On the Equivalence of Bi-Level Optimization and Game-Theoretic Formulations of Invariant Risk Minimization

Kartik Ahuja [1] **Karthikeyan Shanmugam** [1] **Kush R. Varshney** [1] **Amit Dhurandhar** [1]

Abstract

The standard risk minimization paradigm of machine learning is brittle when operating in environments whose test distributions are different from the training distribution due to spurious correlations. Training on data from many environments and finding *invariant* predictors reduces the effect of spurious features by concentrating models on features that have a causal relationship with the outcome. Such invariant risk minimization was posed as a bi-level optimization problem by Arjovsky et al. (2019). In this work, we pose invariant risk minimization as finding the Nash equilibrium of an ensemble game among several environments and show that the set of Nash equilibria for the proposed game are equivalent to the set of invariant predictors obtained by the bi-level optimization even with nonlinear classifiers and transformations.

1. Introduction

Machine learning is rife with embarrassing examples of spurious correlations that fail to hold outside a specific training (and identically distributed test) distribution. For example, Beery et al. (2018) trained a convolutional neural network (CNN) to classify camels from cows. Most of the pictures of cows had green pastures, while most pictures of camels were in deserts. The CNN picked up the spurious correlation: it associated green pastures with cows and failed to classify pictures of cows on sandy beaches correctly.

To address the problem of models inheriting spurious correlations, Arjovsky et al. (2019) show that one can exploit the varying degrees of spurious correlation naturally present in data collected from multiple data sources to learn robust predictors. The authors propose to find a representation Φ such that the optimal classifier given Φ is invariant across training environments. This formulation leads to a challenging bi-level optimization, which the authors relax by fixing a simple linear classifier and learning a representation Φ such that the classifier is "approximately locally optimal" in all the training environments.

In this work, we take a different approach. We create an *ensemble* of classifiers with each environment controlling one component of the ensemble. Each environment uses the entire ensemble to make predictions. We let all the environments play a *game* where each environment's action is to decide its contribution to the ensemble such that it minimizes its risk. Remarkably, we establish that the set of predictors that solve the *ensemble game* is equal to the set of invariant predictors across the training environments; this result holds for a large class of non-linear classifiers. We propose best response dynamics, which has a simple implementation, to solve the game. We do not restrict classifiers to be linear, which was emphasized as a direction for future work by Arjovsky et al. (2019). Broadly speaking, we believe that the game-theoretic perspective herein can open up a new paradigm to address the problem of invariance.

The invariant risk minimization (IRM) formulation of Arjovsky et al. (2019) is the most related work, and is motivated from the theory of causality and causal Bayesian networks (CBNs) (Pearl, 1995). A variable y is caused by a set of non-spurious actual causal factors $x_{\mathrm{Pa}(y)}$ if and only if in all environments where y has not been intervened on, the conditional probability $P(y|x_{\mathrm{Pa}(y)})$ remains invariant. This is called the *modularity condition* (Bareinboim et al., 2012). Related and similar notions are the *independent causal mechanism principle* (Schölkopf et al., 2012; Janzing & Schölkopf, 2010; Janzing et al., 2012) and the *invariant causal prediction principle* (Peters et al., 2016; Heinze-Deml et al., 2018). These principles imply that if all the environments (train and test) are modeled by interventions that do not affect the causal mechanism of target variable y, then a classifier conservatively trained on the transformation that involves the causal factors ($\Phi(x) = x_{\mathrm{Pa}(y)}$) to predict y is robust to unseen interventions.

Statistical machine learning has dealt with the distribution shift between the training distribution and test distribution in a number of ways. Sample weighting attempts to match test and train distributions by reweighting samples. It typically

[1]IBM Research, TJ Watson Research Center, Yorktown Heights, NY. Correspondence to: Kartik Ahuja <kartik.ahuja@ibm.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

assumes that the probability of labels given all covariates does not shift. Domain adaptation tries to find a representation Φ whose distribution is invariant across source and target domains, and is known to have serious limitations even when the marginal distribution of labels shift across environments. When only training data sources are given, robust optimization techniques find the worst case loss over all possible convex combinations of the training sources, assuming that the test distribution is within the convex hull of training distributions, which is not true in many settings.

2. Preliminaries

2.1. Game Theory Concepts

Let $\Gamma = (N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N})$ be the tuple representing a standard normal form game, where N is the finite set of players. Player $i \in N$ takes actions from a strategy set S_i. The utility of player i is $u_i : S \to \mathbb{R}$, where we write the joint set $S = \Pi_{i\in N} S_i$. The joint strategy of all the players is given as $s \in S$, the strategy of player i is s_i and the strategy of the rest of players is $s_{-i} = (s_{i'})_{i'\neq i}$. A strategy s^* is said to be a pure strategy Nash equilibrium (NE) if it satisfies $u_i(s_i^*, s_{-i}^*) \geq u_i(k, s_{-i}^*), \forall k \in S_i, \forall i \in N$.

2.2. Invariant Risk Minimization

Consider datasets $\{(x_i^e, y_i^e)\}_{i=1}^{n_e}$ from multiple training environments $e \in \mathcal{E}_{tr}$. The feature value $x_i^e \in \mathcal{X}$ and the corresponding labels $y_i^e \in \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^k$. Define a predictor $f : \mathcal{X} \to \mathbb{R}^k$. IRM uses these multiple datasets to construct a predictor f that performs well across many unseen environments $\mathcal{E}_{all}$. Define the risk achieved by f in environment e as $R^e(f) = \mathbb{E}_{X^e,Y^e}\left[\ell(f(X^e), Y^e)\right]$, where ℓ is the loss when $f(X)$ is the predicted value and Y is the corresponding label.

Invariant predictor: We say that a data representation $\Phi : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^d$ elicits an invariant predictor $w \circ \Phi$ across environments $e \in \mathcal{E}$ if there is a classifier $w : \mathcal{Z} \to \mathbb{R}^k$ that achieves the minimum risk for all the environments $w \in \arg\min_{\bar{w}\in\mathcal{H}_w} R^e(\bar{w} \circ \Phi)$. The set of all the mappings Φ is given as $\mathcal{H}_\Phi$ and the set of all the classifiers is given as $\mathcal{H}_w$. IRM may be phrased as the following constrained optimization problem (Arjovsky et al., 2019):

$$\begin{aligned} &\min_{\Phi\in\mathcal{H}_\Phi, w\in\mathcal{H}_w} \sum_{e\in\mathcal{E}_{tr}} R^e(w \circ \Phi) \\ &\text{s.t. } w \in \arg\min_{\bar{w}\in\mathcal{H}_w} R^e(\bar{w} \circ \Phi),\ \forall e \in \mathcal{E}_{tr}. \end{aligned} \tag{1}$$

If (Φ, w) satisfies the above constraints, then $w \circ \Phi$ is an invariant predictor across the environments $\mathcal{E}_{tr}$. Define the set of representations and the corresponding classifiers, (Φ, w) that satisfy the constraints in the above problem (1) as $\mathcal{S}^{\mathsf{IV}}$, where IV stands for invariant. Also, define the set of invariant predictors $w \circ \Phi$ as $\hat{\mathcal{S}}^{\mathsf{IV}} = \{w \circ \Phi \mid (\Phi, w) \in \mathcal{S}^{\mathsf{IV}}\}$.

Remark. The sets $\mathcal{S}^{\mathsf{IV}}$, $\hat{\mathcal{S}}^{\mathsf{IV}}$ depend on the choice of classifier class $\mathcal{H}_w$ and representation class $\mathcal{H}_\Phi$. We avoid making this dependence explicit until later sections.

Members of $\mathcal{S}^{\mathsf{IV}}$ are equivalently the solutions to:

$$R^e(w \circ \Phi) \leq R^e(\bar{w} \circ \Phi),\ \forall \bar{w} \in \mathcal{H}_w,\ \forall e \in \mathcal{E}_{tr}. \tag{2}$$

The main result of Arjovsky et al. (2019) states that if $\mathcal{H}_w$ and $\mathcal{H}_\Phi$ are from the class of linear models, i.e., $w(z) = \mathbf{w}^t z$, where $\mathbf{w} \in \mathbb{R}^d$, and $\Phi(x) = \mathbf{\Phi} x$ with $\mathbf{\Phi} \in \mathbb{R}^{d\times n}$, then under certain conditions on the data generation process and training environments $\mathcal{E}_{tr}$, the solution to (2) remains invariant in $\mathcal{E}_{all}$.

3. Ensemble IRM Games

3.1. Game-Theoretic Reformulation

Optimization problem (1) can be quite challenging to solve. We introduce an alternate characterization based on game theory to solve it. We endow each environment with its own classifier $w^e \in \mathcal{H}_w$. We use a simple ensemble to construct an overall classifier $w^{av} : \mathcal{Z} \to \mathbb{R}^k$ defined as $w^{av} = \frac{1}{|\mathcal{E}_{tr}|}\sum_{q=1}^{|\mathcal{E}_{tr}|} w^q$, where for each $z \in \mathcal{Z}$, $w^{av}(z) = \frac{1}{|\mathcal{E}_{tr}|}\sum_{q=1}^{|\mathcal{E}_{tr}|} w^q(z)$. (The av stands for average.) Consider the example of binary classification with two environments $\{e_1, e_2\}$; $w^e = [w_1^e, w_2^e]$ is the classifier of environment e, where each component is the score for each class. We define the component j of the ensemble classifier w^{av} as $w_j^{av} = \frac{w_j^{e_1} + w_j^{e_2}}{2}$. These scores are input to a softmax; the final probability assigned to class j for an input z is $\frac{e^{w_j^{av}(z)}}{e^{w_1^{av}(z)} + e^{w_2^{av}(z)}}$.

We require all the environments to use this ensemble w^{av}. We want to solve the following new optimization problem.

$$\min_{\Phi\in\mathcal{H}_\Phi, w^{av}\in\mathcal{H}_w} \sum_{e\in\mathcal{E}_{tr}} R^e(w^{av} \circ \Phi)$$

$$\text{s.t. } w^e \in \arg\min_{\bar{w}^e\in\mathcal{H}_w} R^e\left(\frac{1}{|\mathcal{E}_{tr}|}\left[\bar{w}^e + \sum_{q\neq e} w^q\right] \circ \Phi\right),\ \forall e \in \mathcal{E}_{tr}$$

We can equivalently restate the above as:

$$\begin{aligned} &\min_{\Phi\in\mathcal{H}_\Phi, w^{av}} \sum_{e\in\mathcal{E}_{tr}} R^e(w^{av} \circ \Phi) \\ &\text{s.t. } R^e\left(\frac{1}{|\mathcal{E}_{tr}|}\left[w^e + \sum_{q\neq e} w^q\right] \circ \Phi\right) \\ &\leq R^e\left(\frac{1}{|\mathcal{E}_{tr}|}\left[\bar{w}^e + \sum_{q\neq e} w^q\right] \circ \Phi\right)\ \forall \bar{w}^e \in \mathcal{H}_w\ \forall e \in \mathcal{E}_{tr} \end{aligned} \tag{3}$$

What are the advantages of this formulation (3)?

- Using the ensemble automatically enforces invariance across environments.
- Each environment is free to select the classifier w^e from the entire set $\mathcal{H}_w$, unlike in (1), where all environments' choices are required to be the same.
- The constraints in (3) are equivalent to the set of pure NE of a game that we define next.

The game is played between $|\mathcal{E}_{tr}|$ players, with each player corresponding to an environment e. The set of actions of the environment e are $w^e \in \mathcal{H}_w$. At the start of the game, a representation Φ is selected from the set $\mathcal{H}_\Phi$, which is observed by all the environments. The utility function for an environment e is defined as $u_e[w^e, w^{-e}, \Phi] = -R^e(w^{av}, \Phi)$, where $w^{-e} = \{w^q\}_{q \neq e}$ is the set of choices of all environments but e. We call this game Ensemble Invariant Risk Minimization (EIRM) and express it as a tuple

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

We represent a pure NE as a tuple $\left(\Phi, \{w^q\}_{q=1}^{|\mathcal{E}_{tr}|}\right)$. Since each pure NE depends on Φ, we include it as a part of the tuple. We define the set of pure NE as $\mathcal{S}^{\mathsf{EIRM}}$. Construct a set of the ensemble predictors constructed from NE as

$$\hat{\mathcal{S}}^{\mathsf{EIRM}} = \left\{ \left[\frac{1}{|\mathcal{E}_{tr}|} \sum_{q=1}^{|\mathcal{E}_{tr}|} w^q \right] \circ \Phi \mid (\Phi, \{w^q\}_{q=1}^{|\mathcal{E}_t|}) \in \mathcal{S}^{\mathsf{EIRM}} \right\}.$$

Members of $\mathcal{S}^{\mathsf{EIRM}}$ are equivalently the solutions to

$$u_e[w^e, w^{-e}, \Phi] \geq u_e[\bar{w}^e, w^{-e}, \Phi], \ \forall w^e \in \mathcal{H}_w, \forall e \in \mathcal{E}_{tr}. \tag{4}$$

If we replace $u_e[w^e, w^{-e}, \Phi]$ with $-R^e(w^{av}, \Phi)$, we obtain the inequalities in (3). So far we have defined the game and given its relationship to the problem in (3).

3.2. Equivalence Between NE and Invariant Predictors

What is the relationship between the predictors obtained from NE $\hat{\mathcal{S}}^{\mathsf{EIRM}}$ and invariant predictors $\hat{\mathcal{S}}^{\mathsf{IV}}$?

Remarkably, these two sets are the same under very mild conditions. Before we show this result, we establish a stronger result and this result will follow from it.

We use the set $\mathcal{S}^{\mathsf{EIRM}}$ to construct a new set. To each tuple $\left(\Phi, \{w^q\}_{q=1}^{|\mathcal{E}_{tr}|}\right) \in \mathcal{S}^{\mathsf{EIRM}}$ augment the ensemble classifier $w^{av} = \frac{1}{|\mathcal{E}_{tr}|} \sum_{q=1}^{|\mathcal{E}_{tr}|} w^q$ to get $\left(\Phi, \{w^q\}_{q=1}^{|\mathcal{E}_{tr}|}, w^{av}\right)$. We call the set of these new tuples $\tilde{\mathcal{S}}^{\mathsf{EIRM}}$.

We use the set $\mathcal{S}^{\mathsf{IV}}$ to construct a new set. Consider an element $(\Phi, w) \in \mathcal{S}^{\mathsf{IV}}$. We define a decomposition for w in terms of the environment-specific classifiers as follows: $w = \frac{1}{|\mathcal{E}_{tr}|} \sum_{q=1}^{|\mathcal{E}_{tr}|} w^q$, where $w^q \in \mathcal{H}_w$. $w^q = w, \forall q \in \mathcal{E}_{tr}$ is one trivial decomposition. We use each such decomposition and augment the tuple to obtain $\left(\Phi, \{w^q\}_{q=1}^{|\mathcal{E}_{tr}|}, w\right)$. We call this set of new tuples $\tilde{\mathcal{S}}^{\mathsf{IV}}$.

Both the sets $\tilde{\mathcal{S}}^{\mathsf{IV}}$ and $\tilde{\mathcal{S}}^{\mathsf{EIRM}}$ consist of tuples of representation, set of environment specific classifiers, and the ensemble classifier. We ask an even more interesting question than the one above. Is the set of representations, environment specific classifiers, and the ensembles found by playing EIRM (4) or solving IRM (2) the same? If these two sets are equal, then equality between $\hat{\mathcal{S}}^{\mathsf{EIRM}}$ and $\hat{\mathcal{S}}^{\mathsf{IV}}$ follows trivially.

We state the only assumption we need.

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 $z \in \mathcal{Z}$, $(w_1 + w_2)(z) = w_1(z) + w_2(z)$*
- *Scalar multiplication: For any $c \in \mathbb{R}$ and $w \in \mathcal{H}_w$, $cw \in \mathcal{H}_w$, where for every $z \in \mathcal{Z}$, $(cw)(z) = c \times w(z)$*

The addition of the functions and scalar multiplication are defined in a standard pointwise manner. Therefore, the class $\mathcal{H}_w$ also forms a vector space. We now state the main result.

Theorem 1. *If Assumption 1 holds, then $\tilde{\mathcal{S}}^{\mathsf{IV}} = \tilde{\mathcal{S}}^{\mathsf{EIRM}}$*

The proofs of all results are in the full paper (Ahuja et al., 2020).

Corollary 1. *If Assumption 1 holds, then $\hat{\mathcal{S}}^{\mathsf{IV}} = \hat{\mathcal{S}}^{\mathsf{EIRM}}$*

Significance of Theorem 1 and Corollary 1

- From a computational standpoint, this equivalence permits tools from game theory to find NE of the EIRM game and, as a result, the invariant predictors.
- From a theoretical standpoint, this equivalence permits to use game theory to analyze the solutions of the EIRM game and understand the invariant predictors.
- In Theorem 9 of Arjovsky et al. (2019), it was shown for linear classifiers and linear representations that the invariant predictors generalize to a large set of unseen environments under certain conditions. Since our result holds for linear classifiers (but is even broader), the generalization result continues to hold for the predictors found by playing the EIRM game.

Role of representation Φ. We investigate the scenario when we fix Φ to the identity mapping; this will motivate one of our approaches. Define the set $\hat{\mathcal{S}}^{\mathsf{EIRM}}(\Phi)$ as the set of ensemble predictors arrived at by playing the EIRM game using a fixed representation representation Φ.[1] Similarly,

[1] $\cup_\Phi \hat{\mathcal{S}}^{\mathsf{EIRM}}(\Phi) = \hat{\mathcal{S}}^{\mathsf{EIRM}}$

we define a set $\hat{\mathcal{S}}^{\mathsf{IV}}(\Phi)$ as the set of invariant predictors derived using the representation Φ. From Theorem 1, it follows that $\hat{\mathcal{S}}^{\mathsf{EIRM}}(\Phi) = \hat{\mathcal{S}}^{\mathsf{IV}}(\Phi)$. We modify some of the earlier notations for results to follow. The set of predictors that result from the EIRM game $\hat{\mathcal{S}}^{\mathsf{EIRM}}$ and the sets of invariant predictors $\hat{\mathcal{S}}^{\mathsf{IV}}$ are defined for a family of maps Φ with co-domain $\mathcal{Z}$. We make the co-domain $\mathcal{Z}$ explicit in the notation. We write $\hat{\mathcal{S}}^{\mathsf{EIRM}}_{\mathcal{Z}}$ for $\hat{\mathcal{S}}^{\mathsf{EIRM}}$ and $\hat{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{Z}}$ for $\hat{\mathcal{S}}^{\mathsf{IV}}$.

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

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

$L^p(\mathcal{Z})$: set of functions $f : \mathcal{Z} \to \mathbb{R}$ s.t. $\int_{\mathcal{Z}} |f|^p d\mu < \infty$

Assumption 3. $\mathcal{H}_w = L^p(\mathcal{Z})$.

Define a subset $\bar{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{Z}} \subseteq \hat{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{Z}}$ consisting of invariant predictors that are in $L^p(\mathcal{X})$, i.e., $\bar{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{Z}} = \{u \mid u \in \hat{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{Z}} \text{ and } u \in L^p(\mathcal{X})\}$. Let $\Phi = \mathsf{I}$, where $\mathsf{I} : \mathcal{X} \to \mathcal{X}$ is the identity mapping. Following the above notation, the set of invariant predictors and the set of ensemble predictors obtained from NE are $\hat{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{X}}(\mathsf{I})$ and $\hat{\mathcal{S}}^{\mathsf{EIRM}}_{\mathcal{X}}(\mathsf{I})$ respectively.

Theorem 2. *If Assumptions 2 and 3 are satisfied and $\bar{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{Z}}$ is non-empty, then $\bar{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{Z}} = \hat{\mathcal{S}}^{\mathsf{IV}}_{\mathcal{X}}(\mathsf{I}) = \hat{\mathcal{S}}^{\mathsf{EIRM}}_{\mathcal{X}}(\mathsf{I})$*

Significance of Theorem 2. If we fix the representation to identity and play the EIRM game, then it is sufficient to recover all the invariant predictors (with bounded L^p norm) that can be obtained using all the representations $\Phi \in \mathcal{H}_\Phi$. Therefore, we can simply fix $\Phi = \mathsf{I}$ and use game-theoretic algorithms for learning equilibria.

3.3. Existence of NE of Γ^{EIRM} and Invariant Predictors

In this section, we first argue that there are many settings when both invariant predictors and the NE exist.

Illustration through generative models. We use a simplified version of the model described by Peters et al. (2016). In each environment e, the random variable $X^e = [X^e_1, ..., X^e_n]$ corresponds to the feature vector and Y^e corresponds to the label. The data for each environment is generated by i.i.d. sampling (X^e, Y^e) from the following generative model. Assume a subset $S^* \subset \{1, ..., n\}$ is causal for the label Y^e. For all the environments e, X^e has an arbitrary distribution and $Y^e = g(X^e_{S^*}) + \epsilon^e$, where $X^e_{S^*}$ is the vector X^e with indices in S^*, $g : \mathbb{R}^{|S^*|} \to \mathbb{R}$ is some underlying function and $\epsilon^e \sim F^e$, $\mathbb{E}[\epsilon_e] = 0$, $\epsilon^e \perp X^e_{S^*}$. Let ℓ be the squared error loss function. We fix the representation $\Phi^*(X^e) = X^e_{|S^*|}$. With Φ^* as the representation, the optimal classifier w among all the functions is $g(X^e_{S^*})$ (this follows from the generative model). If we assume that $g \in \mathcal{H}_w$, then for each environment e, $w^e_* = g$ is the optimal classifier in $\mathcal{H}_w$. Therefore, $w^e_* \circ \Phi^* = g$ is the invariant predictor. If $\mathcal{H}_w$ satisfies affine closure, then any decomposition of g is a pure NE of the EIRM game. We have illustrated existence of NE and invariant predictor when the data is generated as above and when the class $\mathcal{H}_w$ is sufficiently expressive to capture g. Next, we discuss the case when we do not know anything about the underlying data generation process.

Assumption 4.

- *$\mathcal{H}_w$ is a class of linear models, where $w : \mathcal{Z} \to \mathbb{R}$ and $w(z) = \boldsymbol{w}^t z$, where $z \in \mathcal{Z}$. We write $\mathcal{H}_w$ as the set of vectors $\boldsymbol{w}$. $\mathcal{H}_w$ is a closed, bounded and convex. The interior of $\mathcal{H}_w$ is non-empty.*
- *The loss function $\ell(\boldsymbol{w}^t z, Y)$, where $Y \in \mathbb{R}$ is the label, is convex and continuous in $\boldsymbol{w}$. For e.g., if loss is cross-entropy for binary classification or loss is mean squared error for regression, then this assumption is automatically satisfied.*

Theorem 3. *If Assumption 4 is satisfied, then a pure strategy Nash equilibrium of the game Γ^{EIRM} exists. If the weights of all the individual is in the NE are in the interior of $\mathcal{H}_w$, then the corresponding ensemble predictor is an invariant predictor among all the linear models.*

The family $\mathcal{H}_w$ of bounded linear functions does not satisfy affine closure, which is why existence of NE does not immediately imply the existence of invariant predictor (from Theorem 1). However, if the solution is in the interior of $\mathcal{H}_w$, then it is the globally optimal solution among all the linear functions, which in fact actually satisfy affine closure. As a result, in this case the invariant predictor also exists.

Significance of Theorem 3 Our approach is based on finding the NE. Therefore, it is important to understand when the solutions are guaranteed to exist. In the above theorem, we proved the result for linear models only, but there were no assumptions made on the representation class. In the supplement, we show that for a large class of models, pure NE may not exist but mixed NE (a relaxation of pure NE) are guaranteed to exist. Following the sufficient condition for existence of invariant predictors, understanding what conditions cause the NEs to be in the interior or on the boundary of $\mathcal{H}_w$ can help further the theory of invariant prediction.

4. Conclusion

We developed a new framework based on game-theoretic tools to learn invariant predictors. We work with data from multiple environments. In our framework, we set up an ensemble game; we construct an ensemble of classifiers with each environment controlling one portion of the ensemble. Remarkably, the set of solutions to this game is exactly the same as the set of invariant predictors across training environments. We hope this framework opens new ways to address other problems pertaining to invariance in causal inference using tools from game theory.

References

Ahuja, K., Shanmugam, K., Varshney, K. R., and Dhurandhar, A. Invariant risk minimization games. *to appear in International Conference on Machine Learning*, 2020.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bareinboim, E., Brito, C., and Pearl, J. Local characterizations of causal Bayesian networks. In *Graph Structures for Knowledge Representation and Reasoning*, pp. 1–17. Springer, 2012.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

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

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.