Invariant risk minimization

Arjovsky, Bottou, Gulrajani, Lopez-Paz '19

Geelon So
(agso@eng.ucsd.edu)

November 26, 2019

A learning paradigm to estimate invariant predictors

arXiv.org > stat > **arXiv:1907.02893**

Statistics > Machine Learning

Invariant Risk Minimization

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz

(Submitted on 5 Jul 2019 (v1), last revised 1 Sep 2019 (this version, v2))

We introduce Invariant Risk Minimization (IRM), a learning paradigm to estimate invariant correlations across multiple training distributions. To achieve this goal, IRM learns a data representation such that the optimal classifier, on top of that data representation, matches for all training distributions. Through theory and experiments, we show how the invariances learned by IRM relate to the causal structures governing the data and enable out-of-distribution generalization.

Subjects: **Machine Learning (stat.ML)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
Cite as: arXiv:1907.02893 [stat.ML]
(or arXiv:1907.02893v2 [stat.ML] for this version)

`https://arxiv.org/abs/1907.02893`.

Sheep or camel?

Imagine training a learner to distinguish between sheep and camels:






Sheep or camel?

What will the learner predict for this instance?

Sheep or camel?

Green meadow or hump?

Motivating problem

- **Question:** while we can train on data collected from a variety of environments, we want to generalize to *all* environments
  - how to distinguish causation from correlation?

Motivating problem

- **Question:** while we can train on data collected from a variety of environments, we want to generalize to *all* environments
  - how to distinguish causation from correlation?
  - how to avoid spurious correlations?

Motivating problem

- **Question:** while we can train on data collected from a variety of environments, we want to generalize to *all* environments
  - how to distinguish causation from correlation?
  - how to avoid spurious correlations?
  - how to obtain out-of-distribution generalization?

Learning from one environment

In the standard learning problem, given a function class $\mathcal{F}(\mathcal{X};\mathcal{Y})$ and risk functional $R : \mathcal{F}(\mathcal{X};\mathcal{Y}) \to \mathbb{R}$, find an optimizer:

$$\underset{f \in \mathcal{F}(\mathcal{X};\mathcal{Y})}{\arg\min}\ R(f).$$

Learning from one environment

In the standard learning problem, given a function class $\mathcal{F}(\mathcal{X};\mathcal{Y})$ and risk functional $R : \mathcal{F}(\mathcal{X};\mathcal{Y}) \to \mathbb{R}$, find an optimizer:

$$\arg\min_{f \in \mathcal{F}(\mathcal{X};\mathcal{Y})} R(f).$$

For short, define $\mathcal{L} = \big(\mathcal{F}(\mathcal{X};\mathcal{Y}), R\big)$ to be the learning problem.

Representation of data

Given a data representation $\phi : \mathcal{X} \to \tilde{\mathcal{X}}$,

- we obtain a new learning problem over $(\tilde{X}, Y)$

Representation of data

Given a data representation $\phi : \mathcal{X} \to \tilde{\mathcal{X}}$,

- we obtain a new learning problem over $(\tilde{X}, Y)$
- we obtain a new risk functional that evaluates $\tilde{f} : \tilde{\mathcal{X}} \to \mathcal{Y}$

Representation of data, formal

Given a data representation $\phi : \mathcal{X} \to \tilde{\mathcal{X}}$,

- define $\Phi : \mathcal{F}(\tilde{\mathcal{X}}; \mathcal{Y}) \to \mathcal{F}(\mathcal{X}; \mathcal{Y})$ by:

$$\Phi(\tilde{f}) := \tilde{f} \circ \phi$$

Representation of data, formal

Given a data representation $\phi : \mathcal{X} \to \tilde{\mathcal{X}}$,

- define $\Phi : \mathcal{F}(\tilde{\mathcal{X}}; \mathcal{Y}) \to \mathcal{F}(\mathcal{X}; \mathcal{Y})$ by:

$$\Phi(\tilde{f}) := \tilde{f} \circ \phi$$

- define $\Phi^* : \mathcal{F}(\mathcal{X}; \mathcal{Y})^* \to \mathcal{F}(\tilde{\mathcal{X}}; \mathcal{Y})^*$ by:

$$(\Phi^* R)(\tilde{f}) := R(\Phi(\tilde{f}))$$

Representation of data

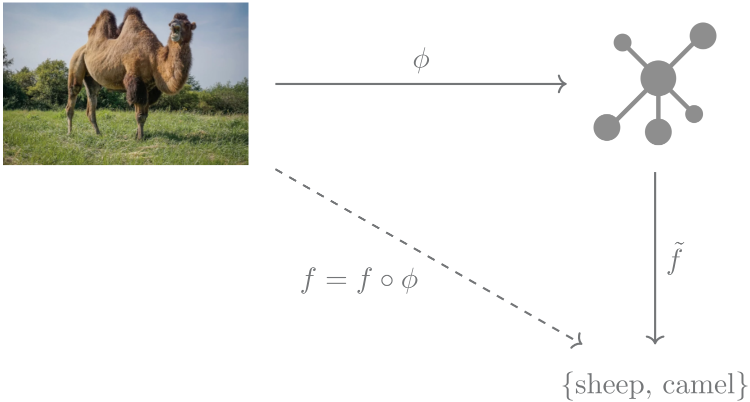


Figure 1: A representation $\phi : \mathcal{X} \to \tilde{\mathcal{X}}$ induces a new risk functional $\Phi^* R$ where $(\Phi^* R)(\tilde{f}) := R(f)$.

Representation of data

Thus, ϕ induces the learning problem $\tilde{\mathcal{L}} = \big(\Phi(\mathcal{F}(\mathcal{X};\mathcal{Y})), \Phi^* R\big)$:

$$\operatorname*{arg\,min}_{f \in \mathcal{F}(\mathcal{X};\mathcal{Y})} R(f) \rightsquigarrow \operatorname*{arg\,min}_{\tilde{f} \in \mathcal{F}(\tilde{\mathcal{X}};\mathcal{Y})} (\Phi^* R)(\tilde{f}).$$

Representation of data

Thus, ϕ induces the learning problem $\tilde{\mathcal{L}} = \big(\Phi(\mathcal{F}(\mathcal{X};\mathcal{Y})), \Phi^* R\big)$:

$$\operatorname*{arg\,min}_{f \in \mathcal{F}(\mathcal{X};\mathcal{Y})} R(f) \rightsquigarrow \operatorname*{arg\,min}_{\tilde{f} \in \mathcal{F}(\tilde{\mathcal{X}};\mathcal{Y})} (\Phi^* R)(\tilde{f}).$$

We hope that:

- we didn't lose too much information:

$$\tilde{f} \text{ solves } \tilde{\mathcal{L}} \quad \Longrightarrow \quad \Phi(\tilde{f}) \text{ approximately solves } \mathcal{L}$$

Representation of data

Thus, ϕ induces the learning problem $\tilde{\mathcal{L}} = \big(\Phi(\mathcal{F}(\mathcal{X};\mathcal{Y})), \Phi^* R\big)$:

$$\underset{f \in \mathcal{F}(\mathcal{X};\mathcal{Y})}{\arg\min}\ R(f) \rightsquigarrow \underset{\tilde{f} \in \mathcal{F}(\tilde{\mathcal{X}};\mathcal{Y})}{\arg\min} (\Phi^* R)(\tilde{f}).$$

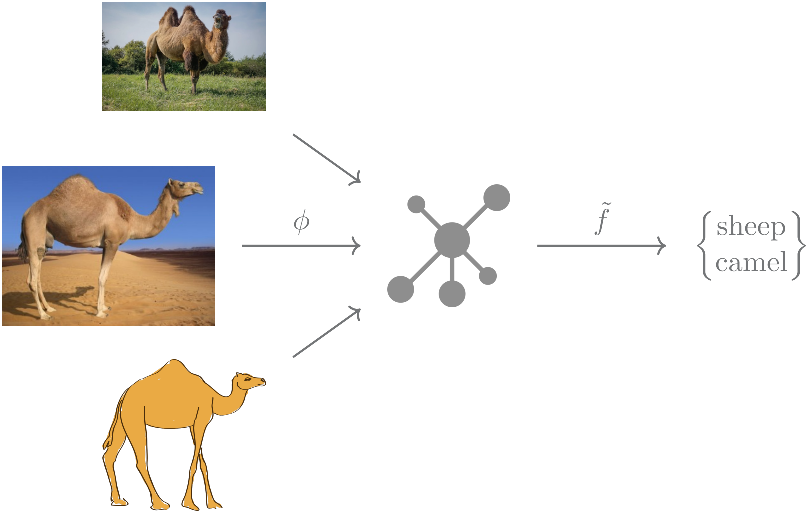
We hope that:

- we didn't lose too much information:

$$\tilde{f} \text{ solves } \tilde{\mathcal{L}} \quad \Longrightarrow \quad \Phi(\tilde{f}) \text{ approximately solves } \mathcal{L}$$

- the new problem is less complex: the function class $\Phi(\mathcal{F})$ is easier to optimize or generalize on

Learning from multiple environments



Learning setting

- **Task:** learn a function f from a function class $\mathcal{F}(\mathcal{X};\mathcal{Y})$

Learning setting

- **Task:** learn a function f from a function class $\mathcal{F}(\mathcal{X};\mathcal{Y})$
  - P is a data distribution over $\mathcal{X} \times \mathcal{Y}$

Learning setting

- **Task:** learn a function f from a function class $\mathcal{F}(\mathcal{X};\mathcal{Y})$
  - P is a data distribution over $\mathcal{X} \times \mathcal{Y}$
  - $R_P(f)$ is the risk associated with f

Learning setting

- **Task:** learn a function f from a function class $\mathcal{F}(\mathcal{X}; \mathcal{Y})$
  - P is a data distribution over $\mathcal{X} \times \mathcal{Y}$
  - $R_P(f)$ is the risk associated with f
- **Setting:** data can be collected from different environments $e \in \mathcal{E}$ with different data distribution P_e and risk $R_e \equiv R_{P_e}$

Learning setting

- **Task:** learn a function f from a function class $\mathcal{F}(\mathcal{X};\mathcal{Y})$
  - P is a data distribution over $\mathcal{X} \times \mathcal{Y}$
  - $R_P(f)$ is the risk associated with f
- **Setting:** data can be collected from different environments $e \in \mathcal{E}$ with different data distribution P_e and risk $R_e \equiv R_{P_e}$
  - **train** on datasets $D_e \sim P_e^{n_e}$ drawn from distributions from training environments $e \in \mathcal{E}_{\mathrm{tr}} \subset \mathcal{E}$

Learning setting

- **Task:** learn a function f from a function class $\mathcal{F}(\mathcal{X};\mathcal{Y})$
  - P is a data distribution over $\mathcal{X} \times \mathcal{Y}$
  - $R_P(f)$ is the risk associated with f
- **Setting:** data can be collected from different environments $e \in \mathcal{E}$ with different data distribution P_e and risk $R_e \equiv R_{P_e}$
  - **train** on datasets $D_e \sim P_e^{n_e}$ drawn from distributions from training environments $e \in \mathcal{E}_{\text{tr}} \subset \mathcal{E}$
  - **test** on all environments $e \in \mathcal{E}$ (i.e. want to generalize to all environments), minimizing the out-of-distribution risk R^{OOD}:

$$R^{\text{OOD}}(f) := \max_{e \in \mathcal{E}} R_e(f).$$

Invariants

Idea. We can recognize a camel anywhere because we can identify invariant features that do not depend on the environment.

Invariants

Idea. We can recognize a camel anywhere because we can identify invariant features that do not depend on the environment.

- A representation of data ϕ admits an invariant predictor $\tilde{f}$ if $\tilde{f}$ solves the new learning problem $\Phi^* R_e$ simultaneously for all environments $e \in \mathcal{E}$.

Invariants

Idea. We can recognize a camel anywhere because we can identify invariant features that do not depend on the environment.

- A representation of data ϕ admits an invariant predictor $\tilde{f}$ if $\tilde{f}$ solves the new learning problem $\Phi^* R_e$ simultaneously for all environments $e \in \mathcal{E}$.
  - **Warning:** $\tilde{f}$ optimizes $\Phi^* R_e$ does not imply that $\Phi(\tilde{f})$ optimizes R_e

Invariant risk minimization

Intuition. In invariant risk minimization (IRM), we'll aim to minimize the risk subject to the constraint that the minimizer is an invariant predictor on all the training environments.

Invariant risk minimization

Intuition. In invariant risk minimization (IRM), we'll aim to minimize the risk subject to the constraint that the minimizer is an invariant predictor on all the training environments.

- If we train on sufficiently distinct environments, then we can generalize to *previously unseen* environments.

Invariant risk minimization

Intuition. In invariant risk minimization (IRM), we'll aim to minimize the risk subject to the constraint that the minimizer is an invariant predictor on all the training environments.

- If we train on sufficiently distinct environments, then we can generalize to *previously unseen* environments.
  - e.g. camel photos to cartoon camels

Roadmap

1. Existing techniques
2. Invariant risk minimization (IRM)
3. Relaxation of optimization problem
4. Open questions

Roadmap

1. Existing techniques
2. Invariant risk minimization (IRM)
3. Relaxation of optimization problem
4. Open questions

Problem statement

- **Given:** data D_e from training environments $e \in \mathcal{E}_{\text{tr}}$
- **Goal:** minimize the out-of-distribution risk,

$$R^{\text{OOD}}(f) = \min_{e \in \mathcal{E}} R_e(f).$$

Existing techniques

1. Perform ERM on merged data
2. Minimize robust learning objective
3. Domain adaptation
4. Invariant causal prediction (ICP)

ERM on merged data

Idea. Merge training data drawn from all training distributions $\mathcal{E}_{\text{tr}}$ and perform empirical risk minimization (ERM):

$$R^{\text{erm}}(f) := \frac{1}{|\mathcal{E}_{\text{tr}}|} \sum_{e \in \mathcal{E}_{\text{tr}}} R_e(f).$$

- **Problem:** training distributions $\mathcal{E}_{\text{tr}} \subset \mathcal{E}$ may not be representative of all distributions, leading to poor out-of-distribution generalization.

Robust learning objective

Idea. Minimize a robust learning objective:

$$R^{\text{rob}}(f) := \max_{e \in \mathcal{E}_{\text{tr}}} R_e(f) - r_e,$$

where r_e is an environment's baseline, which can help prevent noisy environments from dominating the objective.

- **Problem:** under most conditions, this is just a slight generalization of ERM to weighted average risk,

$$\min_f \sum_{e \in \mathcal{E}_{\text{tr}}} \lambda_e R_e(f).$$

Domain adaptation (slide under construction)

Idea. Estimate a data representation $\Phi(X)$ that has the same distribution $(\Phi(X), Y)$ for all environments.

- **Problem:** the distribution of ???

Invariant causal prediction (slide under construction)

Idea. Search for a subset of features...

Roadmap

1. Existing techniques
2. Invariant risk minimization (IRM)
3. Relaxation of optimization problem
4. Open questions

Problem recap

- **Given:** data D_e from training environments $e \in \mathcal{E}_{\mathrm{tr}}$
- **Goal:** minimize the out-of-distribution risk,

$$R^{\mathrm{OOD}}(f) = \min_{e \in \mathcal{E}} R_e(f).$$

Invariant predictor

Definition

A data representation $\phi : \mathcal{X} \to \tilde{\mathcal{X}}$ admits an **invariant predictor** f *across environments $\mathcal{E}$ if there exists some $\tilde{f}$ that simultaneously minimizes $\Phi^* R_e$ for all environments $e \in \mathcal{E}$:*

$$\tilde{f} \in \bigcap_{e \in \mathcal{E}} \operatorname*{arg\,min}_{g \in \mathcal{F}(\tilde{\mathcal{X}};\mathcal{Y})} (\Phi^* R_e)(g),$$

and $f = \Phi(\tilde{f})$. That is, $f = \tilde{f} \circ \phi$.

Trivial example

Let $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, and let $\mathcal{F}(\mathcal{X}; \mathcal{Y})$ be all linear functions.

Trivial example

Let $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, and let $\mathcal{F}(\mathcal{X}; \mathcal{Y})$ be all linear functions.

- Let $\phi : \mathbb{R}^d \to \mathbf{0}$ represent all data as 0.

Trivial example

Let $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, and let $\mathcal{F}(\mathcal{X}; \mathcal{Y})$ be all linear functions.

- Let $\phi : \mathbb{R}^d \to \mathbf{0}$ represent all data as 0.
  - $\mathcal{F}(\mathbf{0}; \mathbb{R})$ linear contains only the map $0 \mapsto 0$, so for all $e \in R_e$,

$$0 \in \underset{g \in \mathcal{F}(\mathbf{0}; \mathbb{R})}{\arg\min}\, (\Phi^* R_e)(g).$$

Trivial example

Let $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, and let $\mathcal{F}(\mathcal{X};\mathcal{Y})$ be all linear functions.

- Let $\phi : \mathbb{R}^d \to \mathbf{0}$ represent all data as 0.
  - $\mathcal{F}(\mathbf{0};\mathbb{R})$ linear contains only the map $0 \mapsto 0$, so for all $e \in R_e$,

$$0 \in \arg\min_{g \in \mathcal{F}(\mathbf{0};\mathbb{R})} (\Phi^* R_e)(g).$$

- For linear models, the trivial representation admits an invariant predictor.

Trivial example

Let $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, and let $\mathcal{F}(\mathcal{X}; \mathcal{Y})$ be all linear functions.

- Let $\phi : \mathbb{R}^d \to \mathbf{0}$ represent all data as 0.
  - $\mathcal{F}(\mathbf{0}; \mathbb{R})$ linear contains only the map $0 \mapsto 0$, so for all $e \in R_e$,

$$0 \in \arg\min_{g \in \mathcal{F}(\mathbf{0};\mathbb{R})} (\Phi^* R_e)(g).$$

- For linear models, the trivial representation admits an invariant predictor.
  - Though invariant, this representation likely does not admit predictors with low risk.

Invariant risk minimization

Problem. Let $\mathcal{E}$ be a collection of environments. Invariant risk minimization is the optimization problem:

$$\min_{\substack{\phi:\mathcal{X}\to\tilde{\mathcal{X}} \\ \tilde{f}:\tilde{\mathcal{X}}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}} R_e(\tilde{f}\circ\phi)$$

subject to the constraint that $\tilde{f} \in \arg\min R_e(\tilde{g}\circ\phi)$ for all $e \in \mathcal{E}$.

Invariant features and stable correlation

Intuition. We'll show that for certain problems, ϕ admits an invariant predictor if and only if the correlation between the representation and target variable is stable across all environments.

Invariant features and stable correlation

Definition

Let $\phi : \mathcal{X} \to \tilde{\mathcal{X}}$ be fixed. A **Bayes' predictor** *on the environment $e \in \mathcal{E}$ is a function $\tilde{f}_e$ that satisfies:*

$$\tilde{f}_e(\tilde{x}) := \mathop{\mathbb{E}}_{(X,Y)\sim P_e} \left[Y | \phi(X) = \tilde{x} \right],$$

for $\tilde{x} \in \mathrm{supp}_{P_e}(\phi(X))$.

- We'll assume that $\tilde{f}_e$ is measurable.

Invariant features and stable correlation

Definition

We say that an objective function R has an **essentially unique** *solution if any optimizer is unique up to a measure zero set.*

- That is, $f_1, f_2 \in \arg\min R(f)$ implies that $f_1 = f_2$ a.e.

Invariant features and stable correlation

Proposition

Suppose that ϕ is a representation such that the Bayes' predictor on e is an essentially unique solution for $\Phi^* R_e$, for all $e \in \mathcal{E}$. The following are equivalent:

Invariant features and stable correlation

Proposition

Suppose that ϕ is a representation such that the Bayes' predictor on e is an essentially unique solution for $\Phi^* R_e$, for all $e \in \mathcal{E}$. The following are equivalent:

- *ϕ admits an invariant predictor across $\mathcal{E}$*

Invariant features and stable correlation

Proposition

Suppose that ϕ is a representation such that the Bayes' predictor on e is an essentially unique solution for $\Phi^* R_e$, for all $e \in \mathcal{E}$. The following are equivalent:

- *ϕ admits an invariant predictor across $\mathcal{E}$*
- *for all $e, e' \in \mathcal{E}$,*

$$\mathop{\mathbb{E}}_{(X,Y)\sim P_e}\left[Y|\phi(X) = \tilde{x}\right] = \mathop{\mathbb{E}}_{(X,Y)\sim P_{e'}}\left[Y|\phi(X) = \tilde{x}\right],$$

Invariant features and stable correlation

Proposition

Suppose that ϕ *is a representation such that the Bayes' predictor on* e *is an essentially unique solution for* $\Phi^* R_e$*, for all* $e \in \mathcal{E}$*. The following are equivalent:*

- ϕ *admits an invariant predictor across* $\mathcal{E}$
- *for all* $e, e' \in \mathcal{E}$*,*

$$\mathbb{E}_{(X,Y)\sim P_e} [Y|\phi(X) = \tilde{x}] = \mathbb{E}_{(X,Y)\sim P_{e'}} [Y|\phi(X) = \tilde{x}],$$

Invariant features and stable correlation

Proposition

Suppose that ϕ is a representation such that the Bayes' predictor on e is an essentially unique solution for $\Phi^* R_e$, for all $e \in \mathcal{E}$. The following are equivalent:

- ϕ *admits an invariant predictor across* $\mathcal{E}$
- *for all* $e, e' \in \mathcal{E}$,

$$\mathop{\mathbb{E}}_{(X,Y)\sim P_e} \left[Y|\phi(X) = \tilde{x}\right] = \mathop{\mathbb{E}}_{(X,Y)\sim P_{e'}} \left[Y|\phi(X) = \tilde{x}\right],$$

for $\tilde{x} \in \mathrm{supp}_{P_e}(\phi(X)) \cap \mathrm{supp}_{P_{e'}}(\phi(X))$.

Invariant features and stable correlation

Proof (forward).

- If $\tilde{f}$ is an invariant predictor, then $\tilde{f}$ is a solution to $\Phi^* R_e$.

Invariant features and stable correlation

Proof (forward).

- If $\tilde{f}$ is an invariant predictor, then $\tilde{f}$ is a solution to $\Phi^* R_e$.
- The Bayes' predictor $\tilde{f}_e$ is also a solution, so $\tilde{f} = \tilde{f}_e$ a.e.

Invariant features and stable correlation

Proof (forward).

- If $\tilde{f}$ is an invariant predictor, then $\tilde{f}$ is a solution to $\Phi^* R_e$.
- The Bayes' predictor $\tilde{f}_e$ is also a solution, so $\tilde{f} = \tilde{f}_e$ a.e.
- Thus, $\tilde{f}_e = \tilde{f}_{e'}$ on the intersection of their supports.

□

Invariant features and stable correlation

Proof (forward).

- If $\tilde{f}$ is an invariant predictor, then $\tilde{f}$ is a solution to $\Phi^* R_e$.
- The Bayes' predictor $\tilde{f}_e$ is also a solution, so $\tilde{f} = \tilde{f}_e$ a.e.
- Thus, $\tilde{f}_e = \tilde{f}_{e'}$ on the intersection of their supports.

□

Reverse direction: stitch $\tilde{f}_e$'s together and read proof backwards.

Least squares regression

Let P_e be a collection of distributions over $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$,

$$X \leftarrow \mathcal{N}(\mu_e, \Sigma_e) \quad Y \leftarrow A_e X + \mathcal{N}(0, \sigma_e^2),$$

where $A_e \in \mathbb{R}^{1 \times d}$.

Least squares regression

Let P_e be a collection of distributions over $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$,

$$X \leftarrow \mathcal{N}(\mu_e, \Sigma_e) \quad Y \leftarrow A_e X + \mathcal{N}(0, \sigma_e^2),$$

where $A_e \in \mathbb{R}^{1 \times d}$. The least squares objective is:

$$R_e(f) := \mathop{\mathbb{E}}_{(x,y) \sim P_e} \left[|f(x) - y|^2 \right].$$

Least squares regression

Let P_e be a collection of distributions over $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$,

$$X \leftarrow \mathcal{N}(\mu_e, \Sigma_e) \quad Y \leftarrow A_e X + \mathcal{N}(0, \sigma_e^2),$$

where $A_e \in \mathbb{R}^{1 \times d}$. The least squares objective is:

$$R_e(f) := \mathop{\mathbb{E}}_{(x,y) \sim P_e} \left[|f(x) - y|^2 \right].$$

Define $\mathcal{F}_k = \{\tilde{f} : \mathbb{R}^k \to \mathbb{R} \ \text{linear}\}$ to the class of k-dimensional linear functionals.

Least squares regression

Let P_e be a collection of distributions over $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$,

$$X \leftarrow \mathcal{N}(\mu_e, \Sigma_e) \quad Y \leftarrow A_e X + \mathcal{N}(0, \sigma_e^2),$$

where $A_e \in \mathbb{R}^{1 \times d}$. The least squares objective is:

$$R_e(f) := \mathop{\mathbb{E}}_{(x,y) \sim P_e} \left[|f(x) - y|^2 \right].$$

Define $\mathcal{F}_k = \{\tilde{f} : \mathbb{R}^k \to \mathbb{R} \ \text{linear}\}$ to the class of k-dimensional linear functionals.

- Let $\phi : \mathbb{R}^d \to \mathbb{R}^k$ be linear. Then, $\mathcal{F}_k$ contains the Bayes' predictor for $\Phi^* R_e$, and it is essentially unique.

Out-of-distribution generalization through causality

- So far, we've shown that finding an invariant predictor $(\phi, \tilde{f})$ can be thought of finding stable correlations between X and Y across all environments in certain classes of problems.

Out-of-distribution generalization through causality

- So far, we've shown that finding an invariant predictor $(\phi, \tilde{f})$ can be thought of finding stable correlations between X and Y across all environments in certain classes of problems.
  - We can use this technique to recover causal features in situations where the only stable correlations arise from an assumed causal structure.

Out-of-distribution generalization through causality

- So far, we've shown that finding an invariant predictor $(\phi, \tilde{f})$ can be thought of finding stable correlations between X and Y across all environments in certain classes of problems.
  - We can use this technique to recover causal features in situations where the only stable correlations arise from an assumed causal structure.
  - This would allow us to generalize to unseen distributions that satisfy the same causal structure.

Structural equation model

Let $X = (X_1, \ldots, X_d)$ be a random vector.

Structural equation model

Let $X = (X_1, \ldots, X_d)$ be a random vector.

Definition

A **structural equation model** $\mathcal{C} := (\mathcal{S}, N)$ *governing* X *is a set of* *structural equations* *such that:*

$$\mathcal{S}_i : X_i \leftarrow f_i(\mathrm{Pa}(X_i), N_i),$$

where $\mathrm{Pa}(X_i)$ *are the* *parents* *of* X_i *and* N_i *are independent noise.*

Structural equation model

Let $X = (X_1, \ldots, X_d)$ be a random vector.

Definition

A **structural equation model** $\mathcal{C} := (\mathcal{S}, N)$ *governing* X *is a set of* *structural equations* *such that:*

$$\mathcal{S}_i : X_i \leftarrow f_i(\mathrm{Pa}(X_i), N_i),$$

where $\mathrm{Pa}(X_i)$ *are the* *parents* *of* X_i *and* N_i *are independent noise.*

- We say that X_i *causes* X_j if $X_i \in \mathrm{Pa}(X_j)$.

Structural equation model

Let $X = (X_1, \ldots, X_d)$ be a random vector.

Definition

A **structural equation model** $\mathcal{C} := (\mathcal{S}, N)$ *governing* X *is a set of* *structural equations such that:*

$$\mathcal{S}_i : X_i \leftarrow f_i(\mathrm{Pa}(X_i), N_i),$$

where $\mathrm{Pa}(X_i)$ *are the* *parents* *of* X_i *and* N_i *are independent noise.*

- We say that X_i *causes* X_j if $X_i \in \mathrm{Pa}(X_j)$.
- A *causal graph* is the graph $G = (V, E)$ where $V = [d]$ and $(i, j) \in E$ if and only if X_i causes X_j.

Structural equation model

Let $X = (X_1, \ldots, X_d)$ be a random vector.

Definition

A **structural equation model** $\mathcal{C} := (\mathcal{S}, N)$ *governing* X *is a set of* *structural equations such that:*

$$\mathcal{S}_i : X_i \leftarrow f_i(\mathrm{Pa}(X_i), N_i),$$

where $\mathrm{Pa}(X_i)$ *are the* *parents* *of* X_i *and* N_i *are independent noise.*

- We say that X_i *causes* X_j if $X_i \in \mathrm{Pa}(X_j)$.
- A *causal graph* is the graph $G = (V, E)$ where $V = [d]$ and $(i, j) \in E$ if and only if X_i causes X_j.
- We assume *acyclic* causal graphs.

Interventions

Definition

Let $\mathcal{C} = (\mathcal{S}, N)$ be a SEM. An **intervention** e *on* $\mathcal{C}$ *consists of replacing one or several of its structural equations to obtain an* *intervened* *SEM* $\mathcal{C}^e = (\mathcal{S}^e, N^e)$*,*

$$S_i^e : X_i^e \leftarrow f_i^e(\mathrm{Pa}^e(X_i^e), N_i^e).$$

Interventions

Definition

Let $\mathcal{C} = (\mathcal{S}, N)$ be a SEM. An **intervention** *e on $\mathcal{C}$ consists of replacing one or several of its structural equations to obtain an intervened SEM $\mathcal{C}^e = (\mathcal{S}^e, N^e)$,*

$$S_i^e : X_i^e \leftarrow f_i^e(\mathrm{Pa}^e(X_i^e), N_i^e).$$

The variable X_i^e is intervened if $S_i \neq S_i^e$ or $N_i \neq N_i^e$.

Valid interventions

Definition

Let $\mathcal{C}$ be a SEM that governs $(X_1, \ldots, X_d, Y)$ for the learning task of predicting Y from X. An intervention e is **valid** *as long as:*

Valid interventions

Definition

Let $\mathcal{C}$ be a SEM that governs $(X_1, \ldots, X_d, Y)$ for the learning task of predicting Y from X. An intervention e is **valid** *as long as:*

(i) *the causal graph remains acyclic*

Valid interventions

Definition

Let $\mathcal{C}$ be a SEM that governs $(X_1, \ldots, X_d, Y)$ for the learning task of predicting Y from X. An intervention e is **valid** *as long as:*

(i) *the causal graph remains acyclic*

(ii) $\mathbb{E}[Y^e|\mathrm{Pa}(Y)] = \mathbb{E}[Y|\mathrm{Pa}(Y)]$,

Valid interventions

Definition

Let $\mathcal{C}$ be a SEM that governs $(X_1, \ldots, X_d, Y)$ for the learning task of predicting Y from X. An intervention e is **valid** *as long as:*

(i) *the causal graph remains acyclic*

(ii) $\mathbb{E}[Y^e | \mathrm{Pa}(Y)] = \mathbb{E}[Y | \mathrm{Pa}(Y)]$,

(iii) $\mathrm{Var}[Y^e | \mathrm{Pa}(Y)] < \infty$.

Valid interventions

Definition

Let $\mathcal{C}$ be a SEM that governs $(X_1, \ldots, X_d, Y)$ for the learning task of predicting Y from X. An intervention e is **valid** *as long as:*

(i) *the causal graph remains acyclic*

(ii) $\mathbb{E}[Y^e | \mathrm{Pa}(Y)] = \mathbb{E}[Y | \mathrm{Pa}(Y)]$,

(iii) $\mathrm{Var}[Y^e | \mathrm{Pa}(Y)] < \infty$.

Valid interventions

Definition

Let $\mathcal{C}$ be a SEM that governs $(X_1, \ldots, X_d, Y)$ for the learning task of predicting Y from X. An intervention e is **valid** *as long as:*

(i) *the causal graph remains acyclic*

(ii) $\mathbb{E}[Y^e|\mathrm{Pa}(Y)] = \mathbb{E}[Y|\mathrm{Pa}(Y)]$,

(iii) $\mathrm{Var}[Y^e|\mathrm{Pa}(Y)] < \infty$.

- Let $\mathcal{E}_{\text{all}}(\mathcal{C})$ be the set of all environments containing the interventional distribution $P(X^e, Y^e)$ indexed by valid interventions e.

Least squares regression

Example. Consider the structural equation model:

$$X_1 \leftarrow \mathcal{N}(0, \sigma^2) \qquad Y \leftarrow X_1 + \mathcal{N}(0, \sigma^2) \qquad X_2 \leftarrow Y + c.$$

Least squares regression

Example. Consider the structural equation model:

$$X_1 \leftarrow \mathcal{N}(0, \sigma^2) \qquad Y \leftarrow X_1 + \mathcal{N}(0, \sigma^2) \qquad X_2 \leftarrow Y + c.$$

- **environments** $e \in \mathcal{E}$

$$\mathcal{E} = \left\{ (\sigma^2, c) : \sigma^2 \in \left[0, \sigma^2_{\max}\right], c \in \mathbb{R} \right\}$$

Least squares regression

Example. Consider the structural equation model:

$$X_1 \leftarrow \mathcal{N}(0, \sigma^2) \qquad Y \leftarrow X_1 + \mathcal{N}(0, \sigma^2) \qquad X_2 \leftarrow Y + c.$$

- **environments** $e \in \mathcal{E}$

$$\mathcal{E} = \left\{(\sigma^2, c) : \sigma^2 \in \left[0, \sigma^2_{\max}\right], c \in \mathbb{R}\right\}$$

- **predictor** parametrized by $\mathbf{w} \in \mathbb{R}^2$, where

$$\hat{Y} = X_1 w_1 + X_2 w_2$$

Least squares regression

Example. Consider the structural equation model:

$$X_1 \leftarrow \mathcal{N}(0, \sigma^2) \qquad Y \leftarrow X_1 + \mathcal{N}(0, \sigma^2) \qquad X_2 \leftarrow Y + c.$$

- **environments** $e \in \mathcal{E}$

$$\mathcal{E} = \left\{ (\sigma^2, c) : \sigma^2 \in \left[0, \sigma^2_{\max}\right], c \in \mathbb{R} \right\}$$

- **predictor** parametrized by $\mathbf{w} \in \mathbb{R}^2$, where

$$\hat{Y} = X_1 w_1 + X_2 w_2$$

- **risk** R_{σ^2} is the mean squared error:

$$R_{\sigma^2}(\mathbf{w}) = \mathop{\mathbb{E}}_{(\mathbf{X},Y)\sim P_{\sigma^2}} \left[\left(\mathbf{X}^\mathsf{T} \mathbf{w} - Y \right)^2 \right]$$

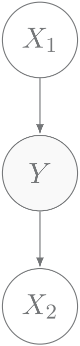
Least squares regression

Observations.

- If $w_2 \neq 0$ (i.e. the predictor uses X_2), then:

$$R^{\text{OOD}}(\mathbf{w}) = \infty,$$

since $c^2 \cdot w_2^2 \leq R_{(\sigma^2, c)}(\mathbf{w})$ and $c \in \mathbb{R}$.



Least squares regression

Observations.

- If $w_2 \neq 0$ (i.e. the predictor uses X_2), then:

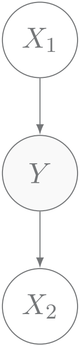
$$R^{\text{OOD}}(\mathbf{w}) = \infty,$$

  since $c^2 \cdot w_2^2 \leq R_{(\sigma^2,c)}(\mathbf{w})$ and $c \in \mathbb{R}$.

- In particular, the best-in-class predictor is:

$$\mathbf{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

  and corresponds to the *causal predictor*.



Preview of Linear IRM

Let $\mathcal{C}$ be the causal graph for $(X_1, \ldots, X_d, Y)$ and $\mathcal{E}(\mathcal{C})$ be the set of valid environments given $\mathcal{C}$. In the linear setting:

Preview of Linear IRM

Let $\mathcal{C}$ be the causal graph for $(X_1, \ldots, X_d, Y)$ and $\mathcal{E}(\mathcal{C})$ be the set of valid environments given $\mathcal{C}$. In the linear setting:

Theorem (Informal)

A predictor is invariant across $\mathcal{E}(\mathcal{C})$

- *if and only if attains optimal R^{OOD}, and*

Preview of Linear IRM

Let $\mathcal{C}$ be the causal graph for $(X_1, \ldots, X_d, Y)$ and $\mathcal{E}(\mathcal{C})$ be the set of valid environments given $\mathcal{C}$. In the linear setting:

Theorem (Informal)

A predictor is invariant across $\mathcal{E}(\mathcal{C})$

- *if and only if attains optimal R^{OOD}, and*
- *if and only if it is the Bayes' predictor using only the direct causal parents of Y to predict.*

Example generalization result

Return to our least squares problem. Making the following assumption:

Assumption 8. *A set of training environments* $\mathcal{E}_{tr}$ *lie in a* linear general position *of degree* r *if* $|\mathcal{E}_{tr}| > d - r + \frac{d}{r}$ *for some* $r \in \mathbb{N}$, *and for all non-zero* $x \in \mathbb{R}^{d\times 1}$*:*

$$\dim\left(\operatorname{span}\left(\left\{\mathbb{E}_{X^e}\left[X^{e\top}X^e\right]x - \mathbb{E}_{X^e,\epsilon^e}\left[X^{e\top}\epsilon^e\right]\right\}_{e\in\mathcal{E}_{tr}}\right)\right) > d - r.$$

Example generalization result

We obtain an upper bound on the number of 'linearly independent' training environments we need to see before we can generalize to all environments:

Theorem 9. *Assume that*

$$Y^e = Z_1^e \cdot \gamma + \epsilon^e, \quad Z_1^e \perp \epsilon^e, \quad \mathbb{E}[\epsilon^e] = 0,$$
$$X^e = (Z_1^e, Z_2^e) \cdot S.$$

Here, $\gamma \in \mathbb{R}^{d \times 1}$*,* Z_1^e *takes values in* $\mathbb{R}^{1 \times d}$*, and* Z_2^e *takes values in* $\mathbb{R}^{1 \times q}$*. Assume that there exists* $\tilde{S} \in \mathbb{R}^{(d+q) \times d}$ *such that* $X^e \tilde{S} = X_1^e$*, for all environments* $e \in \mathcal{E}_{all}$*. Let* $\Phi \in \mathbb{R}^{d \times d}$ *have rank* $r > 0$*. Then, if at least* $d - r + \frac{d}{r}$ *training environments* $\mathcal{E}_{tr} \subseteq \mathcal{E}_{all}$ *lie in a linear general position of degree* r*, we have that*

$$\Phi \, \mathbb{E}_{X^e}\left[X^{e\top} X^e\right] \Phi^\top w = \Phi \, \mathbb{E}_{X^e, Y^e}\left[X^{e\top} Y^e\right] \tag{7}$$

holds for all $e \in \mathcal{E}_{tr}$ *iff* Φ *elicits the invariant predictor* $\Phi^\top w$ *for all* $e \in \mathcal{E}_{all}$*.*

Roadmap

1. Existing techniques
2. Invariant risk minimization (IRM)
3. Relaxation of optimization problem
4. Open questions

IRM objective

Recall the IRM objective:

$$\min_{\substack{\phi:\mathcal{X}\to\tilde{\mathcal{X}} \\ \tilde{f}:\tilde{\mathcal{X}}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\text{tr}}} R_e(\tilde{f}\circ\phi)$$

subject to the constraint that $\tilde{f} \in \arg\min\, R_e(\tilde{g}\circ\phi)$ for all $e \in \mathcal{E}_{\text{tr}}$.

Relaxation of objective

To enable a practical algorithm, define the relaxation:

$$\min_{\phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R_e(w\cdot\phi) + \lambda \cdot \left\|\nabla_{w|w=1} R_e(w\cdot\phi)\right\|^2,$$

where $w \in \mathbb{R}$ is set to 1.

Roadmap

1. Existing techniques
2. Invariant risk minimization (IRM)
3. Relaxation of optimization problem
4. Open questions

Generating practice tasks from SEMs

Question: given knowledge of SEMs, can a learner construct tasks for themselves to perform to learn?

- For example, say you want to learn how to drive a car. You could set up an imaginary environment (imagine lane lines) that you task yourself with, to learn the physics of navigation.

Communicating knowledge through SEMs

Question: it seems that a lot of hard work is done by humans to learn causal relations. Would knowledge of this speed up learning?

- Perhaps this is how multiple learners could communicate knowledge.

Noise model

What about cases where there does not exist a representation that admits an invariant predictor? Or, in other words, here it is assumed that there is zero-mean noise, so that the model is correct. But what about an adversarial noise model/agnostic setting?

Active IRM

Actively work to figure out the underlying SEM through statistical tests?

IRM fundamental question

If the goal is to find a representation that admits an invariant predictor for all environment, how can one prove that this is possible through an estimation procedure? That is, drawing some number of points, from the training environments, how many point/how well do you need to perform ERM to find ϕ that is close to the true ϕ?

Citations

Images.

- https://www.ciwf.org.uk/farm-animals/sheep/
- https://www.thetimes.co.uk/article/adopt-a-sheep-with-sheep-inc-a-new-knitwear-brand-txhmfmgx5
- https://www.morningagclips.com/understanding-agriculture-sheep/
- https://wsbt.com/news/offbeat/truck-stop-camel-prescribed-antibiotics-after-woman-bites-it
- https://www.youtube.com/watch?v=XXQ859dZb24
- http://www.todayifoundout.com/index.php/2010/07/camels-humps-are-not-filled-with-water/
- https://pixabay.com/da/photos/kamel-vilde-kameler-bactrian-2694771/
- https://www.flaticon.com/free-icon/network_148800
- https://www.pngfly.com/png-6zfc7y