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 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

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 $x_{\text{Pa}(y)}$ if and only if in all environments where y has not been intervened on, the conditional probability $P(y|x_{\text{Pa}(y)})$ 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 ($\Phi(x) = x_{\text{Pa}(y)}$) 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 $u_i : \mathcal{S} \to \mathbb{R}$, 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 $e \in \{1, 2\}$, $\boldsymbol{\mu}_e = \mathbf{0}$ 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 $\boldsymbol{\mu}_e = \mathbf{0}$, $\boldsymbol{\Sigma}_e$ 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.*

The above model captures many different settings. If $\boldsymbol{\alpha}_e = \mathbf{0}$ and $\boldsymbol{\Theta}_e \neq \mathbf{0}$, then features $\boldsymbol{X}_e^2$ appear correlated with the label due to the confounder $\boldsymbol{H}_e$. If $\boldsymbol{\alpha_e} \neq \mathbf{0}$ and $\boldsymbol{\Theta}_e = \mathbf{0}$, then features $\boldsymbol{X}_e^2$ are correlated with the label but they are effects or anti-causal. If both $\boldsymbol{\alpha}_e \neq \mathbf{0}, \boldsymbol{\Theta}_e \neq \mathbf{0}$, 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 $(\boldsymbol{\gamma}, \mathbf{0})$ as it performs well across many distribution shifts (modeled by interventions). Hence, the goal is to learn $(\boldsymbol{\gamma}, \mathbf{0})$.

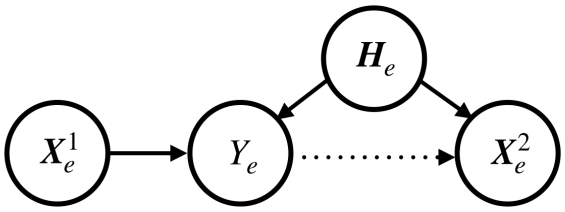

Figure 1: SEM from Assumption 2. We show the link between Y_e and $\boldsymbol{X}_e^2$ with a dotted line because in our theoretical analysis (Proposition 4,5) we assume that the edge does not exist but in the experiments we compare in the more general setting where such an edge exists.

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 $\boldsymbol{w}_e^* = (\boldsymbol{\gamma}, \mathbf{0})$, 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 $\boldsymbol{\alpha}_e = \mathbf{0}$, $\boldsymbol{\Theta}_e \neq 0$, i.e. confounders only setting. Define $\boldsymbol{\Sigma}_{e1} = \mathbb{E}_e\left[\boldsymbol{X}_e^1 \boldsymbol{X}_e^{1,\mathsf{T}}\right]$ and define the variance for the noise vector $\boldsymbol{\zeta}$ as $\boldsymbol{\sigma}_{\boldsymbol{\zeta}_e}^2 = \mathbb{E}_e[\boldsymbol{\zeta}_e \odot \boldsymbol{\zeta}_e]$, 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*

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

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 $\boldsymbol{w}^{\mathsf{inv}}_e = \boldsymbol{\gamma}$ and $\boldsymbol{w}^{\mathsf{var}}_e = \Big(\boldsymbol{\Theta}_e\boldsymbol{\Theta}_e^\mathsf{T} + \mathsf{diag}[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e}]\Big)^{-1}\boldsymbol{\Theta}_e\boldsymbol{\eta}_e$. Observe that the first half $\boldsymbol{w}^{\mathsf{inv}}_e$ is invariant, i.e., it does not depend on the environment, while $\boldsymbol{w}^{\mathsf{var}}_e$ may vary as it depends on the parameters specific to the environment e.g., $\boldsymbol{\Theta}_e, \boldsymbol{\eta}_e$. In general, $\boldsymbol{w}^{\mathsf{var}}_1 \neq \boldsymbol{w}^{\mathsf{var}}_2$ (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 $\boldsymbol{X}^2_e$ thus placing a non-zero weight on the second half corresponding to the features $\boldsymbol{X}^2_e$ 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, $\boldsymbol{X}^1_e$, with invariant coefficients, $\boldsymbol{w}^{\mathsf{inv}}_e$, from spurious features, $\boldsymbol{X}^2_e$, with variant coefficients $\boldsymbol{w}^{\mathsf{var}}_e$ 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.*** $\boldsymbol{w}^{\mathsf{var}}_1 \neq \boldsymbol{w}^{\mathsf{var}}_2$

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 $\mathcal{W} = \big\{\boldsymbol{w}_e \;\big|\; \|\boldsymbol{w}_e\|_\infty \leq w^{\mathsf{sup}}\big\}$, 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 $\boldsymbol{w}^\dagger_1, \boldsymbol{w}^\dagger_2$ be an NE of Γ_c and let $\bar{\boldsymbol{w}}^\dagger$ be the corresponding ensemble predictor, i.e. $\bar{\boldsymbol{w}}^\dagger = \boldsymbol{w}^\dagger_1 + \boldsymbol{w}^\dagger_2$. In the next theorem, we analyze the properties of $\boldsymbol{w}^\dagger_1, \boldsymbol{w}^\dagger_2$ 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 $\boldsymbol{X}_e = (X_{e1}, \ldots, X_{en})$ and the least squares optimal solution in environment e as $\boldsymbol{w}^*_e = (w_{e1}, \ldots, w_{en})$. Divide the features indexed $\{1, \ldots, n\}$ 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 , $\boldsymbol{w}^{\mathsf{inv}}_1 = \boldsymbol{w}^{\mathsf{inv}}_2 = \boldsymbol{\gamma} \implies \mathcal{U} = \{1, \ldots, p\}$, and $\boldsymbol{w}^{\mathsf{var}}_1 \neq \boldsymbol{w}^{\mathsf{var}}_2 \implies \mathcal{V} = \{p+1, \ldots, p+q\}$.

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 $j \in \{1, \ldots, n\}\backslash\{i\}$, 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 $\mathbf{\Theta}_1\boldsymbol{\eta}_1$ and $\mathbf{\Theta}_2\boldsymbol{\eta}_2$ 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 ($\boldsymbol{\alpha}_e = \mathbf{0}$, $\mathbf{\Theta}_e$ an orthogonal matrix), the ensemble predictor constructed from the NE is closer to the ideal OOD solution than ERM. For the second part, set $\mathbf{\Theta}_e = \boldsymbol{I}_q$, 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 $\mathbf{\Theta}_e$ 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 $\mathbf{\Theta}_e$ 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.

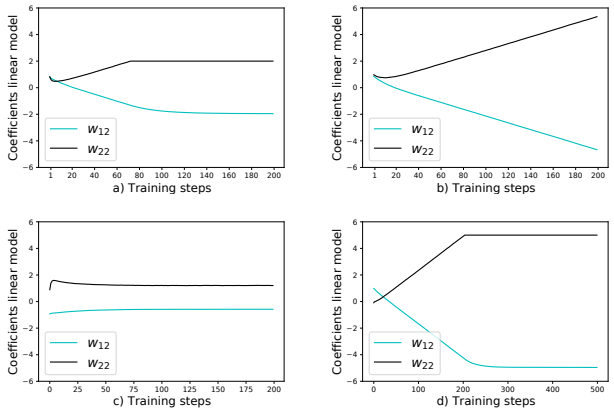

Figure 2: a) C-LRG ($w^{\mathsf{sup}} = 2$), b) U-LRG, c) R_∞-LRG , d) C-LRG ($w^{\mathsf{sup}} = 5$)

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_∞-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 $\left(\mathbf{1}_p, \mathbf{0}_q\right)$. Each component of the confounder $\boldsymbol{H}_e$ is drawn i.i.d. from $\mathcal{N}(0, \sigma^2_{\boldsymbol{H}_e})$. $\sigma_{\boldsymbol{H}_1} = 0.2$, $\sigma_{\boldsymbol{H}_2} = 2.0$. 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 $\mathcal{N}(0, 1)$. $\varepsilon_e \sim \mathcal{N}(0, \sigma^2_{\varepsilon_e})$ and each component of the vector $\boldsymbol{\zeta}_e$ is drawn from $\mathcal{N}(0, \sigma^2_{\boldsymbol{\zeta}_e})$. 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., $\boldsymbol{X}^2_e$ 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,

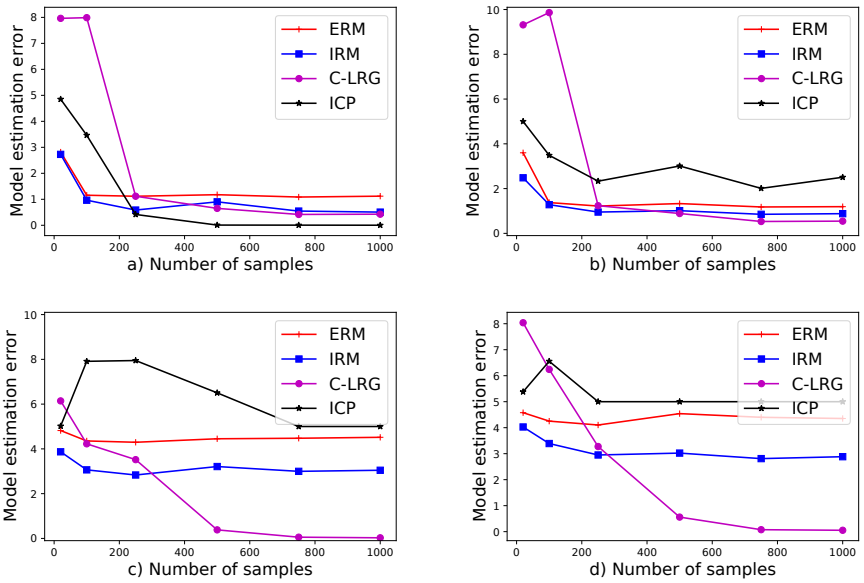

Figure 3: We compare across four settings: a) F-HET, b) P-HET, c) F-HOM and d) P-HOM.

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_∞-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.

7 Appendix

In this section, we provide the proofs to the propositions and theorems, and also provide other details on the experiments. We restate all the propositions and theorems for reader's convenience. In all our results, we use the following notation $\boldsymbol{a}$ is a vector, a_i is the i^{th} component of vector $\boldsymbol{a}$, A is a scalar random variable, $\boldsymbol{A}$ is a vector random variable, A_i is the i^{th} component of the random variable $\boldsymbol{A}$, $\mathcal{A}$ is a set, bold capitalized Greek letters e.g., $\boldsymbol{\Sigma}$ are used for matrices. $\boldsymbol{I}_m$ is a m dimensional identity matrix and $\mathbf{1}_m$ is a m dimensional vector of ones. A bar over a vector $\boldsymbol{w}$, $\bar{\boldsymbol{w}}$, denotes the ensemble predictor (sum of predictor from the two environments).

7.1 Proposition 1

We restate Proposition 1 below.

Proposition 6. *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.*

Proof. We start with latter part of the proposition. Suppose there exists a pair $\boldsymbol{w}_1^\dagger, \boldsymbol{w}_2^\dagger$ which is a NE of U-LRG. Observe that $R_e(\boldsymbol{w}_1, \boldsymbol{w}_2)$ is jointly convex in $\boldsymbol{w}_1, \boldsymbol{w}_2$ ($R_e(\boldsymbol{w}_1, \boldsymbol{w}_2) = \mathbb{E}_e[(Y_e - \boldsymbol{w}_1^\mathsf{T}\boldsymbol{X}_e - \boldsymbol{w}_2^\mathsf{T}\boldsymbol{X}_e)^2]$; loss inside the expectation is convex and expectation is a weighted sum over these losses). Let us compute the gradient of $R_e(\boldsymbol{w}_1, \boldsymbol{w}_2)$ w.r.t $\boldsymbol{w}_e$.

$$\begin{aligned}\nabla_{\boldsymbol{w}_1} R_1(\boldsymbol{w}_1, \boldsymbol{w}_2) &= 2\boldsymbol{\Sigma}_1(\boldsymbol{w}_1 + \boldsymbol{w}_2) - 2\boldsymbol{\rho}_1 \\ \nabla_{\boldsymbol{w}_2} R_2(\boldsymbol{w}_1, \boldsymbol{w}_2) &= 2\boldsymbol{\Sigma}_2(\boldsymbol{w}_1 + \boldsymbol{w}_2) - 2\boldsymbol{\rho}_2\end{aligned} \tag{7}$$

From the definition of pure strategy NE, it follows that $\boldsymbol{w}_1^\dagger$ ($\boldsymbol{w}_2^\dagger$) minimizes $R_1(\cdot, \boldsymbol{w}_2^\dagger)$ ($R_2(\boldsymbol{w}_1^\dagger, \cdot)$). From the convexity of $R_1(\cdot, \boldsymbol{w}_2^\dagger)$ and $R_2(\boldsymbol{w}_1^\dagger, \cdot)$ it follows that $\nabla_{\boldsymbol{w}_1|\boldsymbol{w}_1=\boldsymbol{w}_1^\dagger} R_1(\boldsymbol{w}_1, \boldsymbol{w}_2^\dagger) = 0$ and $\nabla_{\boldsymbol{w}_2|\boldsymbol{w}_2=\boldsymbol{w}_2^\dagger} R_2(\boldsymbol{w}_1^\dagger, \boldsymbol{w}_2) = 0$. Therefore, we have

$$\begin{aligned}&\boldsymbol{\Sigma}_1(\boldsymbol{w}_1^\dagger + \boldsymbol{w}_2^\dagger) - \boldsymbol{\rho}_1 = 0 \implies \\ &\boldsymbol{w}_1^\dagger + \boldsymbol{w}_2^\dagger = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\rho}_1 = \boldsymbol{w}_1^* \\ &\boldsymbol{\Sigma}_2(\boldsymbol{w}_1^\dagger + \boldsymbol{w}_2^\dagger) - \boldsymbol{\rho}_2 = 0 \implies \\ &\boldsymbol{w}_1^\dagger + \boldsymbol{w}_2^\dagger = \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\rho}_2 = \boldsymbol{w}_2^*\end{aligned} \tag{8}$$

In the above equation (8), we use Assumption 1 and the optimal solution defined in Section 4.1, $\boldsymbol{w}_e^* = \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}_e$, for each $e \in \{1, 2\}$. From equations (8) it follows that $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$. Therefore, the existence of NE implies $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$ or in other words if $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$ implies NE does not exist. In the above we learned that $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$ is a necessary condition for NE to exist. In the next part we show that this condition is sufficient as well. Suppose $\boldsymbol{w}_1^* = \boldsymbol{w}_2^* = \boldsymbol{w}^*$. Define any point $\hat{\boldsymbol{w}}_1$ and another point $\hat{\boldsymbol{w}}_2 = \boldsymbol{w}^* - \hat{\boldsymbol{w}}_1$. Compute $\nabla_{\boldsymbol{w}_1|\boldsymbol{w}_1=\hat{\boldsymbol{w}}_1} R_1(\boldsymbol{w}_1, \hat{\boldsymbol{w}}_2)$ and $\nabla_{\boldsymbol{w}_2|\boldsymbol{w}_2=\hat{\boldsymbol{w}}_2} R_2(\hat{\boldsymbol{w}}_1, \boldsymbol{w}_2)$. Using the expression in equation (7) we get

$$\begin{aligned}&\nabla_{\boldsymbol{w}_1|\boldsymbol{w}_1=\hat{\boldsymbol{w}}_1} R_1(\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2) = 2\boldsymbol{\Sigma}_1(\hat{\boldsymbol{w}}_1 + \hat{\boldsymbol{w}}_2) - 2\boldsymbol{\rho}_1 \\ &= 2\boldsymbol{\Sigma}_1\boldsymbol{w}^* - 2\boldsymbol{\rho}_1 = 0 \text{ (From the optimality of } \boldsymbol{w}^* \text{ for } R_1) \\ &\nabla_{\boldsymbol{w}_2|\boldsymbol{w}_2=\hat{\boldsymbol{w}}_2} R_2(\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2) = 2\boldsymbol{\Sigma}_2(\hat{\boldsymbol{w}}_1 + \hat{\boldsymbol{w}}_2) - 2\boldsymbol{\rho}_2 \\ &= 2\boldsymbol{\Sigma}_2\boldsymbol{w}^* - 2\boldsymbol{\rho}_2 = 0 \text{ (From the optimality of } \boldsymbol{w}^* \text{ for } R_2)\end{aligned} \tag{9}$$

From the convexity of R_1 and R_2 it follows that $\hat{\boldsymbol{w}}_1$, $\hat{\boldsymbol{w}}_2$ simultaneously minimize R_1 and R_2. Therefore, every such $\hat{\boldsymbol{w}}_1$ and $\hat{\boldsymbol{w}}_2$ that sum to $\boldsymbol{w}^*$ form a NE. This completes the proof. □

7.2 Proposition 2

We restate Proposition 2 below.

Proposition 7. *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*

$$\boldsymbol{w}_e^* = (\boldsymbol{w}_e^{\mathsf{inv}}, \boldsymbol{w}_e^{\mathsf{var}}) = \left(\boldsymbol{\gamma}, \left(\boldsymbol{\Theta}_e\boldsymbol{\Theta}_e^\mathsf{T} + \mathsf{diag}[\boldsymbol{\sigma}_{\boldsymbol{\zeta}_e}^2]\right)^{-1}\boldsymbol{\Theta}_e\boldsymbol{\eta}_e\right) \tag{10}$$

Proof. We derive the expression for the optimal predictor in the confounder only SEM in Assumption 2. Recall that the general expression for the least squares optimal predictor (defined in Section 4.1) is

$$\boldsymbol{w}_e^* = \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}_e \tag{11}$$

We use the SEM in Assumption 2 to derive an expression for $\boldsymbol{\Sigma}_e$. First observe that from Assumption 2, we have that $\mathbb{E}_e[\boldsymbol{X}_e^1] = 0$ and

$$\mathbb{E}_e\big[Y_e\big] = \boldsymbol{\gamma}^\mathsf{T}\mathbb{E}_e\big[\boldsymbol{X}_e^1\big] + \boldsymbol{\eta}_e^\mathsf{T}\mathbb{E}_e\big[\boldsymbol{H}_e\big] + \mathbb{E}_e\big[\varepsilon_e\big] = 0$$

$$\mathbb{E}_e\big[\boldsymbol{X}_e^2\big] = \boldsymbol{\alpha}_e\mathbb{E}_e\big[Y_e\big] + \boldsymbol{\Theta}_e\mathbb{E}_e\big[\boldsymbol{H}_e\big] + \mathbb{E}_e\big[\boldsymbol{\zeta}_e\big] = \mathbf{0}$$

Therefore

$$\mathbb{E}_e\big[\boldsymbol{X}_e^1\big] = 0, \mathbb{E}_e\big[\boldsymbol{X}_e^2\big] = 0 \tag{12}$$

We divide $\boldsymbol{\Sigma}_e$ into four smaller matrices $\boldsymbol{\Sigma}_{e1} = \boldsymbol{E}_e[\boldsymbol{X}_e^1\boldsymbol{X}_e^{1,\mathsf{T}}]$, $\boldsymbol{\Sigma}_{e2} = \boldsymbol{E}_e[\boldsymbol{X}_e^2\boldsymbol{X}_e^{2,\mathsf{T}}]$, $\boldsymbol{\Sigma}_{e12} = \boldsymbol{E}_e[\boldsymbol{X}_e^1\boldsymbol{X}_e^{2,\mathsf{T}}]$ and $\boldsymbol{\Sigma}_{e21} = \boldsymbol{E}_e[\boldsymbol{X}_e^2\boldsymbol{X}_e^{1,\mathsf{T}}]$.

From Assumption 2, we know $(\boldsymbol{H}_e, \boldsymbol{\zeta}_e) \perp \boldsymbol{X}_e^1$ and $\boldsymbol{X}_e^2 \leftarrow \boldsymbol{\Theta}_e \boldsymbol{H}_e + \boldsymbol{\zeta}_e$, which implies $\boldsymbol{X}_e^2 \perp \boldsymbol{X}_e^1$.

Therefore, from $\boldsymbol{X}_e^2 \perp \boldsymbol{X}_e^1$ and equation (12) it follows that

$$\boldsymbol{\Sigma}_{e21} = \mathbb{E}_e\Big[\boldsymbol{X}_e^2 \boldsymbol{X}_e^{1,\mathsf{T}}\Big] = \mathbb{E}_e\Big[\boldsymbol{X}_e^2\Big]\mathbb{E}_e\Big[\boldsymbol{X}_e^{1,\mathsf{T}}\Big] = \boldsymbol{0}_{q\times p} \tag{13}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{e2} &= \mathbb{E}_e\Big[\boldsymbol{X}_e^2 \boldsymbol{X}_e^{2,\mathsf{T}}\Big] = \boldsymbol{\Theta}_e \mathbb{E}_e\Big[\boldsymbol{H}_e \boldsymbol{H}_e^{\mathsf{T}}\Big]\boldsymbol{\Theta}_e^{\mathsf{T}} + \\ &\boldsymbol{\Theta}_e \mathbb{E}_e\Big[\boldsymbol{H}_e \boldsymbol{\zeta}_e^{\mathsf{T}}\Big] + \mathbb{E}_e\Big[\boldsymbol{\zeta}_e \boldsymbol{H}_e^{\mathsf{T}}\Big]\boldsymbol{\Theta}_e^{\mathsf{T}} + \mathbb{E}\Big[\boldsymbol{\zeta}_e \boldsymbol{\zeta}_e^{\mathsf{T}}\Big] \\ &= \boldsymbol{\Theta}_e \boldsymbol{\Theta}_e^{\mathsf{T}} + \mathsf{diag}[\boldsymbol{\sigma}_{\boldsymbol{\eta}_e}^2]\end{aligned} \tag{14}$$

In the above equation (14), we use $\mathbb{E}_e\Big[\boldsymbol{H}_e \boldsymbol{H}_e^{\mathsf{T}}\Big] = \boldsymbol{I}_s$ and $\mathbb{E}_e\Big[\boldsymbol{H}_e \boldsymbol{\zeta}_e^{\mathsf{T}}\Big] = \boldsymbol{0}_{s\times q}$, which follow from Assumption 2. From Assumption 3, we know that $\boldsymbol{\sigma}_{\boldsymbol{\eta}_e}^2 > \boldsymbol{0}$ and we use this observation in equation (14) to deduce that $\boldsymbol{\Sigma}_{e2}$ is positive definite.

From equation (13) we can simplify $\boldsymbol{\Sigma}_e$ into a block diagonal matrix written as $\mathsf{diag}\Big[\boldsymbol{\Sigma}_{e1}, \boldsymbol{\Sigma}_{e2}\Big]$, where $\boldsymbol{\Sigma}_{e1} = \mathbb{E}_e\Big[\boldsymbol{X}_e^1, \boldsymbol{X}_e^{1,\mathsf{T}}\Big]$ and $\boldsymbol{\Sigma}_{e2} = \mathbb{E}_e\Big[\boldsymbol{X}_e^2, \boldsymbol{X}_e^{2,\mathsf{T}}\Big]$.

From Assumption 3, $\boldsymbol{\Sigma}_{e1}$ is positive definite and we showed above that $\boldsymbol{\Sigma}_{e2}$ is positive definite as well. Therefore, we can write the inverse of $\boldsymbol{\Sigma}_e$ as another block diagonal matrix written as

$$\boldsymbol{\Sigma}_e^{-1} = \mathsf{diag}\Big[\boldsymbol{\Sigma}_{e1}^{-1}, \boldsymbol{\Sigma}_{e2}^{-1}\Big] \tag{15}$$

Next let us simplify $\boldsymbol{\rho}_e = \Big[\mathbb{E}_e\Big[\boldsymbol{X}_e^1 Y_e\Big], \mathbb{E}_e\Big[\boldsymbol{X}_e^2 Y_e\Big]\Big]$.

$$\mathbb{E}_e\Big[\boldsymbol{X}_e^1 Y_e\Big] = \mathbb{E}_e\Big[\boldsymbol{X}_e^1 \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{X}_e^1 + \boldsymbol{\eta}_e^{\mathsf{T}} \boldsymbol{H}_e + \varepsilon_e\Big] = \boldsymbol{\Sigma}_{e1}\boldsymbol{\gamma} \tag{16}$$

$$\begin{aligned}&\mathbb{E}_e\Big[\boldsymbol{X}_e^2 Y_e\Big] = \\ &\mathbb{E}_e\Big[\boldsymbol{X}_e^2(\boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{X}_e^1 + \boldsymbol{\eta}_e^{\mathsf{T}} \boldsymbol{H}_e + \varepsilon_e)\Big] = \mathbb{E}_e[\boldsymbol{X}_e^2 \boldsymbol{\eta}_e^{\mathsf{T}} \boldsymbol{H}^e\Big] \\ &\mathbb{E}_e\Big[\boldsymbol{X}_e^2 \boldsymbol{\eta}_e^{\mathsf{T}} \boldsymbol{H}^e\Big] = \mathbb{E}_e\Big[\boldsymbol{\Theta}_e \boldsymbol{H}_e \boldsymbol{\eta}_e^{\mathsf{T}} \boldsymbol{H}^e\Big] = \boldsymbol{\Theta}_e \mathbb{E}_e\Big[\boldsymbol{H}_e \boldsymbol{H}_e^{\mathsf{T}}\Big]\boldsymbol{\eta}_e \\ &= \boldsymbol{\Theta}_e \boldsymbol{\eta}_e \ \Big(\text{Since } \mathbb{E}_e\Big[\boldsymbol{H}_e \boldsymbol{H}_e^{\mathsf{T}}\Big] = \boldsymbol{I}_s\Big)\end{aligned} \tag{17}$$

Combining equations (11)- (17),

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

This completes the derivation. □

7.3 Theorem 2

We first state a lemma needed for proving Theorem 2.

Lemma 1. *Suppose Assumptions 1 and 5 hold. Consider the case when $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$. In this case, at least one of the predictors in the NE of C-LRG $\boldsymbol{w}_1^\dagger$ or $\boldsymbol{w}_2^\dagger$ has to be on the boundary of the set, i.e. for at least one $e \in \{1, 2\}$, $\|\boldsymbol{w}_e^\dagger\|_\infty = w^{\mathsf{sup}}$. Moreover, if $\|\boldsymbol{w}_1^\dagger\|_\infty < w^{\mathsf{sup}}$ ($\|\boldsymbol{w}_2^\dagger\|_\infty < w^{\mathsf{sup}}$) and $\|\boldsymbol{w}_2^\dagger\|_\infty = w^{\mathsf{sup}}$ ($\|\boldsymbol{w}_1^\dagger\|_\infty = w^{\mathsf{sup}}$) then the ensemble predictor is optimal for environment e, i.e., $\bar{\boldsymbol{w}}^\dagger = \boldsymbol{w}_2^*$ ($\bar{\boldsymbol{w}}^\dagger = \boldsymbol{w}_1^*$).*

Proof. We start with the first part of the above lemma. In the first part, the only case that is excluded is when both the points forming the NE are in the interior, i.e., $\|\boldsymbol{w}_1^\dagger\|_\infty < w^{\mathsf{sup}}$ and $\|\boldsymbol{w}_2^\dagger\|_\infty < w^{\mathsf{sup}}$. Denote $\boldsymbol{w}_{-e}$ as the predictor used by the environment $q \in \{1, 2\} \setminus \{e\}$. We interhangeably use $R_e(\boldsymbol{w}_e, \boldsymbol{w}_{-e})$ and $R_e(\boldsymbol{w}_1, \boldsymbol{w}_2)$. For environment e, from the definition of NE, it follows that $\boldsymbol{w}_e^\dagger$ satisfies $\boldsymbol{w}_e^\dagger \in \arg\min_{\boldsymbol{w}_e \in \mathcal{W}} R_e(\boldsymbol{w}_e, \boldsymbol{w}_{-e}^\dagger)$. Note i) $R_e(\boldsymbol{w}_e, \boldsymbol{w}_{-e}^\dagger)$ is a convex function in $\boldsymbol{w}_e$, and ii) the set $\mathcal{W}$ has a non-empty relative interior (Since $w^{\mathsf{sup}} > 0$). From these two conditions it follows that Slater's constraint qualification is satisfied, which implies strong duality holds [Boyd and Vandenberghe, 2004]. From strong duality, it follows that $\boldsymbol{w}_e^\dagger$ and $\lambda_e^\dagger$, where $\lambda_e^\dagger$ is the dual variable for the constraint $\|\boldsymbol{w}_e\|_\infty \leq w^{\mathsf{sup}}$, satisfy the KKT conditions given as follows

$$\begin{aligned}&\|\boldsymbol{w}_e^\dagger\| \leq w^{\mathsf{sup}} \\ &\lambda_e^\dagger \geq 0 \\ &\lambda_e^\dagger\big(\|\boldsymbol{w}_e^\dagger\| - w^{\mathsf{sup}}\big) = 0 \\ &0 \in \nabla_{\boldsymbol{w}_e^\dagger} R_e(\boldsymbol{w}_e^\dagger, \boldsymbol{w}_{-e}^\dagger) + \lambda_e^\dagger \partial(\|\boldsymbol{w}_e^\dagger\|_\infty)\end{aligned} \tag{18}$$

In the above $\partial(\|\boldsymbol{w}_e^\dagger\|_\infty)$ represents the subdifferential of $\|\cdot\|_\infty$ at $\boldsymbol{w}_e^\dagger$. If $\|\boldsymbol{w}_1^\dagger\|_\infty < w^{\mathsf{sup}}$ and $\|\boldsymbol{w}_2^\dagger\|_\infty < w^{\mathsf{sup}}$, then $\lambda_1^\dagger$ and $\lambda_2^\dagger$ are both zero. As a result, we have for $e \in \{1, 2\}$, $\nabla_{\boldsymbol{w}_e^\dagger} R_e(\boldsymbol{w}_e^\dagger, \boldsymbol{w}_{-e}^\dagger) = 0$. From the expression of the gradients in (7) we have for each $e \in \{1, 2\}$

$$\begin{aligned}&\nabla_{\boldsymbol{w}_e^\dagger} R_e(\boldsymbol{w}_e^\dagger, \boldsymbol{w}_{-e}^\dagger) = 2\boldsymbol{\Sigma}_e(\boldsymbol{w}_e^\dagger + \boldsymbol{w}_{-e}^\dagger) - 2\boldsymbol{\rho}^e = 0 \\ &\implies \boldsymbol{w}_1^\dagger + \boldsymbol{w}_2^\dagger = \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}^e = \boldsymbol{w}_e^*\end{aligned} \tag{19}$$

From equation (19) it follows that $\boldsymbol{w}_1^* = \boldsymbol{w}_2^*$, which contradicts the assumption $\boldsymbol{w}_1^* \neq \boldsymbol{w}_2^*$. This completes the proof for the first part of the Lemma.

Next, we move to the latter part of the proof, which states that if $\|\boldsymbol{w}_e^\dagger\|_\infty < w^{\mathsf{sup}}$ and $\|\boldsymbol{w}_{-e}^\dagger\|_\infty = w^{\mathsf{sup}}$, then the ensemble predictor is optimal for environment e, i.e., $\bar{\boldsymbol{w}}^\dagger = \boldsymbol{w}_e^*$. Since $\|\boldsymbol{w}_e^\dagger\|_\infty < w^{\mathsf{sup}}$, from the KKT conditions above in (18) we have that $\lambda_e^\dagger = 0$, which implies that $\nabla_{\boldsymbol{w}_e^\dagger} R_e(\boldsymbol{w}_e^\dagger, \boldsymbol{w}_{-e}^\dagger) = 0$. Using the expression for gradient in equation (7), we have that $\boldsymbol{w}_e^\dagger + \boldsymbol{w}_{-e}^\dagger = \bar{\boldsymbol{w}}^\dagger = \boldsymbol{w}_e^*$. This completes the proof. □

We restate Theorem 2 for reader's convenience.

Theorem 4. *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*

$$\Big(\boldsymbol{w}_1^* \odot \mathbf{1}_{|\boldsymbol{w}_2^*| \geq |\boldsymbol{w}_1^*|} + \boldsymbol{w}_2^* \odot \mathbf{1}_{|\boldsymbol{w}_1^*| > |\boldsymbol{w}_2^*|}\Big)\mathbf{1}_{\boldsymbol{w}_1^* \odot \boldsymbol{w}_2^* \geq \mathbf{0}} \quad (20)$$

Proof. Recall in Section 4.2, we divided the features $\{1, \ldots, n\}$ into two sets $\mathcal{U}$ and $\mathcal{V}$. Without loss of generality assume that the first k components in $\boldsymbol{X}_e$ belong to $\mathcal{U}$ and the next $n-k$ components to be in $\mathcal{V}$. Therefore, $\mathcal{U} = \{1, \ldots, k\}$ and $\mathcal{V} = \{k+1, \ldots, n\}$. Define $\boldsymbol{X}_{e+} = (X_{e1}, \ldots, X_{ek})$ and $\boldsymbol{X}_{e-} = (X_{e(k+1)}, \ldots, X_{en})$. We divide the weights in $\boldsymbol{w}_e = (w_{e1}, \ldots, w_{en})$ into two parts where the weights associated with the first k components, $\boldsymbol{X}_{e+}$, are $\boldsymbol{w}_{e+} = (w_{e1}, \ldots, w_{ek})$ and the weights associated with the next $n-k$ components, $\boldsymbol{X}_{e-}$, are $\boldsymbol{w}_{e-} = (w_{e(k+1)}, \ldots, w_{en})$. Similarly, we divide the vector $\boldsymbol{\rho}_e$ defined in Section 4.1 into $\boldsymbol{\rho}_{e+}$ and $\boldsymbol{\rho}_{e-}$.

Define $\boldsymbol{\Sigma}_{e+} = \mathbb{E}_e\big[\boldsymbol{X}_{e+}\boldsymbol{X}_{e+}^\mathsf{T}\big]$ and define $\boldsymbol{\Sigma}_{e-} = \mathbb{E}_e\big[\boldsymbol{X}_{e-}\boldsymbol{X}_{e-}^\mathsf{T}\big]$. As a consequence of the Assumption 6, we can simplify the expression for $\boldsymbol{\Sigma}_e$ as follows

$$\boldsymbol{\Sigma}_e = \mathsf{diag}\Big[\boldsymbol{\Sigma}_{e+}, \boldsymbol{\Sigma}_{e-}\Big] \quad (21)$$

For each $e \in \{1, 2\}$, each feature component $i \in \{1, \ldots, n\}$ has a mean zero $\mathbb{E}_e\big[X_{ei}\big] = 0$. Therefore, the variance in each feature component $i \in \{1, \ldots, n\}$ is $\sigma_{ei}^2 = \mathbb{E}_e\big[X_{ei}^2\big]$. We can further simplify $\boldsymbol{\Sigma}_{e-}$. Using Assumption 6, we have that $\boldsymbol{\Sigma}_{e-}$ is a diagonal matrix, which we write as

$$\boldsymbol{\Sigma}_{e-} = \mathsf{diag}[(\sigma_{em}^2)_{m=k+1}^n]\Big] \quad (22)$$

We use equations (21) and (22) and the notation introduced above to simplify the risk as follows.

$$\begin{aligned} &R_e(\boldsymbol{w}_1, \boldsymbol{w}_2) = \\ &(\boldsymbol{w}_1 + \boldsymbol{w}_2)^\mathsf{T}\boldsymbol{\Sigma}_e(\boldsymbol{w}_1 + \boldsymbol{w}_2) - \boldsymbol{\rho}_e^\mathsf{T}(\boldsymbol{w}_1 + \boldsymbol{w}_2) + \mathbb{E}_e\big[Y_e^2\big] = \\ &(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+})^\mathsf{T}\boldsymbol{\Sigma}_{e+}(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) - \boldsymbol{\rho}_{e+}^\mathsf{T}(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) + \\ &\sum_{i=k+1}^n \big((w_{1i} + w_{2i})^2\sigma_{ei}^2 - 2(w_{1i} + w_{2i})\rho_{ei}\big) + \mathbb{E}_e\big[Y_e^2\big] \end{aligned} \quad (23)$$

Recall that $\boldsymbol{w}_e^* = \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\rho}_e$ (defined in Section 4.1). From the above equations (21), (22) and Assumption 1, we get

$$\begin{aligned} \boldsymbol{w}_{e+}^* &= \boldsymbol{\Sigma}_{e+}^{-1}\boldsymbol{\rho}_{e+} \\ \boldsymbol{w}_{e-}^* &= \Big[\frac{\rho_{ei}}{\sigma_{ei}^2}\Big]_{i \in \{k+1, \ldots, n\}} \end{aligned} \quad (24)$$

where $\boldsymbol{w}_{e+}^*$ is the vector of the first k components in $\boldsymbol{w}_e^*$, $\boldsymbol{w}_{e-}^*$ are the next $n-k$ components in $\boldsymbol{w}_e^*$, ρ_{ei}, is the i^{th} component of $\boldsymbol{\rho}_e$ and σ_{ei}^2 is the variance in X_{ei}.

Recall that the first k components comprise the set $\mathcal{U}$, which is defined as the set where the features of the least squares coefficients are the same across environments, i.e.,

$$\boldsymbol{w}_{1+}^* = \boldsymbol{w}_{2+}^* \quad (25)$$

Define

$$\begin{aligned} &R_{e+}(\boldsymbol{w}_{1+}, \boldsymbol{w}_{2+}) = (\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+})^\mathsf{T}\boldsymbol{\Sigma}_{e+}(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) - \\ &\boldsymbol{\rho}_{e+}^\mathsf{T}(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) + \mathbb{E}_e\big[Y_e^2\big] \end{aligned} \quad (26)$$

For each $i \in \mathcal{V} = \{k+1, \ldots, n\}$ define

$$R_{ei}(w_1, w_2) = \big((w_{1i} + w_{2i})^2 - 2(w_{1i} + w_{2i})w_{ei}^*\big) \quad (27)$$

We use the above equations (26) and (27) to simplify the risks as follows

$$\begin{aligned} &\min_{\boldsymbol{w}_e \in \mathcal{W}} R_e(\boldsymbol{w}_1, \boldsymbol{w}_2) = \min_{\boldsymbol{w}_{e+} \in \mathcal{W}_+} R_{e+}(\boldsymbol{w}_{1+}, \boldsymbol{w}_{2+}) + \\ &\sum_{i=k+1}^n \sigma_{ei}^2 \min_{|w_{ei}| \leq w^{\mathsf{sup}}} R_{ei}(w_{1i}, w_{2i}) \end{aligned} \quad (28)$$

In the above $\mathcal{W}_+ = \big\{\boldsymbol{w} \mid \boldsymbol{w} \in \mathbb{R}^k, \|\boldsymbol{w}\|_\infty \leq w^{\mathsf{sup}}\big\}$. From the above expression in equation (28), we see that the the optimization for environment e can be decomposed into separate smaller minimizations, which we analyze separately next. ℓ_∞ norm constraints allows to make the problem in equation (28) separable and for other norms such separability is not possible. Henceforth, we will look at each smaller minimization as a separate game between the environments.

Let us consider the first minimization in equation (28)

$$\min_{\boldsymbol{w}_{e+} \in \mathcal{W}_{1+}} R_{e+}(\boldsymbol{w}_{1+}, \boldsymbol{w}_{2+}) \quad (29)$$

Let us minimize the objective in equation (29) without imposing the constraint that $\boldsymbol{w}_{1+} \in \mathcal{W}_{1+}$

$$\begin{aligned} &\boldsymbol{\Sigma}_{e+}(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) = \boldsymbol{\rho}_{e+} \\ &(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) = \boldsymbol{\Sigma}_{e+}^{-1}\boldsymbol{\rho}_{e+} = \boldsymbol{w}_{e+}^* \text{ (From equation (24))} \end{aligned} \quad (30)$$

Therefore, we have that if $(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) = \boldsymbol{w}^*_{e+}$, then environment e achieves the minimum risk possible and cannot do any better. In fact, from equation (24) since $\boldsymbol{w}^*_{1+} = \boldsymbol{w}^*_{2+}$, if $(\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+}) = \boldsymbol{w}^*_{1+}$, then both environments are at the minimum and cannot do any better. Therefore, we know that all the elements in the set $\mathcal{C} = \{\boldsymbol{w}_{1+}, \boldsymbol{w}_{2+} \mid \boldsymbol{w}_{1+} \in \mathcal{W}^+,\ \boldsymbol{w}_{1+} + \boldsymbol{w}_{2+} = \boldsymbol{w}^*_{1+}\}$ form a NE of C-LRG. From Assumption 5, we know that this set $\mathcal{C}$ is non-empty. Moreover, there are no points outside this set $\mathcal{C}$ which form a NE. If $\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+} \neq \boldsymbol{w}^*_{1+}$, then the gradient will not be zero for either of the environments and both would prefer to move to a point where their gradients are zero. Hence, in every NE, $\boldsymbol{w}_{1+} + \boldsymbol{w}_{2+} = \boldsymbol{w}^*_{1+}$. If we use the expression in equation (6) and compute first k components it returns vector $\boldsymbol{w}^*_{1+}$ (we use the condition $\boldsymbol{w}^*_{1+} = \boldsymbol{w}^*_{2+}$ to simplify the expression in equation (6)). This shows that the expression in equation (6) correctly characterizes the NE for the first k components that make up the set $\mathcal{U}$. We now move to the remaining $n-k$ components that make up the set $\mathcal{V}$.

Consider a component $i \in \mathcal{V} = \{k+1, \ldots, n\}$. Environment e is interested in minimizing R_{ei} defined in equation (26). Let us consider the i^{th} component of the expression in equation (6) in Theorem 2 and rewrite the expression in terms of scalars.

$$\Big(w^*_{1i} 1_{|w^*_{2i}| \geq |w^*_{1i}|} + w^*_{2i} 1_{|w^*_{1i}| > |w^*_{2i}|}\Big) 1_{w^*_{1i} w^*_{2i} \geq 0} \tag{31}$$

We divide the analysis into two cases. In the first case, the signs of w^*_{1i} and w^*_{2i} disagree, which implies $1_{w^*_{1i} w^*_{2i} \geq 0}$ is zero. In the the second case, the signs of w^*_{1i} and w^*_{2i} agree, which implies $1_{w^*_{1i} w^*_{2i} \geq 0}$ is one. Let us start with the first case. Without loss of generality say $w^*_{1i} < 0$ and $w^*_{2i} > 0$. Suppose $\bar{w}^\dagger_i > 0$, where $\bar{w}^\dagger_i$ is the i^{th} component of the NE based predictor

$$\begin{aligned} &\bar{w}^\dagger_i > 0 \implies w^\dagger_{1i} + w^\dagger_{2i} > 0 \\ &\implies w^\dagger_{1i} > -w^\dagger_{2i}\ (w^\dagger_{2i} > -w^\dagger_{1i}) \\ &w^\dagger_{1i} > -w^{\mathsf{sup}}\ (w^\dagger_{2i} > -w^{\mathsf{sup}}) \end{aligned} \tag{32}$$

Observe that

$$\frac{\partial R_{1i}(w_{1i}, w^\dagger_{2i})}{\partial w_{1i}}\Big|_{w_{1i} = w^\dagger_{1i}} = 2(\bar{w}^\dagger_i - w^*_{1i}) > 0$$

Since $w^\dagger_{1i} > -w^{\mathsf{sup}}$ $\big($from equation (32)$\big)$, $w^\dagger_{1i}$ can be decreased and improve the utility for environment 1, which contradicts that $w^\dagger_{1i}$ is NE. Suppose $\bar{w}^\dagger_i < 0$, then from symmetry we can show that one of the environments will be able to increase the weight and improve its utility.

Therefore, the only option that remains $\bar{w}^\dagger_i = 0 \implies w^\dagger_{1i} = -w^\dagger_{2i}$.

Observe that $\frac{\partial R_{1i}(w_{1i}, w^\dagger_{2i})}{\partial w_{1i}}\big|_{w_{1i} = w^\dagger_{1i}} = 2(-w^*_{1i}) > 0$ and if $w^\dagger_{1i} > -w^{\mathsf{sup}}$ environment 1 will want to decrease $w^\dagger_{1i}$.

Observe that $\frac{\partial R_{2i}(w^\dagger_{1i}, w_{2i})}{\partial w_{2i}}\big|_{w_{2i} = w^\dagger_{2i}} = 2(-w^*_{2i}) < 0$ and if $w^\dagger_{2i} < w^{\mathsf{sup}}$ environment 2 will want to increase $w^\dagger_{2i}$.

Hence, the only solution left is for environment 1 to be at $-w^{\mathsf{sup}}$ and environment 2 to be at w^{sup}. Environment 1's (2's) risk decreases (increases) as it moves closer to its optimal point w^*_{1i} (w^*_{2i}). When environment 2 uses w^{sup}, environment 1's best response is to use $-w^{\mathsf{sup}}$ as it brings the environment 1 the closest it can get to w^*_{1i}. Therefore, $(w^{\mathsf{sup}}, -w^{\mathsf{sup}})$ is a NE. This completes the first case, i.e., when the coefficients have opposite signs the coefficient of the NE based ensemble predictor for that component is 0, which is what equation (6) states.

Next, consider the case when the signs of w^*_{1i} and w^*_{2i} agree. Let us consider the case when both have positive signs and the negative sign case will follow from symmetry. Suppose $0 < w^*_{1i} < w^*_{2i}$. From Lemma 1, we know that there are three scenarios possible.

In the first scenario, both $w^\dagger_{1i}$ and $w^\dagger_{2i}$ are on the same side of the boundary, say both are at w^{sup}.

$$\begin{aligned} &\frac{\partial R_{1i}(w_{1i}, w^\dagger_{2i})}{\partial w_{1i}}\Big|_{w_{1i} = w^\dagger_{1i}} = 2(2w^{\mathsf{sup}} - w^*_{1i}) \\ &\frac{\partial R_{2i}(w^\dagger_{1i}, w_{2i})}{\partial w_{2i}}\Big|_{w_{2i} = w^\dagger_{2i}} = 2(2w^{\mathsf{sup}} - w^*_{2i}) \end{aligned} \tag{33}$$

From Assumption 5, $0 < w^*_{1i} < w^*_{2i} \leq w^{\mathsf{sup}}$. Thus for both $e \in \{1, 2\}$, from equation (33) it follows that decreasing $w^\dagger_{ei}$ from the current state would improve the utility thus contradicting that they form a NE. The other possibility is that the two are on the other sides of the boundary, which makes $\frac{\partial R_{1i}(w_{1i}, w^\dagger_{2i})}{\partial w_{1i}}\big|_{w_{1i} = w^\dagger_{1i}}$ and $\frac{\partial R_{2i}(w^\dagger_{1i}, w_{2i})}{\partial w_{2i}}\big|_{w_{2i} = w^\dagger_{2i}}$ negative (from Assumption 5); thus prompting each player on the negative side of the boundary to increase the weight and improve its utility, which contradicts the fact that they form a NE.

The other possibility arising out of Lemma 2 is $w^\dagger_{1i} = w^{\mathsf{sup}}$ and $w^\dagger_{2i} = w^*_{2i} - w^{\mathsf{sup}}$. In this case, the $\frac{\partial R_{1i}(w_{1i}, w^\dagger_{2i})}{\partial w_{1i}}\big|_{w_{1i} = w^\dagger_{1i}}$ is positive implying environment 1 can decrease and improve its utility. Thus this state is not a NE.

Therefore, the only remaining possibility is $w^\dagger_{2i} = w^{\mathsf{sup}}$ and $w^\dagger_{1i} = w^*_{1i} - w^{\mathsf{sup}}$. In this case, the $\frac{\partial R_{2i}(w^\dagger_{1i}, w_{2i})}{\partial w_{2i}}\big|_{w_{2i} = w^\dagger_{2i}}$ is negative, environment 2 cannot increase the weight further as it is already at the

boundary (playing w^{sup} is a best response of environment 2 brings it closest to the desired w^*_{2i}). Hence, this state is a NE and the ensemble predictor is at w^*_{1i}.

If we suppose, $w^*_{2i} < w^*_{1i} < 0$. In this case, we can follow the exact same line of reasoning and arrive at the conclusion that the only NE is $\bar{w}^\dagger = w^*_{1i}$.

We have analyzed all the possible cases when $|w^*_{1i}| < |w^*_{2i}|$ and both w^*_{1i} and w^*_{2i} have the same sign. This completes the proof for the first term in the expression in equation (6)

$$w^*_{1i} 1_{|w^*_{2i}| \geq |w^*_{1i}|}$$

The second term is same as the first term in equation (6) with the roles of environments swapped. Therefore, due to symmetry we do not need to work out the second term separately. This completes the analysis for all the cases in the equation (6) in Theorem 2.

□

7.4 Proposition 4

We restate Proposition 4 below.

Proposition 8. *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 $\boldsymbol{w}^{\mathsf{ERM}} \in \mathcal{S}^{\mathsf{ERM}}$.* [3] *Moreover, if all 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. We first show that the Assumptions made in the above proposition imply that the Assumptions needed for Theorem 2 to be true hold.

We show that Assumptions 2, 3 $\implies$ Assumption 1 holds. $\boldsymbol{X}^e_1$ is zero mean (from Assumption 2) and

$$\mathbb{E}_e\big[Y_e\big] = \boldsymbol{\gamma}^\mathsf{T}\mathbb{E}_e\big[\boldsymbol{X}^1_e\big] + \boldsymbol{\eta}_e^\mathsf{T}\mathbb{E}_e\big[\boldsymbol{H}_e\big] + \mathbb{E}_e\big[\varepsilon_e\big] = 0$$

$$\mathbb{E}_e\big[\boldsymbol{X}^2_e\big] = \boldsymbol{\alpha}_e\mathbb{E}_e\big[Y_e\big] + \boldsymbol{\Theta}_e\mathbb{E}_e\big[\boldsymbol{H}_e\big] + \mathbb{E}_e\big[\boldsymbol{\zeta}_e\big] = \mathbf{0}$$

Thus $\mathbb{E}_e\big[\boldsymbol{X}_e\big] = \mathbb{E}_e\big[(\boldsymbol{X}^1_e, \boldsymbol{X}^2_e)\big] = \mathbf{0}$

In the proof of Proposition 2, we had shown that when the data is generated from SEM in Assumption 2

$$\boldsymbol{\Sigma}_e = \mathsf{diag}\Big[\boldsymbol{\Sigma}_{e1}, \boldsymbol{\Sigma}_{e2}\Big] = \mathsf{diag}\Big[\boldsymbol{\Sigma}_{e1}, \boldsymbol{\Theta}_e\boldsymbol{\Theta}_e^\mathsf{T} + \mathsf{diag}\big[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e}\big]\Big]$$

Since $\boldsymbol{\Theta}$ is an orthogonal matrix we have

$$\boldsymbol{\Sigma}_e = \mathsf{diag}\Big[\Sigma_{e1}, \mathsf{diag}[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e} + \mathbf{1}_q]\Big] \tag{34}$$

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

Both $\boldsymbol{\Sigma}_{e1}$ and $\mathsf{diag}[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e} + \mathbf{1}_q]$ are positive definite as a result $\boldsymbol{\Sigma}_e$ is also positive definite. Therefore, Assumption 1 holds.

The expression for the solution to the least squares optimal solution derived in equation (5) has two parts $\boldsymbol{w}^{\mathsf{inv}}_e$ and $\boldsymbol{w}^{\mathsf{var}}_e$. Recall the definition of sets $\mathcal{U}$ and $\mathcal{V}$ from Section 4.2. The first p components corresponding to $\boldsymbol{w}^{\mathsf{inv}}_e \implies \{1, \ldots, p\} \subseteq \mathcal{U}$. The next q components corresponding to $\boldsymbol{w}^{\mathsf{var}}_e \implies \{p+1, \ldots, p+q\} \supseteq \mathcal{V}$. We showed above that $\boldsymbol{\Sigma}_e$ is a block diagonal matrix and the block corresponding to the feature components $\{p+1, \ldots, p+q\}$ also equalling a diagonal matrix $\mathsf{diag}[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e} + \mathbf{1}_q]$. Therefore, we can see that each feature component in $\{p+1, \ldots, p+q\}$ is uncorrelated with any other feature component. Therefore, Assumption 6 also holds. Hence, all the Assumptions required for Theorem 2 also hold. We write the expression for least squares optimal solution in this case (from equation (5)) as $\boldsymbol{w}^*_e = (\boldsymbol{w}^{\mathsf{inv}}_e, \boldsymbol{w}^{\mathsf{var}}_e) = \Big(\boldsymbol{\gamma}, \Big(\mathsf{diag}[\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e} + \mathbf{1}_q]\Big)^{-1}\boldsymbol{\Theta}_e\boldsymbol{\eta}_e\Big)$.

We divide the NE based ensemble predictor $\bar{\boldsymbol{w}}^\dagger$ into two halves: $\boldsymbol{w}^\dagger_1$ is the vector of first p coefficients of $\bar{\boldsymbol{w}}^\dagger$ and $\boldsymbol{w}^\dagger_2$ is the vector of next q coefficients of $\bar{\boldsymbol{w}}^\dagger$.

From Theorem 2 it follows that

$$\boldsymbol{w}^\dagger_1 = \boldsymbol{\gamma} \tag{35}$$

The next q components in the set $\{1, \ldots, q\}$ are computed as follows. For $k \in \{1, \ldots, q\}$, $(p+k)^{th}$ component of $\boldsymbol{w}^*_e$ is $\frac{\big[\boldsymbol{\Theta}_e\boldsymbol{\eta}_e\big]_k}{\sigma^2_{\boldsymbol{\zeta}_e,k}+1}$, where $\big[\boldsymbol{\Theta}_e\boldsymbol{\eta}_e\big]_k$ is the k^{th} component of $\boldsymbol{\Theta}_e\boldsymbol{\eta}_e$ and $\sigma^2_{\boldsymbol{\zeta}_e,k}$ is the k^{th} component of $\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_e}$.

We first prove the latter part of the above proposition. If $\boldsymbol{\Theta}_1\boldsymbol{\eta}_1$ and $\boldsymbol{\Theta}_2\boldsymbol{\eta}_2$ have opposite signs, then for each $k \in \{1, \ldots, q\}$, the sign of $(p+k)^{th}$ component of $\boldsymbol{w}^*_1$ and $\boldsymbol{w}^*_2$ are opposite. From Theorem 2 it follows that $\bar{w}^\dagger_{p+k} = 0$. This holds for all $k \in \{1, \ldots, q\}$ and as a result we have $\bar{\boldsymbol{w}}^\dagger = \big(\boldsymbol{\gamma}, \mathbf{0}\big)$. Now we move to the former part of the Proposition, which compares the NE based ensemble predictor to ERM's solution.

ERM solves the following optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^{n\times 1}} \pi_1\mathbb{E}_1\big[\big(Y_1 - \boldsymbol{w}^\mathsf{T}\boldsymbol{X}_1\big)^2\big] + (1-\pi_1)\mathbb{E}_2\big[\big(Y_2 - \boldsymbol{w}^\mathsf{T}\boldsymbol{X}_2\big)^2\big] \tag{36}$$

By putting the gradient of the above to zero, we get

$$\begin{aligned}\big(\pi_1\boldsymbol{\Sigma_1} + (1-\pi_1)\boldsymbol{\Sigma}_2\big)\boldsymbol{w}^{\mathsf{ERM}} &= \pi_1\boldsymbol{\rho}_1 + (1-\pi_1)\boldsymbol{\rho}_2 \\ \boldsymbol{w}^{\mathsf{ERM}} &= \big(\pi_1\boldsymbol{\Sigma}_1 + (1-\pi_1)\boldsymbol{\Sigma}_2\big)^{-1}(\pi_1\boldsymbol{\rho}_1 + (1-\pi_1)\boldsymbol{\rho}_2)\end{aligned} \tag{37}$$

Substituting the expression for $\boldsymbol{\Sigma}_e$ from equation (34) into equation (37) we get $\boldsymbol{w}^{\mathsf{ERM}}$ equals

$$(\boldsymbol{w}_1^{\mathsf{ERM}}, \boldsymbol{w}_2^{\mathsf{ERM}}) = \Big(\boldsymbol{\gamma}, (\pi_1 \boldsymbol{\Theta}_1 \boldsymbol{\eta}_1 + (1-\pi_1)\boldsymbol{\Theta}_2\boldsymbol{\eta}_2) \odot \boldsymbol{\xi}\Big) \tag{38}$$

where $\boldsymbol{\xi} = \mathbf{1}_q \oslash \Big(\pi_1\big(\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_1} + \mathbf{1}_q\big) + (1-\pi_1)\big(\boldsymbol{\sigma}^2_{\boldsymbol{\zeta}_2} + \mathbf{1}_q\big)\Big)$ and $\boldsymbol{a} \oslash \boldsymbol{b}$ is elementwise division of the two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and $\boldsymbol{w}_1^{\mathsf{ERM}} = \boldsymbol{\gamma}$ and $\boldsymbol{w}_2^{\mathsf{ERM}} = (\pi_1 \boldsymbol{\Theta}_1 \boldsymbol{\eta}_1 + (1-\pi_1)\boldsymbol{\Theta}_2\boldsymbol{\eta}_2) \odot \boldsymbol{\xi}$. For $k \in \{1, \ldots, q\}$, the k^{th} component of $\boldsymbol{w}_2^{\mathsf{ERM}}$ is given as

$$\frac{\pi_1\big[\boldsymbol{\Theta}_1\boldsymbol{\eta}_1\big]_k + \big(1-\pi_1\big)\big[\boldsymbol{\Theta}_2\boldsymbol{\eta}_2\big]_k}{\pi_1\big(\sigma^2_{\boldsymbol{\zeta}_1,k}+1\big) + (1-\pi_1)\big(\sigma^2_{\boldsymbol{\zeta}_2,k}+1\big)} \tag{39}$$

Based on the ERM predictor (equation (38)) and NE-based ensemble predictor (equation (35)) correctly estimate the causal coefficients $\boldsymbol{\gamma}$, i.e., they match in the first p coefficients. We focus on the latter q coefficients. The distance of ERM and NE based ensemble predictors are written as $\|\boldsymbol{w}^{\mathsf{ERM}} - (\boldsymbol{\gamma}, \mathbf{0})\| = \|\boldsymbol{w}_2^{\mathsf{ERM}}\|$, $\|\bar{\boldsymbol{w}}^{\dagger} - (\boldsymbol{\gamma}, \mathbf{0})\| = \|\bar{\boldsymbol{w}}_2^{\dagger}\|$. Hence, we only need to compare the norm of $\boldsymbol{w}_2^{\mathsf{ERM}}$ and $\bar{\boldsymbol{w}}_2^{\dagger}$. From Assumption 4, we know that $\boldsymbol{w}_1^{\mathsf{var}} \neq \boldsymbol{w}_2^{\mathsf{var}}$, thus the two differ in at least one component. Consider a component m, where the two vectors $\boldsymbol{w}_1^{\mathsf{var}}$ and $\boldsymbol{w}_2^{\mathsf{var}}$ do not match. For simplicity, let us write $\Big[\boldsymbol{\Theta}_e\boldsymbol{\eta}_e\Big]_m = \vartheta_e$. Therefore, $\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1} \neq \frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1}$.

There are two possiblities – i) the signs of $\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$ and $\frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1}$ do not match, and ii) the signs of $\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$ and $\frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1}$ match. In case i), the magnitude of the corresponding coefficient of NE based predictor is 0. The magnitude for the ERM based predictor is given as

$$\Big|\frac{\pi_1\vartheta_1 + \big(1-\pi_1\big)\vartheta_2}{\pi_1\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big) + (1-\pi_1)\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big)}\Big| \tag{40}$$

If $\pi_1\vartheta_1 + \big(1-\pi_1\big)\vartheta_2 = 0$ ($\pi_1 = \frac{\vartheta_2}{\vartheta_2 - \vartheta_1}$), then the coefficient of ERM based solution has same magnitude as NE based predictor, which is equal to zero. Therefore, except for when $\pi_1 = \frac{\vartheta_2}{\vartheta_2-\vartheta_1}$, ERM is strictly worse than NE based ensemble predictor.

In case ii), the the signs of $\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$ and $\frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1}$ match. Let us consider the case when both are positive. Without loss of generality assume that $0 \leq \frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1} < \frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1}$. From Theorem 2, we know that the magnitude of the NE based predictor is equal to $\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$ and the magnitude of ERM based predictor is

$$\frac{\pi_1\vartheta_1 + \big(1-\pi_1\big)\vartheta_2}{\pi_1\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big) + (1-\pi_1)\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big)} \tag{41}$$

We take a difference of the magnitudes of the two and get

$$\frac{\pi_1\vartheta_1 + \big(1-\pi_1\big)\vartheta_2}{\pi_1\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big) + (1-\pi_1)\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big)} - \frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$$
$$\frac{\big(1-\pi_1\big)\Big(\vartheta_2\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big) - \vartheta_1\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big)\Big)}{\Big(\pi_1\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big) + (1-\pi_1)\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big)\Big)\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big)} \tag{42}$$

Since $\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1} < \frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1}$ it follows that if $\pi_1 \in [0,1)$, then the above difference in equation (42) is positive. However, if $\pi_1 = 1$, then the difference is zero. Therefore, except for when $\pi_1 = 1$, ERM is strictly worse than NE based ensemble predictor.

Lastly, the analysis for the case when both coefficients are negative also follows on exactly the above lines. Without loss of generality consider the case, $\frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1} < \frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1} \leq 0$. In this case, NE based predictor will take the value $\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$ (follows from Theorem 2) and its magnitude is $-\frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$. The magnitude of ERM based predictor is

$$-\frac{\pi_1\vartheta_1 + \big(1-\pi_1\big)\vartheta_2}{\pi_1\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big) + (1-\pi_1)\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big)} \tag{43}$$

We take a difference of the magnitudes of the NE based predictor and the ERM based predictor to get

$$\frac{\big(1-\pi_1\big)\Big(\vartheta_1\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big) - \vartheta_2\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big)\Big)}{\Big(\pi_1\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big) + (1-\pi_1)\big(\sigma^2_{\boldsymbol{\zeta}_2,m}+1\big)\Big)\big(\sigma^2_{\boldsymbol{\zeta}_1,m}+1\big)} \tag{44}$$

Since $\frac{\vartheta_2}{\sigma^2_{\boldsymbol{\zeta}_2,m}+1} < \frac{\vartheta_1}{\sigma^2_{\boldsymbol{\zeta}_1,m}+1}$ it it follows that if $\pi_1 \in [0,1)$, then the above difference in equation (42) is positive. However, if $\pi_1 = 1$, then the difference is zero. Therefore, except for when $\pi_1 = 1$, ERM is strictly worse than NE based ensemble predictor. This completes the analysis for all the possible cases. For each component where the least squares optimal solution differ, we showed that ERM based predictor is worse than NE based predictor except over a set of measure zero over the probability π_1. This completes the proof.

□

7.5 Theorem 3

We restate Theorem 3 below.

Theorem 5. *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. From Theorem 2, we know that if Assumptions 1, 5, 6 hold, then the NE based ensemble predictor is given as

$$\Big(\boldsymbol{w}_1^* \odot \mathbf{1}_{|\boldsymbol{w}_2^*| \geq |\boldsymbol{w}_1^*|} + \boldsymbol{w}_2^* \odot \mathbf{1}_{|\boldsymbol{w}_1^*| > |\boldsymbol{w}_2^*|}\Big)\mathbf{1}_{\boldsymbol{w}_1^* \odot \boldsymbol{w}_2^* \geq \mathbf{0}} \quad (45)$$

In Algorithm 1, each environment plays the optimal action given the action of the others. Hence, by best responding to each other we hope the procedure would converge to NE based ensemble predictor in equation (45).

We write a dynamic which seems simpler than the dynamic in Algorithm 1. However, we show that the two are equivalent. We index the iteration by t. Define $\boldsymbol{w}_e^t$ as the predictor for environment e at the end of iteration t. The ensemble predictor is given as $\bar{\boldsymbol{w}}^t = \boldsymbol{w}_1^t + \boldsymbol{w}_2^t$. Each component of e's predictor is given as $\boldsymbol{w}_e^t = (w_{e1}^t, \ldots, w_{en}^t)$ and for the other environment $q \in \{1,2\} \setminus \{e\}$ as $\boldsymbol{w}_{-e}^t = (w_{-e1}^t, \ldots, w_{-en}^t)$. Environment e in its turn sees that the other environment is using a predictor $\boldsymbol{w}_{-e}^{t-1}$; environment e updates the predictor by taking a step such that $\min_{\boldsymbol{w}_e^t \in \mathcal{W}} \left\|\boldsymbol{w}_e^t + \boldsymbol{w}_{-e}^{t-1} - \boldsymbol{w}_e^*\right\|^2$, i.e. the environment moves such that it gets closest to the optimal least squares solution. We can simplify this minimization as

$$\begin{aligned} &\min_{\boldsymbol{w}_e^t \in \mathcal{W}} \left\|\boldsymbol{w}_e^t + \boldsymbol{w}_{-e}^{t-1} - \boldsymbol{w}_e^*\right\|^2 \\ &= \sum_{i=1}^{n} \min_{|w_{ei}^t| \leq w^{\mathsf{sup}}} (w_{ei}^t + w_{-ei}^t - w_{ei}^*)^2 \end{aligned} \quad (46)$$

For $t = 0$, $\boldsymbol{w}_1^t = \mathbf{0}$ and $\boldsymbol{w}_2^t = \mathbf{0}$. The dynamic based on equation (46) is written as follows.

For $t \geq 1$

$$\boldsymbol{w}_1^t = \begin{cases} \boldsymbol{w}_1^{t-1} & t \text{ is even} \\ \Pi_{\mathcal{W}}[\boldsymbol{w}_1^* - \boldsymbol{w}_2^{t-1}] & t \text{ is odd} \end{cases} \quad (47)$$

$$\boldsymbol{w}_2^t = \begin{cases} \boldsymbol{w}_2^{t-1}, & t \text{ is odd} \\ \Pi_{\mathcal{W}}[\boldsymbol{w}_2^* - \boldsymbol{w}_1^{t-1}] & t \text{ is even} \end{cases} \quad (48)$$

$t = t + 1$

In the above equations (47), (48), $\Pi_{\mathcal{W}}$ represents the projection on the set $\mathcal{W} = \{\boldsymbol{w} \text{ s.t.} \|\boldsymbol{w}\|_\infty \leq w^{\mathsf{sup}}\}$.

In each iteration, only one of the environment updates the predictors. In the above dynamic, whenever an environment completes its turn to update the predictor, t is incremented by one. Before showing the convergence of this dynamic, we first need to establish that this dynamic is equivalent to the one stated in Algorithm, where when its the turn of environment e to update it minimizes the following $\min_{\boldsymbol{w}_e^t \in \mathcal{W}} R_e(\boldsymbol{w}_e^t, \boldsymbol{w}_{-e}^{t-1})$.

Equivalence of dynamic in equations (47) and (48) to dynamic in Algorithm 1

Recall from Section 4.2 and proof of Theorem 2, that we divide the feature components $\{1, \ldots, n\}$ into two sets, the first k components are in $\mathcal{U}$ and the next $n-k$ components are in $\mathcal{V}$. The two environments have the same least squares coefficients for components in $\mathcal{U}$ but have differing coefficients for points in $\mathcal{V}$. For each $e \in \{1,2\}$, $\boldsymbol{w}_{e+}^t$ corresponds to the first k coefficient at the end of iteration t and $\boldsymbol{w}_{e-}^t$ corresponds to the next $n-k$ coefficients at the end of iteration t.

Recall the decomposition that we stated in equation (28). To arrive at the equation (28) we used Assumptions 1 and 6. We continue to make these assumptions in this theorem as well. Therefore, we can continue to use the decomposition in equation (28). For environment 2 we can write

$$\begin{aligned} &\min_{\boldsymbol{w}_2^t \in \mathcal{W}} R_2(\boldsymbol{w}_1^{t-1}, \boldsymbol{w}_2^t) = \\ &= \min_{\boldsymbol{w}_{2+}^t \in \mathcal{W}_+} R_{2+}(\boldsymbol{w}_{1+}^{t-1}, \boldsymbol{w}_{2+}^{t-1}) + \\ &\sum_i \sigma_{2i}^2 \min_{|w_{2i}^t| \leq w^{\mathsf{sup}}} (w_{2i}^t + w_{1i}^t - w_{2i}^*)^2 \end{aligned} \quad (49)$$

A decomposition identical to above equation (49) also holds for environment 1. From equation (46) and (49), we gather that for latter $n-k$ components, the update rule in equations (47), (48) and the update rule in Algorithm 1 are equivalent for both the environments. We now show that both the rules are equivalent in the first k components as well.

Consider iteration $t = 1$. If the environment 1 uses the update rule in equation (47), then the ensemble predictor is set to $\boldsymbol{w}_1^*$ (From Assumption $\boldsymbol{w}_1^* \in \mathcal{W}$ is in the interior, the projection will be the point itself). Note that if environment 1 used $\min_{\boldsymbol{w}_1^t \in \mathcal{W}} R_1(\boldsymbol{w}_1^t, \boldsymbol{w}_2^{t-1})$, then as well it will move the ensemble predictor to $\boldsymbol{w}_1^*$ (since $\boldsymbol{w}_1^*$ is the least squares optimal solution).

Consider iteration $t = 2$. Suppose the environment 2 uses the update rule in equation (48). Define $\boldsymbol{\Delta}^* = \boldsymbol{w}_2^* - \boldsymbol{w}_1^*$ and the i^{th} component of $\boldsymbol{\Delta}^*$ as Δ_i^*. Given the environment 1 is at $\boldsymbol{w}_1^*$ the rule dictates that environment 2 should update the predictor to $\Pi_{\mathcal{W}}[\boldsymbol{\Delta}^*]$. The first k components of $\boldsymbol{\Delta}^*$ would be zero as the two environments agree in these coefficients. Therefore, environment 2 will not move its predictor for the first k components and continue to be at $\mathbf{0}$. After this the two environments do not need to update the first k

components as they have already converged. Suppose the environment 2 uses the update rule from Algorithm 1. Consider the first k components in which both the environments agree. Since environment 1 already is using $\boldsymbol{w}_{1+}^*$, which is optimal for environment 2 as well, environment 2 will not move its predictor for first k components and continue to be at $\mathbf{0}$. After this the two environments do not need to update the first k components as they have already converged.

Thus so far we have established that both dynamics in equations (47), (48) and the update rule in Algorithm 1 are equivalent. We have also shown the convergence in the first k components. We now focus on establishing the convergence for the next $n-k$ components that make up the set $\mathcal{V}$.

Convergence of the dynamics in equations (47) and (48). In the previous section, we showed that dynamic in equations (47) and (48) are equivalent to the dynamic in Algorithm 1. While we had shown the convergence of the first k components in the set $\mathcal{U}$, we will repeat the analysis for ease of exposition. In this section, we begin by showing how the dynamic in equations (47) and (48) plays out. Just for the sake of clarity of exposition, in the dynamic we show below we assume that the predictors of the environment continue to be in the interior of the set $\mathcal{W}$.

1. End of $t=1$, $\bar{\boldsymbol{w}}^t = \boldsymbol{w}_1^*$, $e=1$ plays $\boldsymbol{w}_1^t = \boldsymbol{w}_1^*$, $e=2$ plays $\boldsymbol{w}_2^t = \mathbf{0}$.
2. End of $t=2$, $\bar{\boldsymbol{w}}^t = \boldsymbol{w}_2^*$, $e=1$ plays $\boldsymbol{w}_1^t = \boldsymbol{w}_1^*$, $e=2$ plays $\boldsymbol{w}_2^t = \boldsymbol{\Delta}^*$.
3. End of $t=3$, $\bar{\boldsymbol{w}}^t = \boldsymbol{w}_1^*$, $e=1$ plays $\boldsymbol{w}_1^t = \boldsymbol{w}_1^* - \boldsymbol{\Delta}^*$, $e=2$ plays $\boldsymbol{w}_2^t = \boldsymbol{\Delta}^*$.
4. End of $t=4$, $\bar{\boldsymbol{w}}^t = \boldsymbol{w}_2^*$, $e=1$ plays $\boldsymbol{w}_1^t = \boldsymbol{w}_1^* - \boldsymbol{\Delta}^*$, $e=2$ plays $\boldsymbol{w}_2^t = 2\boldsymbol{\Delta}^*$.
5. End of $t=5$, $\bar{\boldsymbol{w}}^t = \boldsymbol{w}_1^*$, $e=1$ plays $\boldsymbol{w}_1^t = \boldsymbol{w}_1^* - 2\boldsymbol{\Delta}^*$, $e=2$ plays $\boldsymbol{w}_2^t = 2\boldsymbol{\Delta}^*$.
6. End of $t=6$, $\bar{\boldsymbol{w}}^t = \boldsymbol{w}_2^*$, $e=1$ plays $\boldsymbol{w}_1^t = \boldsymbol{w}_1^* - 2\boldsymbol{\Delta}^*$, $e=2$ plays $\boldsymbol{w}_2^t = 3\boldsymbol{\Delta}^*$.

In the dynamic displayed above, we assumed that the predictor $\boldsymbol{w}_1^t$ and $\boldsymbol{w}_2^t$ were in the interior just to illustrate that the two sequences $\boldsymbol{w}_1^t$ and $\boldsymbol{w}_2^t$ are monotonic. Observe that if a certain component of $\boldsymbol{\Delta}^*$ say Δ_i^* is non-zero, the two sequences are strictly monotonic in that component. The sequences cannot grow unbounded and at least one of them will first hit the boundary at w^{sup} or $-w^{\mathsf{sup}}$. Recall from the last section, where we already showed that for the first k components of $\boldsymbol{\Delta}^*$ associated with these $\mathcal{U}$ are zero (from equation (25)). Hence, for the first k components the the dynamic $\boldsymbol{w}_1^t$ and $\boldsymbol{w}_2^t$ converges at the end of $t=1$. Therefore, we now only need to focus on the remaining $n-k$ components comprising the set $\mathcal{V}$. Since the update rules in equation (47) and (48) are separable for the different components, we only focus on one of the components say i.

We divide our analysis based on if w_{1i}^* and w_{2i}^* have the same sign or not. Suppose w_{1i}^* and w_{2i}^* have the same sign. Let us consider the case when both are positive (negative case follows from symmetry as the dynamic starts at zero).

- Suppose $0 \leq w_{1i}^* < w_{2i}^*$. If $0 \leq w_{1i}^* < w_{2i}^*$ is plugged into the equation (6), we obtain w_{1i}^*. Our objective is to show convergence to w_{1i}^*. In this case, Δ_i^*, which corresponds to the i^{th} component of $\boldsymbol{\Delta}^*$, is greater than zero. Observe from the dynamic that environment 2 will first hit the boundary in this case and since $\Delta_i^* > 0$, it will hit the positive end, i.e., w^{sup}. The best response of environment 1 is to play $\Pi_{\mathcal{W}}[w_{1i}^* - w^{\mathsf{sup}}]$. Since $w_{1i}^* > 0$, we get that environment 1 uses the predictor $w_{1i}^* - w^{\mathsf{sup}}$, the ensemble predictor takes the value w_{1i}^*. Environment 2 in the next step continues to play $\Pi_{\mathcal{W}}[w_{2i}^* - w_{1i}^* + w^{\mathsf{sup}}] = w^{\mathsf{sup}}$ and environment 1 continues to play $w_{1i}^* - w^{\mathsf{sup}}$. Hence, the predictors stop updating. Thus in this case at convergence, the ensemble predictor achieves the value that we wanted to prove w_{1i}^*. Also, both environments best respond to each other, which implies that the state is a NE.
- Suppose $0 \leq w_{2i}^* < w_{1i}^*$. If $0 \leq w_{2i}^* < w_{1i}^*$ is plugged into the equation (6), we obtain w_{1i}^*. Our objective is to show convergence to w_{2i}^*. Observe from the dynamic that environment 1 will first hit the boundary in this case and since $\Delta_i^* < 0$, it will hit the positive end, i.e., w^{sup}. The best response of environment 2 is to play $\Pi_{\mathcal{W}}[w_{2i}^* - w^{\mathsf{sup}}]$. Since $w_{2i}^* > 0$, we get that environment 2 uses the predictor $w_{2i}^* - w^{\mathsf{sup}}$ and the ensemble predictor takes the value w_{2i}^*. Thus just like the case described above both environments stop updating. Hence, the state $w^{\mathsf{sup}}, w_{2i}^* - w^{\mathsf{sup}}$ is a NE and the final ensemble predictor is at w_{2i}^*, which is what we wanted to prove.

Suppose w_{1i}^* and w_{2i}^* have the opposite sign.

- Consider the case when $w_{1i}^* < 0 < w_{2i}^*$. If $w_{1i}^* < 0 < w_{2i}^*$ is plugged into the equation (6), we obtain 0. In this setting, environment 2 moves towards the w^{sup} and environment 1 moves towards $-w^{\mathsf{sup}}$. Suppose environment 2 hits the boundary. In the next step, the best response from environment 1 is computed as $\Pi_{\mathcal{W}}[w_{1i}^* - w^{\mathsf{sup}}] = -w^{\mathsf{sup}}$. Since

both environments best respond to each other the state $(-w^{\mathsf{sup}}, w^{\mathsf{sup}})$ is NE and the final ensemble predictor is at 0, which is what we wanted to prove. Hence, both the environment continue to stay at the boundary. This is also the case if environment 1 hits the boundary first. The same analysis applies to the case when $w^*_{2i} < 0 < w^*_{1i}$.

We focused on one of the components i and the above analysis applies to all the components $j \in \{k+1, \ldots, n\}$. This completes the proof. Note that the entire analysis is symmetric and it does not matter which environment moves first, the analysis also extends to the case when initialization is not zero. □

7.6 Proposition 5

We restate Proposition 5

Proposition 9. *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 $\bar{\boldsymbol{w}}^+ = (\boldsymbol{\gamma}, 0)$.*

Proof. In the above proposition, we make the same set of assumptions as in Proposition 4. In the proof of Proposition 4, we showed that the set of Assumptions in Proposition 4 imply that the Assumptions 1, 5, 6 hold. Since Assumptions 1, 5, 6 hold, from Theorem 3 it follows that the output of Algorithm 1 is equal to the NE based ensemble predictor given by equation (6). In Proposition 4, we have already shown that this NE based ensemble predictor (equation (6)), which is the output of Algorithm 1, is closer to $(\boldsymbol{\gamma}, \mathbf{0})$ than the solution of ERM (except over a set of measure zero defined in the proof of Proposition 4). We had also shown that the NE based ensemble predictor is equal $(\boldsymbol{\gamma}, \mathbf{0})$ when the signs of $\boldsymbol{\Theta}_1\boldsymbol{\eta}_1$ and $\boldsymbol{\Theta}_2\boldsymbol{\eta}_2$ are opposite.

This completes the proof. □

7.7 Extensions

In this section, we discuss extensions and generalizations of the results presented in the main manuscript.

7.7.1 Other best response dynamics

In this section, we describe a simple signed gradient descent based dynamic. The aim is to show that for simple variations of the dynamic proposed in Algorithm 1 we continue to have convergence guarantees.

We define a signed gradient descent based version of the dynamic in equation (47) and (48) with step length β.

For $t = 0$, $\boldsymbol{w}_1^t = \mathbf{0}$ and $\boldsymbol{w}_2^t = \mathbf{0}$.

For $t \geq 1$

$$\boldsymbol{w}_1^t = \Pi_{\mathcal{W}}\Big[\boldsymbol{w}_1^{t-1} + \beta \mathsf{sgn}[\boldsymbol{w}_1^* - \boldsymbol{w}_2^{t-1}]\Big] \tag{50}$$

$$\boldsymbol{w}_2^t = \Pi_{\mathcal{W}}\Big[\boldsymbol{w}_2^{t-1} + \beta \mathsf{sgn}[\boldsymbol{w}_2^* - \boldsymbol{w}_1^{t-1}]\Big] \tag{51}$$

$$\bar{\boldsymbol{w}}^t = \boldsymbol{w}_1^t + \boldsymbol{w}_2^t$$

$$t = t + 1$$

In the above sgn is the component-wise sign function, which takes a value 1 when the input is positive (including zero) and -1 if the input is negative. $\bar{\boldsymbol{w}}^t$ is the ensemble predictor at the end of iteration t, $\boldsymbol{w}_e^t$ is the predictor for environment e at the end of iteration t. Suppose we want to get within ϵ (per component) distance of the NE based predictor. Divide the components into two sets $\mathcal{E}$ and $\mathcal{F}$ defined as follows. $i \in \mathcal{E}$ if and only if the least squares solution are within ϵ distance, i.e. $|w^*_{1i} - w^*_{2i}| \leq \epsilon$ and $i \in \mathcal{F}$ if and only if the least squares solution are separated by at least epsilon i.e. $|w^*_{1i} - w^*_{2i}| > \epsilon$.

Let $\beta < \epsilon$. Let us analyze the dynamic for a component $i \in \mathcal{F}$. We divide the analysis into two cases – w^*_{1i} and w^*_{2i} have the same sign, and w^*_{1i} and w^*_{2i} have opposite signs. Let us start with same sign case with both w^*_{1i} and w^*_{2i} positive (negative sign case follows from symmetry). Consider the case $0 < w^*_{1i} < w^*_{2i}$. In this case from the expression of NE in equation (6), we would hope the dynamic can eventually achieve an ensemble predictor that stays within ϵ distance of w^*_{1i}. Since the dynamic starts at 0 and $\mathsf{sgn}[w^*_{1i}] = 1$ and $\mathsf{sgn}[w^*_{2i}] = 1$, the ensemble predictor $\bar{w}_i^t$ after some iterations will enter the interval $[w^*_{1i}, w^*_{2i}]$. Once the ensemble predictor enters the interval, the two predictors w^t_{1i} and w^t_{2i} will push the predictor in opposite directions and since the step length is the same the ensemble predictor will not move. This would continue until environment 2 hits the positive boundary w^{sup}. Once the environment 2 hits the boundary it stops updating and environment 1 pushes the ensemble predictor towards w^*_{1i}. Once the predictor is within β distance from w^*_{1i} it continues to oscillate around w^*_{1i}. Consider the case $0 < w^*_{2i} < w^*_{1i}$, the same analysis follows and dynamic eventually oscillates around w^*_{2i}. Next, consider the case when w^*_{1i} and w^*_{2i} have opposite signs. In this case from the expression of NE in (6), we would hope the dynamic can eventually achieve an ensemble predictor that stays within ϵ distance of 0. In this case, since the predictors start at zero, both environments will push in opposite directions. Eventually, both environments hit opposite ends of the boundary and stay there. This results in ensemble predictor coefficient of zero.

Now let us analyze the game for components in $\mathcal{E}$. We again carry out analysis based on the whether the signs agree or not. Consider the case $0 < w^*_{1i} < w^*_{2i}$. In this case from the expression of NE in (6), we would hope the dynamic can eventually achieve an ensemble predictor that stays within ϵ distance of w^*_{1i}. The dynamic starts at 0 and $\mathsf{sgn}[w^*_{1i}] = 1$ and $\mathsf{sgn}[w^*_{2i}] = 1$, the ensemble predictor $\bar{w}^t_i$ may or may not enter the interval $[w^*_{1i}, w^*_{2i}]$. If it enters, then the analysis is identical to the previous case when $i \in \mathcal{F}$. If the ensemble predictor does not enter the interval, then it has to overshoot it and move to the right of the interval, which implies in the next step it will be pulled back to the left of the interval. This sets the ensemble predictor in an oscillation around w^*_{1i}. The analysis for the case when w^*_{1i} and w^*_{2i} have opposite signs is the same as the previous case.

7.7.2 Convergence when Assumption 6 does not hold

In this section, we discuss can we still learn NE if the Assumption 6 is relaxed? We would rely on the results in [Zhou et al., 2017] for our discussion here. In [Zhou et al., 2017], the authors introduced a notion called variational stability. Consider the class of concave games. It was shown that if the set of Nash equilibria satisfy variational stability, then a mirror descent based learning dynamic (described in [Zhou et al., 2017]) converges to the NE. Next, we analyze the variational stability for C-LRG.

Define the gradient of utility of the environment e $\boldsymbol{v}_e(\boldsymbol{w}) = -\nabla_{\boldsymbol{w}_e} R_e(\boldsymbol{w}_e, \boldsymbol{w}_{-e})$, where recall that $\boldsymbol{w}_e$ is action of environment e and $\boldsymbol{w}_{-e}$ is the action of the other environment, R_e is the risk, and $\boldsymbol{w} = (\boldsymbol{w}_e, \boldsymbol{w}_{-e})$. Let us recall a characterization of NE in terms of the gradients ([Zhou et al., 2017]). Suppose $\boldsymbol{w}^\dagger = (\boldsymbol{w}^\dagger_1, \boldsymbol{w}^\dagger_2)$ is a NE of C-LRG. For every $\boldsymbol{w}_e \in \mathcal{W}$, we have

$$\boldsymbol{v}_e(\boldsymbol{w}^\dagger)^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) \leq 0 \tag{52}$$

Next, we show how to relate the gradient at NE, $\boldsymbol{v}_e(\boldsymbol{w}^\dagger)$, to the gradient at any other point $\boldsymbol{v}_e(\boldsymbol{w})$

$$\begin{aligned}
&\boldsymbol{v}_e(\boldsymbol{w}) \\
&= \mathbb{E}_e\Big[\big(Y_e - \boldsymbol{w}_1^\mathsf{T}\boldsymbol{X}_e - \boldsymbol{w}_2^\mathsf{T}\boldsymbol{X}_e\big)\boldsymbol{X}_e\Big] \\
&= \mathbb{E}_e\Big[\Big(Y_e - (\boldsymbol{w}_1 - \boldsymbol{w}^\dagger_1 + \boldsymbol{w}^\dagger_1)^\mathsf{T}\boldsymbol{X}_e - \\
&\quad (\boldsymbol{w}_2 - \boldsymbol{w}^\dagger_2 + \boldsymbol{w}^\dagger_2)^\mathsf{T}\boldsymbol{X}_e\Big)\boldsymbol{X}_e\Big] \\
&\boldsymbol{v}_e(\boldsymbol{w}) = \boldsymbol{v}_e(\boldsymbol{w}^\dagger) - \boldsymbol{\Sigma}_e(\boldsymbol{w}_1 - \boldsymbol{w}^\dagger_1 + \boldsymbol{w}_2 - \boldsymbol{w}^\dagger_2) \\
&\boldsymbol{v}_e(\boldsymbol{w}) = \boldsymbol{v}_e(\boldsymbol{w}^\dagger) - \boldsymbol{\Sigma}_e(\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)
\end{aligned} \tag{53}$$

In the above $\bar{\boldsymbol{w}} = \boldsymbol{w}_1 + \boldsymbol{w}_2$, $\bar{\boldsymbol{w}}^\dagger = \boldsymbol{w}^\dagger_1 + \boldsymbol{w}^\dagger_2$.

For establishing variational stability, we need to show that for each $\boldsymbol{w} \in \mathcal{W} \times \mathcal{W}$ and for each NE $\boldsymbol{w}^\dagger$ the following inequality, i.e., $\sum_e \boldsymbol{v}_e(\boldsymbol{w})^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) \leq 0$, holds.

$$\begin{aligned}
&\sum_{e\in\{1,2\}} \boldsymbol{v}_e(\boldsymbol{w})^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) = \\
&\sum_{e\in\{1,2\}} \boldsymbol{v}_e(\boldsymbol{w}^\dagger)^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) - \sum_{e\in\{1,2\}} (\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)^\mathsf{T}\boldsymbol{\Sigma}_e(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) \\
&= \sum_{e\in\{1,2\}} \boldsymbol{v}_e(\boldsymbol{w}^\dagger)^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) - (\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)^\mathsf{T}\boldsymbol{\Sigma}_1(\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger) \\
&- (\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)^\mathsf{T}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)(\boldsymbol{w}_2 - \boldsymbol{w}^\dagger_2)
\end{aligned} \tag{54}$$

We begin by analyzing the case when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Substitute $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ in equation (54),

$$\begin{aligned}
&\sum_{e\in\{1,2\}} \boldsymbol{v}_e(\boldsymbol{w})^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) = \\
&\sum_{e\in\{1,2\}} \boldsymbol{v}_e(\boldsymbol{w}^\dagger)^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) - (\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)^\mathsf{T}\boldsymbol{\Sigma}_1(\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)
\end{aligned}$$

If we use the condition in equation (52) along with the fact that $\boldsymbol{\Sigma}_1$ is positive definite, then we get that $\sum_{e\in\{1,2\}} \boldsymbol{v}_e(\boldsymbol{w})^\mathsf{T}(\boldsymbol{w}_e - \boldsymbol{w}^\dagger_e) \leq 0$, which implies that the set of NE of C-LRG is variationally stable. We now give an example of when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ is satisfied. Consider the SEM in Assumption 2. If between the two environments the only parameters that vary are $\boldsymbol{\eta}_e$ and the distribution of ε_e, and rest all other parameters in the model are the same, then $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ is satisfied.

We now discuss what happens if we relax the assumption, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, made above.

Consider the eigenvalue decomposition of $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 = \boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Omega}^\mathsf{T}$, where since $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ is a symmetric matrix we know that $\boldsymbol{\Omega}$ can be chosen as an orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. Define the smallest eigenvalue of $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ as $\lambda_{\mathsf{min}}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)$.

Define a transformation of vector $\boldsymbol{w}$ under $\boldsymbol{\Omega}$ as $\tilde{\boldsymbol{w}} = \boldsymbol{\Omega}\boldsymbol{w}$. Since $\boldsymbol{\Omega}$ is orthogonal, $\|\tilde{\boldsymbol{w}}\| = \|\boldsymbol{w}\|$. We now use these relationships to simplify

$$\begin{aligned}
&(\boldsymbol{w}_1 - \boldsymbol{w}^\dagger_1)^\mathsf{T}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)(\boldsymbol{w}_2 - \boldsymbol{w}^\dagger_2) = \\
&(\boldsymbol{w}_1 - \boldsymbol{w}^\dagger_1)^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{\Lambda}\boldsymbol{\Omega}^\mathsf{T}(\boldsymbol{w}_2 - \boldsymbol{w}^\dagger_2) \\
&= (\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}^\dagger_1)\boldsymbol{\Lambda}(\tilde{\boldsymbol{w}}_2 - \tilde{\boldsymbol{w}}^\dagger_2) \\
&\geq \lambda_{\mathsf{min}}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)(\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}^\dagger_1)^\mathsf{T}(\tilde{\boldsymbol{w}}_2 - \tilde{\boldsymbol{w}}^\dagger_2) \\
&\geq -|\lambda_{\mathsf{min}}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)|\|(\tilde{\boldsymbol{w}}_1 - \tilde{\boldsymbol{w}}^\dagger_1)\|\|(\tilde{\boldsymbol{w}}_2 - \tilde{\boldsymbol{w}}^\dagger_2)\| \\
&= -|\lambda_{\mathsf{min}}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)|\|(\boldsymbol{w}_1 - \boldsymbol{w}^\dagger_1)\|\|(\boldsymbol{w}_2 - \boldsymbol{w}^\dagger_2)\|
\end{aligned} \tag{55}$$

In the last two inequalities in equation (55), we used Cauchy-Schwarz inequality and the fact that norms do not change under orthogonal transformations $\|\tilde{\boldsymbol{w}}\| = \|\boldsymbol{w}\|$. Now let us bound the term $(\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)^\mathsf{T}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)(\boldsymbol{w}_2 - \boldsymbol{w}_2^\dagger)$ in equation (54).

$$
\begin{aligned}
&- (\bar{\boldsymbol{w}} - \bar{\boldsymbol{w}}^\dagger)^\mathsf{T}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)(\boldsymbol{w}_2 - \boldsymbol{w}_2^\dagger) \\
&= -(\boldsymbol{w}_2 - \boldsymbol{w}_2^\dagger)^\mathsf{T}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)(\boldsymbol{w}_2 - \boldsymbol{w}_2^\dagger) + \\
&\quad - (\boldsymbol{w}_1 - \boldsymbol{w}_1^\dagger)^\mathsf{T}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)(\boldsymbol{w}_2 - \boldsymbol{w}_2^\dagger) \qquad (56) \\
&\leq -\lambda_{\mathsf{min}}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)\|(\boldsymbol{w}_2 - \boldsymbol{w}_2^\dagger)\|^2 + \\
&\quad |\lambda_{\mathsf{min}}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)|\|(\boldsymbol{w}_1 - \boldsymbol{w}_1^\dagger)\|\|(\boldsymbol{w}_2 - \boldsymbol{w}_2^\dagger)\|
\end{aligned}
$$

In the last inequality in equation (56), we used equation (55). If $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ is positive semi-definite with lowest eigenvalue of zero, then the term in equation (56) is bounded above by zero. If we use this observation in equation (54), the the condition for variational stability is satisfied. Note that the entire analysis is symmetric and we can state the same result for the matrix $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$. Therefore, if one of the matrices $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ or $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$, is positive semi-definite with lowest eigenvalue of zero, then we get variational stability for the NE. Therefore, we can use the convergence results in [Zhou et al., 2017] to guarantee that NE will be learned.

7.7.3 Multiple Environments

In the main body of the paper, we discussed the results when the data is gathered from two environments. What happens if the data were gathered from multiple (more than two) environments?

First let us start with the game U-LRG from Section 4.1. The first result in Proposition 1 states that when least squares optimal solution are not equal, then there is no NE of U-LRG. When we move to multiple environments using same proof techniques it can be shown that if there is any two environments, which do not agree on the least squares optimal solution, then no NE will exist. For a NE to exist all environments will have to have the same least squares optimal solution.

Next, we consider the game C-LRG from Section 4.2. With multiple (more than two environments), we are guaranteed that NE will exist. How does the Theorem 2 change for multiple environments? We extend the Assumption 6 to state that any feature component that does not have the same least squares coefficient across all the environments is uncorrelated with the rest of the features. Suppose the environments are indexed from $\{1, \ldots, r\}$. For this discussion, let us focus on one of the feature components say i. Without loss of generality, assume that these environments are ordered in an increasing order w.r.t the optimal least squares coefficient, i.e. if $e, f \in \{1, \ldots, r\}$ such that $e \leq f$, then $w^*_{ei} \leq w^*_{fi}$.

Let us assume that r is odd. Consider the median environment indexed $m = \frac{r+1}{2}$. Ensemble predictor's coefficient will be equal to the coefficient of the median environment $\bar{w}_i^\dagger = w^*_{mi}$. In this case in the NE, all the environments with index $e > m$ play w^{sup}, all the environments with index $e < m$ play $-w^{\mathsf{sup}}$, and median environment m plays w^*_{mi}.

Let us assume that r is even. Consider the two median environments indexed $m = \frac{r}{2}$ and $m+1$. If w^*_{mi} and $w^*_{(m+1)i}$ have the same sign, then the NE based ensemble predictor is equal to the coefficient with a smaller absolute value. If w^*_{mi} and $w^*_{(m+1)i}$ have the same sign, and say $0 \leq w^*_{mi} \leq w^*_{(m+1)i}$, then in NE the environment m plays $w^*_{mi} - w^{\mathsf{sup}}$ and environment $m+1$ plays w^{sup}. If w^*_{mi} and $w^*_{(m+1)i}$ have the opposite sign, then the NE based ensemble predictor is equal to zero. If w^*_{mi} and $w^*_{(m+1)i}$ have the opposite sign, and say $w^*_{mi} < 0 \leq w^*_{(m+1)i}$, then in NE the environment m plays $-w^{\mathsf{sup}}$ and environment $m+1$ plays w^{sup}. For all the remaining environments other than m and $m+1$ their actions are described as — environments with index $e > m+1$ play w^{sup}, all the environments with index $e < m$ play $-w^{\mathsf{sup}}$.

In Proposition 4, we analyzed linear SEMs and showed that NE based ensemble predictor are closer to the OOD solutions than ERM. Proposition 4 relied on Theorem 2, which we have shown can be appropriately extended to multi-environment setting. Hence, by using the same proof techniques used to prove Proposition 4 and the expression for NE that we discussed above for multiple environments, we can show that the same result extends to multiple environments.

In Theorem 3, we proved the convergence for BRD dynamics to NE. We can show convergence in this setup as well using the same ideas discussed in Section 7.5 and Section 7.7.1. We have shown that Theorem 2, Proposition 4 and Theorem 3 extend to multi-environment setting. We also know that Proposition 5 directly follows from Theorem 2, Proposition 4 and Theorem 3. Therefore, Proposition 5 extends to multi-environment setting.

7.7.4 Theory beyond Assumption 6 and 2

Independent component analysis (ICA) [Hyvärinen and Oja, 2000] based feature extraction assumes that complex datasets such as images are a transformation of (independent and identically distributed) i.i.d. hidden features. Thus for complex datasets we can break the analysis in two parts: extract i.i.d. features and apply our linear method on the extracted features. Suppose the observed data

X comes from a linear transformation of hidden i.i.d. features Z and is given as $X = AZ$, where A is a linear transformation. Features Z take the place of X in the model in Assumption 2. In such a case, X need not follow Assumption 6. If we use linear ICA on X, then ideally, one hopes to recover A (or a permutation or scaled version of it), and extract Z and apply our approach on Z.

7.8 Supplement for Experiments

7.8.1 Computing Environment

The experiments were done on 2.3 GHZ Intel Core i9 processor with 32 GB memory (2400 MHz DDR4). The codes can be found at `https://github.com/IBM/OoD` that would allow the reader to both reproduce the results and use the model class to train their own models to build NE based ensemble predictors.

7.8.2 Model, Hyperparameter, Training details

We use linear models for all the methods. We carried out 10 trials for all the experiments and show the average performance in Figure 3. In the experiments for the 10 dimensional case ($p = q = 5$) shown in Figure 3, we use a bound of $w^{\mathsf{sup}} = 2$. In Figure 2, we had shown that provided the solution is contained in the search space, i.e., realizability assumption (Assumption 5) is satisfied, then the choice of the bound does not impact the solution provided the number of training steps are sufficiently large.

We use a stochastic gradient descent based best response dynamic to learn the NE; this dynamic is very similar to the one described in Section 7.7.1. For each environment $e \in \{1, 2\}$ say the loss for the current batch at the end of iteration t is $\hat{R}_e(\boldsymbol{w}_1^t, \boldsymbol{w}_2^t)$ (sample mean estimate of the loss over the current batch). For each $e \in \{1, 2\}$, say $\boldsymbol{w}_e^t$ is the model used by environment e at the end of iteration t. The two environments alternate to take turns to update the model, i.e. in odd iterations t environment 1 updates the model, and even iterations t environment 2 updates the model. Each environment in its turn takes a step based on gradient of its loss over the batch w.r.t its model parameters. In its turn the environment 1 updates $\boldsymbol{w}_1^t = \Pi_{\mathcal{W}}[\boldsymbol{w}_1^{t-1} - \beta \nabla_{\boldsymbol{w}_1^{t-1}} \hat{R}_1(\boldsymbol{w}_1^{t-1}, \boldsymbol{w}_2^{t-1})]$, while the environment 2 does not update the model, $\boldsymbol{w}_2^t = \boldsymbol{w}_2^t - 1$, and then t is incremented by 1. In the next turn, the same procedure is repeated with the roles of environment 1 and 2 reversed, i.e., environment 2 updates and environment 1 does not. We continue this cycle of updates until a fixed number of epochs. In our experiments, we set $\beta = 0.005$, the batch size was set to 128 and the total number of epochs were set to 200 (each epoch is equal to the size of the training data divide by the batch size).

For the implementation of IRM, we needed to change the cross-validation procedure in the implementation provided by [Arjovsky et al., 2019]. The cross-validation procedure in requires access to data from a separate validation environment with a different distribution. Since we only use two environments, we use the cross-validation procedure called the train-domain validation set procedure (defined in [Gulrajani and Lopez-Paz, 2020]), which requires us to split each train environment into a train portion and a validation portion. It finally requires to combine all the validation splits and use them as one validation split. We use a 4:1 split. Besides this change in cross-validation procedure, the rest of the implementation comes from `https://github.com/facebookresearch/InvariantRiskMinimization/`.

7.9 Further details on Figure 3

Below we provide tables (Table 4, 5, 6, 7) containing numerical values and the standard deviation associated with model estimation error shown in Figure 3. ICP can often be conservative in accepting a covariate as a direct cause, which is the reason we see in some rows the entry against ICP is 5.0 ± 0.0; it does not accept any covariate as cause.

7.10 Extra experiments

In this section, we repeat the experiments shown in Figure 3 for a 100 dimensional setup with 1000 training samples. In Table 3, we show the results for the experiments. We find our approach is consistently better. It is not possible to compare with ICP owing to the computational intractability of the procedure for this setup.

7.11 Details for colored MNIST experiments

7.12 Colored MNIST Digits

We use the exact same environment as in [Arjovsky et al., 2019]. [Arjovsky et al., 2019] propose to create an environment for training to classify digits in MNIST digits data [4], where the images in MNIST are now colored in such a way that the colors spuriously correlate with the labels. The task is to classify whether the digit is less than 5 (not including 5) or more than 5. There are three environments (two training containing 30,000 points each, one test containing 10,000 points) We add noise

[4] `https://www.tensorflow.org/api_docs/python/tf/keras/datasets/mnist/load_data`

Setting	ERM	IRM	LRG	C-LRG
F-HET	12.37	10.75	12.32	4.92
P-HET	13.31	11.24	13.27	5.78
F-HOM	45.01	41.79	45.73	1.92
P-HOM	45.49	42.21	46.19	3.43

Table 3: Comparisons across different linear SEMs for number of training samples = 1000 and feature dimension 100

to the preliminary label ($\tilde{y} = 0$ if digit is between 0-4 and $\tilde{y} = 1$ if the digit is between 5-9) by flipping it with 25 percent probability to construct the final labels. We sample the color id z by flipping the final labels with probability p_e, where p_e is 0.2 in the first environment, 0.1 in the second environment, and 0.9 in the third environment. The third environment is the testing environment. We color the digit red if $z = 1$ or green if $z = 0$. The results reported in Table 3 are averaged over 10 random pairs of training environments.

Architecture for our method: Each training environment corresponds to a player and there is an individual model that is assigned to each environment. We assume that all the environments use the same architecture described as follows. The model used is a simple multilayer perceptron with following parameters.

- Input layer: Input batch $(\mathsf{batch}, \mathsf{len}, \mathsf{wid}, \mathsf{depth}) \rightarrow$ Flatten
- Layer 1: Fully connected layer, output size = 390, activation = ELU, Dropout = 0.75
- Layer 2: Fully connected layer, output size = 390, activation = ELU, Dropout = 0.75
- Output layer: Fully connected layer, output size = 2

Other details: We use Adam optimizer with a learning rate of $5e-4$. We use a batch size of 256. We use a warm start phase as described in [Ahuja et al., 2020] of length 100. We use ℓ_∞ constraint of 0.1. For F-IRM game [Ahuja et al., 2020], we use the same architecture that is described above with all the same parameters (except we do not have ℓ_∞ constraints in F-IRM game).

For IRM, we use the same architecture and optimizer parameters as described in [Arjovsky et al., 2019].

Method	**Samples**	**Error**
IRM	20	$2,72 \pm 0.53$
ICP	20	4.85 ± 0.10
ERM	20	2.82 ± 0.49
C-LRG	20	7.96 ± 0.67
IRM	100	0.96 ± 0.17
ICP	100	3.46 ± 0.37
ERM	100	1.16 ± 0.06
C-LRG	100	7.98 ± 0.59
IRM	250	0.59 ± 0.10
ICP	250	0.42 ± 0.21
ERM	250	1.12 ± 0.05
C-LRG	250	1.11 ± 0.05
IRM	500	0.90 ± 0.09
ICP	500	0.01 ± 0.001
ERM	500	1.17 ± 0.03
C-LRG	500	0.65 ± 0.04
IRM	750	0.54 ± 0.10
ICP	750	0.005 ± 0.0001
ERM	750	1.09 ± 0.03
C-LRG	750	0.42 ± 0.02
IRM	1000	0.51 ± 0.11
ICP	1000	0.002 ± 0.0003
ERM	1000	1.12 ± 0.03
C-LRG	1000	0.43 ± 0.02

Table 4: Comparisons for F-HET

Method	Samples	Error
IRM	20	2.48 ± 0.34
ICP	20	5.00 ± 0.00
ERM	20	3.59 ± 0.51
C-LRG	20	9.31 ± 0.87
IRM	100	1.28 ± 0.20
ICP	100	3.49 ± 0.56
ERM	100	1.37 ± 0.10
C-LRG	100	9.86 ± 1.08
IRM	250	0.95 ± 0.16
ICP	250	2.33 ± 0.69
ERM	250	1.22 ± 0.07
C-LRG	250	1.23 ± 0.08
IRM	500	1.01 ± 0.11
ICP	500	3.01 ± 0.77
ERM	500	1.33 ± 0.09
C-LRG	500	0.89 ± 0.09
IRM	750	0.85 ± 0.15
ICP	750	2.01 ± 0.77
ERM	750	1.18 ± 0.05
C-LRG	750	0.53 ± 0.04
IRM	1000	0.88 ± 0.14
ICP	1000	2.50 ± 0.79
ERM	1000	1.19 ± 0.05
C-LRG	1000	0.55 ± 0.03

Table 5: Comparisons for P-HET

Method	Samples	Error
IRM	20	3.89 ± 0.50
ICP	20	5.02 ± 0.02
ERM	20	4.82 ± 0.57
C-LRG	20	6.14 ± 0.66
IRM	100	3.06 ± 0.12
ICP	100	7.91 ± 0.46
ERM	100	4.35 ± 0.12
C-LRG	100	4.22 ± 0.55
IRM	250	2.83 ± 0.06
ICP	250	7.95 ± 0.47
ERM	250	4.29 ± 0.12
C-LRG	250	3.52 ± 0.24
IRM	500	3.21 ± 0.09
ICP	500	6.50 ± 0.58
ERM	500	4.45 ± 0.05
C-LRG	500	0.38 ± 0.05
IRM	750	2.99 ± 0.04
ICP	750	5.00 ± 0.00
ERM	750	4.47 ± 0.04
C-LRG	750	0.05 ± 0.008
IRM	1000	3.04 ± 0.06
ICP	1000	5.00 ± 0.00
ERM	1000	4.51 ± 0.07
C-LRG	1000	0.03 ± 0.003

Table 6: Comparisons for F-HOM

Method	Samples	Error
IRM	20	4.03 ± 0.41
ICP	20	5.38 ± 0.14
ERM	20	4.57 ± 0.69
C-LRG	20	8.03 ± 0.56
IRM	100	3.39 ± 0.32
ICP	100	6.55 ± 0.52
ERM	100	4.25 ± 0.17
C-LRG	100	6.24 ± 0.77
IRM	250	2.95 ± 0.09
ICP	250	5.00 ± 0.00
ERM	250	4.10 ± 0.18
C-LRG	250	3.27 ± 0.26
IRM	500	3.02 ± 0.09
ICP	500	5.00 ± 0.00
ERM	500	4.54 ± 0.13
C-LRG	500	0.56 ± 0.09
IRM	750	2.81 ± 0.10
ICP	750	5.00 ± 0.00
ERM	750	4.39 ± 0.09
C-LRG	750	0.08 ± 0.009
IRM	1000	2.88 ± 0.06
ICP	1000	5.00 ± 0.00
ERM	1000	4.35 ± 0.12
C-LRG	1000	0.05 ± 0.009

Table 7: Comparisons for P-HOM

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.

[Gulrajani and Lopez-Paz, 2020] Gulrajani, I. and Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

[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.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

[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*.

[Zhou et al., 2017] Zhou, Z., Mertikopoulos, P., Moustakas, A. L., Bambos, N., and Glynn, P. (2017). Mirror descent learning in continuous games. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5776–5783. IEEE.