)} h(\rho^{(k)}, \Psi^{(k)}, \Phi^{(k+1)}) = 0. \end{cases}
\end{aligned} \tag{20}$$

with an arbitrary $p > 1$, the primal-dual algorithm (17) or (18) achieves convergent sequences $\{\Phi^{(k)}\}_{k=1}^{\infty}$, $\{\Psi^{(k)}\}_{k=1}^{\infty}$, or $\{(\rho^{(k)}, \Psi^{(k)})\}_{k=1}^{\infty}$, respectively. Moreover, its computational complexity is $\mathcal{O}([(p-1)\epsilon]^{-\frac{1}{p-1}})$ to achieve a convergence tolerance of $\epsilon > 0$.

Proof. The proof is given in Appendix A.2. □

In fact, (17) or (18) facilitates an adversarial learning to dynamically adjust model parameters (Ψ, Φ) or (ρ, Ψ, Φ) to balancing training loss and OOD generalization. This approach extends the theoretical work of IRM-TV by providing a practical, dynamic implementation of $\lambda(\Psi, \Phi)$ that adapts to the most adverse environments.

Remark: the above OOD-TV-IRM and OOD-TV-Minimax models as well as the primal-dual algorithm can also be applied to the TV-ℓ_2 versions, which are omitted here since they are simpler differentiable cases.

3.4 Implementation with Neural Networks

First, we remind that all the theoretical results in this section hold with any functions satisfying the corresponding conditions (e.g., continuity, differentiability), not limited to neural networks. To provide a specific instance, we implement $\lambda(\Psi, \Phi)$ with neural networks that take Ψ and Φ as network trainable parameters and network inputs, respectively. Besides $\lambda(\Psi, \Phi)$, Φ and ρ can also be developed by neural networks. Specifically, $\Phi : \mathcal{X} \to \mathcal{H}$ is the feature extractor that maps a sample x in the sample space $\mathcal{X}$ to the feature space $\mathcal{H}$. It can be implemented by different kinds of neural networks according to the practical situation, as shown in Table 1. The data sets correspond to different synthetic or real-world tasks, which will be illustrated in Section 4. For example, Φ can be a multilayer perceptron in Simulation, or a convolutional neural network in Landcover. Similarly, $\rho : \mathcal{Z} \to \Delta^E$ is the environment inference operator that maps the auxiliary variables $z \in \mathcal{Z}$ to the probability space $\Delta^E := \{v \in \mathbb{R}^E : v \geqslant 0_{(E)} \text{ and } v \cdot 1_{(E)} = 1\}$, where $\mathbb{R}^E$ denotes the E-dimensional real space (i.e., E environments). Moreover, the architectures of ρ are developed by multilayer perceptrons, which are compatible with the corresponding Φ, as shown in Table 1.

Without ambiguity, we can also use Φ to denote the parameters of the feature extractor. Then $\lambda(\Psi, \Phi)$ can be implemented as a neural network that takes Φ as input and Ψ as network parameters. We provide tractable architectures of λ in Table 1, which should also be compatible with the corresponding Φ. With the primal-dual algorithm in Section 3.3, Φ and Ψ can be learned in an adversarial way to improve OOD generalization, which puts Theorem 2 into practice.

4 Experimental Results

We conduct experiments for the proposed OOD-TV-IRM and OOD-TV-Minimax on 7 simulation and real-world data sets to evaluate their effectiveness in different OOD generalization tasks. These tasks largely follow the settings in the literature of IRM (Lin et al., 2022; Lai & Wang, 2024). The implementation details are provided as codes in the Supplementary Material. The original IRM (Arjovsky et al., 2019), ZIN (Lin et al., 2022), IRM-TV-ℓ_1 and Minimax-TV-ℓ_1 (Lai & Wang, 2024) frameworks are also included in comparisons to verify the effectiveness of the proposed primal-dual optimization and adversarial learning scheme. IRM and ZIN are equivalent to IRM-TV-ℓ_2 and Minimax-TV-ℓ_2, respectively.

Table 1: Network architectures of the invariant feature extractor Φ, the environment inference operator ρ, and the TV penalty strength λ, w.r.t. their corresponding inputs.

Data Set	$x \to \Phi$	$z \to \rho$
Simulation	Linear(15, 1)	Linear(1, 16)→ReLu()→Linear(16, 1) →Sigmoid()
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()
House Price	Linear(15,32) →ReLu() → Linear(32,1)	Linear(1,64)→ ReLu() → Linear(64,4) → Softmax()
Colored MNIST	Conv2d(3,32,3) → MaxPool2d(2,2) →Conv2d(32,64,3) → MaxPool2d(2,2) → Conv2d(64,128,3) → MaxPool2d(2,2) → Linear(128*3*3,128)→ Linear(128,10)	Linear(3,16) → ReLU() → Linear(16,1) → Softmax()
NICO	ResNet-18()	Linear(3,16) → ReLU() → Linear(16,1)→ Softmax()

Data Set	$\Phi \to \lambda$
Simulation	Linear(11, 1) → ReLU() → Linear(1, 1) → Softplus()
CelebA	Linear(8225, 16) → ReLU() → Linear(16, 1) → Softplus()
Landcover	Linear(173574, 32) → ReLU() → Linear(32, 1) → Softplus()
Adult	Linear(977, 16) → ReLU() → Linear(16, 1) → Softplus()
House Price	Linear(545, 32) → ReLU() → Linear(32, 16) → Softplus() → Linear(16,1)
Colored MNIST	Linear(242122,32) → ReLU() → Linear(32,1) → Softplus()
NICO	Linear(11180113,32) → ReLU() → Linear(32,1) → Softplus()

4.1 Simulation Data

The simulation data consists of temporal heterogeneity observations with distributional shift w.r.t time, which is used in (Lin et al., 2022; Tan et al., 2023) to evaluate OOD generalization. Details of generating this data set is provided in Appendix A.3. Among all the parameter settings, the most challenging one is $(p_s^-, p_s^+, p_v) = (0.999, 0.9, 0.8)$, thus we use this setting in the evaluation.

Table 2 (left) shows the mean and worst accuracies of different methods over the four test environments on simulation data. The proposed OOD-TV-based methods outperform their counterparts in all the mean and worst accuracy cases. Moreover, OOD-TV-IRM-ℓ_1 achieves the highest accuracies among all the competitors in all the cases, while OOD-TV-Minimax-ℓ_2 achieves the highest accuracies among all the competitors in all the cases without environment partition. Hence the OOD-TV-based methods achieve the best performance regardless of whether the environment partition is available or not.

Table 2: Accuracies of different methods on simulation data (left) and CelebA (right).

Method	**Mean**	**Worst**
ZIN	0.8356	0.7648
OOD-TV-Minimax-ℓ_2	**0.8621**	**0.8127**
Minimax-TV-ℓ_1	0.7831	0.7615
OOD-TV-Minimax-ℓ_1	**0.8054**	**0.7632**
IRM	0.8336	0.7615
OOD-TV-IRM-ℓ_2	**0.8673**	**0.8191**
IRM-TV-ℓ_1	0.8592	0.8215
OOD-TV-IRM-ℓ_1	**0.8817**	**0.8519**

Method	**Mean**	**Worst**
ZIN	0.7329	0.5840
OOD-TV-Minimax-ℓ_2	**0.7783**	**0.6840**
Minimax-TV-ℓ_1	0.7712	0.6990
OOD-TV-Minimax-ℓ_1	0.7551	0.6934
IRM	0.7815	0.7308
OOD-TV-IRM-ℓ_2	**0.7841**	**0.7519**
IRM-TV-ℓ_1	0.7848	0.7392
OOD-TV-IRM-ℓ_1	**0.7889**	**0.7525**

4.2 CelebA

The CelebA data set (Liu et al., 2015) contains face images of celebrities. The task is to identify smiling faces, which are deliberately correlated with the gender variable. The 512-dimensional deep features of the face images are extracted using a pre-trained ResNet18 model (He et al., 2016), while the invariant features are learned using subsequent multilayer perceptrons (MLPs). ZIN, OOD-TV-

Minimax-ℓ_2, Minimax-TV-ℓ_1 and OOD-TV-Minimax-ℓ_1 take seven additional descriptive variables for environment inference, including *Young*, *Blond Hair*, *Eyeglasses*, *High Cheekbones*, *Big Nose*, *Bags Under Eyes*, and *Chubby*. As for IRM, OOD-TV-IRM-ℓ_2, IRM-TV-ℓ_1, and OOD-TV-IRM-ℓ_1, they use the gender variable as the environment indicator.

Table 2 (right) shows the accuracies of different methods on CelebA. The proposed OOD-TV-based methods outperform their counterparts in 6 out of 8 situations with both mean and worst accuracies. Moreover, OOD-TV-IRM-ℓ_1 achieves the highest accuracies among all the competitors. Hence OOD-TV-IRM improves OOD generalization on IRM-TV.

4.3 Landcover

The Landcover data set consists of time series data and the corresponding land cover types derived from satellite images (Gislason et al., 2006; Russwurm et al., 2020; Xie et al., 2021). The input data have dimensions of 46×8 and is used to identify one of six land cover types. The invariant feature extractor Φ is implemented as a 1D-CNN to process the time series input, following the approaches of Xie et al. (2021) and Lin et al. (2022). For the scenarios where ground-truth environment partitions are unavailable, latitude and longitude are used as auxiliary information for environment inference. All methods are trained on non-African data, and then tested on both non-African (from regions not overlapping with the training data) and African regions. This is a complex and challenging experiment among the 6 experiments of this paper.

Table 3 (left) shows the accuracies of different methods on Landcover. The proposed OOD-TV-based methods outperform their counterparts in all the worst accuracy cases and in 3 out of 4 mean accuracy cases, respectively. Hence the OOD-TV-based methods show competitive performance in this challenging task.

Table 3: Accuracies of different methods on Landcover (left) and adult income prediction (right).

Method	Mean	Worst
ZIN	0.7423	0.6235
OOD-TV-Minimax-ℓ_2	0.6418	**0.6386**
Minimax-TV-ℓ_1	0.6557	0.6547
OOD-TV-Minimax-ℓ_1	**0.6677**	**0.6625**
IRM	0.6447	0.6425
OOD-TV-IRM-ℓ_2	**0.6789**	**0.6787**
IRM-TV-ℓ_1	0.6459	0.6430
OOD-TV-IRM-ℓ_1	**0.6647**	**0.6616**

Method	Mean	Worst
ZIN	0.8261	0.8013
OOD-TV-Minimax-ℓ_2	**0.8354**	**0.8116**
Minimax-TV-ℓ_1	0.8298	0.8062
OOD-TV-Minimax-ℓ_1	**0.8348**	**0.8100**
IRM	0.8264	0.8012
OOD-TV-IRM-ℓ_2	**0.8349**	**0.8121**
IRM-TV-ℓ_1	0.8264	0.8012
OOD-TV-IRM-ℓ_1	**0.8339**	**0.8100**

4.4 Adult Income Prediction

This task uses the Adult data set[1] to predict whether an individual's income exceeds $50K per year based on census data. The data set is split into four subgroups representing different environments based on *race* $\in$ {Black, Non-Black} and *sex* $\in$ {Male, Female}. Two-thirds of the data from the Black Male and Non-Black Female subgroups are randomly selected for training, and the compared methods are verified across all four subgroups using the remaining data. Six integer variables — *Age*, *FNLWGT*, *Education-Number*, *Capital-Gain*, *Capital-Loss*, and *Hours-Per-Week* — are fed into ZIN, OOD-TV-Minimax-ℓ_2, Minimax-TV-ℓ_1 and OOD-TV-Minimax-ℓ_1 for environment inference. Ground-truth environment indicators are provided for IRM, OOD-TV-IRM-ℓ_2, IRM-TV-ℓ_1, and OOD-TV-IRM-ℓ_1. Categorical variables, excluding race and sex, are encoded using one-hot encoding, followed by principal component analysis (PCA), retaining over 99% of the cumulative explained variance. The transformed features are then combined with the six integral variables, resulting in 59-dimensional representations, which are normalized to have zero mean and unit variance for invariant feature learning.

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

Table 3 (right) shows the accuracies of different methods on this income prediction task. The proposed OOD-TV-based methods outperform their counterparts in all the situations with both mean and worst accuracies. Hence they are more effective and robust than the original IRM-TV methods in this prediction task regardless of whether the environment partition is available or not.

4.5 House Price Prediction

The above four experiments are all classification tasks. We also perform a regression task with the House Prices data set[2]. This task uses 15 variables, such as the number of bathrooms, locations, etc., to predict the house price. The training and test sets consist of samples with built year in periods $[1900, 1950]$ and $(1950, 2000]$, respectively. The house price within the same built year is normalized. The built year variable is fed into ZIN, OOD-TV-Minimax-ℓ_2, Minimax-TV-ℓ_1 and OOD-TV-Minimax-ℓ_1 for environment inference. The training set are partitioned into 5 subsets with 10-year range in each subset. Each subset can be seen as having the same environment.

Table 4 (left) provides the mean squared errors (MSE) of different methods in this regression task. The OOD-TV-based methods achieves lower MSEs than their counterparts in the mean result of all the environments and the worst environment. Hence they are more effective and robust than the original IRM-TV methods in the regression task.

Table 4: Left: mean squared errors of different methods in house price prediction. Right: accuracies of different methods on Colored MNIST.

Method	Mean	Worst
ZIN	0.2862	0.4115
OOD-TV-Minimax-ℓ_2	**0.2552**	**0.3649**
Minimax-TV-ℓ_1	0.2961	0.4548
OOD-TV-Minimax-ℓ_1	**0.2476**	**0.3641**
IRM	0.9583	1.3858
OOD-TV-IRM-ℓ_2	**0.5319**	**0.7616**
IRM-TV-ℓ_1	0.4263	0.6281
OOD-TV-IRM-ℓ_1	**0.4021**	**0.5595**

Method	Mean
ZIN	0.9208
OOD-TV-Minimax-ℓ_2	**0.9705**
Minimax-TV-ℓ_1	0.9199
OOD-TV-Minimax-ℓ_1	**0.9699**

4.6 Colored MNIST

We use the Colored MNIST data set[3] to evaluate the performance of our approach in multi-group classification of a more general scenario. It consists of hand-written digits of $0 \sim 9$ with different background colors of red, green or blue. To make it more challenging, we hide the background color information of samples, so that only the mean values of the red, green, and blue channels of a sample can be used as three auxiliary variables for environment inference. Therefore, OOD-TV-Minimax and Minimax-TV should be used in this setting. Their classification accuracies are shown in Table 4 (right). OOD-TV-Minimax achieves a high accuracy in this multi-group classification task, which is significantly higher than Minimax-TV. Thus OOD-TV-Minimax is effective in improving OOD generalization for this complex scenario.

4.7 NICO

The NICO data set (He et al., 2021) is a widely-used benchmark in Non-Independent-and-Identically-Distributed (Non-I.I.D.) image classification with contexts. It is a challenging data set including both correlation shift and diversity shift. We use the *Vehicle* superclass to conduct experiments, which consists of 7 classes of vehicles. Each class has several contexts, while different classes may have different context sets. We randomly select 80% and 20% samples from each context as training and test samples, respectively. Then we hide the context information of samples and unite all the contexts to form a whole class. Therefore, only the mean values of the red, green, and blue channels of a sample can be used as three auxiliary variables for environment inference, and

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

[3] https://www.kaggle.com/datasets/youssifhisham/colored-mnist-dataset

only OOD-TV-Minimax and Minimax-TV can be used in this setting. The classification accuracies are shown in Table 5, which indicates that OOD-TV-Minimax significantly outperforms Minimax-TV. Thus OOD-TV-Minimax is effective in improving OOD generalization with both correlation shift and diversity shift.

Table 5: Accuracies of different methods on NICO Vehicle.

Method	**Mean**
ZIN	0.6357
OOD-TV-Minimax-ℓ_2	**0.7548**
Minimax-TV-ℓ_1	0.6547
OOD-TV-Minimax-ℓ_1	**0.7143**

4.8 Adversarial Learning Process

To visualize the adversarial learning process, we plot the objective values and parameter changes for OOD-TV-IRM and OOD-TV-Minimax on simulation data in Figure A1. Specifically, the overall objectives are the $g(\Psi, \Phi)$ and $h(\rho, \Psi, \Phi)$ defined in (13) and (14), respectively. The penalty terms correspond to the second terms in (13) and (14), respectively. The parameter changes correspond to $\|\Phi^{(k+1)} - \Phi^{(k)}\|_2$ and $\|\Psi^{(k+1)} - \Psi^{(k)}\|_2$, respectively. We follow the annealing strategy of (Lin et al., 2022) in the early epochs, thus the adversarial learning starts from the 2001st epoch. The Adam scheme (Kingma & Ba, 2015) is adopted as the optimizer. Figure A1 indicates that the adversarial learning process becomes stable with 400 epochs. Both of the overall objective and the penalty term converge to form a stable gap, which corresponds to the fidelity term to be optimized. Besides, both $\|\Phi^{(k+1)} - \Phi^{(k)}\|_2$ and $\|\Psi^{(k+1)} - \Psi^{(k)}\|_2$ converge to small values, which indicates that the parameters $\Phi^{(k)}$ and $\Psi^{(k)}$ are rarely updated.

We also demonstrate the adversarial learning process for OOD-TV-Minimax (OOD-TV-IRM is not applicable) on Colored MNIST in Figure A2. The Adam optimizer is used for the primal parameter Φ, while the dual update (18)(20) is used for Ψ to ensure convergence. Results show that the adversarial learning process successfully converges with 600 epochs for OOD-TV-Minimax-ℓ_1 or with 300 epochs for OOD-TV-Minimax-ℓ_2. All these results indicate that the adversarial learning process is feasible.

5 Conclusion and Discussion

We extend the invariant risk minimization based on total variation model (IRM-TV) to a Lagrangian multiplier model OOD-TV-IRM, in order to improve out-of-distribution (OOD) generalization. It is essentially a primal-dual optimization model. The primal optimization reduces the entire invariant risk, in order to extract invariant features. The dual optimization strengthens the autonomous TV penalty, in order to provide an adversarial interference. The whole OOD-TV-IRM objective is to find a semi-Nash-equilibrium to keep a balance between the training loss and OOD generalization. We further develop a convergent primal-dual algorithm to facilitate an adversarial learning for the invariant features and the adverse environments. Experimental results show that the proposed OOD-TV-IRM framework improves effectiveness and robustness in most experimental tasks with both mean and worst accuracies or mean squared errors across different test environments.

One possible future work may be making the most of OOD-TV-IRM by improving the diversity and representation of the training environments. Another approach may be improving effectiveness of the primal-dual optimization scheme. A third approach may be developing new penalties with different variations other than the widely-used TV penalty.

References

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 145–155. PMLR, 13–18 Jul 2020.

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

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Terrence Chen, Wotao Yin, Xiang Zhou, and Dorin Comaniciu. Total variation models for variable lighting face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28: 1519–1524, 10 2006.

Nicolas Dey, Laure Blanc-Feraud, Christophe Zimmer, Pascal Roux, Zvi Kam, Jean-Christophe Olivo-Marin, and Josiane Zerubia. Richardson–lucy algorithm with total variation regularization for 3d confocal microscope deconvolution. *Microscopy research and technique*, 69(4):260–266, 2006.

Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R. Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

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

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 25913–25935. PMLR, 21–27 Jul 2024.

Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35:24529–24542, 2022.

Yong Lin, Lu Tan, Yifan Hao, Honam Wong, Hanze Dong, Weizhong Zhang, Yujiu Yang, and Tong Zhang. Spurious feature diversification improves out-of-distribution generalization, 2023.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6804–6814, 18–24 Jul 2021.

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

David Mumford and Jayant Shah. Boundary detection by minimizing functionals. In *IEEE Conference on computer vision and pattern recognition*, volume 17, pp. 137–154. San Francisco, 1985.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Leonid Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 11 1992.

Marc Russwurm, Sherrie Wang, Marco Korner, and David Lobell. 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.

Xiaoyu Tan, Lin Yong, Shengyu Zhu, Chao Qu, Xihe Qiu, Xu Yinghui, Peng Cui, and Yuan Qi. Provably invariant learning without domain information. In *International Conference on Machine Learning*, pp. 33563–33580. PMLR, 2023.

Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations*, 2021.

Shiji Xin, Yifei Wang, Jingtong Su, and Yisen Wang. On the connection between invariant learning and adversarial training for out-of-distribution generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10519–10527, Jun. 2023.

Mengyue Yang, Zhen Fang, Yonggang Zhang, Yali Du, Furui Liu, Jean-Francois Ton, Jianhong Wang, and Jun Wang. Invariant learning via probability of sufficient and necessary causes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79832–79857, 2023.

Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pp. 27222–27244. PMLR, 2022.

A Appendix

A.1 Proof of Theorem 2

Proof. **Part (a).**

First, we verify that there exists at least one solution to (15). Since $\lambda(\Psi, \Phi)$ is continuous w.r.t. (Ψ, Φ) and (Ψ, Φ) belongs to a bounded and closed set, a solution to $\max_\Psi \lambda(\Psi, \Phi)$ can be obtained for any given Φ by the Weierstrass extreme value theorem, denoted by $\Psi_{max}(\Phi)$. Let $\Lambda(\Phi) := \lambda(\Psi_{max}(\Phi), \Phi)$. We then turn to prove that $\Lambda(\Phi)$ is a continuous function w.r.t. Φ. For $\Phi_1 \neq \Phi_2$, we have

$$\lambda(\Psi_{max}(\Phi_2), \Phi_1) \leqslant \lambda(\Psi_{max}(\Phi_1), \Phi_1). \tag{21}$$

From the continuity of λ, for any given $\epsilon > 0$, there exists some $\delta_1 > 0$, such that for any $\Phi \in \mathrm{B}(\Phi_1, \delta_1)$, we have

$$\lambda(\Psi_{max}(\Phi_2), \Phi) - \epsilon < \lambda(\Psi_{max}(\Phi_2), \Phi_1), \tag{22}$$

where $\mathrm{B}(\Phi_1, \delta_1)$ denotes an open ball centered at Φ_1 with radius δ_1. Letting $\Phi_2 \in \mathrm{B}(\Phi_1, \delta_1)$ in (22) and combining it with (21), we have

$$\lambda(\Psi_{max}(\Phi_2), \Phi_2) - \epsilon < \lambda(\Psi_{max}(\Phi_1), \Phi_1). \tag{23}$$

Following a similar deduction, for the above $\epsilon > 0$, there exists some $\delta_2 > 0$, such that for $\Phi_1 \in \mathrm{B}(\Phi_2, \delta_2)$, we have

$$\lambda(\Psi_{max}(\Phi_1), \Phi_1) - \epsilon < \lambda(\Psi_{max}(\Phi_2), \Phi_2). \tag{24}$$

Combining (23) and (24), for any given $\epsilon > 0$, as long as $\|\Phi_1 - \Phi_2\|_2 < \min\{\delta_1, \delta_2\}$, we have

$$|\lambda(\Psi_{max}(\Phi_1), \Phi_1) - \lambda(\Psi_{max}(\Phi_2), \Phi_2)| < \epsilon. \tag{25}$$

This proves that $\Lambda(\Phi)$ is continuous w.r.t. Φ.

As for $(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2$, the TV operator and the expectation are taken on the variable w, which does not affect the continuity w.r.t. Φ. Define

$$G(\Phi) := g(\Psi_{max}(\Phi), \Phi) = \mathbb{E}_w[R(w \circ \Phi)] + \Lambda(\Phi)(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2. \tag{26}$$

It can be seen that $G(\Phi)$ is the sum, multiplication, and composition of continuous functions w.r.t. Φ, thus it is also continuous w.r.t. Φ. Using the Weierstrass extreme value theorem again, there exists a solution to $\min_\Phi G(\Phi)$, denoted by Φ^*. Let $\Psi^* := \Psi_{max}(\Phi^*)$, then (Ψ^*, Φ^*) is a solution to (15).

Part (b).

Next, we turn to investigate $h(\rho, \Psi, \Phi)$ in (16). From Appendix A.3 of IRM-TV (Lai & Wang, 2024),

$$\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|] = \sum_{i=1}^{E} |\nabla_w R(w \circ \Phi, e_i)| \rho_i, \tag{27}$$

where e_i denotes the i-th training environment in the literature of IRM, and ρ_i denotes the environment inference operator for the i-th training environment. (27) indicates that $\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|]$ is a linear function w.r.t. $\rho \in \Delta^E$, where Δ^E is the E-dimensional simplex defined in Section 3.4. Thus $\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|]$ as well as $(\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2$ is continuous w.r.t. ρ. On the other hand, $\lambda(\Psi, \Phi)$ is continuous w.r.t. Ψ. Define

$$\mathfrak{E}(\rho, \Phi) := (\mathbb{E}_{w \leftarrow \rho}[|\nabla_w R(w \circ \Phi)|])^2. \tag{28}$$

To examine the continuity of $\lambda(\Psi, \Phi)\mathfrak{E}(\rho, \Phi)$ w.r.t. the entire dual variable (ρ, Ψ), we check its values at two different points (ρ_1, Ψ_1) and (ρ_2, Ψ_2):

$$\begin{aligned} &|\lambda(\Psi_1, \Phi)\mathfrak{E}(\rho_1, \Phi) - \lambda(\Psi_2, \Phi)\mathfrak{E}(\rho_2, \Phi)| \\ <&|\lambda(\Psi_1, \Phi)\mathfrak{E}(\rho_1, \Phi) - \lambda(\Psi_1, \Phi)\mathfrak{E}(\rho_2, \Phi)| + |\lambda(\Psi_1, \Phi)\mathfrak{E}(\rho_2, \Phi) - \lambda(\Psi_2, \Phi)\mathfrak{E}(\rho_2, \Phi)| \\ =&|\lambda(\Psi_1, \Phi)| \cdot |\mathfrak{E}(\rho_1, \Phi) - \mathfrak{E}(\rho_2, \Phi)| + |\mathfrak{E}(\rho_2, \Phi)| \cdot |\lambda(\Psi_1, \Phi) - \lambda(\Psi_2, \Phi)|. \end{aligned} \tag{29}$$

When $(\rho_2, \Psi_2) \to (\rho_1, \Psi_1)$, the right hand side of (29) tends to zero, thus the left hand side of (29) also tends to zero. This verifies that $\lambda(\Psi, \Phi)\mathfrak{E}(\rho, \Phi)$ is continuous w.r.t. (ρ, Ψ) for any given Φ.

Using the Weierstrass extreme value theorem, there exists at least one solution to $\max_{\rho,\Psi}[\lambda(\Psi, \Phi)\mathfrak{E}(\rho, \Phi)]$ for any given Φ, denoted by $(\rho, \Psi)_{max}(\Phi)$. Define

$$H(\Phi) := h((\rho, \Psi)_{max}(\Phi), \Phi) = \mathbb{E}_{w\leftarrow\frac{1_{(E)}}{E}}[R(w \circ \Phi)] + \max_{\rho,\Psi}[\lambda(\Psi, \Phi)\mathfrak{E}(\rho, \Phi)], \tag{30}$$

which is continuous w.r.t. Φ. Following similar deductions to Part (a), there exists a solution to $\min_\Phi H(\Phi)$, denoted by Φ^*. Let $(\rho^*, \Psi^*) := (\rho, \Psi)_{max}(\Phi^*)$, then (ρ^*, Ψ^*, Φ^*) is a solution to (16).

Part (c).

Last, we turn to verify that any solutions (Ψ^*, Φ^*) to (15) and (ρ^*, Ψ^*, Φ^*) to (16) are semi-Nash-equilibria. We first fix Φ^*, since $\Psi^* = \Psi_{max}(\Phi^*)$ and $(\rho^*, \Psi^*) = (\rho, \Psi)_{max}(\Phi^*)$, we have $g(\Psi^*, \Phi^*) \geqslant g(\Psi, \Phi^*)$ or $h(\rho^*, \Psi^*, \Phi^*) \geqslant h(\rho, \Psi, \Phi^*)$ for any Ψ or (ρ, Ψ) in the parameter space. Hence Item 1 in Definition 1 is satisfied.

On the other hand, from (26) and (30), we have $G(\Phi^*) \leqslant G(\Phi)$ and $H(\Phi^*) \leqslant H(\Phi)$ in the parameter space. Hence Item 2 in Definition 1 is also satisfied. In summary, Theorem 2 has been proved.

□

A.2 Proof of Theorem 3

Before starting the proof, we compute some gradients or subgradients in place of the conventional gradients at non-differentiable points for g and h. This allows us to maintain the optimization flow with the autograd module of mainstream learning architectures (like Pytorch[4]). First, the subgradient of $|\nabla_w R(w \circ \Phi)|$ w.r.t. Φ is:

$$\partial_\Phi |\nabla_w R(w \circ \Phi)| = \begin{cases} \frac{J_\Phi^\top[\nabla_w R(w\circ\Phi)] * \nabla_w R(w\circ\Phi)}{|\nabla_w R(w\circ\Phi)|} & \text{if} \quad \nabla_w R(w \circ \Phi) \neq 0, \\ 0 & \text{if} \quad \nabla_w R(w \circ \Phi) = 0, \end{cases} \tag{31}$$

where $J_\Phi[\cdot]$ is the Jacobian matrix w.r.t. Φ, and $*$ is the matrix multiplication.

By the chain rule of derivative, the (sub)gradients of g w.r.t. Φ and Ψ are:

$$\begin{aligned} \partial_\Phi g(\Psi, \Phi) =& \mathbb{E}_w[\nabla_\Phi R(w \circ \Phi)] + 2\lambda(\Psi, \Phi)\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|]\mathbb{E}_w[\partial_\Phi |\nabla_w R(w \circ \Phi)|] \\ &+ \nabla_\Phi \lambda(\Psi, \Phi)(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2, \\ \nabla_\Psi g(\Psi, \Phi) =& \nabla_\Psi \lambda(\Psi, \Phi)(\mathbb{E}_w[|\nabla_w R(w \circ \Phi)|])^2. \end{aligned} \tag{32}$$

Similarly, the (sub)gradients of h w.r.t. ρ, Ψ, and Φ are:

$$\begin{aligned} \nabla_\rho h(\rho, \Psi, \Phi) =& 2\lambda(\Psi, \Phi)\mathbb{E}_{w\leftarrow\rho}[|\nabla_w R(w\circ\Phi)|]\nabla_\rho(\mathbb{E}_{w\leftarrow\rho}[|\nabla_w R(w\circ\Phi)|]), \\ \nabla_\Psi h(\rho, \Psi, \Phi) =& \nabla_\Psi \lambda(\Psi, \Phi)(\mathbb{E}_{w\leftarrow\rho}[|\nabla_w R(w \circ \Phi)|])^2, \\ \partial_\Phi h(\rho, \Psi, \Phi) =& \mathbb{E}_{w\leftarrow\frac{1_{(E)}}{E}}[\nabla_\Phi R(w\circ\Phi)] + 2\lambda(\Psi, \Phi)\mathbb{E}_{w\leftarrow\rho}[|\nabla_w R(w \circ \Phi)|]\mathbb{E}_{w\leftarrow\rho}[\partial_\Phi |\nabla_w R(w \circ \Phi)|] \\ &+ \nabla_\Phi \lambda(\Psi, \Phi)(\mathbb{E}_{w\leftarrow\rho}[|\nabla_w R(w \circ \Phi)|])^2. \end{aligned} \tag{33}$$

Proof of Theorem 3. We first investigate the sequence $\{\Phi^{(k)}\}$. Combining (17) and (19), we have

$$\|\Phi^{(k+1)} - \Phi^{(k)}\|_2 \leqslant \frac{1}{k^p}. \tag{34}$$

Summing up both sides of (35) from $k = 2$ to ∞ yields

$$\sum_{k=2}^{\infty} \|\Phi^{(k+1)} - \Phi^{(k)}\|_2 \leqslant \sum_{k=2}^{\infty} \frac{1}{k^p} < \int_1^\infty \frac{1}{\zeta^p}\,\mathrm{d}\zeta = \frac{1}{1-p} \cdot \frac{1}{\zeta^{p-1}}\Big|_1^\infty = \frac{1}{p-1} < \infty. \tag{35}$$

[4] https://pytorch.org/

It means that the sum of infinite series $\sum_{k=2}^{\infty}\|\Phi^{(k+1)}-\Phi^{(k)}\|_2$ is convergent. Then for any given $\epsilon>0$, there exists some N such that $\sum_{k=N}^{\infty}\|\Phi^{(k+1)}-\Phi^{(k)}\|_2<\epsilon$. Hence for any $k>N$ and any $m\in\mathbb{N}_+$,

$$\|\Phi^{(k+m)}-\Phi^{(k)}\|_2 \leqslant \sum_{j=k}^{k+m-1}\|\Phi^{(j+1)}-\Phi^{(j)}\|_2 \leqslant \sum_{j=N}^{\infty}\|\Phi^{(j+1)}-\Phi^{(j)}\|_2 < \epsilon. \tag{36}$$

It indicates that $\{\Phi^{(k)}\}$ is a Cauchy sequence. From the completeness of the parameter space for Φ, $\Phi^{(k)}$ converges to some point $\Phi^{\bullet}$. Following similar deductions, the sequences $\{\Psi^{(k)}\}$ for g, $\{\Phi^{(k)}\}$ and $\{(\rho^{(k)},\Psi^{(k)})\}$ for h also converge.

Next, we turn to analyze the computational complexity of this primal-dual algorithm. Without loss of generality, we consider one iteration as having a complexity of $\mathcal{O}(1)$, since the complexity within one iteration depends on specific forms of g or h and is relatively constant across different (sub)gradient-type algorithms. Suppose the algorithm needs N iterations in order to achieve a convergence tolerance of ϵ. From the above deduction, we have

$$\|\Phi^{(N)}-\Phi^{\bullet}\|_2 \leqslant \sum_{k=N}^{\infty}\|\Phi^{(k+1)}-\Phi^{(k)}\|_2 < \int_{N-1}^{\infty}\frac{1}{\zeta^p}\,\mathrm{d}\zeta = \frac{1}{p-1}\cdot\frac{1}{(N-1)^{p-1}} < \epsilon. \tag{37}$$

The last inequality in (37) indicates that the convergence tolerance ϵ can be achieved as long as

$$N > [(p-1)\epsilon]^{-\frac{1}{p-1}}+1. \tag{38}$$

By taking the smallest integer for N in (38), we obtain the computational complexity $\mathcal{O}([(p-1)\epsilon]^{-\frac{1}{p-1}})$.

□

A.3 Simulation Data Generation

The binary outcome of interest $Y(t)$ with time $t\in[0,1]$ is caused and influenced by the invariant features $X_v(t)\in\mathbb{R}$ and the spurious features $X_s(t)\in\mathbb{R}$:

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

$$Y(t)\sim\begin{cases}\mathrm{sign}(X_v(t)), & \text{w.p. } p_v;\\ -\mathrm{sign}(X_v(t)), & \text{w.p. } 1-p_v.\end{cases}$$

$X_v(t)$ and $X_s(t)$ are further extended to 5 and 10 dimensional sequences by adding standard Gaussian noise, respectively. We control the correlation between $X_v(t)$ and $Y(t)$ by a parameter p_v, keeping it constant, but change the correlation between $X_s(t)$ and $Y(t)$ over t by a parameter $p_s(t)$. Hence, the training data is generated with parameter setting (p_s^-,p_s^+,p_v), where p_s^- and p_s^+ denote the $p_s(t)$ setting for $t\in[0,0.5)$ and $t\in[0.5,1]$, respectively. As for the test data, we set $p_s\in\{0.999,0.8,0.2,0.001\}$ and keep the same p_v. Time t can be exploited as the auxiliary variable in ZIN and Minimax-TV-ℓ_1 for environment inference.

A.4 More Results of Adversarial Learning Process

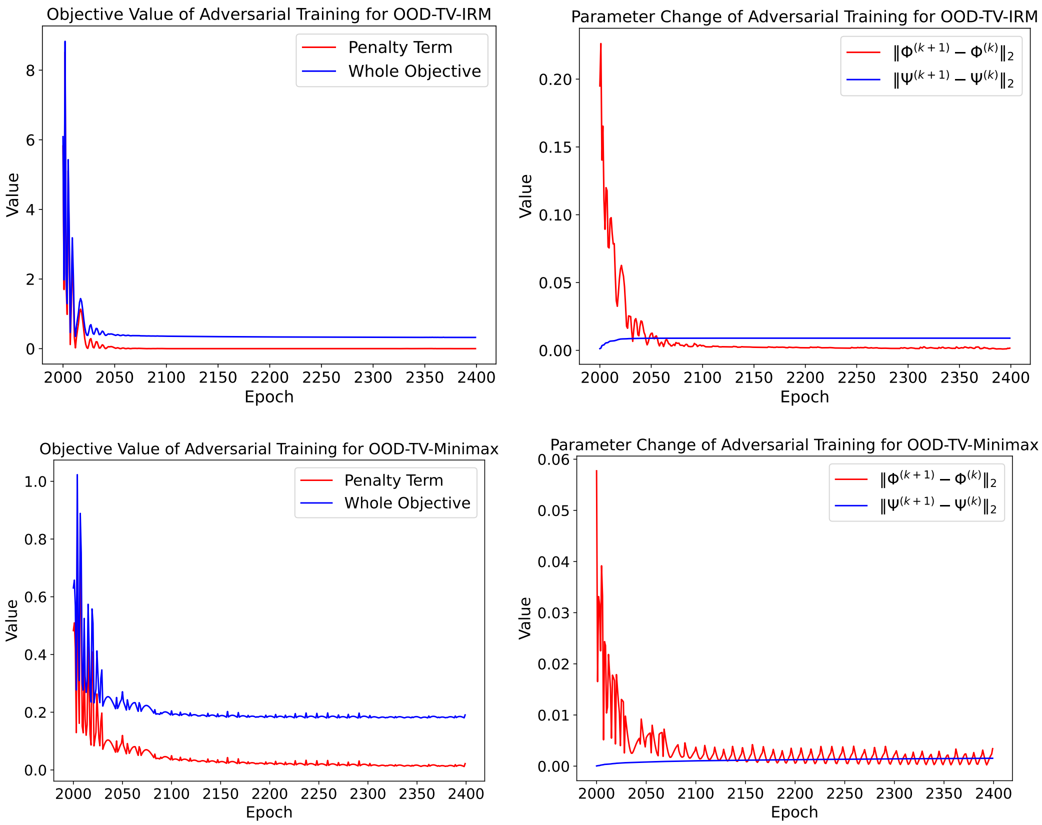

Figure A1: Objective value and parameter change of adversarial learning for OOD-TV-IRM and OOD-TV-Minimax on simulation data.

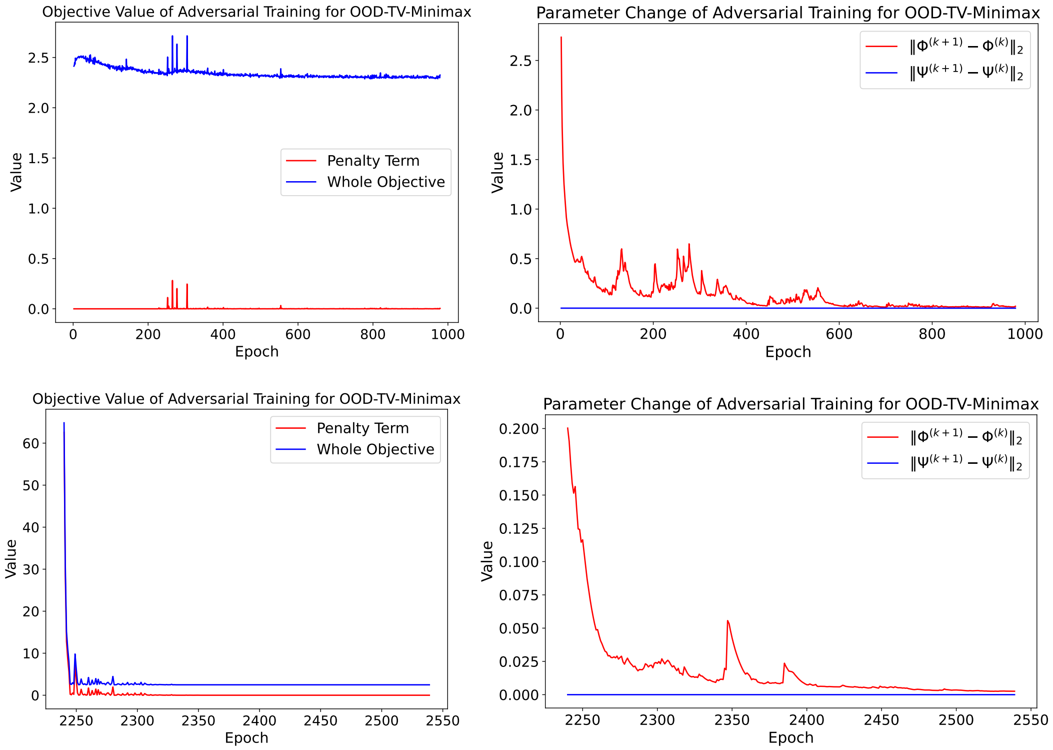

Figure A2: Objective value and parameter change of adversarial learning for OOD-TV-Minimax on the Colored MNIST data set. The top two figures correspond to OOD-TV-Minimax-ℓ_1, while the bottom two figures correspond to OOD-TV-Minimax-ℓ_2.