Linear Regression Games: Convergence Guarantees to Approximate Out-Of-Distribution Solutions

Kartik Ahuja
ahujak@ucla.edu

Karthikeyan Shanmugam
karthikeyan.shanmugam@ibm.com

Amit Dhurandhar
adhuran@us.ibm.com

IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY

Abstract

Recently, invariant risk minimization (IRM) [Arjovsky et al., 2019] was proposed as a promising solution to address out-of-distribution (OOD) generalization. In [Ahuja et al., 2020], it was shown that solving for the Nash equilibria of a new class of "ensemble-games" is equivalent to solving IRM. In this work, we extend the framework in [Ahuja et al., 2020] for linear regressions by projecting the ensemble-game on an ℓ_∞ ball. We show that such projections help achieve non-trivial OOD guarantees despite not achieving perfect invariance. For linear models with confounders, we prove that Nash equilibria of these games are closer to the ideal OOD solutions than the standard empirical risk minimization (ERM) and we also provide learning algorithms that provably converge to these Nash Equilibria. Empirical comparisons of the proposed approach with the state-of-the-art show consistent gains in achieving OOD solutions in several settings involving anti-causal variables and confounders.

1 Introduction

Recent years have witnessed a surge in examples highlighting vulnerabilities of machine learning models [Geirhos et al., 2020]. In an alarming study [DeGrave et al., 2020], it was shown how models trained to detect COVID-19 from chest radiographs used spurious factors such as the source of the data rather than the lung pathology [DeGrave et al., 2020]. In another commonly cited example [Beery et al., 2018] trained a convolutional neural network (CNN) to classify camels from cows and found the model to rely on the background color (green pastures for cows and desert for camels) to carry out classification.

Recently, [Arjovsky et al., 2019] proposed a framework called invariant risk minimization (IRM) to address the problem of models inheriting spurious correlations. They showed that when data is gathered from multiple environments, one can learn to exploit invariant causal relationships, rather than relying on varying spurious relationships, thus learning robust predictors. The authors used the invariance principle based on causality [Pearl, 1995] to construct powerful objects called "invariant predictors". An invariant predictor loosely speaking is a predictor that is simultaneously optimal across all the training environments under a shared representation. In [Arjovsky et al., 2019], it was shown that for linear models with confounders and/or anti-causal variables, learning ideal invariant predictors translates to learning solutions with ideal out-of-distribution (OOD) generalization behavior. However, building efficient algorithms guaranteed to learn these invariant predictors is still a challenge.

The algorithm in [Arjovsky et al., 2019] is based on minimizing a risk function comprising of the standard risk and a penalty term that tries to approximately ensure that predictors learned are invariant. The penalty is non-convex even for linear models and thus the algorithm is not guaranteed to arrive at invariant predictors. Another recent work [Ahuja et al., 2020], proposed a framework called invariant risk minimization games (IRM-games) and showed that solving for the Nash equilibria (NE) of a special class of "ensemble-games" is equivalent to solving IRM for many settings. The algorithm in [Ahuja et al., 2020] has no convergence guarantees to the NE of the ensemble-game. To summarize, building algorithms that are guaranteed to converge to predictors with non-trivial OOD generalization is unsolved even for linear models with confounders and/or anti-causal variables.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

In this work, we take important steps towards this highly sought after goal. As such, we formulate an ensemble-game that is constrained to be in the ℓ_∞ ball. Although this construction might seem to be surprising at first, we show that these constrained ensemble-game based predictors have a good OOD behavior even though that they may not be the exact invariant predictors. We provide efficient algorithms that are guaranteed to learn these predictors in many settings. To the best of our knowledge, our algorithms are the first for which we can guarantee both convergence and better OOD behavior than standard empirical risk minimization. We carry out empirical comparisons in the settings proposed in [Arjovsky et al., 2019], where the data is generated from models that include both causal and anti-causal variables as well as confounders in some cases. These comparisons of our approach with the state-of-the-art depict its promise in achieving OOD solutions in these setups. This demonstrates that searching over the NE of constrained ensemble-games is a principled alternative to searching over invariant predictors as is done in IRM.

2 Related Work

IRM [Arjovsky et al., 2019] has its roots in the theory of causality [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 and Schölkopf, 2010, Janzing et al., 2012] and the *invariant causal prediction principle* (ICP) [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 trained on the transformation that involves the causal factors ($\Phi(x) = x_{\mathrm{Pa}(y)}$) to predict y is an invariant predictor, which is robust to unseen interventions.

In general, for finite sets of environments, there may be other invariant predictors. If one has information about the causal Bayesian network structure, one can find invariant predictors that are maximally predictive using conditional independence tests and other graph-theoretic tools [Magliacane et al., 2018, Subbaswamy et al., 2019]. The above works select subsets of features, primarily using conditional independence tests, that make the optimal classifier trained on the selected features invariant. In IRM [Arjovsky et al., 2019], the authors give an optimization-based reformulation of this invariance that facilitates searching over transformations in a continuous space. Following the original work IRM from [Arjovsky et al., 2019], there have been several interesting works — [Teney et al., 2020, Krueger et al., 2020, Chang et al., 2020, Koyama and Yamaguchi, 2020, Mahajan et al., 2020] is an incomplete representative list — that build new methods inspired from IRM to address OOD generalization. In these works, similar to IRM, the algorithms are not provably guaranteed to converge to predictors with desirable OOD behavior.

3 Background

3.1 Nash Equilibrium and Concave Games

A standard normal form game is written as a tuple $\Omega = (\mathcal{N}, \{u_i\}_{i\in\mathcal{N}}, \{\mathcal{S}_i\}_{i\in\mathcal{N}})$, where $\mathcal{N}$ is a finite set of players. Player $i \in \mathcal{N}$ takes actions from a strategy set $\mathcal{S}_i$. The utility of player i is $u_i : \mathcal{S} \rightarrow \mathbb{R}$, where we write the joint set of actions of all the players as $\mathcal{S} = \Pi_{i\in\mathcal{N}}\mathcal{S}_i$. The joint strategy of all the players is given as $\boldsymbol{s} \in \mathcal{S}$, the strategy of player i is $\boldsymbol{s}_i$ and the strategy of the rest of players is $\boldsymbol{s}_{-i} = (\boldsymbol{s}_{i'})_{i'\neq i}$.

Definition 1. *A strategy $\boldsymbol{s}^\dagger \in \mathcal{S}$ is said to be a pure strategy Nash equilibrium (NE) if it satisfies*

$$u_i(\boldsymbol{s}_i^\dagger, \boldsymbol{s}_{-i}^\dagger) \geq u_i(k, \boldsymbol{s}_{-i}^\dagger), \forall k \in \mathcal{S}_i, \forall i \in \mathcal{N}$$

NE defines a state where each player is using the best possible strategy in response to the rest of the players. A natural question to ask is when does a pure strategy NE exist. In the seminal work of [Debreu, 1952] it was shown that for a special class of games called concave games such a NE always exists.

Definition 2. *A game Ω is called a concave game if for each $i \in \mathcal{S}$*

- *$\mathcal{S}_i$ is a compact, convex subset of $\mathbb{R}^{m_i}$*
- *$u_i(\boldsymbol{s}_i, \boldsymbol{s}_{-i})$ is continuous in $\boldsymbol{s}_{-i}$*
- *$u_i(\boldsymbol{s}_i, \boldsymbol{s}_{-i})$ is continuous and concave in $\boldsymbol{s}_i$.*

Theorem 1. *[Debreu, 1952] For any concave game Ω a pure strategy Nash equilibrium $\boldsymbol{s}^\dagger$ always exists.*

In this work, we only study pure strategy NE and use the terms pure strategy NE and NE interchangeably.

3.2 Invariant Risk Minimization & Invariant Risk Minimization Games

We are given a collection of training datasets $D = \{D_e\}_{e\in\mathcal{E}_{tr}}$ gathered from a set of environments $\mathcal{E}_{tr}$, where $D_e = \{\boldsymbol{x}_e^i, y_e^i\}_{i=1}^{n_e}$ is the dataset gathered from environment $e \in \mathcal{E}_{tr}$ and n_e is the number of points in environment e. The feature value for data point i is $\boldsymbol{x}_e^i \in \mathcal{X}$ and the corresponding label is $y_e^i \in \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}$. Each point $(\boldsymbol{x}_e^i, y_e^i)$ in environment e is drawn i.i.d from a distribution $\mathbb{P}_e$. Define a predictor $f : \mathcal{X} \to \mathbb{R}$. The goal of IRM is to use these collection of datasets D to construct a predictor f that performs well across many unseen environments $\mathcal{E}_{all}$, where $\mathcal{E}_{all} \supseteq \mathcal{E}_{tr}$. Define the risk achieved by f in environment e as $R_e(f) = \mathbb{E}_e\big[\ell(f(\boldsymbol{X}_e), Y_e)\big]$, where ℓ is the square loss when $f(\boldsymbol{X}_e)$ is the predicted value and Y_e is the corresponding label, $(\boldsymbol{X}_e, Y_e) \sim \mathbb{P}_e$ and the expectation $\mathbb{E}_e$ is defined with respect to (w.r.t.) the distribution of points in environment e.

Invariant predictor and IRM optimization: An invariant predictor is composed of two parts a representation $\boldsymbol{\Phi} \in \mathbb{R}^{d \times n}$ and a predictor $\boldsymbol{w} \in \mathbb{R}^{d \times 1}$. We say that a data representation $\boldsymbol{\Phi}$ elicits an invariant predictor $\boldsymbol{w}^\mathsf{T}\boldsymbol{\Phi}$ across the set of environments $\mathcal{E}_{tr}$ if there is a predictor $\boldsymbol{w}$ that achieves the minimum risk for all the environments $\boldsymbol{w} \in \operatorname{argmin}_{\tilde{\boldsymbol{w}} \in \mathbb{R}^{d \times 1}} R_e(\tilde{\boldsymbol{w}}^\mathsf{T}\boldsymbol{\Phi})$, $\forall e \in \mathcal{E}_{tr}$. IRM may be phrased as the following constrained optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\Phi} \in \mathbb{R}^{d \times n}, \boldsymbol{w} \in \mathbb{R}^{d \times 1}} \sum_{e \in \mathcal{E}_{tr}} R_e(\boldsymbol{w}^\mathsf{T}\boldsymbol{\Phi}) \\ &\text{s.t. } \boldsymbol{w} \in \operatorname*{argmin}_{\tilde{\boldsymbol{w}} \in \mathbb{R}^{d \times 1}} R_e(\tilde{\boldsymbol{w}}^\mathsf{T}\boldsymbol{\Phi}),\ \forall e \in \mathcal{E}_{tr} \end{aligned} \tag{1}$$

If $\boldsymbol{w}^\mathsf{T}\boldsymbol{\Phi}$ satisfies the constraints above, then it is an invariant predictor across the training environments $\mathcal{E}_{tr}$. Define the set of invariant predictors $\boldsymbol{w}^\mathsf{T}\boldsymbol{\Phi}$ satisfying the constraints in (1) as $\mathcal{S}^{\mathsf{IV}}$. Informally stated, the main idea behind the above optimization is inspired from invariance principles in causality [Bareinboim et al., 2012][Pearl, 2009]. Each environment can be understood as an intervention. By learning an invariant predictor the learner hopes to identify a representation $\boldsymbol{\Phi}$ that transforms the observed features into the causal features and the optimal model trained on causal representations are likely to be same (invariant) across the environments provided we do not intervene on the label itself. These invariant models can be shown to have a good out-of-distribution performance. Next, we briefly describe IRM-games.

Ensemble-game: Each environment e is endowed with its own predictor $\boldsymbol{w}_e \in \mathbb{R}^{d \times 1}$. Define an ensemble predictor $\bar{\boldsymbol{w}} \in \mathbb{R}^{d \times 1}$ given as $\bar{\boldsymbol{w}} = \sum_{q \in \mathcal{E}_{tr}} \boldsymbol{w}_q$; for the rest of this work a bar on top of vector represents an ensemble predictor. We require all the environments to use this ensemble $\bar{\boldsymbol{w}}$. We want to solve the following new optimization problem.

$$\begin{aligned} &\min_{\boldsymbol{\Phi} \in \mathbb{R}^{d \times n}, \bar{\boldsymbol{w}} \in \mathbb{R}^{d \times 1}} \sum_{e \in \mathcal{E}_{tr}} R_e\Big(\bar{\boldsymbol{w}}^\mathsf{T}\boldsymbol{\Phi}\Big) \\ &\text{s.t. } \boldsymbol{w}_e \in \operatorname*{argmin}_{\tilde{\boldsymbol{w}}_e \in \mathbb{R}^{d \times 1}} R_e\Bigg(\Big[\tilde{\boldsymbol{w}}_e + \sum_{q \in \mathcal{E}_{tr} \setminus \{e\}} \boldsymbol{w}_q\Big]^\mathsf{T}\boldsymbol{\Phi}\Bigg),\ \forall e \in \mathcal{E}_{tr} \end{aligned}$$

For a fixed representation $\boldsymbol{\Phi}$, the constraints in the above optimization (3.2) represent the NE of a game with each environment e as a player with actions $\tilde{\boldsymbol{w}}_e$. Environment e selects $\tilde{\boldsymbol{w}}_e$ to maximize its utility $-R_e\bigg(\Big[\tilde{\boldsymbol{w}}_e + \sum_{q \neq e} \tilde{\boldsymbol{w}}_q\Big]^\mathsf{T}\boldsymbol{\Phi}\bigg)$. Define the set of ensemble-game predictors $\bar{\boldsymbol{w}}^\mathsf{T}\boldsymbol{\Phi}$, i.e. the predictors that satisfy the constraints in (3.2) as $\mathcal{S}^{\mathsf{EG}}$. In [Ahuja et al., 2020] it was shown that the set of ensemble $\mathcal{S}^{\mathsf{EG}} = \mathcal{S}^{\mathsf{IV}}$. Having briefly reviewed IRM and IRM-games (we presented them with linear models but these works are more general), we are now ready to build our framework.

4 Linear Regression Games

4.1 Unconstrained Linear Regression Games

The data is gathered from a set of two environments, $\mathcal{E}_{tr} = \{1, 2\}$. [1] Each data point $(\boldsymbol{X}_e, Y_e)$ in environment e is sampled from $\mathbb{P}_e$. Each environment $e \in \{1, 2\}$ is a player that wants to select a predictor $\boldsymbol{w}_e \in \mathbb{R}^{n \times 1}$ such that it minimizes

$$R_e(\boldsymbol{w}_1, \boldsymbol{w}_2) = \mathbb{E}_e\Big[\big(Y_e - \boldsymbol{w}_1^\mathsf{T}\boldsymbol{X}_e - \boldsymbol{w}_2^\mathsf{T}\boldsymbol{X}_e\big)^2\Big] \tag{2}$$

where $\mathbb{E}_e$ is expectation w.r.t $\mathbb{P}_e$. We write the above as a two player game represented by a tuple $\Gamma = (\{1, 2\}, \{R_e\}_{e \in \{1,2\}}, \mathbb{R}^{n \times 1})$. We refer to Γ as a unconstrained linear regression game (U-LRG). A Nash equilibrium $\boldsymbol{w}^\dagger = (\boldsymbol{w}_1^\dagger, \boldsymbol{w}_2^\dagger)$ of U-LRG is a solution to

$$\begin{aligned} \boldsymbol{w}_1^\dagger &\in \operatorname*{argmin}_{\tilde{\boldsymbol{w}}_1 \in \mathbb{R}^{n \times 1}} \mathbb{E}_1\Big[\big(Y_1 - \tilde{\boldsymbol{w}}_1^\mathsf{T}\boldsymbol{X}_1 - \boldsymbol{w}_2^{\dagger,\mathsf{T}}\boldsymbol{X}_1\big)^2\Big] \\ \boldsymbol{w}_2^\dagger &\in \operatorname*{argmin}_{\tilde{\boldsymbol{w}}_2 \in \mathbb{R}^{n \times 1}} \mathbb{E}_2\Big[\big(Y_2 - \boldsymbol{w}_1^{\dagger,\mathsf{T}}\boldsymbol{X}_2 - \tilde{\boldsymbol{w}}_2^\mathsf{T}\boldsymbol{X}_2\big)^2\Big] \end{aligned} \tag{3}$$

The above two-player U-LRG is a natural extension of linear regressions and we start by analyzing the NE of the above game. Before going further, the above game can be understood as fixing $\boldsymbol{\Phi}$ to identity in the ensemble-game defined in the previous section.

For each $e \in \{1, 2\}$, define the mean of features $\boldsymbol{\mu}_e = \mathbb{E}_e[\boldsymbol{X}_e]$, $\boldsymbol{\Sigma}_e = \mathbb{E}_e\big[\boldsymbol{X}_e\boldsymbol{X}_e^\mathsf{T}\big]$ and the correlation between the feature $\boldsymbol{X}_e$ and the label Y_e as $\boldsymbol{\rho}_e = \mathbb{E}_e\big[\boldsymbol{X}_e Y_e\big]$.

Assumption 1. ***Regularity condition.*** *For each $e \in \{1, 2\}$, $\boldsymbol{\mu}_e = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_e$ is positive definite.*

The above regularity conditions are fairly standard and the mean zero condition can be relaxed by introducing intercepts in the model. When $\boldsymbol{\mu}_e = \boldsymbol{0}$, $\boldsymbol{\Sigma}_e$ is the covariance matrix. For each $e \in \{1, 2\}$, define $\boldsymbol{w}_e^* = \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}_e$, where $\boldsymbol{\Sigma}_e^{-1}$ is the inverse of $\boldsymbol{\Sigma}_e$. $\boldsymbol{w}_e^*$ is the least squares optimal solution for environment e, i.e., it solves $\min_{\tilde{\boldsymbol{w}} \in \mathbb{R}^{n \times 1}} \mathbb{E}_e\Big[\big(Y_e - \tilde{\boldsymbol{w}}^\mathsf{T}\boldsymbol{X}_e\big)^2\Big]$.

[1] Discussion on multiple environments is in the supplement.

Proposition 1. *If Assumption 1 holds and if the least squares optimal solution in the two environments are*

- *equal, i.e., $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$, then the set $\{(\boldsymbol{w}_1^\dagger, \boldsymbol{w}_2^\dagger) \mid \boldsymbol{w}_1^\dagger + \boldsymbol{w}_2^\dagger = \boldsymbol{w}_1^*\}$ describes all the pure strategy Nash equilibrium of U-LRG, Γ.*
- *not equal, i.e., $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$, then U-LRG, Γ, has no pure strategy Nash equilibrium.*

We provide brief proof sketches here and all the detailed proofs are in the Appendix.

Proof sketch: Consider the case when the least squares optimal solution is different for the two environments. Also, assume that the NE of the U-LRG exists. In the NE, the ensemble predictor used will not be the least squares optimal predictor for at least one of the environments. If this is the case, then such an environment can always update its predictor to improve its loss. This contradicts the fact that the two environments are using predictors that form the NE. Therefore, NE cannot exist. □

From the above proposition, it follows that agreement between the environments on least squares optimal solution is both necessary and sufficient for the existence of NE of U-LRG. Next, we describe the family of linear structural equation models (SEMs) in [Arjovsky et al., 2019] and show how the two cases, $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$, and $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$ naturally arise.

4.1.1 Nash Equilibria for Linear SEMs

In this section, we consider linear SEMs from [Arjovsky et al., 2019] and study the NE of U-LRG.

Assumption 2. ***Linear SEM with confounders and anti-causal variables (Figure 1)*** *For each $e \in \{1,2\}$, $(\boldsymbol{X}_e, Y_e)$ is generated from the following SEM*

$$\begin{aligned} Y_e &\leftarrow \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{X}_e^1 + \boldsymbol{\eta}_e^\mathsf{T}\boldsymbol{H}_e + \varepsilon_e, \\ \boldsymbol{X}_e^2 &\leftarrow \boldsymbol{\alpha}_e Y_e + \boldsymbol{\Theta}_e\boldsymbol{H}_e + \boldsymbol{\zeta}_e \end{aligned} \tag{4}$$

The feature vector is $\boldsymbol{X}_e = (\boldsymbol{X}_e^1, \boldsymbol{X}_e^2)$. $\boldsymbol{H}_e \in \mathbb{R}^s$ is a confounding random variable, where each component of $\boldsymbol{H}_e$ is an i.i.d draw from a distribution with zero mean and unit variance. $\boldsymbol{H}_e$ affects both the labels Y_e through weights $\boldsymbol{\eta}_e \in \mathbb{R}^s$ and a subset of features $\boldsymbol{X}_e^2 \in \mathbb{R}^q$ through weights $\boldsymbol{\Theta}_e \in \mathbb{R}^{q\times s}$. $\varepsilon_e \in \mathbb{R}$ is independent zero mean noise in the label generation. Y_e affects a subset of features $\boldsymbol{X}_e^2$ with weight $\boldsymbol{\alpha}_e \in \mathbb{R}^q$, $\boldsymbol{\zeta}_e \in \mathbb{R}^q$ is an independent zero mean noise vector affecting $\boldsymbol{X}_e^2$. $\boldsymbol{X}_e^1 \in \mathbb{R}^p$ are the causal features drawn from a distribution with zero mean and affect the label through a weight $\boldsymbol{\gamma} \in \mathbb{R}^p$, which is invariant across the environments.

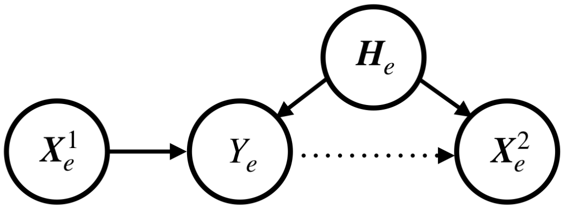

Figure 1: SEM from Assumption 2. We show the link between Y_e and $\boldsymbol{X}_e^2$ with a dotted line because in our theoretical analysis (Proposition 4,5) we assume that the edge does not exist but in the experiments we compare in the more general setting where such an edge exists.

The above model captures many different settings. If $\boldsymbol{\alpha}_e = \boldsymbol{0}$ and $\boldsymbol{\Theta}_e \neq \boldsymbol{0}$, then features $\boldsymbol{X}_e^2$ appear correlated with the label due to the confounder $\boldsymbol{H}_e$. If $\boldsymbol{\alpha_e} \neq \boldsymbol{0}$ and $\boldsymbol{\Theta}_e = \boldsymbol{0}$, then features $\boldsymbol{X}_e^2$ are correlated with the label but they are effects or anti-causal. If both $\boldsymbol{\alpha}_e \neq \boldsymbol{0}, \boldsymbol{\Theta}_e \neq \boldsymbol{0}$, then we are in a hybrid of the above two settings. In all of the above settings it can be shown that relying on $\boldsymbol{X}_e^2$ to make predictions can lead to failures under distribution shifts (modeled by interventions). From [Arjovsky et al., 2019], we know that for the above family of models the ideal OOD predictor is $(\boldsymbol{\gamma}, \boldsymbol{0})$ as it performs well across many distribution shifts (modeled by interventions). Hence, the goal is to learn $(\boldsymbol{\gamma}, \boldsymbol{0})$.

No confounders & no anti-causal variables ($\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$): Consider the SEM in Assumption 2. For each environment $e \in \{1,2\}$, assume $\boldsymbol{\alpha}_e = 0$ and $\boldsymbol{\Theta}_e = 0$, i.e. no confounding and no anti-causal variables. This setting captures the standard covariate shifts [Gretton et al., 2009], where it is assumed that $\mathbb{P}_e(Y_e|\boldsymbol{X}_e = \boldsymbol{x})$ is invariant across environments, here we assume $\mathbb{E}_e(Y_e|\boldsymbol{X}_e = \boldsymbol{x}) = \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}$ is invariant across environments. The least squares optimal solution for each environment is $\boldsymbol{w}_e^* = (\boldsymbol{\gamma}, \boldsymbol{0})$, which implies that $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$. From Proposition 1 we know that a NE exists (any two predictors adding to $\boldsymbol{w}_1^*$ form an NE). In this setting, different methods – empirical risk minimization (ERM), IRM, IRM-games, and methods designed for covariate shifts such as sample reweighting – should perform well.

Confounders only ($\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$): Consider the SEM in Assumption 2. For each environment $e \in \{1,2\}$, assume $\boldsymbol{\alpha}_e = \boldsymbol{0}$, $\boldsymbol{\Theta}_e \neq 0$, i.e. confounders only setting. Define $\boldsymbol{\Sigma}_{e1} = \mathbb{E}_e\Big[\boldsymbol{X}_e^1\boldsymbol{X}_e^{1,\mathsf{T}}\Big]$ and define the variance for the noise vector $\boldsymbol{\zeta}$ as $\boldsymbol{\sigma}_{\zeta_e}^2 = \mathbb{E}_e[\boldsymbol{\zeta}_e \odot \boldsymbol{\zeta}_e]$, where $\odot$ represents element-wise product between two vectors.

Assumption 3. ***Regularity condition for linear SEM in Assumption 2.*** *For each environment $e \in \{1,2\}$, $\boldsymbol{\Sigma}_{e1}$ is positive definite and each element of the* *vector $\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e}$ is positive.*

Assumption 3 is equivalent to Assumption 1 for SEM in Assumption 2 (it ensures $\boldsymbol{\Sigma}_e$ is positive definite).

Proposition 2. *If Assumption 2 holds with $\boldsymbol{\alpha}_e = \mathbf{0}$ for each $e \in \{1,2\}$, and Assumption 3 holds, then the least squares optimal solution for environment e is*

$$\boldsymbol{w}_e^* = (\boldsymbol{w}_e^{\mathsf{inv}}, \boldsymbol{w}_e^{\mathsf{var}}) = \Big(\boldsymbol{\gamma}, \Big(\boldsymbol{\Theta}_e\boldsymbol{\Theta}_e^{\mathsf{T}} + \mathsf{diag}[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e}]\Big)^{-1}\boldsymbol{\Theta}_e\boldsymbol{\eta}_e\Big) \tag{5}$$

Proof sketch: Recall that the least squares optimal solution for environment e is $\boldsymbol{w}_e^* = \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}_e$. We use the structure of the SEM in Assumption 2 and Assumption 3 to simplify $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\rho}_e$ to arrive at the above expression. □

We divide $\boldsymbol{w}_e^*$ into two halves $\boldsymbol{w}_e^{\mathsf{inv}} = \boldsymbol{\gamma}$ and $\boldsymbol{w}_e^{\mathsf{var}} = \Big(\boldsymbol{\Theta}_e\boldsymbol{\Theta}_e^{\mathsf{T}} + \mathsf{diag}[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e}]\Big)^{-1}\boldsymbol{\Theta}_e\boldsymbol{\eta}_e$. Observe that the first half $\boldsymbol{w}_e^{\mathsf{inv}}$ is invariant, i.e., it does not depend on the environment, while $\boldsymbol{w}_e^{\mathsf{var}}$ may vary as it depends on the parameters specific to the environment e.g., $\boldsymbol{\Theta}_e, \boldsymbol{\eta}_e$. In general, $\boldsymbol{w}_1^{\mathsf{var}} \neq \boldsymbol{w}_2^{\mathsf{var}}$ (e.g., $s = q$, $\boldsymbol{\Theta}_e$ is identity $\boldsymbol{I}_q$, $\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e}$ is one $\mathbf{1}_q$, $\boldsymbol{\eta}_1 \neq \boldsymbol{\eta}_2$) and as a result $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$. In such a case, from Proposition 1, we know that NE does not exist. ERM and other techniques such as domain adaptation [Ajakan et al., 2014, Ben-David et al., 2007, Glorot et al., 2011, Ganin et al., 2016], robust optimization [Mohri et al., 2019, Hoffman et al., 2018, Lee and Raginsky, 2018, Duchi et al., 2016], would tend to learn a model which tends to exploit information from the spuriously correlated $\boldsymbol{X}_e^2$ thus placing a non-zero weight on the second half corresponding to the features $\boldsymbol{X}_e^2$ and not recovering $(\boldsymbol{\gamma}, \mathbf{0})$.

IRM based methods are designed to tackle these problems. These works try to learn representations that filter out causal features, $\boldsymbol{X}_e^1$, with invariant coefficients, $\boldsymbol{w}_e^{\mathsf{inv}}$, from spurious features, $\boldsymbol{X}_e^2$, with variant coefficients $\boldsymbol{w}_e^{\mathsf{var}}$ and learn a classifier on top resulting in the invariant predictor $(\boldsymbol{\gamma}, \mathbf{0})$. However, the current algorithms that search for these representations in IRM and IRM-games are based on gradient descent over non-convex losses and non-trivial best response dynamics respectively, both of which are not guaranteed to converge to the ideal OOD predictor $(\boldsymbol{\gamma}, \mathbf{0})$. We formally state the assumption underlying these methods, which we also use later.

Assumption 4. ***Spurious features have varying coefficents across environments.*** $w_1^{\mathsf{var}} \neq \boldsymbol{w}_2^{\mathsf{var}}$

4.2 Constrained Linear Regression Games

In U-LRG, Γ, the utility of environment 1 (2) is $-R_1(\boldsymbol{w}_1, \boldsymbol{w}_2)\big(-R_2(\boldsymbol{w}_1, \boldsymbol{w}_2)\big)$. For each environment $e \in \{1,2\}$, $-R_e$ is continous and concave in $\boldsymbol{w}_e$. For each e in the game Γ, the set of actions it can take is in $\mathbb{R}^{n\times 1}$, which is not a compact set. If the set of actions for each environment were compact and convex, then we can use Theorem 1 to guarantee that a NE always exists. Let us constraint the predictors to be in the set $\mathcal{W} = \big\{\boldsymbol{w}_e \mid \|\boldsymbol{w}_e\|_\infty \leq w^{\mathsf{sup}}\big\}$, where $\|\cdot\|_\infty$ is the ℓ_∞ norm and $0 < w^{\mathsf{sup}} < \infty$. We define the constrained linear regression game (C-LRG) as $\Gamma_c = \big(\mathcal{E}_{tr}, \{-R^e\}_{e\in\mathcal{E}_{tr}}, \mathcal{W}\big)$.

Proposition 3. *A pure strategy Nash equilibrium always exists for C-LRG, Γ_c.*

Proof sketch: $\mathcal{W}$ is a closed and bounded subset in the Euclidean space, which implies it is also a compact set. $\mathcal{W}$ is also a convex set as ℓ_∞ norm is convex. From Definition 2, C-LRG, Γ_c, is concave. Therefore from Theorem 1 it follows that a NE always exists for Γ_c. □

Unlike the game Γ, a NE always exists for the game Γ_c. Let $\boldsymbol{w}_1^\dagger, \boldsymbol{w}_2^\dagger$ be an NE of Γ_c and let $\bar{\boldsymbol{w}}^\dagger$ be the corresponding ensemble predictor, i.e. $\bar{\boldsymbol{w}}^\dagger = \boldsymbol{w}_1^\dagger + \boldsymbol{w}_2^\dagger$. In the next theorem, we analyze the properties of $\boldsymbol{w}_1^\dagger, \boldsymbol{w}_2^\dagger$ but before that we state some assumptions.

Assumption 5. ***Realizability.*** *For each $e \in \{1,2\}$ the least squares optimal solution $\boldsymbol{w}_e^* \in \mathcal{W}$.*

We write the feature vector in environment e as $\boldsymbol{X}_e = (X_{e1}, \dots, X_{en})$ and the least squares optimal solution in environment e as $\boldsymbol{w}_e^* = (w_{e1}, \dots, w_{en})$. Divide the features indexed $\{1, \dots, n\}$ into two sets $\mathcal{U}$ and $\mathcal{V}$. $\mathcal{U}$ is defined as: $i \in \mathcal{U}$ if and only if the weight associated with i^{th} component in the least squares solution is equal in the two environments, i.e., $w_{1i}^* = w_{2i}^*$. $\mathcal{V}$ is defined as: $i \in \mathcal{V}$ if and only if the weight associated with i^{th} component in the least squares solution is not equal in the two environments, i.e., $w_{1i}^* \neq w_{2i}^*$. For an example of these sets, consider the least squares solution to the confounded only SEM in equation (5) under Assumption 4 , $\boldsymbol{w}_1^{\mathsf{inv}} = \boldsymbol{w}_2^{\mathsf{inv}} = \boldsymbol{\gamma} \implies \mathcal{U} = \{1, \dots, p\}$, and $\boldsymbol{w}_1^{\mathsf{var}} \neq \boldsymbol{w}_2^{\mathsf{var}} \implies \mathcal{V} = \{p+1, \dots, p+q\}$.

Assumption 6. ***Features with varying coefficients across environments are uncorrelated.*** *For each $i \in \mathcal{V}$ the corresponding feature X_{ei} is uncorrelated with every other feature $j \in \{1, \dots, n\}\backslash\{i\}$, i.e., $\mathbb{E}[X_{ei}X_{ej}] = \mathbb{E}[X_{ei}]\mathbb{E}[X_{ej}]$.*

The above assumption says that any feature component whose least squares optimal solution coefficient varies across environments is not correlated with the rest of the features. We use the above assumption to derive an analytical expression for the NE of Γ_c next.

For a vector $\boldsymbol{a}$, $|\boldsymbol{a}|$ represents the vector of absolute values of all the elements. Element-wise product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is written as $\boldsymbol{a} \odot \boldsymbol{b}$. Define an indicator function $\mathbf{1}_{\boldsymbol{a}\geq\boldsymbol{b}}$; it carries out an element-wise comparison of $\boldsymbol{a}$ and $\boldsymbol{b}$ and it outputs a vector of ones and zeros, where a one at component i indicates that i^{th} component of $\boldsymbol{a}$, a_i, is greater than or equal to the i^{th} component of $\boldsymbol{b}$, b_i. Recall that the ensemble predictor constructed from NE is $\bar{\boldsymbol{w}}^{\dagger} = \bar{\boldsymbol{w}}_1^{\dagger} + \bar{\boldsymbol{w}}_2^{\dagger}$.

Theorem 2. *If Assumptions 1, 5, 6 hold, then the ensemble predictor, $\bar{\boldsymbol{w}}^{\dagger}$, constructed from the Nash equilibrium, $(\boldsymbol{w}_1^{\dagger}, \boldsymbol{w}_2^{\dagger})$, of Γ_c is equal to*

$$\Big(\boldsymbol{w}_1^* \odot \mathbf{1}_{|\boldsymbol{w}_2^*|\geq|\boldsymbol{w}_1^*|} + \boldsymbol{w}_2^* \odot \mathbf{1}_{|\boldsymbol{w}_1^*|>|\boldsymbol{w}_2^*|}\Big)\mathbf{1}_{\boldsymbol{w}_1^*\odot\boldsymbol{w}_2^*\geq\mathbf{0}} \quad (6)$$

Proof sketch: In order to prove the above theorem, we first establish an intermediate result in the form of a lemma. In the lemma, we show that if the least squares optimal solution in the two environments are different, i.e., $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$, then the NE predictor for at least one of the environments $\boldsymbol{w}_e^{\dagger}$ is at the boundary of the constraint set $\mathcal{W}$. We use Karush-Kuhn-Tucker (KKT) conditions [Boyd and Vandenberghe, 2004] for subdifferentiable convex functions to arrive at this lemma.

Building on this lemma, we use the Assumption 6 and the ℓ_∞ norm constraint to arrive at a component-wise separability for feature components in set $\mathcal{V}$ (defined in Assumption 6). This separability enables us to analyze the NE independently in a component-wise fashion. We discuss two main cases in which the component-wise analysis of NE is divided. Say we are looking at one of the components $k \in \mathcal{V}$. The least squares optimal coefficient for the component k are w_{1k}^* and w_{2k}^* for the two environments. Consider the case when $0 \leq w_{1k}^* < w_{2k}^*$. In this case, the $w_{1k}^{\dagger} = -w^{\mathsf{sup}} + w_{1k}^*$ and $w_{2k}^{\dagger} = w^{\mathsf{sup}}$ form the NE. In this state, the first environment has no incentive to deviate as the total weight for component k is w_{1k}^*, which is the optimal choice for environment 1 for component k. Since the second environment's optimal weight is larger than the first environment, it has an incentive to increase its weight but it cannot as it is already using the largest weight possible w^{sup}. Consider another case when $w_{1k}^* < 0 < w_{2k}^*$. In this case, the $w_{1k}^{\dagger} = -w^{\mathsf{sup}}$ and $w_{2k}^{\dagger} = w^{\mathsf{sup}}$ corresponds to the NE. In this state, the total weight for component k is 0, environment 1 will want to decrease the weight further to push it closer to w_{1k}^* but it cannot as it is already using the smallest weight possible $-w^{\mathsf{sup}}$. Similarly, environment 2 wants to increase the weight but it cannot as it is already using the largest weight possible w^{sup}.

□

Casewise analysis of NE in equation (6)

- $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$: Similar to Proposition 1 $\big\{(\boldsymbol{w}_1^{\dagger}, \boldsymbol{w}_2^{\dagger}) \mid \boldsymbol{w}_1^{\dagger} \in \mathcal{W}, \boldsymbol{w}_2^{\dagger} \in \mathcal{W}, \boldsymbol{w}_1^{\dagger} + \boldsymbol{w}_2^{\dagger} = \boldsymbol{w}_1^*\big\}$ is the set of NE of C-LRG
- $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$: We analyze this case under two categories
  - □ **Opposite sign coefficients:** If the i^{th} component of $\boldsymbol{w}_1^*$ and $\boldsymbol{w}_2^*$ have opposite signs, then the i^{th} component of the ensemble predictor, $\bar{\boldsymbol{w}}^{\dagger}$, constructed from the NE of Γ_c, is zero, i.e., $\bar{w}_i^{\dagger} = \big[\mathbf{1}_{\boldsymbol{w}_1^*\odot\boldsymbol{w}_2^*\geq\mathbf{0}}\big]_i = 0$. In this case, the coefficient of the environments' predictors in the NE, $w_{1i}^{\dagger}$ and $w_{2i}^{\dagger}$, have exact opposite signs and both are at the boundary one at w^{sup} and other at $-w^{\mathsf{sup}}$. This case shows that when the features have a large variation in their least squares coefficients across environments, they can be spurious (see Proposition 4) and the ensemble predictor filters them by assigning a zero weight to them.
  - □ **Same sign coefficients:** If the i^{th} component of $\boldsymbol{w}_1^*$ and $\boldsymbol{w}_2^*$ have same signs, then the i^{th} component of ensemble predictor, $\bar{\boldsymbol{w}}^{\dagger}$, constructed from the NE of Γ_c, is set to the least squares coefficient with a smaller absolute value, i.e., $\bar{w}_i^{\dagger} = w_{1i}^*$, where $|w_{1i}^*| \leq |w_{2i}^*|$. Suppose $0 < w_{1i}^* < w_{2i}^*$, the coefficient of the environments' predictors in the NE, $w_{1i}^{\dagger}$ and $w_{2i}^{\dagger}$ have opposite signs, i.e., $w_{2i}^{\dagger} = w^{\mathsf{sup}}$ and $w_{1i}^{\dagger} = w_1^* - w^{\mathsf{sup}}$. This shows that ensemble predictor is conservative and selects the smaller least squares coefficient. This property is useful to identifying predictors that are robust (see Proposition 4). Lastly, only when the least square coefficients are the same, i.e., $w_{1i}^* = w_{2i}^*$, the coefficient of the environments' predictors in the NE can be in the interior, i.e., $|w_{1i}^{\dagger}| < w^{\mathsf{sup}}$ and $|w_{2i}^{\dagger}| < w^{\mathsf{sup}}$.

4.2.1 Nash Equilibria for Linear SEMs

Suppose for each environment $e \in \{1, 2\}$ the data is generated from SEM in Assumption 2. We study if the NE of C-LRG, Γ_c, achieves or gets close to the ideal OOD predictor $(\boldsymbol{\gamma}, 0)$. We compare the ensemble predictors $\bar{\boldsymbol{w}}^{\dagger}$ constructed from the NE of Γ_c to the solutions of ERM (Theorem 2 enables this comparison). In ERM, the data from both the environments is combined and the overall least squares loss is minimized. Define the probability that a point is from environment e as π_e ($\pi_2 = 1 - \pi_1$). The set of ERM solutions for all distributions, $\{\pi_1, \pi_2\}$, is $\mathcal{S}^{\mathsf{ERM}}$ given as

$$\Big\{\boldsymbol{w} \,\big|\, \pi_1 \in [0,1], \boldsymbol{w} \in \operatorname*{argmin}_{\tilde{\boldsymbol{w}}\in\mathbb{R}^{n\times 1}} \sum_{e\in\{1,2\}} \pi_e \mathbb{E}_e\big[\big(Y_e - \tilde{\boldsymbol{w}}^{\mathsf{T}}\boldsymbol{X}_e\big)^2\big]\Big\}$$

Proposition 4. *If Assumption 2 holds with $\boldsymbol{\alpha}_e = \mathbf{0}$ and $\boldsymbol{\Theta}_e$ an orthogonal matrix for each $e \in \{1, 2\}$, and Assumptions 3, 4, 5 hold, then $\|\bar{\boldsymbol{w}}^{\dagger} - (\boldsymbol{\gamma}, 0)\| < \|\boldsymbol{w}^{\mathsf{ERM}} - (\boldsymbol{\gamma}, 0)\|$ holds for all $\boldsymbol{w}^{\mathsf{ERM}} \in \mathcal{S}^{\mathsf{ERM}}$.* [2] *Moreover, if all*

[2] Exception occurs over measure zero set over probabilities π_1. If least squares solution are strictly ordered, i.e., $\forall i \in \{1, \ldots, n\}, 0 < w_{1i}^* < w_{2i}^*$ and $\pi_1 = 1$, then

the components of two vectors $\boldsymbol{\Theta}_1\boldsymbol{\eta}_1$ *and* $\boldsymbol{\Theta}_2\boldsymbol{\eta}_2$ *have opposite signs, then* $\bar{\boldsymbol{w}}^{\dagger} = (\boldsymbol{\gamma}, 0)$.

Proof sketch: The assumptions in the above proposition imply that Assumptions 1, 5, 6 hold. Therefore, we can use Theorem 2 to derive the expression for the NE based ensemble predictor. From Proposition 2 we can derive the expression for the ERM based predictor. We use these expressions to compare the distance of the NE based ensemble predictor and the ERM based predictor from the ideal OOD predictor to arrive at the above result □

From the first part of the above we learn that for many confounder only models ($\boldsymbol{\alpha}_e = \mathbf{0}$, $\boldsymbol{\Theta}_e$ an orthogonal matrix), the ensemble predictor constructed from the NE is closer to the ideal OOD solution than ERM. For the second part, set $\boldsymbol{\Theta}_e = \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ is identity matrix. Suppose the signs of all the components of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ disagree. As a result, the signs of latter half of least squares solution $\boldsymbol{w}_e^{\mathsf{var}}$ (in equation (5)) disagree. From Theorem 2, we know that if the signs of the coefficients in least squares solution disagree, then the corresponding coefficient in the ensemble predictor is zero, which implies $\bar{\boldsymbol{w}}^{\dagger} = (\boldsymbol{\gamma}, 0)$.

Remark. In Proposition 4, besides the regularity conditions, the main assumption is $\boldsymbol{\Theta}_e$ is orthogonal. This assumption ensures that the the spurious features $\boldsymbol{X}_e^2$ are uncorrelated (Assumption 6). For confounder only models this seems reasonable. However, in the models involving anti-causal variables, i.e., $\boldsymbol{\alpha}_e \neq 0$, the spurious features can be correlated and one may wonder how does the ensemble predictor behave in such setups? In experiments, we show that ensemble predictors perform well in these settings as well. Extending the theory to anti-causal models is a part of future work.

Insights from Theorem 2, Proposition 4

Suppose the data comes from the SEM in Assumption 2. For this SEM, [Arjovsky et al., 2019] showed that if the number of environments grow linearly in the total number of features, then the solution to non-convex IRM optimization recovers the ideal OOD predictor. We showed that for many confounder only SEMs ($\boldsymbol{\alpha}_e = 0$ and $\boldsymbol{\Theta}_e$ orthogonal) NE based ensemble predictor gets closer to the OOD predictor than ERM and sometimes recovers it exactly with just *two environments*, while *no such guarantees* exist for IRM. Next, we show how to learn these NE based ensemble predictor.

4.3 Learning NE of C-LRG

In this section, we show how we can use best response dynamics (BRD) [Fudenberg et al., 1998] to learn the NE. Each environment takes its turn and finds the best possible model given the choice made by the other environment. This procedure (Algorithm 1) is allowed to run until the environments stop updating their models. In the next theorem, we make the same set of Assumptions as in Theorem 2 and show that Algorithm 1 converges to the NE derived in Theorem 2.

```
Algorithm 1: Best response based learning
  Initialize: $\tilde{\boldsymbol{w}}_1 = \mathbf{0}$, $\tilde{\boldsymbol{w}}_2 = \mathbf{0}$, $p = 0$
  while $w_1^{\mathsf{diff}} > 0$ or $w_2^{\mathsf{diff}} > 0$ do
      $\tilde{\boldsymbol{w}}_1^{\mathsf{cur}} = \tilde{\boldsymbol{w}}_1$, $\tilde{\boldsymbol{w}}_2^{\mathsf{cur}} = \tilde{\boldsymbol{w}}_2$
      $\tilde{\boldsymbol{w}}_1 = \min_{\boldsymbol{w}_1 \in \mathcal{W}} R_1(\boldsymbol{w}_1, \tilde{\boldsymbol{w}}_2)$
      $\tilde{\boldsymbol{w}}_2 = \min_{\boldsymbol{w}_2 \in \mathcal{W}} R_2(\tilde{\boldsymbol{w}}_1, \boldsymbol{w}_2)$
      $w_1^{\mathsf{diff}} = \|\tilde{\boldsymbol{w}}_1^{\mathsf{cur}} - \tilde{\boldsymbol{w}}_1\|$, $w_2^{\mathsf{diff}} = \|\tilde{\boldsymbol{w}}_2^{\mathsf{cur}} - \tilde{\boldsymbol{w}}_2\|$
  end
  Output: $\bar{\boldsymbol{w}}^{+} = \tilde{\boldsymbol{w}}_1 + \tilde{\boldsymbol{w}}_2$
```

Theorem 3. *If Assumption 1, 5, 6 hold, then the output of Algorithm 1,* $\bar{\boldsymbol{w}}^{+}$*, is*

$$\Big(\boldsymbol{w}_1^* \odot \mathbf{1}_{|\boldsymbol{w}_2^*| \geq |\boldsymbol{w}_1^*|} + \boldsymbol{w}_2^* \odot \mathbf{1}_{|\boldsymbol{w}_1^*| > |\boldsymbol{w}_2^*|}\Big)\mathbf{1}_{\boldsymbol{w}_1^* \odot \boldsymbol{w}_2^* \geq \mathbf{0}}$$

Proof sketch. We illustrate the dynamic of one of the cases to provide some insight into the convergence. Consider the i^{th} component of the predictors $\tilde{w}_{1i}$ and $\tilde{w}_{2i}$ from Algorithm 1. Suppose $w_{1i}^* > w_{2i}^*$ and $|w_{1i}^*| > |w_{2i}^*|$. The two environments push the ensemble predictor, $\tilde{w}_{1i} + \tilde{w}_{2i}$, in opposite directions during their turns, with the first environment increasing its weight, $\tilde{w}_{1i}$, and the second environment decreasing its weight, $\tilde{w}_{2i}$. Eventually, the environment with a higher absolute value ($e = 1$ since $|w_{1i}^*| > |w_{2i}^*|$) reaches the boundary ($\tilde{w}_{1i} = w^{\mathsf{sup}}$) and cannot move any further due to the constraint. The other environment ($e = 2$) best responds. It either hits the other end of the boundary ($\tilde{w}_{2i} = -w^{\mathsf{sup}}$), in which case the weight of the ensemble for component i is zero, or gets close to the other boundary while staying in the interior ($\tilde{w}_{2i} = w_{2i}^* - w^{\mathsf{sup}}$), in which case the weight of the ensemble for component i is w_{2i}^*. □

BRD a sequence of convex minimizations. In Algorithm 1, we assumed that at each time step each environment can do an exact minimization operation. The minimization for each environment is a simple least squares regression, which is a convex quadratic minimization problem. There can be several ways of solving it – gradient descent for R_e and solving for gradient of R_e equals zero directly, which is a linear system of equations. We provide a simple bound for the total number of convex minimizations (or turns for each environment) in Algorithm 1 next. For each $i \in \mathcal{V}$ (defined in Section 4.2), compute the distance between the least square coefficients in the two environments

$\boldsymbol{w}^{\mathsf{ERM}} = \bar{\boldsymbol{w}}^{\dagger} = \boldsymbol{w}_1^*$. In general, $\boldsymbol{w}_1^*, \boldsymbol{w}_2^*$ are not ordered and $\pi_1 \in (0, 1)$, thus C-LRG improves over ERM.

$|w_{1i}^* - w_{2i}^*|$ and find the least distance over the set $\mathcal{V}$ given as $\Delta_{\mathsf{min}} = \min_{i \in \mathcal{V}} |w_{1i}^* - w_{2i}^*|$ (following the definition of $\mathcal{V}$ this distance is positive). The bound on number of minimizations is $\frac{2w^{\mathsf{sup}}}{\Delta_{\mathsf{min}}}$.

4.3.1 Learning NE of C-LRG: Linear SEMs

Suppose the data is generated from SEM in Assumption 2. Next, we show the final result that the NE based predictor, which we proved in Proposition 4 is closer to the OOD solution, is achieved by Algorithm 1.

Proposition 5. *If Assumption 2 holds with $\boldsymbol{\alpha}_e = \mathbf{0}$ and $\boldsymbol{\Theta}_e$ an orthogonal matrix for each $e \in \{1, 2\}$, and Assumptions 3, 4, 5 hold, then the output of Algorithm 1, $\bar{\boldsymbol{w}}^+$ obeys $\|\bar{\boldsymbol{w}}^+ - (\boldsymbol{\gamma}, 0)\| < \|\boldsymbol{w}^{\mathsf{ERM}} - (\boldsymbol{\gamma}, 0)\|$ for all $\boldsymbol{w}^{\mathsf{ERM}} \in \mathcal{S}^{\mathsf{ERM}}$ except over a set of measure zero (see footnote 2). Moreover, if all the components of vectors $\boldsymbol{\Theta}_1\boldsymbol{\eta}_1$ and $\boldsymbol{\Theta}_2\boldsymbol{\eta}_2$ have opposite signs, then $\bar{\boldsymbol{w}}^+ = (\boldsymbol{\gamma}, 0)$.*

We use Theorem 3 to arive at the above result. We have shown through Theorem 2, Proposition 4, Theorem 3 and Proposition 5 that the NE based ensemble predictor of Γ_c has good OOD properties and it can be learned by solving a sequence of convex quadratic minimizations.

Extensions: In the supplement, we extend the Theorem 3 to other BRD that are commonly used. We also discuss how to extend the theory to settings beyond Assumption 6. The entire analysis is for linear SEMs. In the experiments section, we show how the method performs when we use non-linear models and analysis for non-linear models is left to future work.

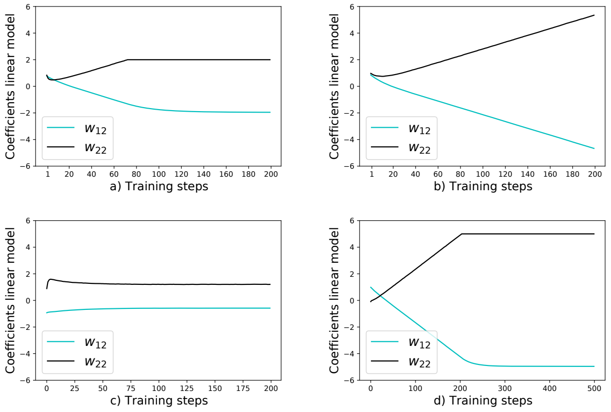

Figure 2: a) C-LRG ($w^{\mathsf{sup}} = 2$), b) U-LRG, c) R$_\infty$-LRG , d) C-LRG ($w^{\mathsf{sup}} = 5$)

5 Experiments

5.1 Linear SEM experiments

In this section, we first run the regression experiments described in [Arjovsky et al., 2019]. We use the SEM in Assumption 2 with following configurations.

Method	Solution	Error
Oracle	$(1.0, 0.0)$	0.0
U-LRG	$(0.34, 0.67)$	0.88
C-LRG ($w^{\mathsf{sup}} = 2$)	$(0.95, 0.05)$	**0.005**
C-LRG ($w^{\mathsf{sup}} = 5$)	$(0.95, 0.04)$	**0.005**
R$_\infty$-LRG	$(0.33, 0.65)$	0.87
R$_2$-LRG	$(0.33, 0.63)$	0.83
ERM	$(0.34, 0.67)$	0.88
IRM	$(0.63, 0.44)$	0.33
ICP	$(0.0, 0.0)$	1.0

Table 1: Comparing variants of LRG, IRM, ICP, and ERM.

• $\boldsymbol{\gamma}$ is a vector of ones with p dimensions, $\mathbf{1}_p$, which makes the ideal OOD model $\big(\mathbf{1}_p, \mathbf{0}_q\big)$. Each component of the confounder $\boldsymbol{H}_e$ is drawn i.i.d. from $\mathcal{N}(0, \sigma_{\boldsymbol{H}_e}^2)$. $\sigma_{\boldsymbol{H}_1} = 0.2$, $\sigma_{\boldsymbol{H}_2} = 2.0$. We consider two configurations for $\boldsymbol{\Theta}_e$ and $\boldsymbol{\eta}_e$. i) $\boldsymbol{\Theta}_e = \mathbf{0}, \boldsymbol{\eta}_e = \mathbf{0}$, thus there is full observability (F) as there are no confounding effects, ii) each component of $\boldsymbol{\Theta}_e$ and $\boldsymbol{\eta}_e$ is drawn i.i.d. from $\mathcal{N}(0, 1)$ thus there is partial observability (P) as there are confounding effects.

• Each component of $\boldsymbol{\alpha}_e$ is drawn i.i.d from $\mathcal{N}(0, 1)$. $\varepsilon_e \sim \mathcal{N}(0, \sigma_{\varepsilon_e}^2)$ and each component of the vector $\boldsymbol{\zeta}_e$ is drawn from $\mathcal{N}(0, \sigma_{\boldsymbol{\zeta}_e}^2)$. We consider two settings for the noise variances – Homoskedastic (HOM) $\sigma_{\varepsilon_1} = 0.2$ and $\sigma_{\varepsilon_2} = 2.0$, $\sigma_{\boldsymbol{\zeta}_1} = \sigma_{\boldsymbol{\zeta}_2} = 1.0$ and Heteroskedastic (HET) $\sigma_{\boldsymbol{\zeta}_1} = 0.2$ and $\sigma_{\boldsymbol{\zeta}_2} = 2.0$, $\sigma_{\varepsilon_1} = \sigma_{\varepsilon_2} = 1.0$.

From the above, we gather that there are four possible combination of settings in which comparisons will be carried out – F-HOM, P-HOM, F-HET, P-HET. We use the following benchmarks in our comparison. IRM from [Arjovsky et al., 2019], ICP from [Peters et al., 2015], and standard ERM. Note in each of the cases we use a linear model. The code for our experiments can be found at `https://github.com/IBM/OoD`. All other implementation details can be found in the supplement. The performance is measured in terms of the model estimation error, i.e., the square of the distance from the ideal model $(\mathbf{1}_p, \mathbf{0}_q)$.

Before we discuss a comparison in all these settings, we look at a two dimensional experiment where $p = q = 1$ and the parameters are set to F-HOM. We carry out this comparison to illustrate several points. Firstly, we want to show why is ℓ_∞ constraint very important. Secondly, we want to show that the works when $\boldsymbol{\alpha}_e$ is non-zero, i.e., $\boldsymbol{X}_e^2$ is anti-causal (in the theory we had assumed $\boldsymbol{\alpha}_e = 0$). We compare with following variants of the linear regression game (LRG) i) no constraints,

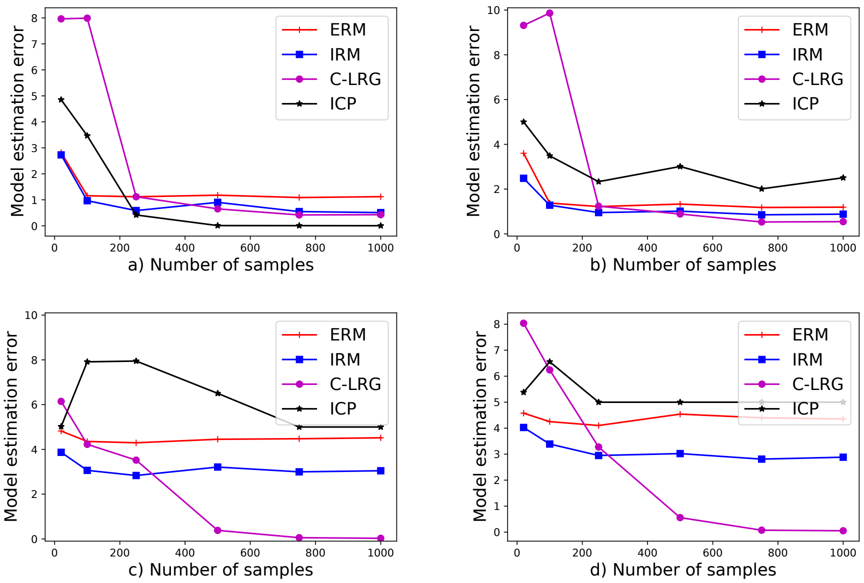

Figure 3: We compare across four settings: a) F-HET, b) P-HET, c) F-HOM and d) P-HOM.

Method	Test accuracy
Oracle	75
ERM	17.1 ± 0.60
IRM [Arjovsky et al., 2019]	66.90 ± 2.50
F-IRM game [Ahuja et al., 2020]	65.21 ± 1.56
Ours	$\mathbf{66.99 \pm 1.37}$

Table 2: Comparing test accuracies on colored MNIST.

which is the game U-LRG (Section 4.1), ii) regularize each R_e with ℓ_∞ penalty (R$_\infty$-LRG), and iii) regularize each R_e with ℓ_2 penalty (R$_2$-LRG). In Table 1, we show the estimated model against the respective method and the estimation error. Observe that C-LRG was able to outperform other variants of LRG. Moreover, C-LRG performed better than the other existing methods as well. w_{12} (w_{22}) are the coefficients that model 1 (2) associates with feature 2, which is spuriously correlated. We plot the trajectories of the coefficients w_{12} (w_{22}) of the models of each of the environments for the spurious features as the best response dynamics based training proceeds in Figure 2. Observe how the ℓ_∞ constrained models saturate on opposite ends of the boundary and as a result they cancel the spurious factors out. In contrast for other models, we do not see such an effect. Lastly, see if we choose a larger bound $w^{\mathsf{sup}} = 5$ the coefficients reach the boundary they just take more steps than $w^{\mathsf{sup}} = 2$.

Next, we move to a more elaborate comparison for the 10 dimensional setting from [Arjovsky et al., 2019] (we also show results for 100 dimensional setting in supplement). In Figure 3a, 3b, we show the model estimation error for F-HET and P-HET settings. In Figure 3c, 3d, we show the model estimation error as a function of the training samples for F-HOM and P-HOM settings. Observe that in each of the settings C-LRG performs better than the rest or is close to the best when the number of samples is more than 400.

5.2 Colored MNIST experiments

The entire discussion so far has been focused on linear SEMs. We now to move non-linear setups and carry out the colored MNIST (CMNIST) classification experiment from [Arjovsky et al., 2019]. In CMNIST the task is to classify the digits while ensuring the model does not rely on the background color. We use the ensemble-model construction from [Ahuja et al., 2020]. Each environment uses its own neural network (NN) and the ensemble model averages the logits from the different NNs. We use an ℓ_∞ constraint on the weights of the last layer of the NN. In Table 2, we show the comparisons of the different methods in terms of the test accuracy. We defer other details to the supplement.

6 Conclusion

In this work, we developed a new game-theoretic approach to learn OOD solutions for linear regressions. To the best of our knowledge, we have provided the first algorithms for which we can guarantee both convergence and better OOD behavior than standard empirical risk minimization. Experimentally too we see the promise of our approach as it is either competitive or outperforms the state-of-the-art by a margin.

Acknowledgements

We would like to thank Dr. Kush R. Varshney for the valuable discussions in the initial stages of this work.

References

[Ahuja et al., 2020] Ahuja, K., Shanmugam, K., Varshney, K., and Dhurandhar, A. (2020). Invariant risk minimization game. In *International Conference on Machine Learning, 2020.*

[Ajakan et al., 2014] Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. (2014). Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446.*

[Arjovsky et al., 2019] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893.*

[Bareinboim et al., 2012] Bareinboim, E., Brito, C., and Pearl, J. (2012). Local characterizations of causal Bayesian networks. In *Graph Structures for Knowledge Representation and Reasoning*, pages 1–17. Springer.

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

[Ben-David et al., 2007] Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization.* Cambridge university press.

[Chang et al., 2020] Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. S. (2020). Invariant rationalization. In *International Conference on Machine Learning, 2020.*

[Debreu, 1952] Debreu, G. (1952). A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893.

[DeGrave et al., 2020] DeGrave, A. J., Janizek, J. D., and Lee, S.-I. (2020). Ai for radiographic covid-19 detection selects shortcuts over signal. *medRxiv.*

[Duchi et al., 2016] Duchi, J., Glynn, P., and Namkoong, H. (2016). Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425.*

[Fudenberg et al., 1998] Fudenberg, D., Drew, F., Levine, D. K., and Levine, D. K. (1998). *The theory of learning in games*, volume 2. MIT press.

[Ganin et al., 2016] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

[Geirhos et al., 2020] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

[Glorot et al., 2011] Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the International Conference on Machine Learning*, pages 513–520.

[Gretton et al., 2009] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5.

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

[Hoffman et al., 2018] Hoffman, J., Mohri, M., and Zhang, N. (2018). Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pages 8246–8256.

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

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

[Koyama and Yamaguchi, 2020] Koyama, M. and Yamaguchi, S. (2020). Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883.*

[Krueger et al., 2020] Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Priol, R. L., and Courville, A. (2020). Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688.*

[Lee and Raginsky, 2018] Lee, J. and Raginsky, M. (2018). Minimax statistical learning with Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 2687–2696.

[Magliacane et al., 2018] Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856.

[Mahajan et al., 2020] Mahajan, D., Tople, S., and Sharma, A. (2020). Domain generalization using causal matching. *arXiv preprint arXiv:2006.07500*.

[Mohri et al., 2019] Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. *arXiv preprint arXiv:1902.00146*.

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

[Pearl, 2009] Pearl, J. (2009). *Causality*. Cambridge university press.

[Peters et al., 2015] Peters, J., Bühlmann, P., and Meinshausen, N. (2015). Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*.

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

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

[Subbaswamy et al., 2019] Subbaswamy, A., Chen, B., and Saria, S. (2019). Should I include this edge in my prediction? Analyzing the stability-performance tradeoff. *arXiv preprint arXiv:1905.11374*.

[Teney et al., 2020] Teney, D., Abbasnejad, E., and Hengel, A. v. d. (2020). Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*.