Invariant Risk Minimization Games with a twist - A Game Theory mini project

Adarsh Shah, adarshshah@iisc.ac.in

Abstract

In many real-world applications, we need to learn models that can perform well on new domains or environments that are different from the ones used for training. This is known as domain generalization or out-of-distribution generalization. One way to achieve this is to learn feature representations that are invariant across different domains or environments. Invariant risk minimization games [1] is a framework that formulates this problem as a game between multiple training environments, where each environment tries to find a classifier that minimizes its own risk while being consistent with the others. The final predictor is an ensemble of the classifiers learned by each environment. However, invariant risk minimization games may not always succeed in finding the true causal features of the target variable. In this project, we propose Bayesian invariant risk minimization games, which incorporate prior knowledge and uncertainty into the learning process. We show that Bayesian invariant risk minimization games are not truly domain invariant but can improve the performance and robustness of domain generalization models.

1 Introduction

The initial approach to address most of the problems in machine learning is empirical risk minimization (ERM) [2]. The standard methodology is to develop two datasets, namely, $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$. A hypothesis class $\mathcal{H}_w$ is chosen, and the optimal hypothesis h_{w^*} is obtained by minimizing the empirical risk with appropriate loss function l defined as follows.

$$h_{w^*} \in \underset{h_w \in \mathcal{H}_w}{argmin}\, \mathbb{E}_{X,Y\sim\mathcal{D}_{train}}[l(Y, h_w(X))] \tag{1}$$

The hope is that the optimal hypothesis h_{w^*} works well on $\mathcal{D}_{test}$. If the samples (X, Y) in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are sampled in i.i.d. fashion from the same distribution, h_{w^*} works well for $\mathcal{D}_{test}$ almost surely. However, in practice, this is not the case.

A popular case is presented in 'Recognition in Terra Incognita'[3]. $\mathcal{D}_{train}$ consists of cows and camels images. The objective is to learn a hypothesis function that successfully classifies cow images from camels. Also, there exists a selection bias in the training dataset. Images of cows are taken in green pastures, whereas camel images mainly consist of deserts. Once trained on this dataset, the optimal model fails to classify images of cows taken on sandy beaches correctly. This is mainly because the optimal hypothesis treats green pastures as cows and brown sandy deserts as camels. These kinds of correlations present in the dataset are called *spurious correlations*.

Invariant risk minimization [4] introduces an approach to address the issues caused by spurious correlations. Multiple sources of data have varying degrees of spurious correlations. IRM proposes to exploit this and obtain a domain-invariant feature representation Φ. A bi-level optimization problem is introduced in which a linear classifier w is learned on top of invariant feature representation Φ to obtain domain invariant feature extractor $w \circ \Phi$.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Invariant risk minimization games [1] is a game theoretic formulation of invariant risk minimization. This paper proposes a game where the players are training environments $\mathcal{E}_{tr}$, and their strategy set is the set of classifiers $\mathcal{H}_w$. The invariant predictor is obtained by taking an ensemble of each environment's classifier. Using best response dynamics [5], it can be shown that when the players/environments reach Nash equilibrium, the optimal predictor obtained is invariant across training environments $\mathcal{E}_{tr}$.

Achieving invariance across training environments $\mathcal{E}_{tr}$ does not guarantee invariance across all environments $\mathcal{E}_{all}$.

IRM is not a panacea and has some limitations and drawbacks. The paper "The Risks of Invariant Risk Minimization" [6] analyzes IRM objective formally. It shows the settings in which IRM fails to achieve invariance, and its performance is not better than Empirical Risk Minimization (ERM). The paper [6] provides exact conditions in which the IRM does not perform better than ERM in linear settings. In this project, we discuss this work and highlight that the same failure settings also hold for invariant risk minimization games, which are a generalization of IRM that allows for multiple players with different objectives.

One of the main reasons why optimal predictors fail to achieve domain invariance is when optimal predictors overfit the training data $\mathcal{E}_{tr}$. The theoretical and empirical analysis supporting this is argued in Bayesian Invariant Risk Minimization [7] (BIRM). BIRM is an extension of IRM where posterior distribution over each environment's classifier is estimated instead of a single classifier. This approach prevents overfitting to some extent. In this project, we extend invariant risk minimization games to bayesian settings using the BIRM approach. We develop a synthetic dataset to demonstrate a setting in which IRM games perform as poorly as ERM, but BIRM performs better.

2 Related Works

2.1 Domain Adaptation

Domain adaptation [8], [9], [10] and domain generalization are two related but distinct problems in machine learning that deal with the challenge of generalizing to out-of-distribution (OOD) data.

Domain adaptation assumes that both the source domain (the training data) and the target domain (the OOD data) are available during training, but labels for the target domain are not always available. The goal is to learn a model that can leverage the information from the source domain and adapt it to the target domain, minimizing the domain shift. Domain adaptation methods can be categorized into three types: supervised, unsupervised, and semi-supervised, depending on whether some, none, or all of the target data have labels.

Domain generalization assumes that only the source domain is available during training, and the target domain is unknown. The goal is to learn a model that can generalize well to any unseen domain, without requiring any adaptation. Domain generalization methods can be categorized into four types: domain alignment, meta-learning, data augmentation, and ensemble learning, depending on how they try to achieve domain-invariant or robust representations.

2.2 Invariant Risk Minimization

Invariant Risk Minimization [4] proposes a framework to prevent optimal predictors from learning spurious correlations. The main idea is that the optimal predictor will be truly invariant if it classifies on the invariant descriptions of objects under consideration. The invariant descriptions of objects arise from the causal factors of the objects themselves [11]. Hence, invariance and causation are closely related to each other, as shown in [12], [13]. IRM is a novel learning technique to estimate domain invariant and causal predictors from multiple training environments $\mathcal{E}_{tr}$.

Definition 1. *A data representation $\Phi : X \to H$ elicits an invariant predictor $w \circ \Phi$ across environments $\mathcal{E}$ if there is a classifier $w : H \to Y$ simultaneously optimal for all environments, that is, $w \in argmin_{\bar{w}:H\to Y} R^e(\bar{w} \circ \Phi)$ for all $e \in \mathcal{E}$.*

Following [12], let $\mathcal{D}_e = \{(x_e, y_e)_{i=1}^{n_e}\}$ be the dataset for each training environment $e \in \mathcal{E}_{tr}$. Every sample (x_e, y_e) from $\mathcal{D}_e$ is assumed to sampled in i.i.d. fashion from environment's underlying distribution $P(X^e, Y^e)$. The main goal of IRM is to learn invariant predictor $Y \approx f(X)$, which performs well across all environments $\mathcal{E}_{all} \supset \mathcal{E}_{tr}$. Mathematically, we want to achieve out-of-distribution risk minimization as follows.

$$R^{OOD}(f) = \underset{e \in \mathcal{E}_{all}}{max} R^e(f) \tag{2}$$

Here, $R_e(f) = \mathbb{E}_{X_e,Y_e}[l(f(X_e), Y_e)]$ is the risk under environment e.

The following bi-leveled optimization problem is introduced in IRM to capture invariant predictors across training environments $\mathcal{E}_{tr}$.

$$\begin{gathered} \underset{\substack{\Phi:X\to H \\ w:H\to Y}}{\texttt{min}} \sum_{e\in\mathcal{E}_{tr}} R^e(w \circ \Phi) \\ \texttt{subject to } w{\in}\underset{\bar{w}:H\to Y}{\texttt{arg min}}\ R^e(\bar{w}{\circ}\Phi)\texttt{, for all } e{\in}\mathcal{E}_{tr} \end{gathered} \tag{3}$$

This problem is difficult to optimize because each constraint is also an optimization problem. The practical version of the above is proposed in [4] as follows:

$$\underset{\Phi:X\to Y}{min} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda \cdot ||\nabla_{w|w=1.0} R^e(w \cdot \Phi)||^2 \tag{4}$$

2.3 Domain generalization and causality

The theory of causation provides insights about conditions in which invariance across all environments $\mathcal{E}_{all}$ is achieved [4]. We will revisit a few concepts from causal theory[14] to understand the same.

Definition 2. *A Structural Equation Model (SEM) $C := (S, N)$ governing the random vector $X = (X_1, \ldots, X_d)$ is a set of structural equations:*

$$S_i : X_i \leftarrow f_i(Pa(X_i), N_i),$$

where $Pa(X_i) \subseteq X_1, ..., X_d$ X_i are called the parents of $_Xi$, and the N_i are independent noise random variables. We say that X_i causes X_j if $X_i \in Pa(X_j)$. We call the causal graph of X to the graph obtained by drawing i) one node for each X_i, and ii) one edge from X_i to X_j if $X_i \in Pa(X_j)$. We assume acyclic causal graphs.

Definition 3. *Consider a SEM $C := (S, N)$. An intervention e on C consists of replacing one or several of its structural equations to obtain an intervened SEM $C^e = (S^e, N^e)$ with structural equations:*

$$S^e_i : X^e_i \leftarrow f^e_i(Pa^e(X^e_i), N^e_i),$$

The variable X_e is intervened if $S_i \neq S^e_i$ or $N_i \neq N^e_i$.

The distribution associated with SEM C is called *observational distribution* $P(X, Y)$. Sampling from the observational distribution $P(X, Y)$ is the same as sampling a random vector by tracing the C's structural equations following the causal graph's topological order. Similarly, the distribution associated with intervened SEM C^e is *intervened distribution* $P(X^e, Y^e)$. An environment e, with distribution $P(X^e, Y^e)$, is generated by intervening in the original SEM C. The interventions which do not distort Y a lot are considered as *valid interventions*. These valid interventions form the set of all environments $\mathcal{E}_{all}$. [12] considers the interventions which do not modify the structural equation of Y as valid interventions.

Theorem 1. *Consider a SEM C governing the random vector $(X_1, ..., X_d, Y)$, and the learning goal of predicting Y from X. Then, the set of all environments $E_{all}(C)$ indexes all the interventional distributions $P(X_e, Y_e)$ obtainable by valid interventions e. An intervention $e \in E_{all}(C)$ is valid as long as (i) the causal graph remains acyclic, (ii) $\mathbb{E}[Y^e|Pa(Y)] = \mathbb{E}[Y|Pa(Y)]$, and (iii) $\mathbb{V}[Y^e|Pa(Y)]$ remains within a finite range.*

A predictor $v : X \to Y$ is domain invariant if and only if $v(x)$ uses Y's direct causal parents, i.e., $v(x) := \mathbb{E}[f_Y(Pa(Y), N_Y)]$. It also implies that $v(x)$ attains out-of-distribution i.e., $R^{OOD}(v)$, is minimized. This is a fundamental link between invariance and causation. If the underlying SEM structure is known in advance, invariant predictors can be obtained using multiple techniques [15] [16], but finding underlying SEM is itself highly non-trivial.

2.4 Domain generalization using class conditional independence

The distribution of a truly invariant data representation Φ is independent across all environments $\mathcal{E}_{all}$ as evident from previous sections. Inspired by this, a line of work [17] [18] began to explore representations by assuming class conditional distribution $P(\Phi(X)|Y)$ is independent of the environments. The insufficiency of this approach is proved using a simple counter-example in Section (3.1, [19]).

3 Main Ideas

3.1 Ensemble Invariant Risk Minimization Games (EIRM)

EIRM is a game theoretic formulation of IRM optimization problem (3). The players of EIRM are training environments $e \in \mathcal{E}_{tr}$. Each player e has a classifier $w_e \in \mathcal{H}_w$ associated with it. Hence, $\mathcal{H}_w$ is player e's strategy set. The ensemble classifier is defined as the average of all player's classifiers, i.e., $w_{av} = \frac{1}{|\mathcal{E}_{tr}|} \sum\limits_{e \in \mathcal{E}_{tr}} w_e$. The IRM optimization problem (3) is modified for EIRM as follows

$$\begin{gathered} \min_{\substack{w_{av} \in \mathcal{H}_w \\ \Phi \in \mathcal{H}_\Phi}} \sum_{e \in \mathcal{E}_{tr}} R^e(w_{av} \circ \Phi) \\ \texttt{s.t.} \quad w_e \in \texttt{arg}\min_{\bar{w}_e \in \mathcal{H}_w} R^e\Big(\tfrac{1}{|\mathcal{E}_{tr}|}[\bar{w}_e + \sum_{q \neq e} w_q] \circ \Phi\Big), \forall e \in \mathcal{E}_{tr} \end{gathered} \tag{5}$$

The above problem can be reformulated below

$$\begin{gathered} \min_{\substack{w_{av} \in \mathcal{H}_w \\ \Phi \in \mathcal{H}_\Phi}} \sum_{e \in \mathcal{E}_{tr}} R^e(w_{av} \circ \Phi) \\ \texttt{s.t.} \quad R^e\Big(\tfrac{1}{|\mathcal{E}_{tr}|}[w_e + \sum_{q \neq e} w_q] \circ \Phi\Big) \leq R^e\Big(\tfrac{1}{|\mathcal{E}_{tr}|}[\bar{w}_e + \sum_{q \neq e} w_q] \circ \Phi\Big), \forall \bar{w}_e \in \mathcal{E}_{tr} \end{gathered} \tag{6}$$

The advantages of the EIRM problem (6) are described as follows.

- Each environment e optimizes the common ensemble classifier w_{av}.
- Each environment e can independently select w_e from entire $\mathcal{H}_w$, thereby decoupling environments.
- The set of feasible solutions to the given problem is the same as the set of pure Nash equilibrium (NE) of the EIRM game defined below.

$$\Gamma^{EIRM} = (\mathcal{E}_{tr}, \mathcal{H}_\Phi, \{\mathcal{H}_w\}_{q=1}^{|\mathcal{E}_{tr}|}, \{u_e\}_{e \in \mathcal{E}_{tr}}) \tag{7}$$

A representation $\Phi \in H_\Phi$ is fixed before the game starts. Each environment e's utility function is defined as $u_e[w_e, w_{-e}, \Phi] = -R^e(w_{av}, \Phi)$ where $w_{-e} = \{w_q\}_{q \neq e}$.

3.2 Theoretical Insights

Suppose, a tuple $(\Phi, \{w_q\}_{q=1}^{|\mathcal{E}_{tr}|})$ represents a pure NE of the given game. Let S^{EIRM} be the set of all pure NE. We define the set of all ensemble predictors from S^{EIRM} as follows.

$$\bar{S}^{EIRM} = \{[\frac{1}{|\mathcal{E}_{tr}|} \Sigma_{q \in \mathcal{E}_{tr}} w_q] \circ \Phi | (\Phi, w_q{}_{q=1}^{|\mathcal{E}_{tr}|}) \in S^{EIRM}\} \tag{8}$$

Similarly, let S^{IV} be the set of pairs (Φ, w) satisfying the constraints in problem (3) and $\bar{S}^{IV} = \{w \circ \Phi | (\Phi, w) \in S^{IV}\}$ be the set of invariant predictors. It can be shown that the sets $\bar{S}^{EIRM}$ and $\bar{S}^{IV}$ are the same, i.e., the pure NE of the EIRM game is the same as feasible solutions of IRM objective (3) under mild conditions. We prove the same in theorem 2. The theorems 3, 4, and 5 provide insights in specific settings and are not as significant as theorem 2.

Assumption 1. ***Affine closure:*** *The class of functions $\mathcal{H}_w$ is closed under the following operations.*

- *Finite sum: If $w_1 \in \mathcal{H}_w$ and $w_2 \in \mathcal{H}_w$, then $w_1 + w_2 \in \mathcal{H}_w$, where for every $z \in Z, (w_1 + w_2)(z) = w_1(z) + w_2(z)$*

- *Scalar multiplication: For any* $c \in \mathbb{R}$ *and* $w \in \mathcal{H}_w$, $c_w \in \mathcal{H}_w$, *where for every* $z \in Z, (xw)(z) = x \times w(z)$

The above assumption states that $\mathcal{H}_w$ is a vector space.

Theorem 2. *If Assumption 1 holds, then* $\bar{S}^{EIRM} = \bar{S}^{IV}$.

Proof. First, we prove that $\bar{S}^{EIRM} \subseteq \bar{S}^{IV}$.

For a fixed data representation Φ and any pure NE of EIRM game $(\Phi, \{w_q\}_{q=1}^{|\mathcal{E}_{tr}|})$, the corresponding predictor $w_{av} \circ \Phi \in \bar{S}^{EIRM}$. According to EIRM constraints (6),

$$\begin{aligned} &R^e(w_{av} \circ \Phi) \leq R^e(w \circ \Phi), \forall w \in \mathcal{H}_w, \forall e \in \mathcal{E}_{tr} \\ \therefore\ & w_{av} \in \underset{\bar{w}: H \to Y}{\texttt{arg min}}\ R^e(\bar{w} \circ \Phi), \forall e \in \mathcal{E}_{tr} \\ \therefore\ & w_{av} \circ \Phi \in \bar{S}^{IV} \to \bar{S}^{EIRM} \subseteq \bar{S}^{IV} \end{aligned}$$

Next, we prove that $\bar{S}^{IV} \subseteq \bar{S}^{EIRM}$.

For a fixed data representation Φ and any $(\Phi, w) \in S^{IV}$, $w \circ \Phi \in \bar{S}^{IV}$. Suppose, a pure NE of EIRM game $(\Phi, \{w_q\}_{q=1}^{|\mathcal{E}_{tr}|})$ is such that $w \circ \Phi \notin \bar{S}^{EIRM}$,i.e., $R^e(w \circ \Phi) < R^e(w_{av} \circ \Phi)$. Under the affine closure assumption (1), there exists $\{w_q\}_{q=1}^{|\mathcal{E}_{tr}|}$ such that $\frac{1}{|\mathcal{E}_{tr}|}\Sigma_{q=1}^{|\mathcal{E}_{tr}|} w_q = w$. This is a contradiction since w_{av} violates EIRM's constraints (6). Hence, proved. □

The theorem 2 implies that it is *computationally feasible* to navigate along the 3 constraints using EIRM. For a given data representation Φ, the NE of the EIRM game satisfies the 3 constraints.

Theorem 3. *For each environmente* $\in \mathcal{E}_{all}$ *we assume*

$$\begin{aligned} &Y^e \leftarrow Z_1^{eT} + \epsilon^e, Z_1^e \perp \epsilon^e, \mathbb{E}[\epsilon^e] = 0 \\ &X^e \leftarrow S(Z_1^e, Z_2^e) \end{aligned}$$

Here, $\gamma \in \mathbb{R}^c, Z_1^e \in \mathbb{R}^c, Z_2^e \in \mathbb{R}^q, S \in \mathbb{R}^{n \times (c+q)}$. *Assume that* Z_1 *is invertible component of* S, *i.e.,* $\exists \bar{S} \in \mathbb{R}^{c \times n}$ *such that* $\bar{S}(S(z_1, z_2)) = z_1$ *for all* $z_1 \in \mathbb{R}^c$ *and* $z_2 \in \mathbb{R}^d$. *Let* $\Phi \in \mathbb{R}^{n \times n}$ *have rank r. If at least* $n - r + \frac{n}{r}$ *training environments* $\mathcal{E}_{tr} \subseteq \mathcal{E}_{all}$ *lie in the linear general position of degree* r, *then any predictor obtained from EIRM game over the training environments in* $\bar{S}^{EIRM}$ *is invariant across all the testing environments* $\mathcal{E}_{all}$.

The theorem 3 states a linear regime and extends the generalization guarantees provided in (9, Theorem [4]) to EIRM. Using theorem 2 and proof of Theorem 9 in [4], the proof of theorem 3 is straightforward.

Assumption 2. $\Phi \in \mathcal{H}_\Phi$ *satisfies the following*

- *Bijective:* $\exists \Phi^{-1} : Z \to X$ *such that* $\forall x \in X, (\Phi^{-1} \circ \Phi)(x) = x$ *and* $\forall z \in Z (\Phi \circ \Phi^{-1})(z) = z$. *Both* X *and* Z *are subsets of* $\mathbb{R}^n$.
- Φ *is differentiable and Lipschitz continuous.*

Assumption 3. $L^p(Z)$: *set of functions* $f : Z \to \mathbb{R}$ *s.t.* $\int_Z |f|^p d_\mu < \infty$. $\mathcal{H}_w = L^p(Z)$.

Let $\bar{S}^{IV}(\Phi) = \{w \circ \Phi | w \circ \Phi \in \bar{S}^{IV}\} \subseteq \bar{S}^{IV}$. Therefore, $\bigcup_\Phi \bar{S}^{IV}(\Phi) = \bar{S}^{IV}$. Similarly, $\bar{S}^{EIRM}(\Phi)$ is defined.

Theorem 4. *If assumptions 2 and 3 are satisfied and* $\bar{S}^{IV}$ *is non empty,* $\bar{S}^{IV} = \bar{S}^{IV}(I) = \bar{S}^{EIRM}(I)$.

The theorem 4 discusses the role of representation Φ in EIRM. EIRM game is played for a fixed representation Φ only. EIRM games do not aim to find an invariant representation Φ but aim to find an invariant predictor $w \circ \Phi$ across $\mathcal{E}_{tr}$ for a given Φ. For the settings defined by assumptions 2 and 3, *fixing the representation to identity* I is enough to obtain invariant predictors using EIRM games.

Algorithm Best Response Training

```
Input: Data for each environment and combined data
Initialize: Randomly initialize $\{w^e_{cur}\}_{e=1}^{|\mathcal{E}_{tr}|}$ and $\Phi_{cur}$ from $\mathcal{H}_w$ and $\mathcal{H}_\Phi$ respectively.
while $\texttt{iter} \leq \texttt{iter}_{\texttt{max}}$ do
    if F-IRM then
        $\Phi_{cur} = I$
    end if
    if V-IRM then
        $\Phi_{cur} = \texttt{SGD}[\Sigma_e R^e(w^{av}_{cur} \circ \Phi)]$, update using stochastic gradient descent
    end if
    for $p \in \{1, \ldots, K\}$ do
        for $e \in \{1, \ldots, |\mathcal{E}_{tr}|\}$ do
            $w^e_{curr} = \texttt{SGD}[\Sigma_e R^e(w^{av}_{cur} \circ \Phi)]$, update using stochastic gradient descent
            $w^{av}_{curr} = \frac{1}{|\mathcal{E}_{tr}|} \Sigma_e w^e_{cur}$
        end for
        $\texttt{iter} = \texttt{iter} + 1$
    end for
end while
```

Assumption 4.

- *$\mathcal{H}_w$ is a class of linear models, i.e., $w \in \mathcal{H}_w \subseteq \mathbb{R}^{d\times 1}$ and classifier output for input z is $w^T z$. $\mathcal{H}_w$ is a closed, bounded, and convex set. The interior of $\mathcal{H}_w$ is non-empty.*
- *The loss function $l(w^T z, Y)$, where $Y \in \mathbb{R}$ is the label, is convex and continuous in w. For e.g., if the loss is cross-entropy for binary classification or loss is mean squared error for regression, then this assumption is automatically satisfied.*

Theorem 5. *If assumption 4 is satisfied, then a pure strategy Nash equilibrium of the game Γ^{EIRM} exists, i.e., S^{EIRM} is non-empty. Suppose there exists a $(\Phi, \{w^q\}_{q=1}^{|\mathcal{E}_{tr}|})$ such that $\forall q \in \mathcal{E}_{tr}$, w_q is in the interior of $\mathcal{H}_w$, then the corresponding ensemble predictor $\frac{1}{|\mathcal{E}_{tr}|\Sigma_{q=1}^{|\mathcal{E}_{tr}|}} w_q \circ \Phi$ is invariant across all the training environments $\mathcal{E}_{tr}$.*

The theorem 5 guarantees the existence of invariant predictors across training environments when the classifiers are linear and no assumption is made on the data representation.

4 EIRM Algorithm

The paper [1] describes two algorithms for EIRM games. First, F-IRM algorithm stands for playing EIRM with fixed data representation ($\Phi = I$), and second, V-IRM algorithm with learnable data representation. Both algorithms are described in a combined fashion in 4

5 Few applications

5.1 FL Games: A federated learning framework for distribution shifts [20]

This paper applies ensemble invariant risk minimization games in a federated environment where the data is distributed across clients, and the server coordinates clients to learn effective predictors. It addresses the client heterogeneity problem where since each client is assumed to be different, the data distributed across clients are not i.i.d. This causes distribution shifts across data and makes it challenging to learn invariant predictors. The other problems arise due to the sequential nature of F-IRM and V-IRM algorithms, the large oscillations observed during training, and the slow convergence speed in the case of many clients. The following modifications are made to F-IRM and V-IRM algorithms.

$$\underset{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w^{av}:\mathcal{H}\to\mathcal{Y}}}{min} \sum_{e\in\mathcal{E}_{tr}} R^e(w^{av} \circ \phi) \texttt{ subject to}$$

$$w \in \underset{w_e:\mathcal{H}\to\mathcal{Y}}{argmin} R^e\left(\frac{1}{|\mathcal{E}_{tr}|}(w_e + \underset{\substack{p\in\mathcal{E}_{tr}\\ p\neq e}}{\Sigma} w^p + \Sigma_{\substack{j=1\\ p\in\mathcal{E}_{tr}\\ p\neq e}}^{|\mathcal{B}_p|} \frac{1}{|\mathcal{B}_p|} w_j^p) \circ \phi\right); \forall e \in \mathcal{E}_{tr}$$

A two-way ensemble approach is introduced where historical models are also considered to reduce oscillations during training. Additionally, gradient update steps are parallelized over batched data to avoid problems arising due to the original algorithm's sequential nature.

5.2 Linear Regression Games: Convergence Guarantees to Approximate Out-Of-Distribution Solutions [21]

This paper is also an application of Invariant Risk Minimization Games. Consider an IRM game with two environments $\mathcal{E}_{tr} = \{1, 2\}$ where strategy set $\mathbf{w}_i$ is the set of all linear regressors and the utility is defined using the risk function $R_e(\mathbf{w}_1, \mathbf{w}_2) = \mathbb{E}_e[(Y_e - \mathbf{w}_1^T\mathbf{X}_e - \mathbf{w}_2^T\mathbf{X}_e)^2]$. Suppose, $\mathbf{X}_e \in \mathbb{R}^{n\times 1}$, the strategic form game is defined as follows:

$$\Gamma = (\{1, 2\}, \mathbb{R}^{n\times\ 1}, \{R_e\}_{e\in\{1,2\}})$$

With sufficient regularity conditions on $\mathbf{X}_e$ and $\mathbf{Y}_e$, i.e. $\mu_e = \mathbb{E}_e[\mathbf{X}_e] = 0$, $\mathbf{\Sigma}_e = \mathbb{E}[\mathbf{X}_e\mathbf{X}_e^T]$ as positive definite and $\rho_e = \mathbb{E}[\mathbf{X}_eY_e]$, the least square solution is $\mathbf{w}_e^* = \mathbf{\Sigma}_e^{-1}\rho_e$. If $\mathbf{w}_1^* = \mathbf{w}_2^*$, the set $\{(\mathbf{w}_1, \mathbf{w}_2); \mathbf{w}_1 + \mathbf{w}_2 = \mathbf{w}_1^*\}$ represent all PSNE of Γ otherwise the PSNE does not exist. It also discusses closed-form solutions of optimal invariant linear regressors in case of the presence/absence of confounders and anti-causal variables. The strategy set in this game is unconstrained and hence not compact. The predictors are constrained to be in $\mathbb{W} = \{\mathbf{w}_e | ||\mathbf{w}_e||_\infty \leq w^{sup}\}$. A new game is defined with this strategy set, and it is shown that Nash equilibrium always exists for the new game.

6 Limitations of Invariant Risk Minimization Games

In this section, we discuss results from multiple sources, establish a connection between these results and EIRM games, and frame the limitations of EIRM games concretely.

Consider an SEM model with a set of endogenous variables x_c, x_e, and y. x_e variables are dependent on environments and x_c variables are direct causal parents of y. Let f be a true unknown function such that $y = f(x_c, x_e)$ and $x_c \in \mathbb{R}^{d_e}$. The risks of invariant risk minimization [6] state the following three main theorems informally.

Theorem 6. *For linear f, consider solving the IRM objective to learn a linear Φ with an invariant optimal classifier w. If $\mathcal{E}_{tr} > d_e$, then Φ, w is precisely the optimal invariant predictor; it uses only invariant features and generalizes to all environments with minimax-optimal risk. If $\mathcal{E}_{tr} \leq d_e$, then Φ, w relies upon non-invariant features.*

The theorem 6 describes the number of training environments $|\mathcal{E}_{tr}|$ required to achieve domain invariance using IRM in settings where the predicted variable y is a linear transform of its causal parents x_e in the underlying unknown SEM model. d_e is the number of y's causal parents, which is also unknown.

Theorem 7. *For linear f and $\mathcal{E}_{tr} \leq d_e$, there exists a linear predictor Φ, w which uses only environmental features yet achieves lower risk than the optimal invariant predictor.*

The theorem 7 implies that in linear settings when IRM fails to achieve domain invariance, its performance is the same as empirical risk minimization. Hence, an important point to note is that there exist conditions in which IRM performs as badly as ERM.

In both the theorems 6 and 7, we assume the feature extractor Φ is restricted to a linear transform. The following theorem describes settings in which IRM fails for nonlinear feature extractors Φ.

Theorem 8. *For arbitrary f, there exists a non-linear predictor Φ, w, which is nearly optimal under the penalized objective and is nearly identical to the optimal invariant predictor on the training distribution. However, for any test environment with a mean sufficiently different from the training means, this predictor will be equivalent to the ERM solution on nearly all test points. For test distributions where the environmental feature correlations with the label are reversed, this predictor has almost 0 accuracy.*

The theorem 2 establishes the equivalence of IRM with EIRM games. Hence, the settings described in theorems 6, 7, and 8 also hold for EIRM games.

Another line of work [7] argues that one of the prominent reasons IRM degenerates to emprical risk minimization (ERM) is overfitting. Using appropriate assumptions, we describe the main result as follows.

Assumption 5. *(Finite Sample size) The number of training environments and samples are finite:* $|\mathcal{E}_{tr}| < \infty$ *and* $|\mathcal{D}_e| = n^e < \infty, \forall e \in \mathcal{E}_{tr}$.

Assumption 6. *(Sufficient Capacity) The hypothesis classes* $\mathcal{H}_w$ *and* $\mathcal{H}_\Phi$ *have sufficient capacity to fit the training data: There exists* $w \in \mathcal{H}_w$ *and* $\Phi \in \mathcal{H}_\Phi$*, such that* $\forall e \in \mathcal{E}_{tr}, R^e(w \circ \Phi) = 0$.

Definition 4. *The overfitting region* Ω*, is the collection of* w *and* Φ *such that assumption 6 is satisfied:*

$$\Omega := \{(w, \Phi) | R^e(w \circ \Phi) = 0, \forall e \in \mathcal{E}_{tr}\}$$

Theorem 9. *Under assumptions 5 and 6, IRM degenerates to ERM in* Ω*. Furthermore, any element in* Ω *is a solution of IRM defined in Eq. 3.*

Proof. We first prove that every element of Ω is a fixed point of 3. For any $(w, \Phi) \in \Omega$, by definition 4, $R^e(w \circ \Phi) = 0, \forall e \in \mathcal{E}_{tr}$ Also, $R^e(w \circ \Phi) \geq 0, \forall w, \Phi$. Hence,

$$w \in \underset{w \in \mathcal{H}_w}{\texttt{arg min}} \; R^e(w \circ \Phi), \forall e \in \mathcal{E}_{tr}$$

Since all elements of Ω satisfy 3 constraints and $R^e(w \circ \Phi) = 0, \forall e \in \mathcal{E}_{tr}$, all elements of Ω are fixed point of 3.

To show that IRM degenerates to ERM, consider the following collection of (w, Φ):

$$\Omega' = \{(w, \Phi) | w \in \underset{w \in \mathcal{H}_w}{\texttt{arg min}} \; R^e(w \circ \Phi), \forall e \in \mathcal{E}_{tr} \texttt{ and } (w, \Phi) \notin \Omega\}$$

For any $(w, \Phi) \in \Omega$ and $(w', \Phi') \in \Omega'$, there exists training environment $e \in \mathcal{E}_{tr}$ such that

$$R^e(w' \circ \Phi') > R^e(w \circ \Phi)$$
$$\therefore \Sigma_e R^e(w' \circ \Phi') > \Sigma_e R^e(w \circ \Phi)$$

The above shows that any member of Ω' will not be as good as any member of Ω. Hence, IRM degenerates to ERM, which is the Ω region. □

Using theorem 9 and 2, we conclude that overfitting causes EIRM games to degenerate to ERM.

7 Bayesian Invariant Risk Minimization

Bayesian invariant risk minimization [7] adopts tools from Bayesian inference to address the problems caused by overfitting. Bayesian inference is shown to prevent overfitting [22], [23].

Suppose the finite dataset corresponding to each training environment e is denoted as $\mathcal{D}_e = \{(x_i^e, y_i^e)_{i=1}^{n_e}\}$. Corresponding to each training environment e, we define a class of distributions $q_\Phi^e(w_e) \in \mathcal{Q}_w$ over the classifier's hypothesis class $\mathcal{H}_w$. We define an additional environment with training data $\mathcal{D} = \bigcup_{e \in \mathcal{E}_{tr}} \mathcal{D}_e$ with a class of distributions $q_\Phi(w) \in \mathcal{Q}_w$.

The main idea of BIRM is to assume a posterior distribution $q_\Phi^e(w^e) \in \mathcal{Q}_w$ over hypothesis class $\mathcal{H}_w$ for each environment $e \in \mathcal{E}_{tr}$ instead of selecting $w \in \mathcal{H}_w$ pointwise as done in IRM 3. The main objective of bayesian invariant risk minimization is to obtain a feature representation Φ by minimizing the following.

$$\underset{\Phi}{\texttt{min}} \; \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{p(w|\mathcal{D},\Phi)}[R^e(w \circ \Phi)] + \lambda(\mathbb{E}_{p(w|\mathcal{D},\Phi)}[R^e(w \circ \Phi)] - \mathbb{E}_{p(w|\mathcal{D}_e,\Phi)}[R^e(w \circ \Phi)]) \tag{9}$$

where λ is a tunable parameter In equation 9, the first term is the reconstruction loss of $q_\Phi(w)$ across all training environments $\mathcal{E}_{tr}$. The second term within λ represents the deviation in distributions $q_\Phi(w)$ and $q_\Phi^e(w_e)$.

The main observation is if Φ is truly domain invariant, then the distribution of data $\{\Phi(x), y\}$ across training environments will be the same. As a result, the optimal posteriors on $\mathcal{H}_w$ will be the same across distributions, i.e., $p(w|\mathcal{D}_{e_1}, \Phi) \approx p(w|\mathcal{D}_{e_2}, \Phi)$. Moreover, $p(w|\mathcal{D}, \Phi) \approx p(w|\mathcal{D}_e, \Phi), \forall e \in \mathcal{E}_{tr}$. In this condition, the first term as well as the second term in equation 9 tend to 0, thereby minimizing it.

8 My Work: Bayesian Invariant Risk Minimization Games

This section proposes Bayesian Invariant Risk Minimization (BIRM) games as a game theoretic formulation of Bayesian Invariant Risk Minimization [7]. The main motivation of this extension is to address one of the reasons, mainly overfitting, that cause Invariant Risk Minimization games to fail. There is no theoretical proof that Bayesian Invariant Risk Minimization achieves true domain invariance. Still, empirically, it is shown to perform better than Invariant Risk Minimization in some practical cases.

8.1 Formulation

Given finite number of training environments $\mathcal{E}_{tr} \supset \mathcal{E}_{all}$ each with finite sample set $\mathcal{D}_e = \{x_i^e, y_i^e\}_{i=1}^{n_e}$, assume an additional training environment $\bar{e}$ such that $\mathcal{D}_{\bar{e}} = \bigcup_{e \in \mathcal{E}_{tr}} \mathcal{D}_e$. In this game, there are $|\mathcal{E}_{tr}| + 2$ players, composed of training environments and additional players, described as follows.

$$N := < \{e | e \in \mathcal{E}_{tr}\} \cup \{\bar{e}, F\} >$$

where $\bar{e}$ and F are the players representing the union of all training environments and the feature extractor, respectively, and $< \cdots >$ represents the ordered set. We define the strategy set of each player as follows.

- For players $e \in \mathcal{E}_{tr} \cup \{\bar{e}\}$, strategy set is $S_e = \mathcal{Q}_w$.
- For player F, strategy set is $S_F = \mathcal{H}_\Phi$.

Here, $\mathcal{Q}_w$ is the distribution class, i.e., a set of all distributions over $\mathcal{H}_w$. We define the strategy set of each player as follows.

1. For players $e \in \mathcal{E}_{tr} \cup \{\bar{e}\}$, utility function is
$$u_e[q_e, q_{-e}, \Phi] = \mathbb{E}_{q_e(w)}[-R^e(w \circ \Phi)] - D_{KL}(q_e(w)||p_0(w))$$
where $-e = \mathcal{E}_{tr} \cup \{\bar{e}\} \setminus \{e\}$ and $p_0 \in \mathcal{Q}_w$ is a fixed prior distribution.
2. For feature extractor f, utility function is
$$u_F[q_e, q_{-e}, \Phi] = \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{q_{\bar{e}}}[-R^e(w \circ \Phi)] + \lambda(\mathbb{E}_{q_{\bar{e}}}[-R^e(w \circ \Phi)] - \mathbb{E}_{q_e}[-R^e(w \circ \Phi))])$$

We define the BIRM game as a finite strategic form game as follows.

$$\Gamma_{BIRM} := < N, S_{e \in N}, u_{e \in N} > \tag{10}$$

8.2 Nash Equilibrium of Γ_{BIRM}

To understand the pure NE of the proposed game, we establish the following proposition. This proposition is equivalent to 2 of IRM games.

Proposition 1. *The set of pure NE of Γ_{BIRM} is $\{p(w|\mathcal{D}_e, \Phi^*), \Phi^*\}$ and the set of feature extractors $\{\Phi^*\} \subset \mathcal{H}_\Phi$ is the solution set of BIRM equation 9.*

Proof. Suppose, (q^*, Φ^*) is the Nash equilibrium of Γ_{BIRM}. Using Best Response Dynamics, for any player $e \in \mathcal{E}_{tr} \cup \{\bar{e}\}$,

$$\begin{aligned} q_e^* &= \underset{q_e \in \mathcal{Q}_w}{\texttt{arg max}} \; u_e[q_e, q_{-e}^*, \Phi^*] \\ &= \underset{q_e \in \mathcal{Q}_w}{\texttt{arg max}} \; \mathbb{E}_{q_e(w)}[-R^e(w \circ \Phi^*)] - D_{KL}(q_e(w)||p_0(w)) \\ &= p(w|\mathcal{D}_e, \Phi^*) \end{aligned} \tag{11}$$

In variational inference [24], $\mathbb{E}_{q_e(w)}[-R^e(w \circ \Phi^*)] - D_{KL}(q_e(w)||p_0(w))$ is called $ELBO$ and it is shown that $ELBO \leq p(w|\mathcal{D}, \Phi^*)$. For player F, using Best Response Dynamics,

$$\begin{aligned}
\Phi^* &= \underset{\Phi \in \mathcal{H}_\Phi}{\texttt{arg max}}\ u_F[q_e^*, q_{-e}^*, \Phi] \\
\Phi^* &= \underset{\Phi \in \mathcal{H}_\Phi}{\texttt{arg max}}\ \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{q_{\bar{e}}^*}[-R^e(w \circ \Phi)] + \lambda(\mathbb{E}_{q_{\bar{e}}^*}[-R^e(w \circ \Phi)] - \mathbb{E}_{q_e^*}[-R^e(w \circ \Phi))]) \\
&= \underset{\Phi \in \mathcal{H}_\Phi}{\texttt{arg min}}\ \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{p(w|\mathcal{D},\Phi^*)}[R^e(w \circ \Phi)] + \lambda(\mathbb{E}_{p(w|\mathcal{D},\Phi^*)}[R^e(w \circ \Phi)] - \mathbb{E}_{p(w|\mathcal{D}_e,\Phi^*)}[R^e(w \circ \Phi)])
\end{aligned} \tag{12}$$

Comparing equation 12 with BIRM equation 9, we conclude the proof. □

The proposition 1 implies that playing BIRM games is the same as performing BIRM.

8.3 Theoretical Analysis

There is no formal analysis in the literature addressing whether bayesian invariant risk minimization can/cannot achieve domain invariance. We address this question using tools from statistical learning theory in this section.

Let $\bar{S}_{BIRM} = \{(q^*, \Phi^*)|(q^*, \Phi^*) \in$ `pure NE of` $\Gamma_{BIRM}\}$.

Proposition 2. *There can exist $(q^*, \Phi^*) \in \bar{S}_{BIRM}$ such that optimal strategies of all the training environments are not the same, i.e.,*

$$\exists e_1, e_2 \in \mathcal{E}_{tr} \cup \{\bar{e}\}; q_{e_1}^*(w) = p(w|\mathcal{D}_{e_1}, \Phi^*) \neq q_{e_2}^*(w) = p(w|\mathcal{D}_{e_2}, \Phi^*)$$

Proof. The risk R^e is the negative log-likelihood of dataset $\mathcal{D}_e$

$$R^e(w \circ \Phi) = \mathbb{E}_{(x,y)\sim\mathcal{D}_e}[-log(y|w \circ \Phi(x))] \tag{13}$$

For any $(q^*, \Phi^*) \in \bar{S}_{BIRM}$, since $R^e(w \circ \Phi) \geq 0$, the following holds

$$\mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)] \geq 0 \tag{14}$$

For player $\bar{e}$,

$$\begin{aligned}
\mathbb{E}_{q_{\bar{e}}^*(w)}[R^{\bar{e}}(w \circ \Phi^*)] &= \mathbb{E}_{q_{\bar{e}}^*(w)}[\Sigma_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi^*)] \\
&= \mathbb{E}_{q_{\bar{e}}^*(w)}[R^{e'}(w \circ \Phi^*)] + \mathbb{E}_{q_{\bar{e}}^*(w)}[\Sigma_{e \in \mathcal{E}_{tr} \setminus \{e'\}} R^e(w \circ \Phi^*)] \\
&\geq (\underset{q_e \in \mathcal{Q}_w}{\texttt{min}} \mathbb{E}_{q_e(w)}[R^{e'}(w \circ \Phi^*)]) + \mathbb{E}_{q_{\bar{e}}^*(w)}[\Sigma_{e \in \mathcal{E}_{tr} \setminus \{e'\}} R^e(w \circ \Phi^*)] \\
&\geq \mathbb{E}_{q_{e'}^*(w)}[R^{e'}(w \circ \Phi^*)] + \mathbb{E}_{q_{\bar{e}}^*(w)}[\Sigma_{e \in \mathcal{E}_{tr} \setminus \{e'\}} R^e(w \circ \Phi^*)] \\
&\geq \dots \\
&\geq \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)]
\end{aligned}$$

Therefore,

$$\mathbb{E}_{q_{\bar{e}}^*(w)}[R^{\bar{e}}(w \circ \Phi^*)] - \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)] \geq 0 \tag{15}$$

$$\mathbb{E}_{q_{\bar{e}}^*(w)}[R^e(w \circ \Phi^*)] - \mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)] \geq 0 \tag{16}$$

Proposition 1 implies that the pure NE of Γ_{BIRM} minimizes equation 9. Rewriting the equation 9 in appropriate form,

$$\begin{aligned}
&\Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{q_{\bar{e}}^*(w)}[R^e(w \circ \Phi^*)] + \lambda(\mathbb{E}_{q_{\bar{e}}^*(w)}[R^e(w \circ \Phi^*)] - \mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)]) \\
=&\Sigma_{e \in \mathcal{E}_{tr}} (1 + \lambda)\mathbb{E}_{q_{\bar{e}}^*(w)}[R^e(w \circ \Phi^*)] - \mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)]) \\
=&(1 + \lambda)\mathbb{E}_{q_{\bar{e}}^*(w)}[\Sigma_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi^*)] - \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)]) \\
=&(1 + \lambda)\mathbb{E}_{q_{\bar{e}}^*(w)}[R^{\bar{e}}(w \circ \Phi^*)] - \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{q_e^*(w)}[R^e(w \circ \Phi^*)] \\
=&(1 + \lambda)\mathbb{E}_{p(w|\mathcal{D}_{\bar{e}},\Phi^*)}[R^{\bar{e}}(w \circ \Phi^*)] - \Sigma_{e \in \mathcal{E}_{tr}} \mathbb{E}_{p(w|\mathcal{D}_e,\Phi^*)}[R^e(w \circ \Phi^*)] \\
\geq&0 \textit{ (due to 15)}
\end{aligned}$$

Since we set $\lambda > 0$, due to inequality 16, when the above is minimized, the following holds

$$\mathbb{E}_{p(w|\mathcal{D}_{\bar{e}},\Phi^*)}[R^e(w \circ \Phi^*)] = \mathbb{E}_{p(w|\mathcal{D}_e,\Phi^*)}[R^e(w \circ \Phi^*)]; \forall e \in \mathcal{E}_{tr} \tag{17}$$

The result 17 is the sufficient condition at pure NE to hold between any training environment $e \in \mathcal{E}_{tr}$ and $\bar{e}$. It means that the following can hold.

$$\mathbb{E}_{p(w|\mathcal{D}_{e_i},\Phi^*)}[R^{e_i}(w \circ \Phi^*)] \neq \mathbb{E}_{p(w|\mathcal{D}_{e_j},\Phi^*)}[R^{e_j}(w \circ \Phi^*)]; \;\; \texttt{for} \;\; e_i \neq e_j$$

Hence, it is possible that $(q_e^*, \Phi^*) \in \bar{S}_{BIRM}$ where $q_{e_i}^* \neq q_{e_j}^*$ for some training environments e_i, e_j.

□

Proposition 3. *If the data representation Φ^* is truly invariant, then*

$$p(w|\mathcal{D}_{\bar{e}}, \Phi^*) = p(w|\mathcal{D}_e, \Phi^*); \forall e \in \mathcal{E}_{tr}$$

Proof. Given Φ^* is truly invariant, the distribution $p(\Phi^*(x), y)$ is the same across $\mathcal{E}_{tr} \cup \{\bar{e}\}$. Therefore, for any environment $e_i \in \mathcal{E}_{tr}$, for any $w \sim q(w)$,

$$R^{e_i}(w \circ \Phi^*) = R^{\bar{e}}(w \circ \Phi^*)$$

From 11,

$$\begin{aligned} q_e^*(w) &= \underset{q_e \in \mathcal{Q}_w}{\texttt{arg max}} \;\; u_e[q_e, q_{-e}^*, \Phi^*] \\ &= \underset{q_e \in \mathcal{Q}_w}{\texttt{arg max}} \;\; \mathbb{E}_{q_e(w)}[-R^e(w \circ \Phi^*)] - D_{KL}(q_e(w)||p_0(w)) \\ &= \underset{q_e \in \mathcal{Q}_w}{\texttt{arg max}} \;\; \mathbb{E}_{q_e(w)}[-R^{\bar{e}}(w \circ \Phi^*)] - D_{KL}(q_e(w)||p_0(w)) \\ &= p(w|\mathcal{D}_{\bar{e}}, \Phi^*) \end{aligned} \tag{18}$$

□

Proposition 4. *If proposition 2 holds, Γ_{BIRM} may not achieve domain invariance.*

The proposition 2 describes the nature of $\bar{S}_{BIRM}$. The proposition 3 describes the condition on pure NE of Γ_{BIRM}, which is necessary to hold when Φ^* is truly domain invariant. We conclude in proposition 2 that there can exist pure NE such that the necessary conditions, as mentioned in 3, are not satisfied. Hence, we formally showed that BIRM games may not achieve true domain invariance for settings mentioned in proposition 2. This also holds for BIRM as an implication of the proposition 1. This is the case in many practical scenarios.

8.4 Algorithm

In experiments, we assume that $\mathcal{Q}_w = \mathcal{N}(\mu, diag(\sigma))$ and $p_0(w) = \mathcal{N}(0, I)$. The BIRM games algorithm is described in 8.4

8.5 Experiment using synthetic dataset

The main aim of this experiment is to demonstrate the situation in which IRM games degenerate to empirical risk minimization (ERM), but BIRM games perform better than both. We generate the dataset by tracing the following SEM model.

$$\begin{aligned} X_1 &\leftarrow \mathcal{N}(0, \sigma^2 I) \\ Y &\leftarrow \mathbf{1}^T X_1 + \mathcal{N}(0, \sigma^2 I) \\ X_2 &\leftarrow Y \cdot \mathbf{1} + \mathcal{N}(0, (\rho^e \sigma)^2 I) \end{aligned}$$

Here, $X_1, X_2 \in \mathbb{R}^2$ and they form feature vector $z = [x_1, x_2]^T \in \mathbb{R}^4$. Since X_1 is the causal feature of Y, the optimal invariant predictor is $w^* = [1, 1, 0, 0]$. ρ^e is environment dependent variable and is varied across different environments. The larger the value of ρ^e is, the weaker the correlation between Y and X_2. We use two training environments with ρ^e set to 1.4 and 1.5 and

BIRM Games Best Response Training

```
Input: $\lambda$, Data for each environment and combined data
Initialize: Randomly initialize $q_e$ from $\mathcal{Q}_w$ for $e \in \mathcal{E}_{tr} \cup \{\bar{e}\}$ and $\Phi$ from $\mathcal{H}_\Phi$
for epochs in $\{1, \dots, \texttt{max\_epochs}\}$
    for e in $\mathcal{E}_{tr} \cup \{\bar{e}\}$
        for k in $\{1, \dots, K\}$
            $q_e := \texttt{SGD}[\mathbb{E}_{q_e(w)}[-R^e(w \circ \Phi^*)] - D_{KL}(q_e(w)||p_0(w))]$
        end for
    end for
    $\texttt{loss} := 0$
    for k in $\{1, \dots, K\}$
        for e in $\mathcal{E}_{tr} \cup \{\bar{e}\}$
            $\texttt{loss} := \texttt{loss} + (1+\lambda)\mathbb{E}_{q_{\bar{e}}(w)}[-R^e(w \circ \Phi^*)] - \mathbb{E}_{q_e(w)}[-R^e(w \circ \Phi^*)]$
        end for
    end for
    $\Phi := \texttt{SGD}[\texttt{loss}]$
end for
```

Samples	100		250		500		1000	
Method	**train**	**test**	**train**	**test**	**train**	**test**	**train**	**test**
ERM	0.534	13.426	0.477	11.784	0.525	10.476	0.489	13.239
EIRM games	0.846	14.616	1.058	9.4	1.0496	11.613	1.013	12.327
BIRM games	0.676	**8.077**	0.584	**6.85**	0.717	**6.336**	0.675	**5.98**

Table 1: Experiments comparing ERM, EIRM games, and BIRM games on the synthetic dataset with different numbers of samples in each training environment.

one test environment with ρ^e set to 9.9. The closer the value of ρ^e is, the more EIRM games will be incentivized to adopt spurious correlations generated by X_2.

For data representation Φ, we set hypothesis class to $\Phi \in \mathbb{R}^{4\times 2}$. Hence, $\Phi(z) \in \mathbb{R}^2$. The invariant data representation in our synthetic dataset is $\Phi^*(Z) = X_1$. We set hypothesis class $\mathcal{H}_w$ to $\mathbb{R}^2$ and distribution class $\mathcal{Q}_w$ to $\mathcal{N}(\mu, \texttt{diag}(\sigma))$ where $\mu, \sigma \in \mathcal{R}^2$. We evaluate the performance in terms of MSE loss. The results are summarised in the table 1. The relevant code is also submitted.

8.6 Conclusion

In this project, we propose bayesian invariant risk minimization games as a game theoretic formulation of bayesian invariant risk minimization [7]. BIRM games improve upon IRM games by addressing the problems caused when IRM games overfit the training environments. We analyzed the pure Nash equilibrium of BIRM games and established formally that BIRM games fail to achieve domain invariance when certain conditions are not fulfilled. We generate a synthetic dataset and demonstrate the superiority of BIRM games over IRM games when IRM games fail due to overfitting.

References

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

[2] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[3] Sara Beery, Grant van Horn, and Pietro Perona. Recognition in terra incognita, 2018.

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

[5] Drew Fudenberg, Fudenberg Drew, David K Levine, and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[6] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[7] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16021–16030, June 2022.

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

[9] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

[10] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[11] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images, 2017.

[12] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, 2015.

[13] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models, 2018.

[14] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[15] Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

[16] A Subbaswamy, B Chen, and S Saria. Should i include this edge in my prediction? analyzing the stability-performance tradeoff. *arXiv preprint arXiv:1905.11374*, 2019.

[17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Meta-learning the invariant representation for domain generalization. In *Proceedings of the 38th International Conference on Machine Learning*, 2018.

[18] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[19] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

[20] Sharut Gupta, Kartik Ahuja, Mohammad Havaei, Niladri Chatterjee, and Yoshua Bengio. Fl games: A federated learning framework for distribution shifts. *arXiv preprint arXiv:2205.11101*, 2022.

[21] Kartik Ahuja, Karthikeyan Shanmugam, and Amit Dhurandhar. Linear regression games: Convergence guarantees to approximate out-of-distribution solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 1270–1278. PMLR, 2021.

[22] Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2007.

[23] Guillaume Lecue. Suboptimality of penalized empirical risk minimization in classification. In *International Conference on Computational Learning Theory*, pages 142–156. Springer, 2007.

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.