Optimization-based Causal Estimation from Heterogenous Environments

Mingzhang Yin [1] **Yixin Wang** [2] **David M. Blei** [1]

Abstract

This paper presents an optimization approach to causal estimation. In classical machine learning, the goal of optimization is to maximize predictive accuracy. However, some covariates might exhibit non-causal association to the outcome. Such spurious associations provide predictive power for classical ML, but prevent us from interpreting the result causally. This paper proposes CoCo, an optimization algorithm that bridges the gap between pure prediction and causal inference. CoCo leverages the recently-proposed idea of environments. Given datasets from multiple environments—and ones that exhibit enough heterogeneity—CoCo maximizes an objective for which the only solution is the causal solution. We describe the theoretical foundations of this approach and demonstrate its effectiveness on simulated and real datasets. Compared to classical ML and the recently-proposed IRMv1, CoCo provides more accurate estimates of the causal model.

1. Introduction

Consider a classical machine learning (ML) problem. We observe a dataset of $\boldsymbol{x}_i, y_i$ pairs; $\boldsymbol{x}$ contains p covariates and y is an outcome. In classical ML, our goal is to be able to predict y from $\boldsymbol{x}$. We want to learn a model $y = f(\boldsymbol{\alpha}^\top \boldsymbol{x})$, inferring the unknown coefficients $\boldsymbol{\alpha}$ from the training data.

Suppose our goal is not purely predictive, but is one of *causal estimation*. Consider that the outcome y is drawn from a true data generating process (DGP) that involves *direct causes*, a subset of the p covariates. For simplicity we assume linearity, that the outcome is truly drawn from $y_i = g(\boldsymbol{\beta}^\top \boldsymbol{x}) + \epsilon$, where $\beta_k \neq 0$ for the direct causes, $\beta_k = 0$ for the other covariates, and ϵ is independent noise. The causal goal is to estimate the true *causal coefficients* $\boldsymbol{\beta}$, including both its structure and its value.

Classical empirical risk minimization (ERM) cannot reliably solve this problem. In its search for finding the best predictor, it will capitalize on spurious (non-causal) associations between the components of $\boldsymbol{x}$ and the outcome y, such as due to confounding or conditioning on a collider. Consequently, $\hat{\boldsymbol{\alpha}}$, the resulting estimate of the coefficients, will be a biased estimate of the causal coefficients $\boldsymbol{\beta}$.

We develop CoCo, an optimization-based approach to estimate causal coefficients. We derive CoCo in two steps. In the first step, we posit a risk-based objective for which the causal coefficients are one of several optima. In the second step, we use the idea of *invariant environments*—multiple training datasets that leave the targeted causal coefficients intact—to whittle down the number of optima and leave only the causal coefficients. The result is a practical algorithm that analyzes data from multiple environments to produce an estimate of the causal coefficients. On synthetic and real-world data, relative to ERM and invariant risk minimization (IRM) [1], CoCo better estimates the causal coefficients and predicts more robustly on new data.

2. Causal estimation as optimization

Assumptions. Consider an observed dataset of n datapoints $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{Y} \in \mathbb{R}^n$ is the outcome variable, $\mathbf{X} = [X_1, \cdots, X_p] \in \mathbb{R}^{n \times p}$ are the covariates. Each column $X_j \in \mathbb{R}^n,\ j \in \{1, 2, \cdots, p\}$ is the observations of the j-th covariate on the n units.

Assume the underlying DGP of the outcome variable follows a linear SEM [11] as

$$y \leftarrow \mu + \boldsymbol{\beta}^\top \boldsymbol{x} + \epsilon, \qquad \boldsymbol{x} \sim P(x_1, \cdots, x_p), \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the *causal coefficients* [12]. (We absorb the intercept term into $\boldsymbol{x}$ and $\boldsymbol{\beta}$.) Eq. (1) assumes linearity; we extend to nonlinear causal models in Appendix A.

The causal coefficient might contain some zeros, indicating covariates that are not causally connected to the outcome. The indices of non-zero coefficients are the *support set* of $\boldsymbol{\beta}$, denoted $S \subset \{1, 2, \cdots, p\}$, where $\beta_j \neq 0$ for $j \in S$ and $\beta_j = 0$ for $j \notin S$. The support set represents the set of direct causes of the outcome.

[1] Columbia University [2] University of Michigan. Correspondence to: Mingzhang Yin <mzyin11@gmail.com>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

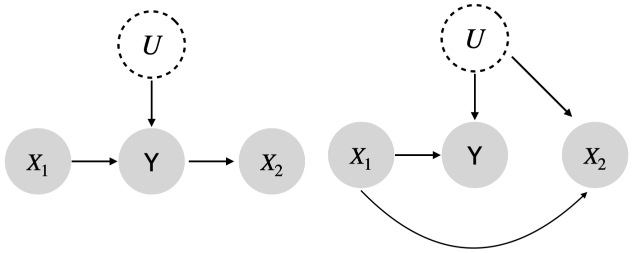


Figure 1. Example of spurious associations. X_1 is a cause of the outcome Y and X_2 is a spurious variable. Spurious association might be due to the observed descendants (left) and the unobserved common causes (right).

Assume the noise is zero-mean $\mathbb{E}[\epsilon] = 0$, the covariates and noise have finite variance Var$[x_j]$, Var$[\epsilon] < \infty$ for all j. We do not specify the SEM for the covariates $\boldsymbol{x}$ and allow their joint distribution to be arbitrary. Assume the causal covariates $\boldsymbol{x}_S$ are independent of the noise, i.e. $\boldsymbol{x}_S \perp\!\!\!\perp \epsilon$. This independence assumption implies that there is no unmeasured confounding between the true causes and the outcome. Note that we allow unmeasured confounding between noncauses $\boldsymbol{x}_{\setminus S}$ (or spurious variables) and the outcome. Fig. 1 shows the examples of endogenous covariates.

Pure prediction is biased. With data from Eq. (1) our goal is to estimate $\boldsymbol{\beta}$. We minimize the L_2 risk as a function of $\boldsymbol{\alpha}$

$$R(\boldsymbol{\alpha}; y, \hat{y}) = \mathbb{E}[(1/2)(\hat{y}(\boldsymbol{x}, \boldsymbol{\alpha}) - y)^2] \tag{2}$$

with linear predictions

$$\hat{y}(\boldsymbol{x}, \boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \boldsymbol{x}, \quad \boldsymbol{\alpha} \in \mathbb{R}^p. \tag{3}$$

When the form of $\hat{y}$ is clear in context, we abbreviate the risk function notation as $R(\boldsymbol{\alpha})$. Under the potential existence of spurious variables, directly optimizing Eq. (2) will produce biased estimates of the causal coefficients.

Idealized causal optimization. Suppose we do not know the coefficients $\boldsymbol{\beta}$, but we do know which covariates are direct causes, the set S. A key observation, though a simple one, is that among the models that share the true causal structure, the causal model has the best predictive accuracy.

Lemma 1 (Causal Optimality)**.** *The causal model with $\boldsymbol{\alpha} = \boldsymbol{\beta}$ is the optimal solution of the following constrained optimization problem*

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} R(\boldsymbol{\alpha}; y, \hat{y} = \boldsymbol{\alpha}^\top \boldsymbol{x}) \\ &\text{s.t. } \alpha_j = 0, \quad j \notin S. \end{aligned} \tag{4}$$

The proof is in Appendix C.

Lemma 1 is conceptually straightforward, but it provides a direct connection between optimization and causal estimation. Of course, in practice, we do not know which covariates are causal and which are not. We will build on this idealized optimization problem to construct a tractable objective for causal estimation.

3. Relaxed optimization for causal estimation

In this section, we use directional derivative to create a new optimization objective for causal estimation.

Directional derivatives and feasible directions. We first review the ideas of directional derivatives and its application in constrained optimization [16]. Consider a unit-length vector $\boldsymbol{v}$, where $\|\boldsymbol{v}\|_2 = 1$. The directional derivative in the direction $\boldsymbol{v}$ is denoted as operator $\mathbf{D}_{\boldsymbol{v}}$ and is defined to be the rate of change of a function in that direction. The directional derivative, as a scalar, can be computed as the inner product of the gradient and the direction vector

$$\mathbf{D}_{\boldsymbol{v}} R(\boldsymbol{\alpha}) := \lim_{t \to 0} \frac{R(\boldsymbol{\alpha} + t\boldsymbol{v}) - R(\boldsymbol{\alpha})}{t} = \langle \nabla R(\boldsymbol{\alpha}), \boldsymbol{v} \rangle,$$

where we denote an inner product as $\langle \cdot, \cdot \rangle$.

In Eq. (4), denote the constraints as $g_j(\boldsymbol{\alpha}) = \alpha_j = 0$ for $j \notin S$. (Recall S is the support set of $\boldsymbol{\beta}$, the indices of the non-zero causal coefficients.) Given a parameter $\boldsymbol{\alpha}$ and the optimization problem Eq. (4), the directions that violate the constraints at the maximum rate are the gradient direction of the constraint function,

$$dg_j(\boldsymbol{\alpha})/d\boldsymbol{\alpha} = \mathbf{e}_j, \quad j \notin S. \tag{5}$$

The feasible directions are defined as the directions orthogonal to the gradient of the constraints [10]. The feasible directions for Eq. (4) form a linear space $\mathcal{U} = \text{span}\{\mathbf{e}_j : j \in S\}$.

The first-order condition for a point $\boldsymbol{\alpha}$ to be an optima is that the directional derivative in the feasible directions vanish [10; 16]. For problem Eq. (4), this condition means $\mathbf{D}_{\boldsymbol{v}} = 0$ for each $\boldsymbol{v} \in \mathcal{U}$. The first order condition can be guaranteed by

$$\mathbf{D}_{\mathbf{e}_j} R(\boldsymbol{\alpha}) = \langle \nabla R(\boldsymbol{\alpha}), \mathbf{e}_j \rangle = 0, \text{ for } j \in S, \tag{6}$$

because $\boldsymbol{v} \in \mathcal{U}$ is a linear combination of the basis $\{\mathbf{e}_j\}_{j \in S}$. More compactly, these conditions can be written as

$$\|\nabla R(\boldsymbol{\alpha}) \circ \boldsymbol{\beta}\|_2 = 0 \tag{7}$$

with $\circ$ as Hadamard product because $\beta_j \neq 0$ for $j \in S$.

Relaxing the causal optimization. Lemma 1 states that $\boldsymbol{\alpha} = \boldsymbol{\beta}$ is an optimal solution of the problem in Eq. (4), which means that it satisfies the first-order condition of Eq. (7). Plugging $\boldsymbol{\alpha} = \boldsymbol{\beta}$ into this condition reveals that $\|\nabla R(\boldsymbol{\beta}) \circ \boldsymbol{\beta}\|_2 = 0$. This fact, in turn, means that the causal coefficients $\boldsymbol{\beta}$ is an optimum of the following optimization problem,

$$\min_{\boldsymbol{\alpha}} \|\nabla R(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2 . \tag{8}$$

Notice that Eq. (8) is an optimization problem that is entirely a function of the observed data; it does not require knowing the set S of non-zero causal coefficients. Yet the true causal coefficient $\boldsymbol{\beta}$ is an optimum of this problem.

We call the set of points that minimizes Eq. (8) the *plausible set* $\mathcal{F}$. However, the plausible set does not only contain the causal coefficients. By Eq. (8), the plausible set also contains the all-zero vector $\mathbf{0}$, the OLS solution, and the points "in-between" these two solutions, those that zero a subset of the covariates and zero the gradient of the remainder. Our next step is to whittle down the plausible set until the causal coefficients become the unique solution.

4. Optimization with multiple environments

In this section, we describe the invariance property of causal coefficients under interventions and how it helps restore the identifiability of the causal model via optimization.

Environments and invariance. As discussed in § 3, the causal model in general is non-identifiable by optimization with i.i.d. data. To obtain identifiability, we turn to the settings where data from multiple environments are available.

Denote $\mathcal{E}$ as a set of environments. An environment $e \in \mathcal{E}$ specifies a DGP. For linear model, data from multiple environments is generated by an SEM similar to Eq. (1):

$$y^e \leftarrow \boldsymbol{\beta}^\top \boldsymbol{x}^e + \epsilon^e, \quad \boldsymbol{x}^e \sim P^e(x_1^e, \cdots, x_p^e). \tag{9}$$

The independence assumptions and moments conditions in § 2 apply to all environments. Across environments, the set of observed variables, the set of direct causes of the outcome, and the causal coefficients remain the same.

The environments are a set of DGPs. The environments are heterogeneous if the DGPs are different. Heterogeneous environments can be constructed by (hard) interventions that fix a variable at a specific value. They can also be DGPs where the joint distributions of variables are different in nature, also known as soft interventions [6]. For example, when studying the effect of health measurements on the probability of cancer, the environments can be different hospitals where the data are collected from [17].

The key property that environments enjoy is the invariance [1; 12]. Invariance assumes conditional on the same value of direct causes, the expectation of the outcome is the same across environments, i.e.

$$\mathbb{E}[y^e|\text{Pa}(y^e) = \mathbf{c}] = \mathbb{E}[y^{e'}|\text{Pa}(y^{e'}) = \mathbf{c}], \tag{10}$$

for all $e, e' \in \mathcal{E}$. Note that we ask that the distribution of covariates $\boldsymbol{x}^e$ and noise ϵ^e changes across $e \in \mathcal{E}$, which makes spurious associations vary with environments.

Narrowing down the optima set by environments. The invariance property motivates us to aggregate the optimization problems and get the CoCo objective:

$$\min_{\boldsymbol{\alpha}} f_{\mathcal{E}}(\boldsymbol{\alpha}) := \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \big(\|\nabla R^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2 \big). \tag{11}$$

Denote the solutions of each environment as $\mathcal{F}^e := \arg\min_{\boldsymbol{\alpha}} \|\nabla R^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2$. The solutions of the CoCo objective Eq. (11) is the intersection of all $\mathcal{F}^e$s,

$$\mathcal{F}^{\mathcal{E}} := \arg\min_{\boldsymbol{\alpha}} f_{\mathcal{E}}(\boldsymbol{\alpha}) = \bigcap_{e \in \mathcal{E}} \mathcal{F}^e, \tag{12}$$

as long as the intersection is not empty; this fact is guaranteed by the invariance assumption with $\boldsymbol{\beta} \in \mathcal{F}^e$ for all e. Because of the intersection, the size of $\mathcal{F}^{\mathcal{E}}$ shrinks with the increasing number of environments, i.e. $|\mathcal{F}^{\mathcal{E}_1}| \leq |\mathcal{F}^{\mathcal{E}_2}|$ if $\mathcal{E}_2 \subset \mathcal{E}_1$. The multiple environments and heterogeneity therein induce differences among the set of solutions and, as a result, narrow down the solution set of CoCo objective Eq. (11). The plausible sets are visualized with examples in Fig. 4 in Appendix F. Yet, as a last step, we need to remove the all-zero vector from the solutions.

Removing non-informative solution from the optima set. We propose two modifications to remove the zero vector from the objective solutions.

Suppose there is prior knowledge of the underlying causal graph. Specifically, suppose a set $\mathcal{C}$ of covariates that are known to be independent of the unobserved noise ϵ of the outcome (e.g. some non-descendants of the outcome). We modify the CoCo objective Eq. (11) to be

$$\min_{\boldsymbol{\alpha}} \sum_{e \in \mathcal{E}} \|\nabla R^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}]\|_2\,, \tag{13}$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \circ (\mathbf{1} - \mathbf{1}_{\mathcal{C}}) + \mathbf{1}_{\mathcal{C}}$. For the risk function Eq. (2), $\nabla R(\boldsymbol{\beta})_j = \mathbb{E}[x_j \epsilon] = 0$ for $j \in \mathcal{C}$. So the causal coefficient $\boldsymbol{\beta}$ remains as a solution to the modified objective Eq. (13). The solution set of Eq. (13) is a subset of that of Eq. (11) with the all-zero vector being removed. The algorithm is summarized in Alg. 1. The theoretical results on causal identification is presented in Appendix D. CoCo objective is related to the invariant risk minimization (IRM) [1]; we illustrate the mathematical connections in Appendix B.

If there is no prior knowledge of the graph, we can add the risk function as a regularization term to Eq. (11) as

$$\min_{\boldsymbol{\alpha}} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \big\{ \|\nabla R^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2 + \lambda_r R^e(\boldsymbol{\alpha}) \big\}, \tag{14}$$

where $\lambda_r \geq 0$ controls the regularization strength. The regularization encourages the causal solution $\boldsymbol{\beta}$ against the zero vector because the zero vector has lower predictive accuracy than the causal model by the DGP. Solving Eq. (14)

Algorithm 1 CoCo for Causal Inference

```
input : Data $\mathbf{D}^e = \{\mathbf{Y}^e, \mathbf{X}^e\}$, $\mathbf{X}^e \in \mathbb{R}^{n^e \times p}$; the risk function $R^e$ for each environment $e \in \mathcal{E}$; the set of known non-descendant variables $\mathcal{C}$; the predictor $f(\cdot)$.
output : Coefficient estimation $\boldsymbol{\alpha}$ with causal interpretation.
Initialize $\boldsymbol{\alpha}$ randomly
while not converged do
    for e in $\mathcal{E}$ do
        Compute the gradient of the empirical risk: $\boldsymbol{g}^e(\boldsymbol{\alpha}) = (1/n_e)\frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{i=1}^{n_e} R^e(\boldsymbol{\alpha}; y_i^e, \hat{y}_i^e),\ \hat{y}_i^e = f(\boldsymbol{x}_i^e; \boldsymbol{\alpha})$
        $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \circ (\mathbf{1} - \mathbf{1}_{\mathcal{C}}) + \mathbf{1}_{\mathcal{C}}$
        Compute the loss: $\mathcal{L}^e(\boldsymbol{\alpha}) = \|\boldsymbol{g}^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}\|_2$
    end
    Update $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{e \in \mathcal{E}} \mathcal{L}^e(\boldsymbol{\alpha})$ with step size $\eta$
end
```

Algorithm 2 CoCo for Robust Prediction

```
input : Data $D^e = \{\mathbf{Y}^e, \mathbf{X}^e\}$, $\mathbf{X}^e \in \mathbb{R}^{n^e \times p}$, the risk function $R^e$ for each environment $e \in \mathcal{E}$; predictor $f_{\boldsymbol{\alpha}}(\cdot)$; regularizer coefficients $\lambda_r$, $\lambda_w$.
output : Predictor $f_{\boldsymbol{\alpha}}(\cdot)$ that is robust to interventions
Initialize $\boldsymbol{\alpha}$ randomly
while not converged do
    for e in $\mathcal{E}$ do
        Compute the gradient of the empirical risk: $\boldsymbol{g}^e(\boldsymbol{\alpha}) = (1/n_e)\frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{i=1}^{n_e} R^e(\boldsymbol{\alpha}|y_i^e, \hat{y}_i^e),\ \hat{y}_i^e = f(\boldsymbol{x}_i^e; \boldsymbol{\alpha})$
        Compute: $\mathcal{L}^e(\boldsymbol{\alpha}) = \|\boldsymbol{g}^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}\|_2$
        (Optional step:) add weak condition $\mathcal{L}^e(\boldsymbol{\alpha})$ += $\lambda_w(\langle \boldsymbol{g}^e(\boldsymbol{\alpha}), \boldsymbol{\alpha} \rangle)^2$
        Add risk function as regularization: $\mathcal{L}^e(\boldsymbol{\alpha})$ += $\lambda_r(1/n_e)(\sum_{i=1}^{n_e} R^e(\boldsymbol{\alpha}|y_i^e, \hat{y}_i^e))$
    end
    Update $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{e \in \mathcal{E}} \mathcal{L}^e(\boldsymbol{\alpha})$ with step size $\eta$
end
```

produces a model that enjoys distributional robustness under interventions [13]. The algorithm is summarized in Alg. 2.

Generalizing to nonlinear model. We extend CoCo to predictors that is a nonlinear mapping of linear combinations of covariates, i.e. $\hat{y} = f(\mathbf{A}\boldsymbol{x})$, in Appendix A. It includes fully connected neural network as a special case. The key is to build a constrained optimization problem similar to Eq. (4) and show that it admits the causal coefficient as an optimum. The analysis presented in §§ 3 and 4 can then be applied to such nonlinear models.

5. Empirical Studies

We study CoCo on simulated and real data. Across datasets, we find that CoCo produces an unbiased estimate of the causal structure and coefficients. CoCo can generalize its predictive ability from observed environments to new environments. The details about the data generation and algorithms implementation in this section are in Appendix E.

5.1. Linear Synthetic Data

We study causal inference with optimization-based methods. The data are generated from 5 different graphs in Fig. 3 in Appendix E, each including a spurious variable.

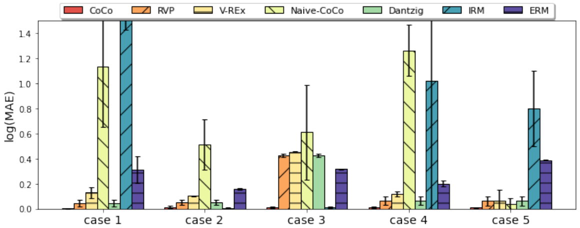


Figure 2. The mean absolute error of the estimations for causal parameters $\boldsymbol{\beta}$. CoCo estimation has small bias across data.

The five graphs test different scenarios: (1) independent causes; (2) observed mediator; (3) observed confounder and mediator; (4) observed confounder and unobserved mediator; (5) collider. We evaluate with mean absolute error (MAE) between the estimation $\boldsymbol{\alpha}$ and true coefficients $\boldsymbol{\beta}$.

The results are summarized in Fig. 2. It shows when the covariates have spurious associations with the outcome, the estimate of ERM is biased. IRM with proper hyper-parameter performs well in cases 2,3 while it has a large error in other cases. CoCo has small estimation error in all cases. As an ablation study, we replace the strong penalty in CoCo objective Eq. (13) with the weak penalty of Eq. (20) and minimize $\sum_{e \in \mathcal{E}}(\langle \nabla R^e(\boldsymbol{\alpha}), \tilde{\boldsymbol{\alpha}} \rangle)^2$. This method is labeled Naive-CoCo. The comparison between Naive-CoCo with CoCo in cases 1-4 shows that it is crucial to design the objective based on strong condition Eq. (8) instead of weak condition Eq. (20).

5.2. Colored MNIST (CMNIST)

In this section, we study whether CoCo produce a nonlinear model that uses causal variables to make predictions. The covariates are the colored digit image and the binary label y^e corresponds to the digit being odd or even. By the data construction, the relationship between the digit shape and y^e is genuinely causative while the relationship between the color and y^e is spurious. A desired predictor is one that make prediction based on the digit shape rather than the color. A predictor using color information can neither accurately predict the label y^e at testing.

Empirical results. The results are shown in Fig. 8 in Appendix, and Table 1. In Table 1, ERM has the lowest accuracy in both training and testing. The reason might be it largely depends on the color information to predict rather than on the digit shape, whereas the label is generated from the digit shape. The testing accuracy for IRM increases in the early stage of training but drops in the later stage. We

Table 1. Predictive accuracy in training and testing environments for CMNIST, and Wildlife data. For CMNIST the prediction accuracy is reported for both clean and noised labels. The Oracle is the same predictor but trained on grey-scale images with ERM.

	CMNIST			Wildlife	
	Training ($\tilde{y}$)	Testing ($\tilde{y}$)	Testing (y^e)	Training	Testing
ERM	75.8	44.4	31.8	99.6	58.4
IRM	81.4	70.3	46.5	83.4	**84.9**
CoCo	93.0	**92.9**	**74.7**	86.1	**85.2**
Random guess	50	50	50	50	50
Oracle	99.3	97.9	74.8	-	-

hypothesize that the model at first improves the prediction by utilizing all information including the digit shape, but later it relies more on the color information to boost the predictive accuracy, which reduces the accuracy at the test time.

5.3. Natural Image Classification

In this example, following Cloudera [4], we adapt the iWildCam 2019 dataset [2] that contains wildlife images taken in the wild. The goal is to classify coyote and raccoon in images. The image background, such as plants and rocks, might be predictive to the species but in a spurious way. This spurious association changes across locations hence we consider the images taken at different locations as coming from different environments. Based on the setting of Cloudera [4], we use images from two locations as the training data, images from one location as the validation data, and images from another location as the test data,

The results are summarized in Table 1 and Fig. 8 in Appendix G. ERM has high accuracy in training but low accuracy at testing. CoCo performs on par with or slightly better than IRM. Comparing to ERM, both methods have a slight drop in training accuracy but significantly higher testing accuracy. CoCo has a small performance gap between training and testing, indicating that it is not predicting animal labels via information from image backgrounds, i.e., information that varies across environments.

6. Conclusion

This paper formulated causal estimation as an optimization problem. Using directional derivatives, we proposed the CoCo objective, a computationally tractable optimization method for estimating causal coefficients with datasets from multiple environments. We discussed the mathematical connection between CoCo and IRM. In empirical studies, we found that CoCo produces accurate causal estimation and distributionally robust predictions. CoCo is applicable to high dimensional data, and to linear and nonlinear models.

References

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

[2] Beery, S., Morris, D., and Perona, P. The iWildCam 2019 challenge dataset. *arXiv preprint arXiv:1907.07617*, 2019.

[3] Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. A causal framework for distribution generalization. *arXiv e-prints*, pp. arXiv–2006, 2020.

[4] Cloudera. Causality for machine learning, 2020. URL https://ff13.fastforwardlabs.com.

[5] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

[6] Eberhardt, F. and Scheines, R. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

[7] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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

[9] Kamath, P., Tangella, A., Sutherland, D. J., and Srebro, N. Does invariant risk minimization capture invariance? *arXiv preprint arXiv:2101.01134*, 2021.

[10] Marban, J. A. *Directional derivatives in classical optimization*. PhD thesis, University of Florida, 1969.

[11] Pearl, J. *Causality*. Cambridge University Press, 2009.

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

[13] Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[14] Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[15] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[16] Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[17] Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

[18] Yin, M., Tucker, G., Zhou, M., Levine, S., and Finn, C. Meta-learning without memorization. In *International Conference on Learning Representations*, 2019.

Supplementary Material for "Optimization-based Causal Estimation from Heterogenous Environments"

A. Extension to Non-linear Models

In the main paper §§ 2 to 4, we focus on linear SEMs and linear predictors. Here we generalize these results to nonlinear models where the outcome is generated by a nonlinear function of a linear transformation of direct causes and the noise is additive. The key is to build a constrained optimization problem similar to Eq. (4) and show that it admits the causal coefficient as an optimum. The analysis presented in § 3 can then be applied to nonlinear models.

Suppose we have a collection of environments $\mathcal{E}$, and for each $e \in \mathcal{E}$, we observed i.i.d. data for variables $(\boldsymbol{x}^e, y^e)$, $\boldsymbol{x}^e \in \mathbb{R}^p$, $y^e \in \mathbb{R}$. Suppose the underlying DGP is

$$y^e \leftarrow f(\mathbf{A}\boldsymbol{x}^e_S; \boldsymbol{\gamma}^*) + \epsilon^e \tag{15}$$

where $S \subset \{1, 2, \cdots, p\}$, $\epsilon^e \perp\!\!\!\perp \boldsymbol{x}^e_S$ and $\mathbb{E}[\epsilon^e] = 0$. $f : \mathbb{R}^K \to \mathbb{R}$ is an arbitrary function mapping with parameters $\boldsymbol{\beta} = (\mathbf{A}, \boldsymbol{\gamma}^*)$ where $\mathbf{A} \in \mathbb{R}^{K \times |S|}$ and $\boldsymbol{\gamma}^* \in \mathbb{R}^M$. When $K = 1$ and $f(\cdot)$ is an identity mapping, Eq. (15) reduces to the linear SEM. Eq. (15) can represent a process when the outcome is generated through a deep neural network (DNN), where K and $\mathbf{A}$ are the width and weights of the first hidden layer respectively.

Assume the nonlinear predictor is

$$\hat{y}^e = f(\mathbf{B}\boldsymbol{x}^e; \boldsymbol{\gamma}), \tag{16}$$

where $\mathbf{B} \in \mathbb{R}^{K \times p}$, $\boldsymbol{\gamma} \in \mathbb{R}^M$ and $\boldsymbol{\alpha} = (\mathbf{B}, \boldsymbol{\gamma})$ are the parameters to optimize. We can re-write $\mathbf{A}\boldsymbol{x}^e_S = \mathbf{A}\Lambda\boldsymbol{x}^e$ where $\Lambda \in \mathbb{R}^{|S| \times p}$ has the i-th row as $\mathbf{e}_i^\top$ if $i \in S$ and as $\mathbf{0}_p^\top$ if $i \notin S$. Let $\mathbf{B}^* = \mathbf{A}\Lambda$ where the j-th column of $\mathbf{B}^*$ is $\mathbf{0}_K$ if $j \notin S$. Then for square error $R^e(\boldsymbol{\alpha})$ we have the following proposition.

Proposition 1 (Causal Optimality, nonlinear)**.** *The causal model $\mathbf{B} = \mathbf{B}^*$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ is an optima of the following problem*

$$\begin{aligned} &\min\ R^e(\mathbf{B}, \boldsymbol{\gamma}; y^e, \hat{y}^e) \\ &s.t.\ B_{kj} = 0 \textit{ if } B^*_{kj} = 0,\ \ 1 \le k \le K,\ 1 \le j \le p \\ &\qquad \gamma_m = 0 \textit{ if } \gamma^*_m = 0,\ \ 1 \le m \le M. \end{aligned} \tag{17}$$

Proposition 1 greenlights the analysis in § 3. It implies that the CoCo objective Eq. (11) and extension Eq. (14) can be used for nonlinear models. When $\mathbf{B} = \mathbf{B}^*$, the multiplication $\mathbf{B}\boldsymbol{x}^e$ zeros out non-causal covariates $\boldsymbol{x}^e_{\setminus S}$, which become independent of prediction $\hat{y}^e$.

In the nonlinear regime, due to high flexibility, identification can be difficult. Different parameterizations can represent similar mappings on the training data. Hence we expect the identification is up to an equivalent class [3; 8].

Another challenge is that a sufficiently flexible model may memorize all outcome labels and become the ERM optimum for each environment, but this model is not causative. To avoid such a solution, one approach is to collect overlapped environments, for example, those sharing the same domain of inputs [18].

B. Connection to invariant risk minimization

In this section, we discuss the connections and distinctions between CoCo and IRM. Arjovsky et al. [1] introduces IRM that can learn robust representation in the presence of spurious associations between covariates and the outcome. In particular, IRM considers a predictor $f(\boldsymbol{x}; \boldsymbol{\alpha}) : \mathbb{R}^p \mapsto \mathbb{R}$ with parameter $\boldsymbol{\alpha}$. In a setting similar to CoCo , it considers a set of heterogeneous environments $\mathcal{E}$ and for each $e \in \mathcal{E}$, a risk function $R^e(\boldsymbol{\alpha}; y, \hat{y})$. Based on the intuition that invariant predictor induces invariant features, IRM proposes the following objective to find an invariant model

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, w} \sum_{e \in \mathcal{E}} R^e(\boldsymbol{\alpha}; y^e, w(f(\boldsymbol{x}^e_i))) \\ &s.t.\ w \in \arg\min_{\bar{w}}\ R^e(\boldsymbol{\alpha}; y^e, \bar{w}(f(\boldsymbol{x}^e_i))),\ \text{for all } e \in \mathcal{E}, \end{aligned} \tag{18}$$

where $w(\cdot)$ is a mapping from the range of $f(\cdot)$ to $\hat{y}$.

For tractable computation, Arjovsky et al. [1] further introduces the IRMv1 objective:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \sum_{e \in \mathcal{E}} \Big[&\underbrace{R^e(\boldsymbol{\alpha}; y^e_i, f(\boldsymbol{x}^e_i))}_{\text{Empirical risk}} + \\ &\lambda \underbrace{||\nabla_{w|w=1.0} R^e(\boldsymbol{\alpha}; y^e_i, wf(\boldsymbol{x}^e_i))||_2^2}_{\text{Invariant risk}} \Big], \end{aligned} \tag{19}$$

where $\lambda > 0$ and w is simplified as a dummy scalar variable. The IRMv1 objective consists of an empirical risk term and an invariant risk term.

We make the connection between IRM and the constrained optimization in Lemma 1. In § 3, we obtain the first-order optimality condition Eq. (7) from the directional derivative in the feasible directions $\{\mathbf{e}_j\}_{j \in S}$. In fact, any vector in the space $\mathcal{U} = \text{span}\{\mathbf{e}_j : j \in S\}$ is a feasible direction. Specially, the causal parameter $\boldsymbol{\beta} \in \mathcal{U}$ is a feasible direction which implies the optima should have zero directional derivative in this direction, i.e. $\langle \nabla R(\boldsymbol{\alpha}), \boldsymbol{\beta} \rangle = 0$.

By Lemma 1, plugging $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ we get $\langle \nabla R(\boldsymbol{\beta}), \boldsymbol{\beta} \rangle = 0$, yielding another objective

$$\min_{\boldsymbol{\alpha}}\ (\langle \nabla R(\boldsymbol{\alpha}), \boldsymbol{\alpha} \rangle)^2 \tag{20}$$

that $\boldsymbol{\beta}$ satisfies. Similarly, any partition $\mathcal{P}$ of the set $\{1, 2, \cdots, p\}$ gives a necessary condition that admits causal model as an extreme point

$$\min_{\boldsymbol{\alpha}} \sum_{A\in\mathcal{P}} (\langle \nabla R(\boldsymbol{\alpha})_A, \boldsymbol{\alpha}_A\rangle)^2. \tag{21}$$

When the outcome model is Linear-Gaussian or Linear-Bernoulli, minimizing the invariant risk term in Eq. (19) is equivalent to Eq. (20). Suppose the DGP and the predictor are linear as in Eq. (9), and L_2 risk function R^e, then

$$\begin{aligned}\left\|\nabla_{w|w=1.0} R^e(\boldsymbol{\alpha}; y^e, w\boldsymbol{\alpha}^\top \boldsymbol{x}^e)\right\|_2^2 &= (\mathbb{E}[(y^e - \hat{y}^e)\boldsymbol{\alpha}^\top \boldsymbol{x}^e])^2\\ &= (\langle \nabla R^e(\boldsymbol{\alpha}; y, \hat{y}), \boldsymbol{\alpha}\rangle)^2, \quad (22)\end{aligned}$$

where the left side is the invariant risk term and the right side is objective in Eq. (20).

Similarly, suppose the outcome is generated by $y^e \leftarrow$ Bernoulli$(\sigma(\boldsymbol{\beta}^\top \boldsymbol{x}^e))$, the predictor is $\hat{y}^e = \sigma(\boldsymbol{\alpha}^\top \boldsymbol{x}^e)$ where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, and the risk function is the cross entropy loss $R^e(\boldsymbol{\alpha}; y^e, \hat{y}^e) = -\mathbb{E}[y^e \log(\hat{y}^e) + (1 - y^e)\log(1 - \hat{y}^e)]$, then

$$\begin{aligned}\left\|\nabla_{w|w=1.0} R^e(\boldsymbol{\alpha}; y^e, \sigma(w\boldsymbol{\alpha}^\top \boldsymbol{x}^e))\right\|_2^2 &= (\mathbb{E}[(\hat{y}^e - y^e)\boldsymbol{\alpha}^\top \boldsymbol{x}^e])^2\\ &= (\langle \nabla R^e(\boldsymbol{\alpha}; y, \hat{y}), \boldsymbol{\alpha}\rangle)^2. \quad (23)\end{aligned}$$

The connections between Eqs. (20), (22) and (23) explain the mechanism behind IRMv1 for linear Gaussian or Bernoulli models, as the causal coefficient belongs to the optima set of the invariant risk term.

The connection also indicates the sub-optimality of the IRMv1 objective. The invariance risk term, rewritten as the inner product between the gradient and parameter vectors, only considers a single feasible direction for the constrained optimization problem Eq. (4), among all feasible directions that form a $(p - |S|)$-dimensional linear space.

The spectrum between CoCo and invariant risk term, as shown in Eq. (21), tells that the finer the partition is, the smaller the optima set of Eq. (21) becomes. This means, among all conditions in the form of Eq. (21), the one given by CoCo as Eq. (8) is the strongest and the one given by IRMv1 as Eq. (20) is the weakest. Since the ultimate goal is to identify the causal coefficient, we prefer the strong condition that gives a small set of solutions in a single environment. See Appendix F for a case study comparing ERM, IRMv1 and CoCo analytically.

Because of an excessive number of solutions of the invariant risk term, IRMv1 puts high requirement on the number of environments and sufficiency of heterogeneity. In practice, there can be multiple parameters that minimize the IRMv1 objective, including that of non-causal models. By simulations in § 5, we will show that optimizing the IRMv1 objective can fail to produce robust predictions, especially when the outcome is generated neither from Linear-Gaussian nor Linear-Bernoulli models. Similar failure modes of IRM are studied in cases of a two-bit model [9] and a nonlinear classification model [14].

Lastly, we notice that adding any general condition in Eq. (21) to the strong condition Eq. (8) does not change the optima set while in practice it may improve the smoothness of the optimization landscape.

C. Proofs

In this section, we present proofs for the results in the main paper. The following proof is for Lemma 1.

Proof. Let the random vector $\boldsymbol{x} = (x_1, \cdots, x_p)^\top$ denote the covariates. The expected mean square error is

$$\begin{aligned}&\mathbb{E}[(y - \hat{y})^2]\\ =&\mathbb{E}[(\boldsymbol{\alpha}^\top \boldsymbol{x} - \boldsymbol{\beta}^\top \boldsymbol{x} - \epsilon)^2]\\ =&(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top](\boldsymbol{\alpha} - \boldsymbol{\beta}) - 2\mathbb{E}[(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \boldsymbol{x}\epsilon] + \mathbb{E}[\epsilon^2].\end{aligned}$$

Since $supp(\boldsymbol{\alpha}) = supp(\boldsymbol{\beta})$, the $(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \boldsymbol{x}$ is a linear combination of the true causes as $\sum_{j\in supp(\boldsymbol{\beta})}(\alpha_j - \beta_j)x_j$ which is independent of ϵ by the SEM, thus $\mathbb{E}[(\boldsymbol{\alpha} - \boldsymbol{\beta})^\top \boldsymbol{x}\epsilon] = 0$. Since $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ is assumed to be positive definite, the unique optima of the square error is $\boldsymbol{\alpha} = \boldsymbol{\beta}$. □

The following proof is for Proposition 1.

Proof. By the construction of Λ, $\mathbf{B}^* = \mathbf{A}\Lambda$ is a matrix where the j-th column $B^*_j = \mathbf{0}$ if $j \notin S$. Similar to the proof of Lemma 1, we can compute the L_2 risk as

$$\begin{aligned}&\mathbb{E}[(y - \hat{y})^2]\\ =&\mathbb{E}[(f_{\boldsymbol{\gamma}}(\mathbf{B}\boldsymbol{x}) - f_{\boldsymbol{\gamma}^*}(\mathbf{B}^*\boldsymbol{x}) - \epsilon)^2]\\ =&\mathbb{E}[(f_{\boldsymbol{\gamma}}(\mathbf{B}\boldsymbol{x}) - f_{\boldsymbol{\gamma}^*}(\mathbf{B}^*\boldsymbol{x}))^2]\\ &- 2\mathbb{E}[((f_{\boldsymbol{\gamma}}(\mathbf{B}\boldsymbol{x}) - f_{\boldsymbol{\gamma}^*}(\mathbf{B}^*\boldsymbol{x}))\epsilon] + \mathbb{E}[\epsilon^2].\end{aligned}$$

Due to the constraints, $B_j = B^*_j = \mathbf{0}$, $\mathbf{B}\boldsymbol{x} \perp\!\!\!\perp \epsilon$, $\mathbf{B}^*\boldsymbol{x} \perp\!\!\!\perp \epsilon$, therefore the second term is zero. Then the L_2 risk reaches its minimum as $\mathbb{E}[\epsilon^2]$ when $\mathbf{B} = \mathbf{B}^*$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$. □

D. Identification with Heterogeneous Environments

We establish causal identification for CoCo. Causal identification involves writing the causal quantity of interest as a functional of the observed data distribution; this functional is also known as the causal identification strategy. In the context of CoCo, we consider the functional that maps the joint distributions $p(\boldsymbol{x}^e, y^e)$ over a set of environments to the risk function, then to the optima of CoCo objective

Eq. (13). Causal identification for CoCo thus amounts to proving any optima of the CoCo objective must coincide with the causal coefficient of interest.

Since all solutions of Eq. (8) can be characterized analytically (see § 3), we are able to define what the general effective and ineffective interventions are. To keep notation consistent, denote $\mathcal{C}$ as the set of known non-descendants of the outcome and S as the unknown set of direct causes. For any set H with $\mathcal{C} \subset H \subset \{1, 2, \cdots, p\}$, we fit a regression model on X_H^e in each environment, and collect the regression coefficients as $\{\hat{\boldsymbol{\alpha}}_H^e\}_{e\in\mathcal{E}}$. We call the set H an *invariant set*, if the estimations

$$\hat{\boldsymbol{\alpha}}_H^e = \hat{\boldsymbol{\alpha}}_H^{e'} := \hat{\boldsymbol{\alpha}}_H, \quad \forall e, e' \in \mathcal{E}. \tag{24}$$

If H is an invariant set, we define a length p vector as an *invariant vector* by equating it to $\hat{\boldsymbol{\alpha}}_H$ when restricting to the set H and padding it with zeros at other elements. When there is more than one invariant vector, we call the interventions that construct the environments as *ineffective interventions*.

Back to the linear SEM and linear predictor. Denote $\mathcal{E}$ as a set of environments, $R^e(\boldsymbol{\alpha}) = \mathbb{E}[(1/2)(\hat{y}^e - y^e)]^2$ as the risk, and $\mathbf{W}^e := \mathbb{E}[\boldsymbol{x}^e(\boldsymbol{x}^e)^T] \in \mathbb{R}^{p\times p}$ as the Gram matrix which is assumed to be positive definite [15]. With all this in place, the causal relationship and causal effects can be identified by CoCo, as long as the interventions are valid and effective.

Theorem 1. *For the linear SEM in Eq. (9) and predictor in Lemma 1, assume $\mathbf{W}^e \succ 0$ for all $e \in \mathcal{E}$, and assume the following conditions hold:*

(A1) Validity: $\exists\, S \subset \{1, 2, \cdots, p\}$, $\boldsymbol{x}_S^e = Pa(y^e)$, and $\mathbb{E}[y^e|\boldsymbol{x}_S^e = \mathbf{c}] = \mathbb{E}[y^{e'}|\boldsymbol{x}_S^{e'} = \mathbf{c}]$ for all $\mathbf{c} \in \mathbb{R}^{|S|}$, $e, e' \in \mathcal{E}$.

(A2) Effectiveness: exploring all the sets H with $\mathcal{C} \subset H \subset \{1, 2, \cdots, p\}$, there are no distinct invariant vectors (defined in Eq. (24)).

Then the causal coefficients $\boldsymbol{\beta}$ are identifiable, and are given by

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \|\nabla R^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}]\|_2\,, \tag{25}$$

which is the solution of Eq. (13).

Proof. Let $s_j^e = \mathbb{E}[X_j^e \epsilon] = cov(X_j^e, \epsilon)$, $\mathbf{s}^e = (s_1^e, \cdots, s_p^e)^T$. By the data generating process, $s_j^e = 0$ for $j \in \{1, \cdots, K\}$. Let

$$g^e(\boldsymbol{\alpha}) = \|\nabla R^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}]\|_2\,, \;\; f(\boldsymbol{\alpha}) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} g^e(\boldsymbol{\alpha}). \tag{26}$$

where $f(\boldsymbol{\alpha})$ is CoCo objective. Direct computation shows

$$\nabla R^e(\boldsymbol{\alpha}) = W^e(\boldsymbol{\alpha} - \boldsymbol{\beta}) - \mathbf{s}^e \tag{27}$$

Notice $f(\boldsymbol{\alpha}) \geq 0$ and by the structual equation model, due to independence of the exogenous noise ϵ and causes $\text{Pa}(Y)$, we have $\mathbf{s}^e \circ \boldsymbol{\beta} = \mathbf{0}$. Hence for $\boldsymbol{\alpha}^* = \boldsymbol{\beta}$, $f(\boldsymbol{\alpha}^*) = 0$. This guarantees the existence of a solution as causal coefficient $\boldsymbol{\beta}$. To prove the identification, it is sufficient to prove that for all $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, $f(\boldsymbol{\alpha}) > 0$. We use proof by contradiction.

Let $H = \text{supp}(\tilde{\boldsymbol{\alpha}})$ and H^c as its compoment set in $\{1, 2, \cdots, p\}$. We assume $f(\boldsymbol{\alpha}) = 0$ and $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ and deduce a contradiction. Since $f(\boldsymbol{\alpha}) = 0$, for all e, $\|g^e(\boldsymbol{\alpha})\| = 0$. Since $g^e(\boldsymbol{\alpha}) = \nabla R^e(\boldsymbol{\alpha}) \circ \tilde{\boldsymbol{\alpha}}$, it means $\nabla R^e(\boldsymbol{\alpha})_H = \mathbf{0}$, for all e. However, by the characterization of the plausible set in Section 3, Assumption A2) implies that there does not exist $\boldsymbol{\alpha} \in \mathbb{R}^p$, $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$, such that $\nabla R^e(\boldsymbol{\alpha})_H = \mathbf{0}, \forall\, e \in \mathcal{E}$. Otherwise, the set H is an invariant set and both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are invariant estimations, which violates Assumption A2). Hence for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$, there exists an environment $e' \in \mathcal{E}$ with $\nabla R^{e'}(\boldsymbol{\alpha})_H \neq \mathbf{0}$. This yields a contradiction. □

E. Data generation and implementation details for § 5

Synthetic data. we generate data for the case 1 according to Fig. 3.

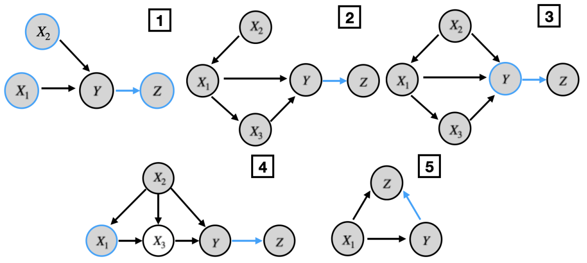


Figure 3. The graph for the simulation study in § 5.1. The case ID of each graph is in the rectangle box. The shaded nodes are the variables that are observed. z denotes a descendant but we do not know this during estimation.

The DGPs for case 1 is

$$\begin{aligned} x_2^e \leftarrow & \mathcal{N}(1, (\frac{1}{2})^2) \\ x_1^e \leftarrow & U(-1, 1) \\ y^e \leftarrow & 2x_1^e + x_2^e + \mathcal{N}(0, 1) \\ z^e \leftarrow & \gamma^e y^e + \mathcal{N}(0, 1). \end{aligned} \tag{28}$$

Case 2-3 with confounder and mediator are generated by the following SEM with $c_1 = 1, c_2 = 0$ for case 2, $c_1 =$

$1, c_2 = 1.5$ for case 3

$$\begin{aligned} x_2^e \leftarrow& \mathcal{N}(1, (\frac{1}{2})^2) \\ x_1^e \leftarrow& c_1 x_2^e + U(-1,1) \\ x_3^e \leftarrow& \sin(x_1^e) + \mathcal{N}(0, (\frac{1}{2})^2) \\ y^e \leftarrow& 2x_1^e + x_2^e + c_2 x_3^e + \mathcal{N}(0,1) \\ z^e \leftarrow& \gamma^e y^e + \mathcal{N}(0,1). \end{aligned} \tag{29}$$

Case 4 with an unobserved direct cause x_3^e is generated with

$$\begin{aligned} x_2^e \leftarrow& \mathcal{N}(1, (\frac{1}{2})^2) \\ x_1^e \leftarrow& x_2^e + U(-1,1) \\ x_3^e \leftarrow& x_1^e + x_2^e + \mathcal{N}(0, (\frac{1}{2})^2) \\ y^e \leftarrow& x_2^e + 2x_3^e + \mathcal{N}(0,1) \\ z^e \leftarrow& \gamma^e y^e + \mathcal{N}(0,1). \end{aligned} \tag{30}$$

Case 5 with a collider is generated with

$$\begin{aligned} x_1^e \leftarrow& \mathcal{N}(1, \gamma^e) \\ y^e \leftarrow& x_1^e + \mathcal{N}(0,1) \\ z^e \leftarrow& 2y^e + x_1^e + \mathcal{N}(0, \gamma^e). \end{aligned} \tag{31}$$

To generate data from different environments, we set the parameter γ^e in DGP (Eqs. (28), (30) and (31)) by $\gamma^e \in \{0.5, 2.0\}$; as required, this leaves the causal effect invariant. For IRM, we minimize IRMv1 objective Eq. (19) and report the hyper-parameter $\lambda \in \{2, 20, 200\}$ that gives the lowest MAE. The function mapping from the causes to the outcome is linear with additive noise. We specify x_1 as a known pre-outcome variable (for use of the method in Eq. (13)) and run CoCo , IRM, and ERM to estimate the causal coefficients.

CMNIST data. CMNIST is a semi-synthetic data set for binary classification, first introduced in Arjovsky et al. [1]. Based on the MNIST data set, the image of hand-written digits 0-4 and 5-9 are labels as $\tilde{y} = 0$ and $\tilde{y} = 1$ respectively. For each environment, the outcome y^e is generated with 0.75 probability as $\tilde{y}$ and with 0.25 probability as $1 - \tilde{y}$. We call $\tilde{y}$ the *clean labels* and y^e the *noised labels*. The digit is colored green with probability p^e if $y^e = 1$ and $1 - p^e$ if $y^e = 0$, otherwise it is colored red. The DGP across environments differs in p^e. Environments are constructed for training with $p^e \in \{0.1, 0.2\}$, for validation $p^e = 0.5$ and for testing $p^e = 0.9$.

The predictor is a fully connected neural network with two hidden layers. For CoCo, we use objective Eq. (14) to optimize the predictor. The risk penalty weight λ_r is chosen on the validation environment and is reduced by a factor of 10 when the parameters are sufficiently away from $\mathbf{0}$. For IRM, we use a learning rate as 10^{-4} to ensure stability over long iterations and use other hyper-parameters and annealing strategy provided by the author's code. [1]

Natural image data. The predictor is a fully connected neural network with one hidden layer of size 10. The inputs are 512-dimensional features extracted from ResNet18 [7], a pre-trained model on the ImageNet dataset [5].

In this example, we find for CoCo objective (14), adding the weak penalty (20) with weight λ_w improves convergence. It is possibly due to the smoothed landscape as discussed in Appendix B. We set the weight $\lambda_w = 10^4$. For both CoCo and IRM, we reduce the weight of risk regularization by a factor of 10^5 after 100 epochs. The parameters are selected on the validation environment. Here we find annealing the risk necessary for both methods otherwise minimizing the risk term often forces the predictor to use spurious associations after long iterations. The need of risk term being small may be due to a limited number of training environments.

F. Illustrative Example

In this section, we study the connection and distinction between ERM, IRM and CoCo by studying a specific example, which is adapted from the "minimal coding implementation" in [1] Appendix Section D. The DGP is:

$$\begin{aligned} &[x_1^e, x_2^e] \leftarrow [\mathcal{N}(0, e^2), \mathcal{N}(0, e^2)] \\ &[\epsilon_1^e, \epsilon_2^e] \leftarrow [\mathcal{N}(0, e^2), \mathcal{N}(0, e^2)] \\ &y^e \leftarrow x_1^e + x_2^e + \epsilon_1^e + \epsilon_2^e \\ &[z_1^e, z_2^e] \leftarrow [x_1^e + \epsilon_1^e + \mathcal{N}(0,1), x_2^e + \epsilon_2^e + \mathcal{N}(0,1)] \end{aligned} \tag{32}$$

The predictive model is $\hat{y}^e = \boldsymbol{\alpha}^\top \boldsymbol{x}^e$, where input $\boldsymbol{x}^e = (x_1^e, x_2^e, z_1^e, z_2^e)^\top$, parameter $\boldsymbol{\alpha} \in \mathbb{R}^4$. The risk function for the environment e is the mean square error, i.e. $R^e(\boldsymbol{\alpha}; y^e, \hat{y}^e) = \mathbb{E}[(1/2)(\hat{y}^e - y^e)^2]$. Variable $\boldsymbol{z} = (z_1, z_2)$ are associated with the outcome spuriously. Assume the number of training environments is $|\mathcal{E}| = K$. Direct computation gives

$$\begin{aligned} R^e(\boldsymbol{\alpha}) = \frac{1}{2}[&(\alpha_1 + \alpha_3 - 1)^2 e^2 + (\alpha_2 + \alpha_4 - 1)^2 e^2 \\ &+ (\alpha_3 - 1)^2 e^2 + (\alpha_4 - 1)^2 e^2 + \alpha_3^2 + \alpha_4^2] \end{aligned} \tag{33}$$

$$\begin{aligned} \nabla R^e(\boldsymbol{\alpha}) = \Big(&(\alpha_1 + \alpha_3 - 1)e^2, (\alpha_2 + \alpha_4 - 1)e^2, \\ &(\alpha_1 + \alpha_3 - 1)e^2 + (\alpha_3 - 1)e^2 + \alpha_3, \\ &(\alpha_2 + \alpha_4 - 1)e^2 + (\alpha_4 - 1)e^2 + \alpha_4\Big)^\top \end{aligned} \tag{34}$$

From Eq. (33), the optimum for ERM in environment e is

$$\hat{\boldsymbol{\alpha}}_{\text{ERM-e}} = (\frac{1}{1+e^2}, \frac{1}{1+e^2}, \frac{e^2}{1+e^2}, \frac{e^2}{1+e^2}), \tag{35}$$

[1] https://github.com/facebookresearch/InvariantRiskMinimization

and the optimum for the ERM over K environments is

$$\hat{\boldsymbol{\alpha}}_{\text{ERM}} =(\frac{K}{K+\sum_{e\in\mathcal{E}} e^2}, \frac{K}{K+\sum_{e\in\mathcal{E}} e^2}, \tag{36}$$

$$\frac{\sum_{e\in\mathcal{E}} e^2}{K+\sum_{e\in\mathcal{E}} e^2}, \frac{\sum_{e\in\mathcal{E}} e^2}{K+\sum_{e\in\mathcal{E}} e^2}) \tag{37}$$

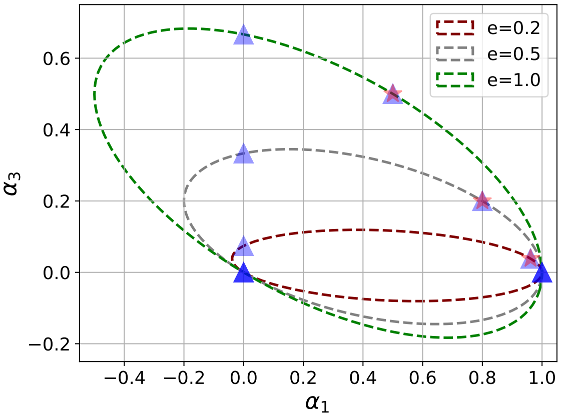


Figure 4. The solutions of ERM, IRM and CoCo for SEM Eq. (32). The ellipses represent the optimums for IRM in each environment, the triangular and star points on each ellipse are the optimum of CoCo and ERM for each environment respectively. Due to symmetry, the figure remains the same when replacing (α_1, α_3) with (α_2, α_4). The solutions of CoCo contains the causal model and the minimal number of non-causal models.

The outcome is generated from a Linear-Gaussian model, so the invariant risk term in IRMv1 objective Eq. (19) equals to Eq. (20) as shown in Appendix B. Therefore, we can solve $(\langle\nabla R^e(\boldsymbol{\alpha}; y^e, \hat{y}^e), \boldsymbol{\alpha}\rangle)^2 = 0$ and $||\nabla R^e(\boldsymbol{\alpha}) \circ \boldsymbol{\alpha}||_2 = 0$ and get the optima set of the invariant risk term and CoCo respectively.

As shown in Fig. 4, the solutions for the invariant risk term in environment e form an ellipse in the space of (α_1, α_3) and (α_2, α_4), which are infinite points. The optimum of CoCo is a strict subset of solutions of the invariant risk term, which has size of 2^4. For a single environment, CoCo solutions contain the ERM solution and the causal coefficients. With data from three environments, the causal coefficient $\boldsymbol{\beta} = (1, 1, 0, 0)^\top$ belongs to the solutions of the aggregated invariant risk term and CoCo, but does not belong to the solution of the aggregated ERM.

G. Additional Simulation Study

This section contains experimental results in addition to the simulations in § 5 in the main paper.

G.1. Model misspecification

We further study the implication of model misspecification using data from § 5.1 case 5. We compare two predictors, one is a linear model and the other is a nonlinear neural network. The data is generated linearly, so the linear predictor correctly specifies the model and the neural network misspecifies it. Both models are trained with ERM and CoCo. In case 5, the variable x is the cause and z is a predictive but non-causal covariate. We study how the two models, trained by the two methods, can generalize its predictive accuracy to new observation (x, z) respectively.

The results are shown in Appendix Fig. 5. When the model is correctly specified, ERM estimation cannot generalize to new values of z, while CoCo estimation can predict on any (x, z) accurately. When the model is misspecified, the model trained by ERM can only interpolate between the training points, while the model trained by CoCo can generalize to new z though not to new x.

This study demonstrates that if the model is not misspecified, CoCo can learn the causal model; if model is misspecified, CoCo learns a model that can generalize to new environments where the spurious association differs from that in the observed environments. In both scenarios, model learned by CoCo has better generalization performance than that learned by ERM.

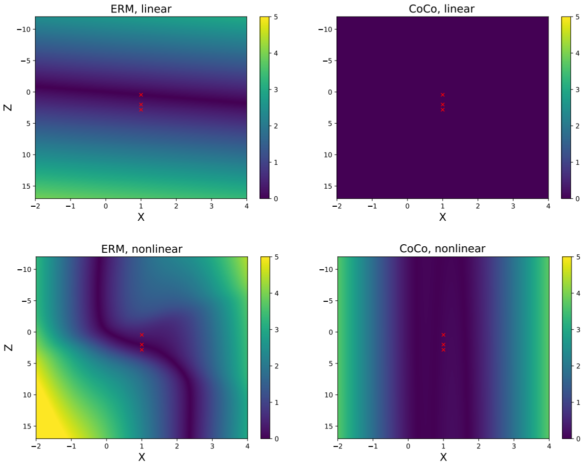


Figure 5. Prediction accuracy for CoCo and ERM, for linear (top row) and nonlinear (bottom row) predictors. The heatmap is the square error $(\hat{y} - \mathbb{E}[y|x])^2$, the x-axis, y-axis are the values of input x and z respectively. The red points are $(\mathbb{E}[x^e], \mathbb{E}[z^e])$ for three environments. CoCo exhibits better out of sample generalization with wider low error region.

G.2. Gaussian mixture example

In this section, we study a multi-class classification problem when the inputs contain non-causal covariates. We modify GMM to simulate the data set. The observed covariates are $(\boldsymbol{x}^e, \boldsymbol{z}^e)$ and the outcome is y^e, where e is the environment index. For each environment e, the data are generated with SEM

$$\begin{aligned} \boldsymbol{x}^e &\leftarrow \textstyle\sum_{k=1}^K \frac{1}{K}\mathcal{N}(\boldsymbol{x}^e; \boldsymbol{\mu}_k, \mathbf{I}) \\ y^e &\leftarrow \text{Categorical}(p_1, \cdots, p_K) \\ \boldsymbol{z}^e &\leftarrow (1-p^e)\delta_{\boldsymbol{u}^e_{y^e}} + p^e \delta_{\boldsymbol{u}^e_{k_1}}, \end{aligned} \tag{38}$$

where $p_k = \mathcal{N}(\boldsymbol{x}^e; \boldsymbol{\mu}_k, \mathbf{I}) / \sum_{k'=1}^K \mathcal{N}(\boldsymbol{x}^e; \boldsymbol{\mu}_{k'}, \mathbf{I})$, $k_1 \sim$ Multinomial$(1/K, \cdots, 1/K)$.

Among the covariates, the mapping from $\boldsymbol{x}^e$ to the label y^e is invariant across all e while $\boldsymbol{z}^e$ is predictive to y^e due to spurious associations. We aim to learn a model that makes predictions based on the causal covariates $\boldsymbol{x}^e$.

In Eq. (38), $\boldsymbol{x}^e$ are generated from GMM with the component centers $\boldsymbol{\mu}_k = \sqrt{1.5K}\mathbf{e}_k \in \mathbb{R}^K$. To generate the non-causal covariates $\boldsymbol{z}^e$, we first generate K random vectors $\{\boldsymbol{u}^e_k\}_{k=1}^K$ with $\boldsymbol{u}^e_k \sim \prod_{i=1}^{[k/2]} U(0,1)$ for environment e. Then for a data point in the component y^e, $\boldsymbol{z}^e$ equals $\boldsymbol{u}^e_{y^e}$ with probability $1 - p^e$ and equals a random vector from $\{\boldsymbol{u}^e_k\}_{k=1}^K$ otherwise. By doing so, $\boldsymbol{z}^e$ is associated with y^e but the association varies across environments when $\boldsymbol{u}^e_{1:K}$ change with e.

The DGPs that generate the environments is characterized by the values of $\boldsymbol{u}^e_{1:K}$ and p^e. We set the training environments with $K = 5$ and $p^e \in \{0.01, 0.02, \cdots, 0.05\}$ by Eq. (38). For a validation/test environment f we generate a new set of $\{\boldsymbol{u}^f_k\}_{k=1}^K$ and set $p^f = 0$. We evaluate the test performance by averaging the accuracy over 10 testing environments. If the predictor learns to predict based on the causes $\boldsymbol{x}^e$ instead of $\boldsymbol{z}^e$, it can accurately predict y^e in both training and testing environments.

The predictor is a fully connected neural network with two hidden layers. For CoCo, we use objective Eq. (14) to optimize the predictor. The penalty weight λ_r is chosen on the validation environment and is reduced to 0 when the parameters are sufficiently away from $\mathbf{0}$. For IRM, we choose λ in Eq. (19) and step size on the validation environment.

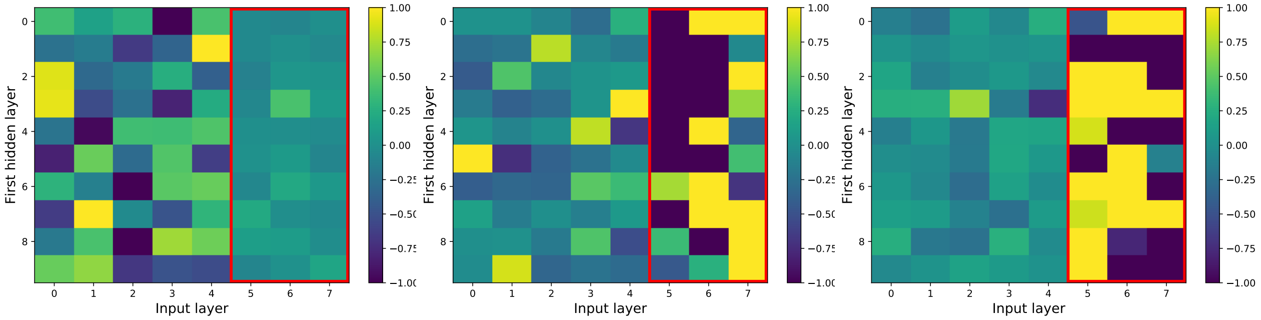
The results are shown in Figs. 6 to 8, and Table 2. Fig. 8 (a) is the trace plot for the predictive accuracy in the testing environments. The testing accuracy increases for all methods in the early stage of training but drops in the later stage for ERM and IRM. We hypothesize that ERM and IRM at first improves the prediction by utilizing all covariates including the causal ones. But in the later stage of training, it relies more heavily on the spurious associations to boost the predictive accuracy, which reduces the accuracy at the test time.

Table 2. Predictive accuracy in training and testing environments for GMM. The Oracle results are obtained by predicting with covariates $\boldsymbol{x}^e$ instead of $(\boldsymbol{x}^e, \boldsymbol{z}^e)$.

	GMM	
	Training	Testing
ERM	99.4	51.0
IRM	95.9	75.9
CoCo	91.9	**91.6**
Random guess	20	20
Oracle	92.3	91.8

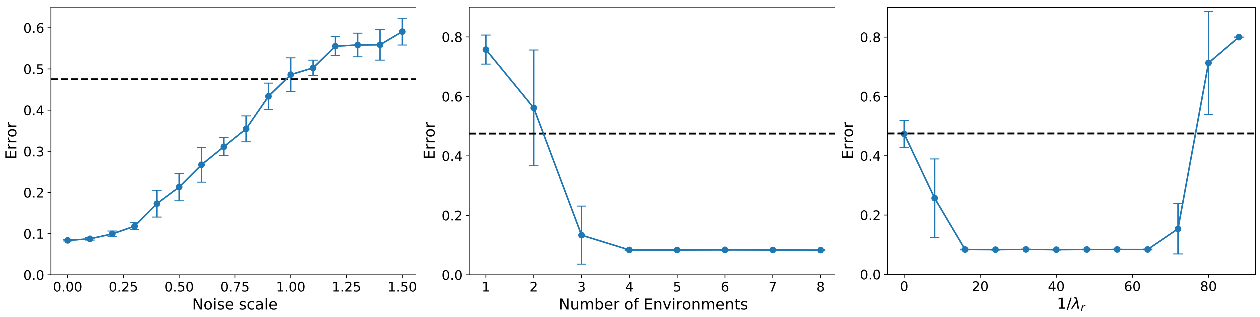
We provide evidence to this hypothesis in Fig. 6, by plotting the weight matrix that connects the input and the first hidden layer. The model trained by CoCo manages to set the weights associated with the non-causes $\boldsymbol{z}$ (the right block) close to zero, aligned with the analysis in Proposition 1. In comparison, these weights obtained by IRM and ERM are mostly non-zero, passing information from non-causal $\boldsymbol{z}$ to the subsequent hidden layers and outputs.

In Fig. 7, we study how CoCo performs if the invariance is violated and how sensitive it is with different hyperparameters. In panel (a), we construct the training environments by changing the cluster centers $\{\boldsymbol{\mu}_k\}_{k=1}^K$ in Eq. (38) to $\{\boldsymbol{\mu}_k + \boldsymbol{\epsilon}^e_k\}_{k=1}^K$, $\boldsymbol{\epsilon}^e_k \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_K)$. The noise $\boldsymbol{\epsilon}^e_k$ changes the mapping from the covariates $\boldsymbol{x}^e$ to the label y^e across the environments, and the noise scale σ^2 reflects the magnitude of change. The panel (a) shows the testing predictive accuracy increases as the invariance tends to hold. In panel (b), we compute the test accuracy when using different number of environments M in training. we construct training environment e by Eq. (38) with vectors $\boldsymbol{u}^e_k \sim \prod_{i=1}^{[k/2]} U(0,1)$ for all k and $p^e \in \{0.01, 0.02, \cdots, 0.01M\}$. We find a growing number of environments reduces the testing error monotonically which might due to the increased heterogeneity in data. In panel (c), we study how the testing error changes with the penalty weights λ_r in CoCo objective Eq. (14). When λ_r is large, the objective is close to the empirical risk and the test error is high; when λ_r is small, the parameters often collapse to $\mathbf{0}$. Between the two extremes, CoCo with a wide range of λ_r can learn a model that makes robust prediction in new environments.



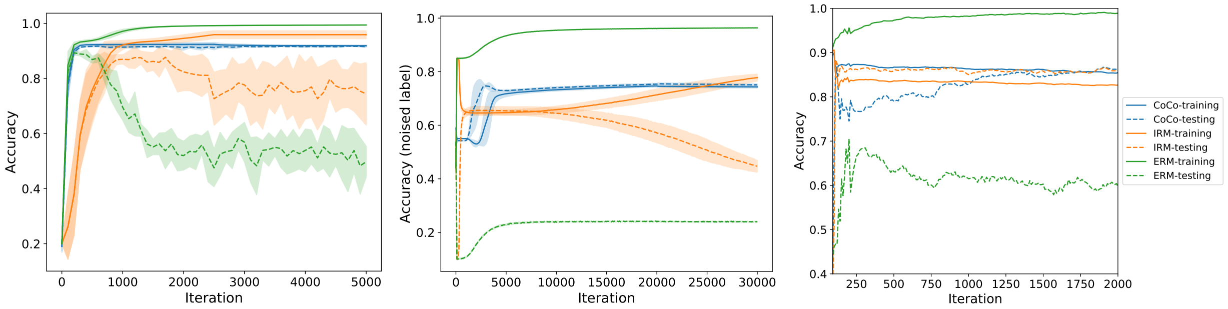
(a) CoCo (b) IRM (c) ERM

Figure 6. The heatmap for the first layer weight matrix of the neural networks trained by CoCo, IRM and ERM. The matrix dimension is 10×8 where the input dimension is 8 and the first hidden layer dimension is 10. In the input, the first five elements are $\boldsymbol{x}$ and the last three elements are $\boldsymbol{z}$. CoCo sets the weights related to non-causal $\boldsymbol{z}$ (the right block) close to 0.



(a) Noise scale (b) # of environments (c) Regularization strength

Figure 7. The change of testing prediction error with different levels of invariance, number of environments and the hyperparameter of CoCo. The dashed line is the ERM error rate. The error bar is the standard deviation over 5 independent trials.



(a) GMM (b) Color-MNIST (noised label) (c) Wildlife

Figure 8. Trace plot of training and testing accuracy for CoCo, IRM and ERM on GMM, Color-MNIST and Wildlife data. In panel (b), the accuracy is measured on predicting the *noised label* y. CoCo trades-off training accuracy slightly for significantly higher testing accuracy.