S1. Proofs

S1.1. Proof of Theorem 1

Restatement of Theorem 1 *If $P_{\mathrm{s}} \in \mathcal{P}$, then $\mathcal{H}_{\mathrm{s}} \subset \mathcal{H}_*$.*

Proof. It suffices to prove that for any $h_{\mathrm{s}} \in \mathcal{H}_{\mathrm{s}}$, we have

$$h_{\mathrm{s}}(\cdot) \in \arg\min_{h} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h(X), Y)]. \tag{1}$$

To prove (1), we only need to show that for any $h(\cdot)$ and $P \in \mathcal{P}$, there exists $Q \in \mathcal{P}$ such that

$$\mathbb{E}_Q[\mathcal{L}(h(X), Y)] \geq \mathbb{E}_P[h_{\mathrm{s}}(X), Y)], \tag{2}$$

and hence

$$\sup_{Q \in \mathcal{P}} \mathbb{E}_Q[\mathcal{L}(h(X), Y)] \geq \sup_{P \in \mathcal{P}} \mathbb{E}_P[\mathcal{L}(h_{\mathrm{s}}(X), Y)].$$

Recall that

$$\mathcal{H}_{\mathrm{s}} = \left\{ (\phi \circ g)(\cdot) \;\middle|\; \phi(w) \in \arg\min_{z} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w], \quad a.s. \right\}.$$

Since $h_{\mathrm{s}}(\cdot) \in \mathcal{H}_{\mathrm{s}}$, there is some $\phi_{\mathrm{s}}(\cdot)$ satisfying $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot)$ and $\phi_{\mathrm{s}}(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w]$ for almost every w. Suppose $(X, \eta) \sim P_X \times F$ and $(m(g(X), \eta), X) \sim Q$ where P_X is the marginal distributions of X under P. Let $\mathcal{U}$ be the support of noise η. Then

$$\begin{aligned}
\mathbb{E}_Q[\mathcal{L}(h(X), Y) \mid X = x] &= \int_{\mathcal{U}} \mathcal{L}(h(x), m(g(x), u)) P_\eta(du) \\
&\geq \int_{\mathcal{U}} \mathcal{L}(\phi_{\mathrm{s}}(g(x)), m(g(x), u)) P_\eta(du) \\
&= \mathbb{E}_P[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = g(x)] \quad a.s.
\end{aligned}$$

Here the first equation follows from the fact that X and η are independent under Q. The inequality is from the fact that

$$\int_{\mathcal{U}} \mathcal{L}(h(x), m(g(x), u)) P_\eta(du) = \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(x), Y) \mid g(X) = g(x)]$$

and

$$\phi_{\mathrm{s}}(w) \in \arg\min_{z} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w] = \arg\min_{z} \int_{\mathcal{U}} \mathcal{L}(z, m(w, u)) P_\eta(du)$$

for almost every w. The last equation is due to $P \in \mathcal{P}$. Then equation (2) follows by taking expectation and the law of iterated expectation. □

S1.2. Proof of Theorem 2

To begin with, we establish two useful lemmas regarding CITs. The first lemma states that $g(\cdot)$ is determined up to an invertible transformation by the transformation that it is invariant to.

For a given function $h(\cdot)$, let $\mathcal{T}_h = \{T(\cdot) : (h \circ T)(\cdot) = h(\cdot)\}$. Then we have the following lemma.

Lemma 1. *For any $h_1(\cdot)$ and $h_2(\cdot)$, $\mathcal{T}_{h_1} \subset \mathcal{T}_{h_2}$ if and only if there exists a function $v(\cdot)$ such that $h_2(\cdot) = (v \circ h_1)(\cdot)$, and $\mathcal{T}_{h_1} = \mathcal{T}_{h_2}$ if and only if there is an invertible function $v(\cdot)$ such that $h_2(\cdot) = (v \circ h_1)(\cdot)$.*

Proof. We only prove the former statement as the latter can be obtained as a corollary of the former. The "if" direction is obvious.

Here we prove the "only if" direction. Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be the range of $h_1(\cdot)$ and $h_2(\cdot)$, respectively. For any $w_1 \in \mathcal{R}_1$ and $w_2 \in \mathcal{R}_2$, define $\mathcal{D}_{h_1, w_1} = \{x : h_1(x) = w_1\}$ and $\mathcal{D}_{h_2, w_2} = \{x : h_2(x) = w_2\}$. Then $h_2(\cdot) = (v \circ h_1)(\cdot)$ if and only if for any $w_2 \in \mathcal{R}_2$, there is some $w_1 \in \mathcal{R}_1$ such that $\mathcal{D}_{h_1, w_1} \subset \mathcal{D}_{h_2, w_2}$. Thus, the former claim holds if we can show the following: $\mathcal{T}_{h_1} \subset \mathcal{T}_{h_2}$ implies that there is some $w_2 \in \mathcal{R}_2$ such that $\mathcal{D}_{h_1, w_1} \subset \mathcal{D}_{h_2, w_2}$ for any $w_1 \in \mathcal{R}_1$. We will prove this by contraction.

Suppose there exists w_1 such that $\mathcal{D}_{h_1,w_1} \not\subset \mathcal{D}_{h_2,w_2}$ for any $w_2 \in \mathcal{R}_2$. Because $\bigcup_{w_2\in\mathcal{R}_2} \mathcal{D}_{h_2,w_2}$ constitutes the whole space, there is some w_2 such that $\mathcal{D}_{h_1,w_1} \bigcap \mathcal{D}_{h_2,w_2} \neq \varnothing$ and $\mathcal{D}_{h_1,w_1} \not\subset \mathcal{D}_{h_2,w_2}$. Thus, $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2} \neq \varnothing$. Let $x^\dagger$ denote a point in $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2}$ and let x' a point in $\mathcal{D}_{h_2,w_2} \bigcap \mathcal{D}_{h_1,w_1}$. Define T_* as the transformation such that $T_*(x') = x^\dagger$, $T_*(x^\dagger) = x'$ and $T_*(x) = x$ for $x \neq \{x', x^\dagger\}$. Then it is straightforward to verify that $T_* \in \mathcal{T}_{h_1}$ but $T_* \notin \mathcal{T}_{h_2}$, which is a contradiction. □

Thus $g(\cdot)$ can be characterized by $\mathcal{T}_g$ up to an invertible transformation. Define $\mathcal{C}_g = \{g'(\cdot) : g'(\cdot) = (v \circ g)(\cdot)$ for some invertible transformation $v(\cdot)\}$. For any $g' \in \mathcal{C}_g$, by defining

$$\mathcal{H}'_{\mathrm{s}} = \left\{ (\phi \circ g)(\cdot) \;\middle|\; \phi(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g'(X) = w], \quad a.s. \right\},$$

similar arguments as in the proof of Theorem 1 can show $\mathcal{H}'_{\mathrm{s}} \subset \mathcal{H}_*$. To train a model that generalizes well on all the data distributions following the same causal mechanism, any $g'(\cdot) \in \mathcal{C}_g$ is sufficient. Thus, if $\mathcal{T}_g$ is known, to find a model belongs to $\mathcal{H}_*$, one may firstly find an invariant feature map $g'(\cdot)$ such that $\mathcal{T}_{g'} = \mathcal{T}_g$ and then obtain the model according to Theorem 1. However, finding a $g'(\cdot)$ such that $\mathcal{T}_{g'} = \mathcal{T}_g$ is sometimes still a hard task.

For any function $h(\cdot)$, define $\mathcal{I}_h$ in the same way as $\mathcal{I}_g$ with $g(\cdot)$ replaced by $h(\cdot)$ in the definition. We then have the following lemma.

Lemma 2. *For any $h_1(\cdot)$ and $h_2(\cdot)$, if $\mathcal{I}_{h_1} \subset \mathcal{T}_{h_2}$, then there exists a function $v(\cdot)$ such that $h_2(\cdot) = (v \circ h_1)(\cdot)$.*

Proof. Like in the proof of Lemma (1), it suffices to show that $\mathcal{I}_{h_1} \subset \mathcal{T}_{h_2}$ implies for any $w_1 \in \mathcal{R}_1$, there is some $w_2 \in \mathcal{R}_2$ such that $\mathcal{D}_{h_1,w_1} \subset \mathcal{D}_{h_2,w_2}$. We prove this by contraction.

Suppose there is some w_1 such that $\mathcal{D}_{h_1,w_1} \not\subset \mathcal{D}_{h_2,w_2}$ for any $w_2 \in \mathcal{R}_2$. Because $\bigcup_{w_2\in\mathcal{R}_2} \mathcal{D}_{h_2,w_2}$ is the whole space, there is some w_2 such that $\mathcal{D}_{h_1,w_1} \bigcap \mathcal{D}_{h_2,w_2} \neq \emptyset$ and $\mathcal{D}_{h_1,w_1} \not\subset \mathcal{D}_{h_2,w_2}$. Thus, $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2} \neq \emptyset$. Let $x^\dagger$ be a point in $\mathcal{D}_{h_1,w_1} \setminus \mathcal{D}_{h_2,w_2}$ and let x' be a point in $\mathcal{D}_{h_2,w_2} \bigcap \mathcal{D}_{h_1,w_1}$. According to the definition of essential invariant subset, because $h_1(x_1) = h_2(x_2)$, there are finite transformations $T_1(\cdot), \ldots, T_K(\cdot) \in \mathcal{I}_g$ such that $\bar{T}(x') = x^\dagger$ where $\bar{T}(\cdot) = (T_1 \circ \cdots \circ T_K)(\cdot)$. It can be verified that $\mathcal{T}_{h_2}$ is closed with respect to function composition. Hence, $\bar{T}(\cdot) \in \mathcal{T}_{h_2}$. However, $h_2(\bar{T}(x')) = h_2(x^\dagger) \neq w_2 = h_2(x')$, which is a contradiction. □

Restatement of Theorem 2 *If $P_{\mathrm{s}} \in \mathcal{P}$, then*

$$\mathcal{H}_{\mathrm{s}} \subset \arg\min_h \sup_{T\in\mathcal{T}_g} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(T(X)), Y)],$$

where $\mathcal{H}_{\mathrm{s}}$ is defined in (3).

Proof. It suffices to show that for all $h_{\mathrm{s}}(\cdot) \in \mathcal{H}_{\mathrm{s}}$, we have

$$h_{\mathrm{s}}(\cdot) \in \arg\min_h \sup_{T\in\mathcal{T}_g} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(T(X)), Y)]. \tag{3}$$

Note that $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot)$ for some $\phi_{\mathrm{s}}(\cdot)$ and hence is invariant to any transformation $T(\cdot) \in \mathcal{T}_g$. We then have $\sup_{T\in\mathcal{T}_g} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)] = \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(T(X)), Y)]$. Thus, it suffices to prove that for all $h(\cdot)$, there exists $T(\cdot) \in \mathcal{T}_g$ such that

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(T(X)), Y)] \geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)]. \tag{4}$$

According to axiom of choice, there is a choice function a such that $a(w) \in \mathcal{D}_{g,w}$ for almost every w. Define $\tilde{T}$ to be a transformation such that $\tilde{T}(x) = a(w)$ for $x \in \mathcal{D}_{g,w}$. Then $\tilde{T}(\cdot) \in \mathcal{T}_g$ and we have

$$\begin{aligned} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(\tilde{T}(X)), Y) \mid g(X) = w] &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(a(w)), Y) \mid g(X) = w] \\ &\geq \mathbb{E}_{P_{\mathrm{s}}}[\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\ &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s. \end{aligned} \tag{5}$$

By taking expectation on both sides, we can obtain equation (4). □

S2. Proof of Theorem 3

Restatement of Theorem 3 *If $P_{\mathrm{s}} \in \mathcal{P}$, then*

$$\mathcal{H}_{\mathrm{s}} = \arg\min_{h} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] \quad \text{subject to} \;\; h(\cdot) = (h \circ T)(\cdot), \; \forall\, T(\cdot) \in \mathcal{I}_g. \tag{6}$$

where $\mathcal{I}_g$ is any causal essential set of $g(\cdot)$ and $\mathcal{H}_{\mathrm{s}}$ is defined in (3).

Proof. We first show

$$\mathcal{H}_{\mathrm{s}} \subset \arg\min_{h} \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] \quad \text{subject to} \;\; h(\cdot) = (h \circ T)(\cdot), \; \forall\, T(\cdot) \in \mathcal{I}_g.$$

Note that the restriction in (6) is equivalent to $\mathcal{I}_g \subset \mathcal{T}_h$. It suffices to show that

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y)] \geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)] \tag{7}$$

for any $h(\cdot)$ with $\mathcal{I}_g \subset \mathcal{T}_h$ and for any $h_{\mathrm{s}}(\cdot) \in \mathcal{H}_{\mathrm{s}}$. If $\mathcal{I}_g \subset \mathcal{T}_h$, according to Lemma 2, there exists $v(\cdot)$ such that $h(\cdot) = (v \circ g)(\cdot)$. By the definition of $h_{\mathrm{s}}(\cdot)$, there also exists $\phi_{\mathrm{s}}(\cdot)$ satisfying $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot)$ and $\phi_{\mathrm{s}}(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w]$ for almost every w. Thus, we have

$$\begin{aligned}
\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h(X), Y) \mid g(X) = w] &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(v(w), Y) \mid g(X) = w] \\
&\geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\
&\geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s.
\end{aligned}$$

Then (7) follows by taking expectation.

Next we show the opposite inclusion to prove (6). Suppose $h_*(\cdot)$ is a solution to the optimization problem in (6). Then according Lemma 2, there is some $v_*(\cdot)$ such that $h_*(\cdot) = (v_* \circ g)(\cdot)$. Let $h_{\mathrm{s}}(\cdot) = (\phi_{\mathrm{s}} \circ g)(\cdot) \in \mathcal{H}_{\mathrm{s}}$. Then

$$\begin{aligned}
\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_*(X), Y) \mid g(X) = w] &= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(v_*(w), Y) \mid g(X) = w] \\
&\geq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\
&= \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s.,
\end{aligned} \tag{8}$$

by definition. Because $h_*(\cdot)$ is a solution to the minimization problem, we have

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_*(X), Y)] = \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y)].$$

Combining this with (8), we have

$$\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_*(X), Y) \mid g(X) = w] \leq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(h_{\mathrm{s}}(X), Y) \mid g(X) = w] \quad a.s. \tag{9}$$

This implies

$$\begin{aligned}
\mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(v_*(w), Y) \mid g(X) = w] &\leq \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(\phi_{\mathrm{s}}(w), Y) \mid g(X) = w] \\
&= \min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w] \quad a.s.
\end{aligned}$$

Thus, we conclude that $v_*(w) \in \arg\min_z \mathbb{E}_{P_{\mathrm{s}}}[\mathcal{L}(z, Y) \mid g(X) = w]$. □

S3. More Experimental Results

S3.1. Toy Example and Simulation

In the following toy example, we are able to construct an explicit formulation of the causal essential invariant set.

Example 1. Let X be a non-singular 2×2 matrix and $X^{(j)}$ be the j-th column of X for $j = 1, 2$. Suppose that $g(X)$ is the area of the triangle formed by the two points $X^{(1)}$, $X^{(2)}$ and the origin. Then it is not hard to show that

$\{T_{R,\theta}(\cdot), T_{S,a}(\cdot), T_M(\cdot), T_P(\cdot), T_I(\cdot) \mid \theta \in [0, \pi/4],\ a \in [2/3, 3/2]\}$ is an essential invariant set of $g(\cdot)$, where

$$T_{R,\theta}(X) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} X, \quad T_{S,a}(X) = X \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix},$$
$$T_M(X) = X \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad T_P(X) = X \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$
$$T_I(X) = X \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Here $T_{R,\theta}(\cdot)$ rotates the triangle with θ degree clockwise, and $T_{S,a}(\cdot)$ scales the two edges (one connects $X^{(1)}$ to the origin and the other connects $X^{(2)}$ to the origin) of the triangle with a and a^{-1} times, respectively. $T_M(\cdot)$ mirrors the triangle with respect to the x-axis. $T_P(\cdot)$ transforms the triangle to another triangle with same base and height, and $T_I(\cdot)$ transforms the the triangle to another one that is symmetric with respect to the origin. All these transformations are known to keep the triangle area unchanged based on elementary geometry.

Now we verify the effectiveness of the proposed method in the main body using this example.

Data. We consider the following data generation process:

$$\begin{aligned} &X^{(1)} \sim N(0, I_2),\ X^{(2)} \sim N(0, 2I_2),\ X = (X^{(1)}, X^{(2)}), \\ &\epsilon \sim N(0,1),\ \eta = \frac{a\Phi^{-1}(\pi^{-1}\alpha) + \epsilon}{\sqrt{a^2+1}}, \\ &Y = |\det(X)| + \eta, \end{aligned} \tag{10}$$

where I_2 is the identity matrix of order 2, $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. In this data generation process, $|\det(X)|$ is the area of the triangle formed by $X^{(1)}$, $X^{(2)}$ and the origin, and is the causal feature in this example. Here α is the angle between $(X^{(1)} + X^{(2)})/2$ and x-axis, and is correlated with Y in certain domains, with a a parameter that reflects this correlation. However, this correlation is a spurious correlation that changes across domains, i.e., a is set to be different in different domains. In the training population, we pick $a = -3$. We then generate i.i.d. samples of size $1,000$, denoted by $\{(Y_i, X_i)\}_{i=1}^{1000}$, and train a model $h(X, \beta)$ with parameter β to predict Y based on these generated samples.

Model. For any 2×2 matrix

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix},$$

let

$$\begin{aligned} v(X) = (1, &X_{11}, X_{21}, X_{12}, X_{22}, X_{11}^2, X_{21}^2, X_{12}^2, X_{22}^2, \\ &X_{11}X_{21}, X_{11}X_{12}, X_{11}X_{22}, X_{21}X_{12}, X_{21}X_{22}, X_{12}X_{22})^{\mathrm{T}}. \end{aligned}$$

The model is

$$h_\beta(X) = \mathrm{ReLU}(\beta_{[1]}^{\mathrm{T}} v(X)) + \beta_{[2]}^{\mathrm{T}} v(X),$$

where $\beta = (\beta_{[1]}^{\mathrm{T}}, \beta_{[2]}^{\mathrm{T}})^{\mathrm{T}}$ is the model parameter. We pick this model because we have known that $|\det(X)|$ is a function of $v(X)$, and there is some β^* such that $h_{\beta^*}(X) = |\det(X)|$.

Method. Based on the essential invariant set given in Example 1, we define five invariant transformations

$$T_1(X) = \begin{pmatrix} \cos\frac{\pi}{12} & -\sin\frac{\pi}{12} \\ \sin\frac{\pi}{12} & \cos\frac{\pi}{12} \end{pmatrix} X, \quad T_2(X) = X \begin{pmatrix} 1.1 & 0 \\ 0 & 1.1^{-1} \end{pmatrix},$$
$$T_3(X) = X \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad T_4(X) = X \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$
$$T_5(X) = X \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

For ease of notation, we let $T_0(X) = X$ be the identity transformation. We learn the model parameter by minimizing four different loss functions, namely, the empirical risk

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\beta(X_i))^2,$$

the average risk over different transformations

$$\frac{1}{n}\sum_{k=0}^{5}\sum_{i=1}^{n}(Y_i - h_\beta(T_k(X_i)))^2,$$

the maximal risk over different transformations

$$\max_{k=0,\ldots,5}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\beta(T_k(X_i)))^2\right\},$$

and the RICE loss function

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\beta(X_i))^2 + \lambda \max_{k=0,\ldots,5}\left\{\frac{1}{n}\sum_{i=1}^{n}(h_\beta(X_i) - h_\beta(T_k(X_i))^2\right\},$$

where $n = 1000$. In the implementation of RICE, for the given quantities $l_0, \cdots, l_5$, we replace the maximum $\max_{k=0,\ldots,5}\{l_k\}$ in the above losses with the softmax weighting quantity $\sum_{k=0}^{5}\exp(0.2l_k)l_k/\sum_{k=0}^{5}\exp(0.2l_k)$, for ease of computation.

Results. The resulting model is evaluated on i.i.d. sample generated following the data generation process (10) with different a. The following figure plots the squared prediction error of the four methods on test data with different values of a. Each reported value is the average over 200 simulations.

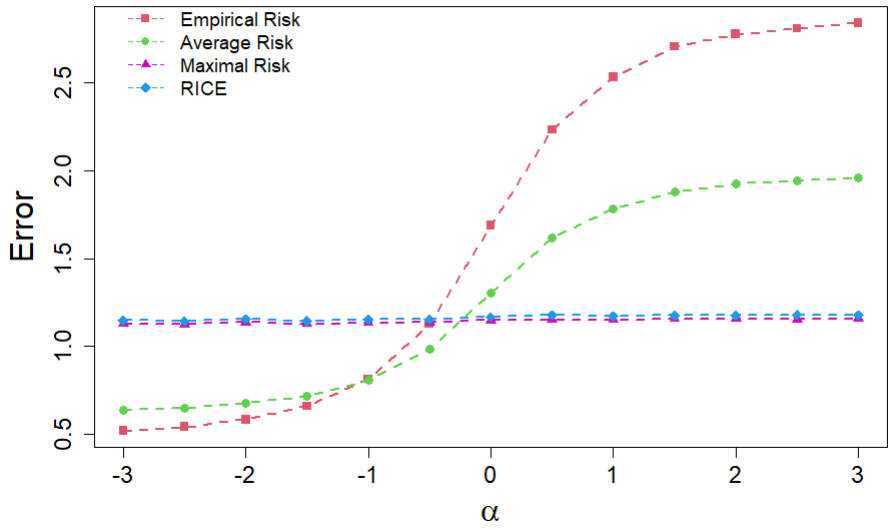


Figure S1. Squared prediction error on test data from distributions with different values of a.

It can be seen that when the test distribution has similar spurious correlations as the training population, minimizing the empirical risk performs the best among the four methods. However, it performs the worst if an opposite spurious correlation appears in the test population. The RICE algorithm has the best worst-case performance, which is consistent with our theoretical analysis. Moreover, the RICE algorithm seems successfully capture the invariant causal mechanism across different environments, as its prediction errors under different test distributions are stable and close to the variance of the intrinsic error η.

Table S1. Hyperparameters of the proposed RICE on C-MNIST, PACS, and VLCS.

Dataset	C-MNIST	PACS	VLCS
Learning Rate	0.1	5e-5	5e-5
Batch Size	128	32	32
Weight Decay	5e-4	0	0
Drop Out	0	0.1	0.1
Epoch	20	20	20
λ_0	0.25	0.5	0.5
β_1	0.9	0.9	0.9
β_2	0.999	0.999	0.999

S3.2. Hyperparameters

We summarize the hyperparameters of the proposed RICE for C-MNIST, PACS, and VLCS datasets in Table S1. The learning rate is decayed by 0.2 at epoch $6, 12$, and 20.

S3.3. Ablation Study

In Section 5, for the experiments on PACS and VLCS, we collect training data from several domains for the proposed RICE. However, our theory in Section 3.3 requires only a single domain. Thus, in this subsection, we study the performance of RICE with single domain training data.

Our experiments are conducted on both PACS and VLCS. All the hyperparameters are set to be same with those in Section 5, except the number of training domains—we only use single domain data and hence less training samples for each single experiment. For example, for PACS, if the test domain is sketch, then we run RICE on training data from one of the three other domains (photo, art and cartoon) and report the accuracy on the test domain. To run RICE, the data generated by CycleGAN are used as augmented data and in the regularization term. For a fair comparison, we do not use the CycleGAN that transfer from training domain to test domain and adopt similar experimental settings for ERM.

The results are summarized in Figure S2. We can see that RICE performs much better than the baseline method ERM, which verifies our theoretical conclusions in Theorem 3. Besides, the test accuracy on the target domain can be quite high even when the model is trained using data from a single domain. For example, on VLCS dataset, when test data is from SUN09 domain, the model trained on VOC2007 domain even exhibits a better OOD generalization than the model trained on data from three domains. This implies that, for OOD generalization problem, the number of domains may not be crucial to the performance as long as some representative CITs are available.

S3.4. Generated Data

Our experiments in the main body involve generating causally invariant images. In this subsection, we present visualizations of some generated images for a better understanding of the proposed algorithm.

C-MNIST Figure S3 shows some C-MNIST images. As seen from the training set, there exist spurious correlations between the colors of the foreground or background and the category. However, the correlation disappears in the test set, as the foreground and background colors are randomly assigned.

PACS We also present some transformed data from PACS dataset generated by CycleGAN. The CycleGAN is used to simulate CITs as we have clarified the main body of this paper. As the data in PACS come from 7 categories, for each category we pick 4 pictures respectively from domains {photo, art, cartoon, sketch}. The transformed images are shown in Figure S4, where the columns correspond to the styles of {photo, art, cartoon, sketch}, respectively.

Let us look at these generated data over different domains. For the generated images of the photo domain (the first column), the trained CycleGAN tends to alter its color of foreground and add a background, especially when the original images are from the cartoon and sketch domains. Similar trends exhibit in the generated data of the art domain (the second column). In contrast to the two aforementioned domains, the generated cartoon data in the third column remove the background (if exists) while keep or alter the color of the foreground. The generated sketch data (the fourth column) are more likely to be a grayscale view of the original images. However, for each generated image, the shape of its foreground (i.e., the casual feature to decide the category) does not change when we vary the domains.

The proposed algorithm RICE regularizes the model to encourage the model to be invariant under the CITs, i.e., invariant to the changes of spurious features. This enables the model to be robust to the misleading signal from spurious features and to make predictions via the casual feature. For example, for the dog images in the last row of Figure S4c, which are generated from the images of cartoon style (the third column), the generated dog image of photo style (the first column) has a grass background. However, RICE requires the model to exhibit similar outputs for the two images, hence breaking the spurious correlation between dog and grass.

VLCS Similar to PACS, we present some of the domain transformed data from VLCS dataset generated by CycleGAN. We pick 4 pictures respectively from domains {VOC2007, LabelMe, Caltech101, SUN09} for each of the 5 categories in VLCS. Then we vary the domains of these picked data using the trained CycleGAN models. The transformed data are visualized in Figure S5.

The generated VLCS images exhibit similar behaviors as PACS. Specifically, for a given image from a certain domain, the CycleGAN model tends to deterministically vary the color of the background according to the domains. Thus, the reasoning about the effectiveness of RICE on PACS also applies here.

S3.5. Benchmark algorithms

- Empirical Risk minimization (ERM) pools together the data from all the domains and then minimizes the empirical loss to train the model. Notice that here an ImageNET pre-trained model is used.
- Marginal Transfer Learning [2] use the mean embedding of the feature distribution in each domain as an input of the classifier.
- Group Distributionally Robust Optimization (GroupDRO) [4] minimizes the largest loss across different domains.
- Domain-Adversarial Neural Networks (DANN) [3] use adversarial networks to match the feature distribution in different domains.
- Invariant Risk Minimization (IRM) [1] learns a feature representation such that the optimal classifiers on top of the representation is the same across the domains.

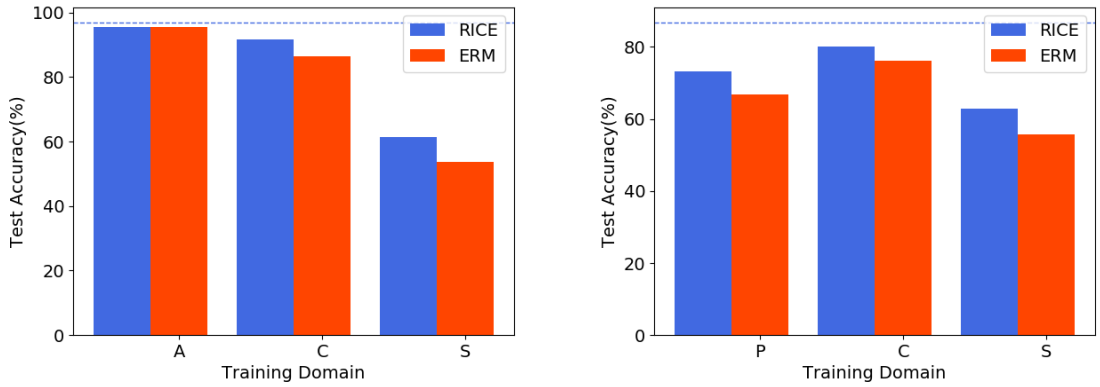
References

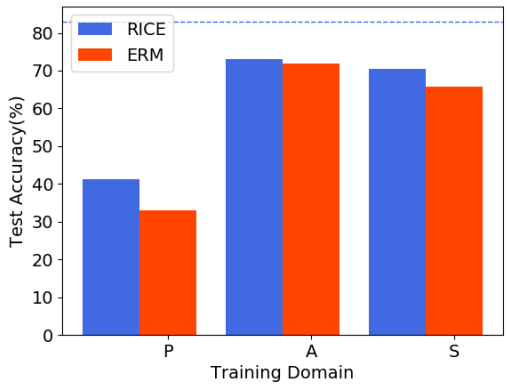
[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. Preprint arXiv:1907.02893, 2019. 7

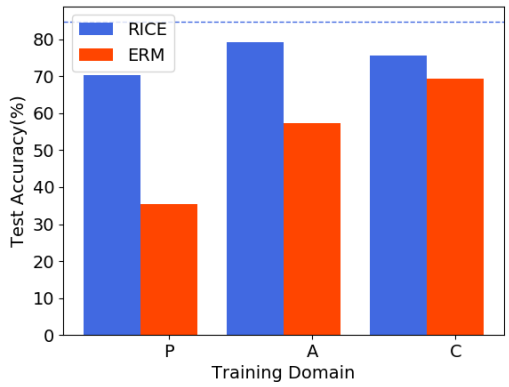
[2] Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22:1–55, 2021. 7

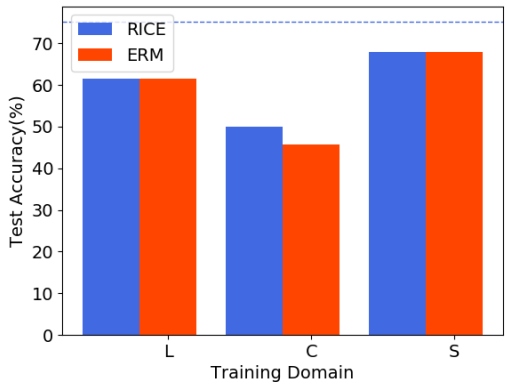
[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 7

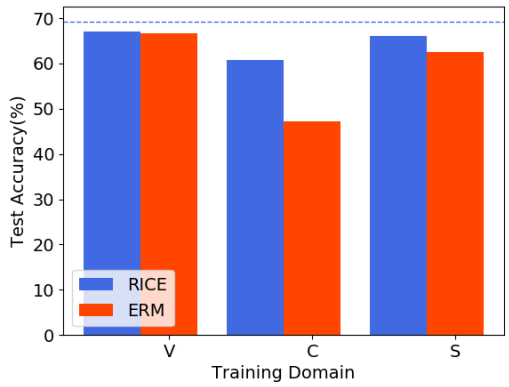
[4] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 7

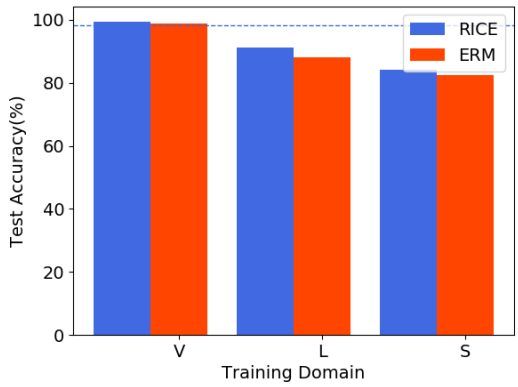




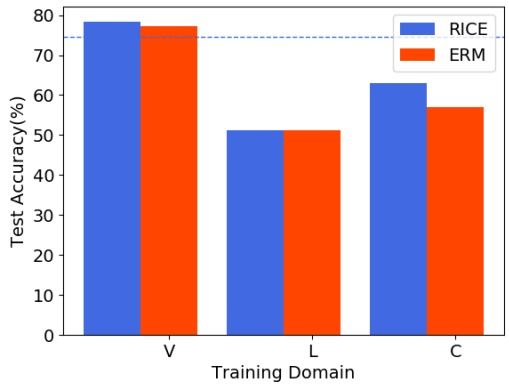






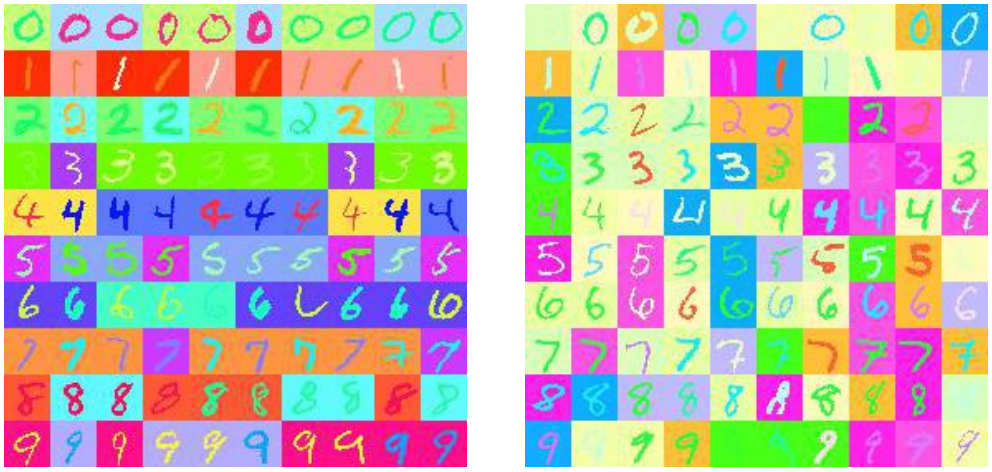


(g) Caltech101



(h) SUN09

Figure S2. Performance of RICE and ERM on the PACS (a-d) and VLCS (e-h) datasets with training data from single domains. Figure title indicates the test domain, and the blue dashed line represents the test accuracy when the training data are from three domains, as reported in Section 5.

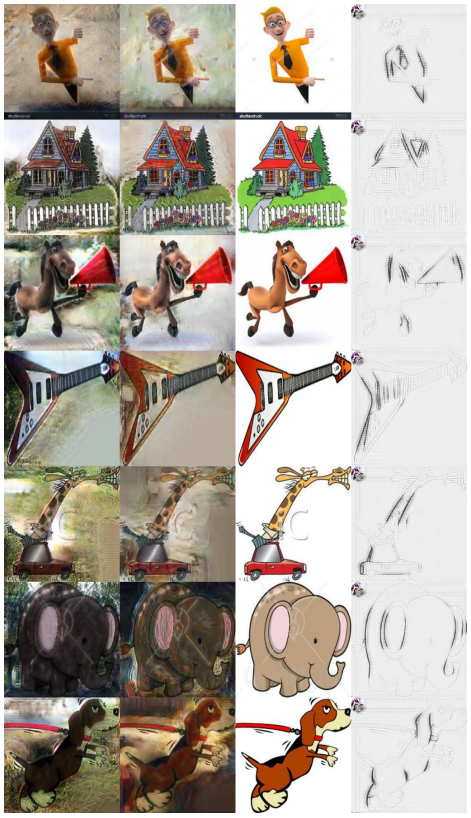


(a) Training data in C-MNIST (b) Test data in C-MNIST

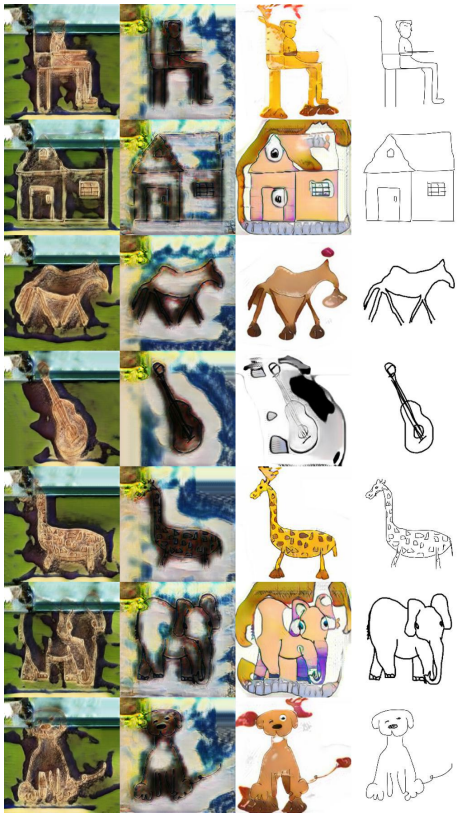
Figure S3. Images of the C-MNIST dataset.

(a) original domain: photo

(b) original domain: art

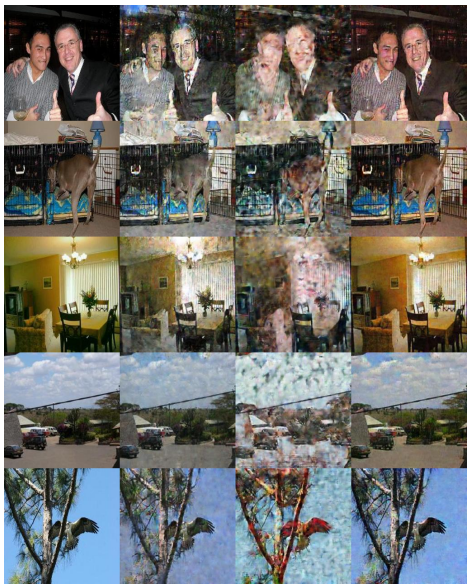


(c) original domain: cartoon



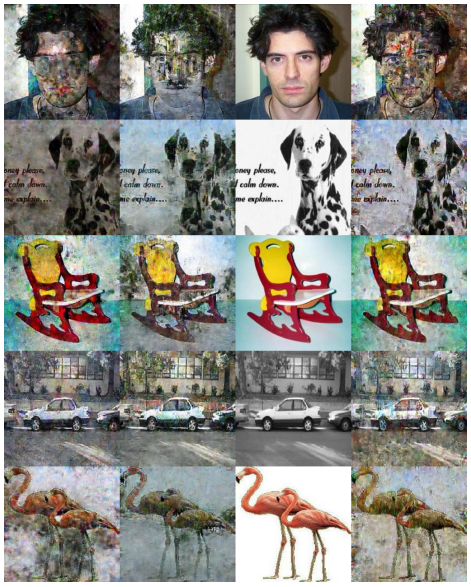
(d) original domain: sketch

Figure S4. Synthetic data of PACS generated by CycleGAN. Columns from left to right correspond to domains of {photo, art, cartoon, sketch}, respectively. Figure title indicates the domain of original data, based on which the data of the rest domains in the figure are generated by CycleGAN.



(a) original domain: VOC2007

(b) original domain: LabelMe



(c) original domain: Caltech101

(d) original domain: SUN09

Figure S5. Synthetic data of `VLCS` generated by CycleGAN. Columns from left to right correspond to domains of {VOC2007, LabelMe, Caltech101, SUN09}, respectively. Figure title indicates the domain of original data, based on which the data of the rest domains in the figure are generated by CycleGAN.