Learning Causal Effects via Weighted Empirical Risk Minimization

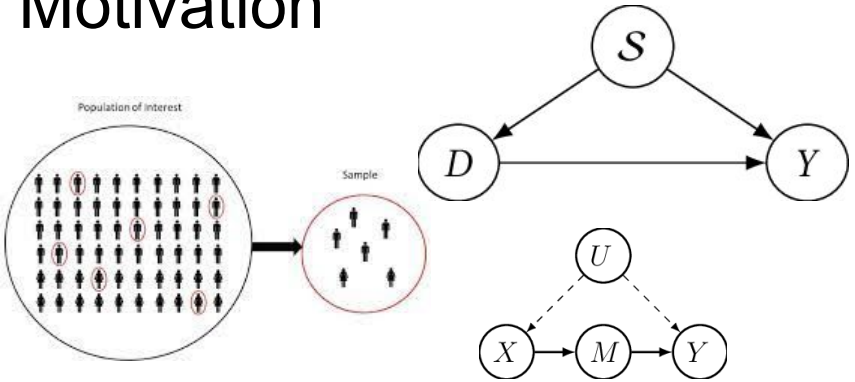
Paper by: Yonghan Jung, Jin Tian, Elias Bareinboim
Some slides inspired by those of Yonghan Jung at NeurIPS-20
Slides also inspired by Causality

Shakil Jiwa

Contents

- Motivation
- Identification and Estimation of Causal Effects
- Empirical Risk Minimization
- Paper Contribution
- Algorithm Description
- Experimental Results

Motivation

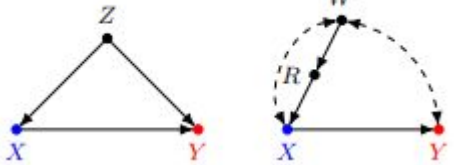


??

P(y|do(x))
*in a way that we can *estimate*

- Can a causal distribution be uniquely computed from a combination of the observational distribution P(V) and the causal graph G.
- In general what do we do? Parametric model of conditional probabilities, suffers on computationally high dimensional data.
- Effective estimators developed for when back door holds (known as ignorability in statistics) and further. Ie g-formula

Identification -> Estimation Of Causal Effects

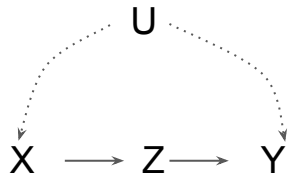


$$P(y|do(x)) = (\sum_{w} P(y,x|r,w)P(w))/(\sum_{w} P(x|r,w)P(w))$$

BD Identification Algorithm, fundamentals?

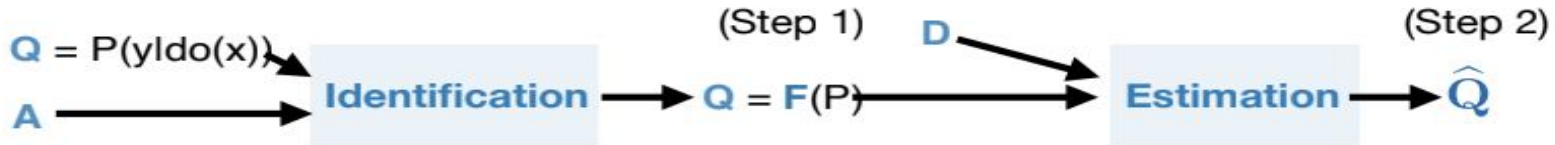
- Problem: No systematic way of estimating arbitrary identifiable functionals that are both computationally and statistically attractive.
- Why is it hard to compute the RHS of the above equation given knowledge of a sample population?
- How is this problem of estimation different than that of identifiability?
- What are some ways of doing this currently?

Estimation of Causal Effects: Towards Parametric Forms



$$P(Y|\mathrm{do}(X)) = \sum_{z}\sum_{x'} P(Y|Z,X')\,P(X')\,P(Z|X)$$

- Ways to estimate? Why not just do it empirically?
- In general estimating Q is NP-Hard in Bayesian Belief Networks, O(2^n) where n is the number of nodes.
- Parametric approximation?
- Since P(Y|do(X)) = Rzx*Ryzx*x (single door criterion) the query can estimated by estimating regression coefficients. These coefficients can be estimated using RMSE.

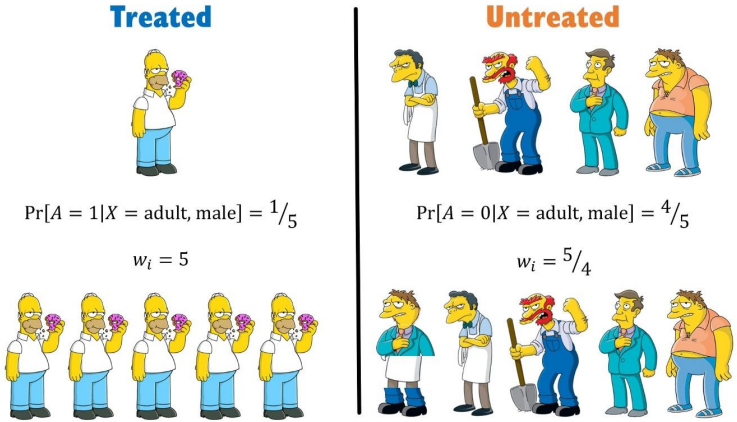


Estimation of Causal Effects: Propensity Score method (Rosenbaum and Rubin, 1983)

BD Example: it is often more advantageous to use the unfactored form:

$$P(y|do(x)) = \sum_z P(y|x,z)P(z) = \sum_z \frac{P(x,y,z)}{P(x|z)}$$

In order to estimate the query, weigh individual samples by the inverse of P(x|z) and treat weighted samples as if they were drawn from the post intervention distribution. This method is known as inverse probability weighting and is an estimation technique in the learning of causal effects.



How can we learn these causal effects in general?

Empirical Risk Minimization

With two spaces of objects X and Y, we would like to learn a hypothesis h: X->Y

From this we can define risk as: R(h) = E(L(h(x),y)) where L is some likelihood function.

The ultimate goal of risk minimization is to find:

$$h^* = \arg\min_{h \in \mathcal{H}} R(h).$$

Since in general P(x) is unknown to the learning algorithm, we can compute an approximation by averaging the loss function on the training set:

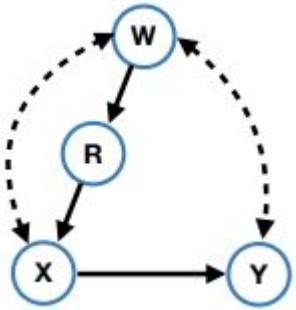
$$R_{\text{emp}}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i) \qquad \hat{h} = \arg\min_{h \in \mathcal{H}} R_{\text{emp}}(h).$$

Empirical risk minimization has been used to estimate functionals in domain adaptation, can these methods be extended to causal inference?

Weighted Empirical Risk Minimization

$$\widetilde{\mathcal{R}}_{w^*,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_i^* \ell(\theta, Z_i')$$

- Where have we seen something similar to this?
- Propensity scoring and other methods exist. OK, great that is it?
- Consider the following graph and query:



$$P(y|do(x)) = \frac{\sum_w P(y,x|r,w)P(w)}{\sum_w P(x|r,w)P(w)}$$

Is the causal effect identifiable?

Is the causal effect in the form of a WERM estimator?

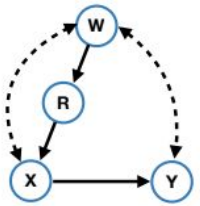
Current Status of Causal Inference

Strengths:

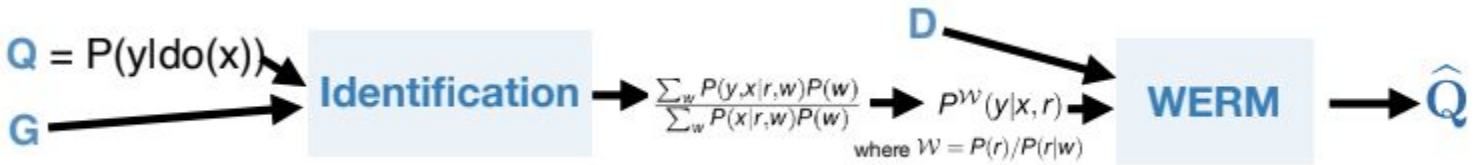
- There exist sound and complete identification algorithms for determining identifiability of a causal query Q
- When Q=F(P) is given as a weighted distribution, empirical risk minimization based estimators have been established.

Weakness:

- Estimation has mainly been done on the backdoor (ignorability) criterion.
- When Q=F(P) is given as a weighted distribution, empirical risk minimization based estimators have been established.



Transformation from query to weighted distribution:



Paper Contribution

- ***(Gap in causal inference)*** Even if sound and complete algorithms for causal identification have been developed, it's still unknown (neither obvious) how to estimate the causal effect sample & time-efficiently.

- ***(Gap in learning theory)*** Even if (weighted)-ERM based estimators have been established when the query is given as a weighed distribution, it's not clear how to use WERM based-estimators.

- ***(Contribution)*** In this paper, we fill the gap between the causal inference and the ERM learning theory by
  - Developing the sound and complete procedure for representing any causal functional into weighted distributions (Algo 1); and
  - Formulating the causal estimation as the WERM optimization, by providing generalization bound and theoretical learning guarantees. (Sec 4)

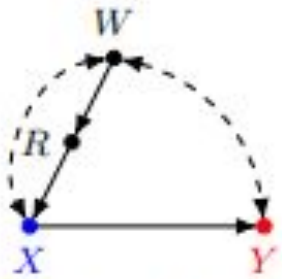
Preliminaries

- Structural Causal Model (SCM) M associated with graph G over a set of variables V which induces a distribution P(v).
- Weighted distribution: given P(v) and a weight function 0<W(v)<inf such that E[W(V)] = 1 and E[W^2(V)] < inf, a weighted distribution is given by W(v)P(v)
- Causal effects identification: P(y|do(x)) is identifiable if it is computable from the observed P(v) in any SCM that induces G.
- In a causal graph H, two variables are said to be in the same confounded component (C-component) iff they are connected by a bi-directed path.
- For any C (as a subset of) V, the quantity Q[C] or the C-factor is defined as the post-intervention distribution of C under an intervention on V\C: P(c|do(v\c))
- The joint distribution P(v) can be factorized by C factors where Si are the set of C components of G.

$$P(\mathbf{v}) = \prod_i Q[\mathbf{S}_i],$$

Confounding Components

What are the c-components of the following graph? What would the c-factors be defined as?



(b) Example 1

Weighted Distribution Identification

Lemma 1 (Computing C-factors as weighted distributions). *Let a topological order over* $\mathbf{V}$ *be* $V_1 < V_2 < \cdots < V_n$. *Suppose* $Q[\mathbf{A}]$ *is given by* $Q[\mathbf{A}] = P^{\mathcal{W}}(\mathbf{a}|\mathbf{r})$ *for some* $\mathbf{R} \subseteq \mathbf{V}$ *and weight function* $\mathcal{W}$.

1. *If* $\mathbf{W}$ *is a C-component of* $G_{\mathbf{A}}$, *then* $Q[\mathbf{W}] = P^{\mathcal{W}\times\mathcal{W}'}(\mathbf{w}|\mathbf{r}')$, *where* $\mathbf{R}' \equiv \mathbf{R} \cup ((\mathbf{A}\backslash\mathbf{W}) \cap An(\mathbf{W}))$ *and* $\mathcal{W}' \equiv \frac{P^{\mathcal{W}}((\mathbf{a}\backslash\mathbf{w})\cap An(\mathbf{w})|\mathbf{r})}{\prod_{V_i\in(\mathbf{A}\backslash\mathbf{W})\cap An(\mathbf{W})} P^{\mathcal{W}}\left(v_i|\mathbf{v}^{(i-1)}\cap\mathbf{a}\cap An(\mathbf{w}),\mathbf{r}\right)}$.

2. *If* $\mathbf{W} \subseteq \mathbf{A}$ *satisfies* $\mathbf{W} = An(\mathbf{W})_{G_{\mathbf{A}}}$, *then* $Q[\mathbf{W}] = P^{\mathcal{W}}(\mathbf{w}|\mathbf{r})$.

- C-Factors can be recursively computed in terms of weighted distributions.
- WERM requires c factorization of the causal query, standard forms usually don’t work.

Algorithm 1: wID $(\mathbf{x}, \mathbf{y}, G, P)$

Input: $\mathbf{x}, \mathbf{y}, G, P$

Output: Expression of $P(\mathbf{y}|do(\mathbf{x}))$ as a weighted distribution; or FAIL if $P(\mathbf{y}|do(\mathbf{x}))$ is unidentifiable.

1 Let $\mathbf{V} \leftarrow An(\mathbf{Y})$; $P(\mathbf{v}) \leftarrow P(An(\mathbf{Y}))$; and $G \leftarrow G_{An(\mathbf{Y})}$.

2 Find the C-components of G: $\mathbf{S}_1, \cdots, \mathbf{S}_k$.

3 Let $Q[\mathbf{S}_i] = P^{\mathcal{W}_{\mathbf{s}_i}}(\mathbf{s}_i|\mathbf{r}_{\mathbf{s}_i})$ where $(\mathcal{W}_{\mathbf{s}_i}, \mathbf{r}_{\mathbf{s}_i})$ are derived from Lemma 1.

4 Let $\mathbf{D} \equiv An(\mathbf{Y})_{G_{\mathbf{V}\backslash\mathbf{X}}}$.

5 Find the C-component of $G_{\mathbf{D}}$: $\mathbf{D}_1, \cdots \mathbf{D}_K$.

6 For each $\mathbf{D}_i \in \mathbf{S}_j$ for some (i, j), let
$Q[\mathbf{D}_i] = \texttt{wIdentify}(\mathbf{D}_i, \mathbf{S}_j, Q[\mathbf{S}_j], \mathbf{r}_{\mathbf{S}_j}, \mathcal{W}_{\mathbf{S}_j}) \equiv P^{\mathcal{W}_{\mathbf{d}_i}}(\mathbf{d}_i|\mathbf{r}_{\mathbf{d}_i})$.

7 **if** $K = 1$ **then**
  return $P(\mathbf{y}|do(\mathbf{x})) = P^{\mathcal{W}_{\mathbf{d}_1}}(\mathbf{y}|\mathbf{r}_{\mathbf{d}_1})$.
end

8 Let $\mathcal{W} \equiv \prod_{i=1}^{K} P^{\mathcal{W}_{\mathbf{d}_i}}(\mathbf{d}_i|\mathbf{r}_{\mathbf{d}_i}) / P(\mathbf{d}|\mathbf{r})$ where $\mathbf{R} \equiv \mathbf{V}\backslash\mathbf{D}$.

9 **return** $P(\mathbf{y}|do(\mathbf{x})) = P^{\mathcal{W}}(\mathbf{y}|\mathbf{r})$

Procedure $\texttt{wIdentify}(\mathbf{C}, \mathbf{T}, Q[\mathbf{T}], \mathbf{r}, \mathcal{W})$

  Input: $\mathbf{T}, Q[\mathbf{T}] = P^{\mathcal{W}}(\mathbf{t}|\mathbf{r})$

  Output: $Q[\mathbf{C}]$ for $\mathbf{C} \subseteq \mathbf{T}$ as a weighted distribution.

a.1 Let $\mathbf{A} \equiv An(\mathbf{C})_{G_{\mathbf{T}}}$, then $Q[\mathbf{A}] = P^{\mathcal{W}}(\mathbf{a}|\mathbf{r})$ by Lemma 1.

a.2 **if** $\mathbf{A} = \mathbf{C}$ **then**
    return $Q[\mathbf{C}] = P^{\mathcal{W}}(\mathbf{a}|\mathbf{r})$
  end

a.3 **if** $\mathbf{A} = \mathbf{T}$ **then**
    return *FAIL*
  end

a.4 **else**

a.5 Let $\mathbf{S}$ denote the C-component in $G_{\mathbf{A}}$ such that $\mathbf{C} \subseteq \mathbf{S}$.

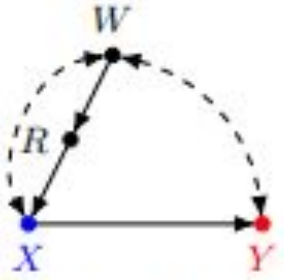
a.6 Compute $Q[\mathbf{S}] = P^{\mathcal{W} \times \mathcal{W}'}(\mathbf{s}|\mathbf{r}')$ where $(\mathcal{W}', \mathbf{r}')$ are derived by Lemma 1.

a.7 **return** $\texttt{wIdentify}(\mathbf{C}, \mathbf{S}, Q[\mathbf{S}], \mathbf{r}', \mathcal{W} \times \mathcal{W}')$

  end

- An(Y) is union with ancestors
- Recursive
- How can we move from this to estimation?

WERM Identification Procedure



(b) Example 1

of Algo. 1 using the model in Fig. 1b, where $P(y|do(x))$ is identified as given in Eq. (1) (i.e., not in the weighting-form). The graph has two C-components $\mathbf{S}_1 = \{W, X, Y\}$ and $\mathbf{S}_2 = \{R\}$ (Line 2). We have $Q[\mathbf{S}_1] = P^{\mathcal{W}_1}(\mathbf{s}_1|r)$ where $\mathcal{W}_1 = P(r)/P(r|w)$, and $Q[\mathbf{S}_2] = P(r|w)$ by Lemma 1 as discussed previously (Line 3). Let $\mathbf{D} = An(Y)_{G_{\mathbf{V}\setminus X}} = \{Y\}$ (Line 4). Run $\texttt{wIdentify}(Y, \mathbf{S}_1, Q[\mathbf{S}_1], r, \mathcal{W}_1)$ (Line 6). In Procedure `wIdentify()`, let $\mathbf{A} = An(Y)_{G_{\mathbf{S}_1}} = \{X, Y\}$, then $Q[\mathbf{A}] = P^{\mathcal{W}_1}(\mathbf{a}|r)$ (Line a.1). In $G_{\mathbf{A}} = G_{\{X,Y\}}$, let $\mathbf{S} \equiv \{Y\}$ denote the C-component containing Y (Line a.5). Then, $Q[\mathbf{S}] = Q[Y] = P^{\mathcal{W}_1 \times \mathcal{W}'}(y|\mathbf{r}')$ where $\mathbf{R}' = \{R, X\}$ and $\mathcal{W}' = P^{\mathcal{W}}(x|r)/P^{\mathcal{W}}(x|r) = 1$ by Lemma 1 (with $\mathbf{W} = \mathbf{S} = Y$) (Line a.6). Line a.7 returns $Q[Y] = \texttt{wIdentify}(Y, \mathbf{S}, Q[\mathbf{S}], r', \mathcal{W}_1) = P^{\mathcal{W}_1}(y|x, r)$. Finally we obtain $P(y|do(x)) = P^{\mathcal{W}_1}(y|x, r)$ (Line 7).

Learning Causal Effects via WERM: Generalization Bound

In the WERM setting we are trying to learn a function h(r) that approximates E[Y|r]. We seek here to minimize the expected loss on the weighted probability distribution function: the *weighted risk*, using finite samples D

$$\widehat{R}^{\mathcal{W}^*}(h) \equiv \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}^*(\mathbf{V}_{(i)})\ell\left(h(\mathbf{R}_{(i)}), Y_{(i)}\right).$$

If the optimization suffers due to high variance and low sample size (which could lead to data overfitting) we can introduce a new weight function intended to have lower variance. The difference in these functions is given for a delta in (0,1):

$$|R^{\mathcal{W}^*}(h) - \widehat{R}^{\mathcal{W}}(h)| \leq \underbrace{\mathbb{E}_P\left[|\mathcal{W}^*(\mathbf{V}) - \mathcal{W}(\mathbf{V})|\right]}_{(a)} + 2^{5/4} \max \underbrace{\left(\sqrt{\mathbb{E}_P\left[\mathcal{W}^2\ell_h^2\right]}, \sqrt{\mathbb{E}_{\hat{P}}\left[\mathcal{W}^2\ell_h^2\right]}\right)}_{(b)} \underbrace{F(p,m,\delta)}_{(c)}, \tag{5}$$

$$\textit{where } F(p,m,\delta) \equiv \left((p\log(2me/p) + \log(4/\delta))^{3/8}\right)/(m^{3/8}).$$

b is the second moment of variance and c is the pseudo-dimension,

Side note: Pseudo-dimension

Definition 4.1. *Let (X, S) be a given measurable space, and let $\mathcal{A} \subseteq S$. A set $S = \{x_1, \ldots, x_n\} \subseteq X$ is said to be* **shattered** *by $\mathcal{A}$ if, for every subset $B \subseteq S$, there exists a set $A \in \mathcal{A}$ such that $S \cap A = B$. The* **Vapnik-Chervonenkis dimension** *of $\mathcal{A}$, denoted by $VC\text{-}dim(\mathcal{A})$, equals the largest integer n such that there exists a set of cardinality n that is shattered by $\mathcal{A}$.*

Pollard's pseudo-dimension is a generalization of the VC dimension to real-values functions.

Learning Causal Effects via WERM: Learning Objective

The following learning objective is proposed to simultaneously learn a hypothesis h that minimizes risk and a low-variance weight-function W*

$$\mathcal{L}(\mathcal{W}, h) \equiv \underbrace{\widehat{\mathcal{R}}^{\mathcal{W}}(h) + \frac{\lambda_h}{m} C(h)}_{\mathcal{L}_h(h, \mathcal{W}, \lambda_h)} + \underbrace{\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\mathcal{W}(\mathbf{V}_{(i)}) - \mathcal{W}^*(\mathbf{V}_{(i)})\right)^2 + \frac{\lambda_{\mathcal{W}}}{m}\|\mathcal{W}\|_2^2}}_{\mathcal{L}_{\mathcal{W}}(\mathcal{W}, \lambda_{\mathcal{W}}; \mathcal{W}^*)}, \quad (6)$$

where $\mathcal{L}_h(h, \mathcal{W}, \lambda_h)$ consists of the WER $\widehat{R}^{\mathcal{W}}(h)$ and a regularizer $C(h)$ of h, such as L_1 or L_2 regularization for the parameters of h; $\mathcal{L}_{\mathcal{W}}(\mathcal{W}, \lambda_{\mathcal{W}}; \mathcal{W}^*)$ measures the deviance of $\mathcal{W}$ from $\mathcal{W}^*$ with L_2 regularization to penalize the variance of $\mathcal{W}$; and $(\lambda_h, \lambda_{\mathcal{W}})$ are hyperparameters.

Learning Causal Effects via WERM: Learning Guarantee

Theorem 2 (Learning guarantee). *Let* $h^* \equiv \arg\min_{h \in \mathcal{H}} \mathcal{R}^{\mathcal{W}^*}(h)$, *and* $(\mathcal{W}_m, h_m) \equiv \arg\min_{\mathcal{W} \in \mathcal{H}_\mathcal{W}, h \in \mathcal{H}} \mathcal{L}(\mathcal{W}, h)$, *where* $\mathcal{H}_\mathcal{W}$ *is the model hypotheses class for* $\mathcal{W}$. *Suppose* $\mathcal{H}_\mathcal{W}$ *is correctly specified such that* $\mathcal{W}^* \in \mathcal{H}_\mathcal{W}$. *Then,* h_m *converges to* h^* *with a rate of* $O_p(m^{-1/4})$. *Specifically,* $\mathcal{R}^{\mathcal{W}^*}(h_m) - \mathcal{R}^{\mathcal{W}^*}(h^*) \leq O_p(m^{-1/4})$.

In words, the theorem ascertains that the hypothesis h_m that minimizes the objective function $\mathcal{L}(\mathcal{W}, h)$ in Eq. (6) converges to the hypothesis h^* that minimizes the target weighted risk $\mathcal{R}^{\mathcal{W}^*}(h)$ in the limit of infinite samples.

With this knowledge we can piece together an algorithm to learn causal effects on limited data.

Learning Causal Effects via WERM: Algorithm

Algorithm 2: WERM-ID-R($\mathcal{D}, G, \mathbf{x}, y$)

Output: An estimate of $\mathbb{E}[Y|do(\mathbf{x})]$ from data sample $\mathcal{D}$

1 Run wID $(\mathbf{x}, y, G, P)$ and derive $(\mathcal{W}^*, \mathbf{R})$ such that $P(y|do(\mathbf{x})) = P^{\mathcal{W}^*}(y|\mathbf{r})$.

2 Evaluate $\widehat{\mathcal{W}^*}$ from $\mathcal{D}$.

3 Learn $\mathcal{W} \equiv \arg\min_{\mathcal{W}' \in \mathcal{H}_{\mathcal{W}}} \mathcal{L}_{\mathcal{W}}(\mathcal{W}', \lambda_{\mathcal{W}}; \widehat{\mathcal{W}^*})$.

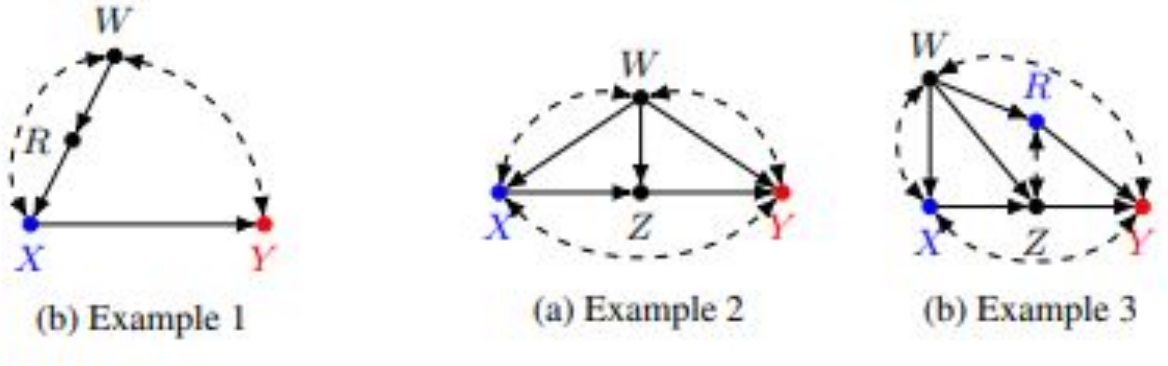
4 Learn $h \equiv \arg\min_{h' \in \mathcal{H}} \mathcal{L}_h(h', \mathcal{W}, \lambda_h)$.

return $\widehat{\mathbb{E}}[Y|do(\mathbf{x})] \equiv h(\mathbf{r})$

- This algorithm is a combination of those previously discussed
- Time complexity is dependant on sub-algorithms.
- This paper provides a heuristic procedure that works well in practice to learn W and h through minimizing an objective function.

Theorem 3 (Time complexity of Algo. 2). *Let $m \equiv |\mathcal{D}|$ and $n \equiv |\mathbf{V}|$. Assume all weights satisfy $0 < \mathcal{W} < c$ for some constant $c > 0$. Let $T_1(m)$ denote the time complexity of estimating $\widehat{P}(v_i|\cdot)$ from sample $\mathcal{D} \sim P(\mathbf{v})$ for $V_i \in \mathbf{V}$. Let K denote the number of C-components in $G_{\mathbf{D}}$ (in Algo. 1). Let $T_2(m)$ denote the time complexity of minimizing $\mathcal{L}_{\mathcal{W}}$ and $\mathcal{L}_h$. Then, Algo. 2 runs in $O(poly(n) + nK(mc + nT_1(m)) + T_2(m))$ time, where $O(poly(n))$ is for running Algo. 1, $O(nK(mc + nT_1(m)))$ for evaluating $\widehat{\mathcal{W}^*}$.*

Experimental Setup



In order to evaluate the performance of WERM for learning causal effects, different SCMs are evaluated with different queries.

Example 1: X is CD4 cell counts, Y is progression of HIV, R denotes a treatment that affects CD4 counts.

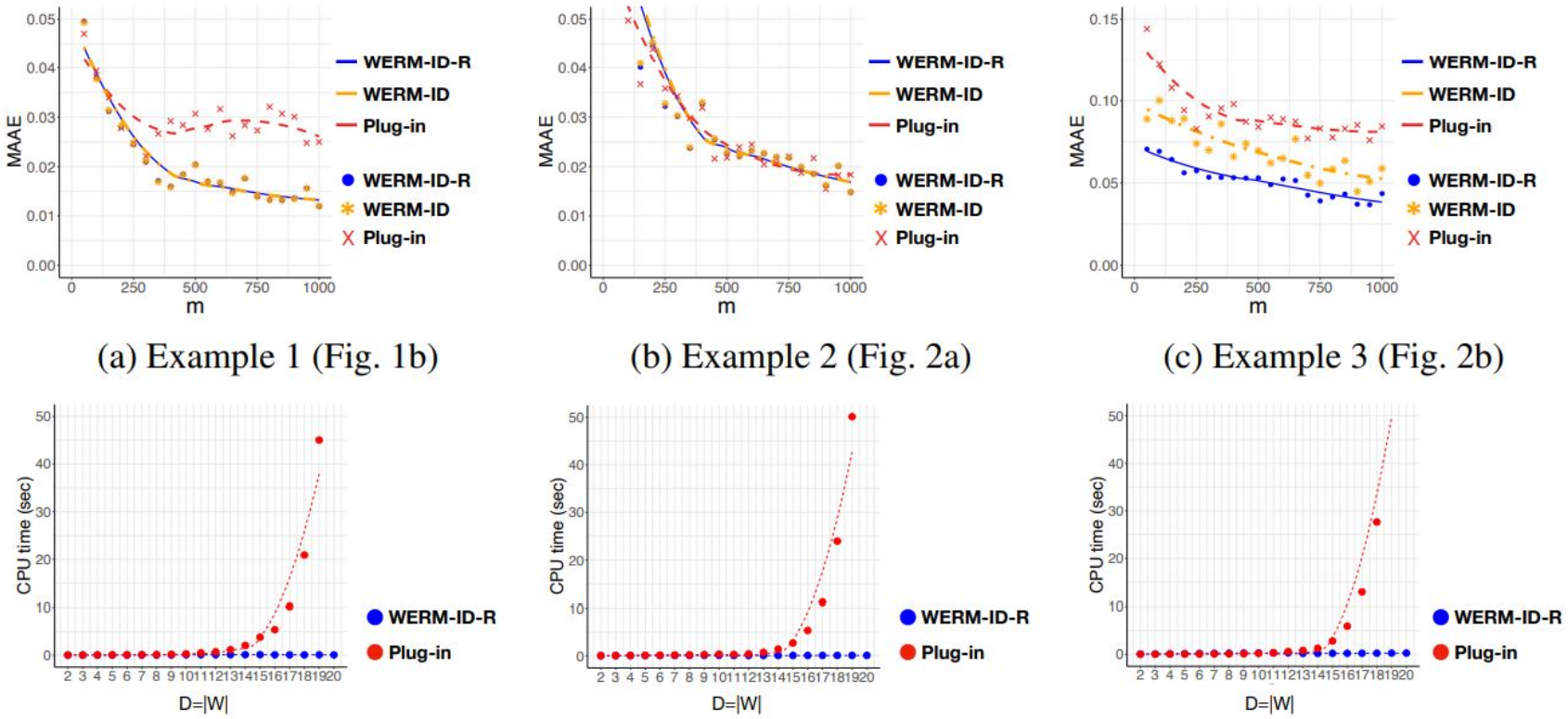
The method is compared to plug-in estimator (parametric estimation of conditionals).

Our estimator is denoted WERM-ID-R and a simpler variant WERM-ID is also used in the study, which directly minimizes WER R after evaluating W* from D

Experimental Setup: Accuracy Measure

Accuracy Measure. Given a data set $\mathcal{D}$ with m samples, let $\hat{\mu}_{\text{IDR}}(\mathbf{x})$, $\hat{\mu}_{\text{ID}}(\mathbf{x})$, and $\hat{\mu}_{\text{plug}}(\mathbf{x})$ be the estimated $\mathbb{E}[Y|do(\mathbf{x})]$ using the WERM-ID-R, WERM-ID, and Plug-in estimators. For each $\hat{\mu} \in \{\hat{\mu}_{\text{IDR}}, \hat{\mu}_{\text{ID}}, \hat{\mu}_{\text{plug}}\}$, we compute the average absolute error (AAE) as $|\mu(\mathbf{x}) - \hat{\mu}(\mathbf{x})|$ averaged over $\mathbf{x}$. We generate 100 datasets for each sample size m. We call the median of the 100 AAEs the *median average absolute error*, or MAAE, and its plot vs. the sample size m, the *MAAE plot*.

Ground truths are estimated by generating 10^7 samples Dint from the model induced by the intervention P(Y|do(x)) and computing the mean of Y in Dint.



(a) Example 1 (Fig. 1b) (b) Example 2 (Fig. 2a) (c) Example 3 (Fig. 2b)

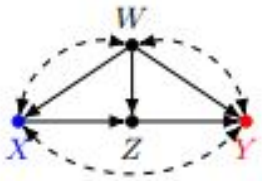
(d) Example 1 (Fig. 1b) (e) Example 2 (Fig. 2a) (f) Example 3 (Fig. 2b)

Figure 3: **(Top)** MAAE plots comparing proposed WERM based estimators (WERM-ID and WERM-ID-R) with Plug-in. **(Bottom)** Plots comparing the running time of WERM-ID-R versus Plug-in.

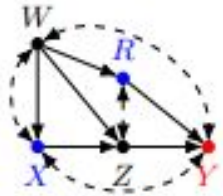
Discussion

The results show that the accuracy of WERM-based estimators are never worse, and mostly superior (both computationally and accuracy-wise) to plug-in based methods.

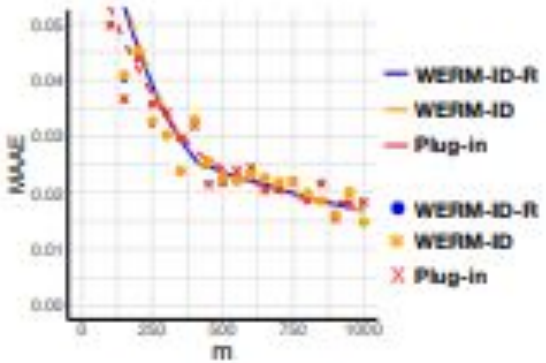
In example 2 WERM methods are on par with plug-in methods. In example 3 WERM methods greatly outperform plug-in methods. Propose a reason for this discrepancy ie. graph-based data-based.



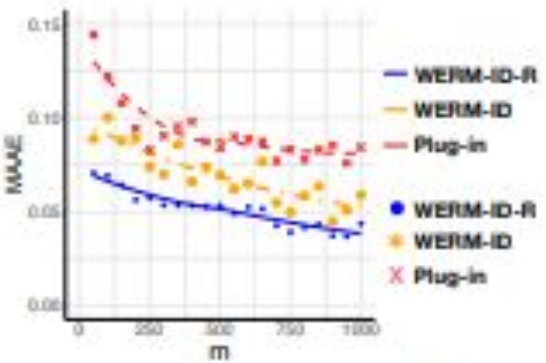
(a) Example 2



(b) Example 3



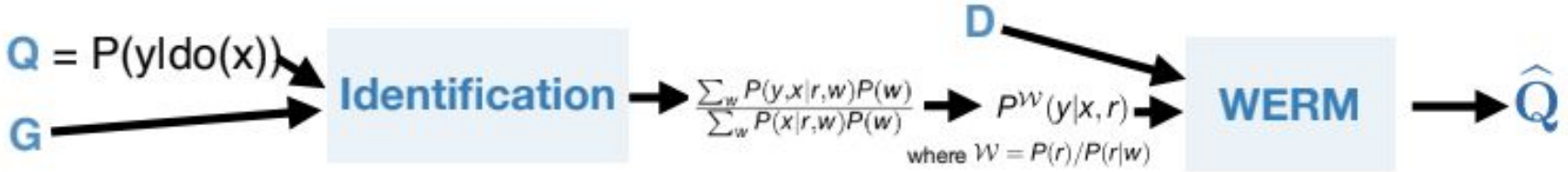
(b) Example 2 (Fig. 2a)



(c) Example 3 (Fig. 2b)

Discussion

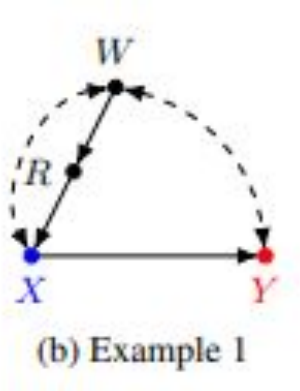
- In this paper, the gap is filled from causal identification to causal estimation using a powerful learning framework that draws from causal identification theory and empirical risk minimization methods.
- This paper provides a learning objective based on the WERM theory and a practical learning algorithm that can estimate causal effects from finite samples.
- The effectiveness of the proposed methods were corroborated with experimental studies.
- Learning guarantees and generalization bounds are provided by this method.



Conclusion

This work brings together two prominent fields in machine learning: ERM and causal inference, where the former is suitable to estimate high-dimensional functionals, and the latter is useful to determine which functional should be estimated that attains causal semantics.

Homework: Weighted Distribution Identification Example



(b) Example 1

Given the graph G find a query form that can be optimized using the WERM algorithm.

- Will be of the form W*P(V)
- C components?
- S1 = {W,X,Y}, S2 = {R}
- Derive Q[S1] = W*P(s1|r), what is W?
- W = P(r)/P(r|w)
- Derive Q[S2]
- …