Learning Causal Effects via Weighted Empirical Risk Minimization

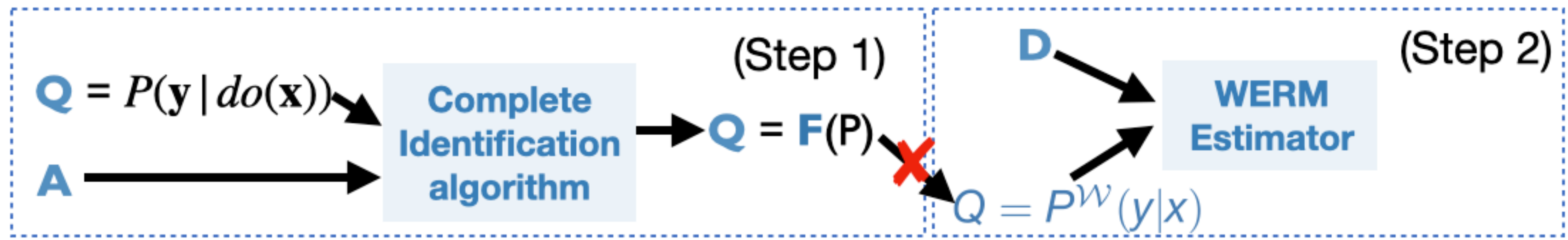
Yonghan Jung — Purdue University

Jin Tian — Iowa State University

Elias Bareinboim — Columbia University in the City of New York

Neural Information Processing Systems

Motivation — Gap between causal inference and machine learning



- (Step 1) There are complete identification algorithm for representing for determining whether a causal query Q can be represented as a functional of P (i.e., Q=F(P)) from a given causal graph.
- (Step 2) Estimation has been mainly done on backdoor/ignorability' assumption. For general identifiable estimands, it's not known (neither obvious) how to estimate the causal effect sample & time-efficiently.
- (Step 2) Weighted ERM (WERM) provides sample and time efficient estimators when the estimand is given weighted distributions $P^{\mathscr{W}}(\mathbf{v}) \equiv \mathscr{W}(\mathbf{v})P(\mathbf{v})$ (e.g., Causal effect $Q = P(y \mid do(x))$ when the back-door holds).
- (Step 1) When Q=P(y|do(x)) is identifiable, but the estimand is not of a weighted form, it's not clear how to use WERM based-estimators.

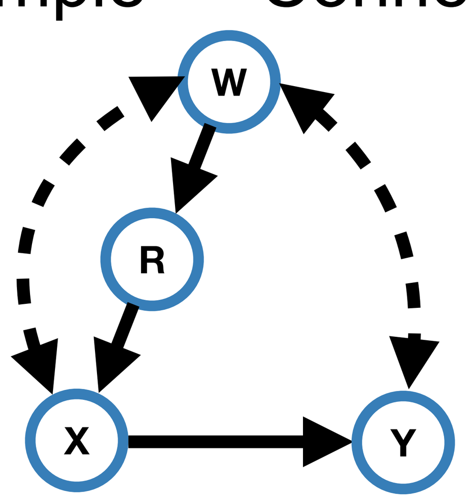
Example — WERM when Back-door holds

Observation $P(z,x,y) = P(z)P(x \mid z)P(y \mid x,z)$

Interventional $P^{\mathscr{W}}(z,x,y) = P(z)P(x)P(y \mid x,z)$

- When the back-door criterion holds, then , $Q = P(y \mid do(x)) = P^{\mathscr{W}}(y \mid x)$, where $P^{\mathscr{W}}(z,x,y) \equiv \mathscr{W}P(z,x,y)$ for $\mathscr{W} = P(x)/P(x \mid z)$, a weighted distribution.
- Then, the weighted-ERM estimators (e.g., Counterfactual risk minimization [1], Re-weighted risk minimization [2]) are available.

Example — Connecting Identifiability Theory & WERM



[Fig. 1] Example graph

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

[Eq. 1]

- Consider Fig. 1. The causal effect is identifiable, and the causal effect is given as in Eq. 1. However, the estimand $P(y \mid do(x))$ is not in a form of an WERM estimator.
- Still, the quantity Q=P(y|do(x)) can be represented as a conditional distribution of the weighted distribution. Specifically, for $\mathscr{W} = P(r)/P(r \mid w)$, the causal effect is written as

$$P(y \mid do(x)) = P^{\mathscr{W}}(y \mid x, r).$$

- **Question**: Can we use weighted ERM based estimator for general identifiable estimands?

wID — Representing causal functional into weighted distribution

Theorem 1: Soundness and completeness of wID (in Algo. 1)

A causal effect $P(\mathbf{y} \mid do(\mathbf{x}))$ is identifiable if and only if wID$(\mathbf{x}, \mathbf{y}, G, P)$ (Algo. 1) returns $P^{\mathscr{W}}(\mathbf{y} \mid \mathbf{r})$ such that $P(\mathbf{y} \mid do(\mathbf{x})) = P^{\mathscr{W}}(\mathbf{y} \mid \mathbf{r})$

- Thm 1. states that *any identifiable causal functional* could be represented as a weighted distribution, a proper input for WERM.

Learning causal effect using WERM

- The ***weighted risk*** $R^{\mathscr{W}^*}(h) \equiv \mathbb{E}_{P^{\mathscr{W}^*}}[\ell(h(\mathbf{R}), Y)] = \mathbb{E}_P[\mathscr{W}^*(\mathbf{V})\ell(h(\mathbf{R}), Y)]$, for the loss function $\ell(h(\mathbf{R}), Y)$ of the hypothesis $h(\cdot)$ and the weight $\mathscr{W}^*$ (where $\mathscr{W}^*$ with * mark is the weight s.t. $P(y \mid do(\mathbf{x})) = P^{\mathscr{W}^*}(y \mid \mathbf{r})$). The ***weighted empirical risk*** is given as $\widehat{R}^{\mathscr{W}^*}(h) \equiv \frac{1}{N}\sum_{i=1}^{N} \mathscr{W}^*(\mathbf{V}_{(i)})\ell(h(\mathbf{R}_{(i)}), Y_{(i)})$.

Proposition 1: Generalization bound for weighted risk [3]

Let p denote the Pollard's pseudo-dimension of loss function $\ell_h \equiv \ell(h(\mathbf{v}), \mathbf{y})$ and $\widehat{P}$ denote the empirical distribution of P. Then, for any $\delta \in (0,1)$, with probability at least $(1-\delta)$, the following holds:

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

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

- Based on the generalization bound in Prop. 1, the ***learning objective*** based on structural risk minimization principle is:

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

Theorem 2: Learning guarantee

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

- That is, the hypothesis h_m that minimizes the objective function $\mathscr{L}(\mathscr{W}, h)$ converges to h^*, the target minimizer.
- Then, Algo. 2 provides the end-to-end procedure to causal effect estimation by combining Algo. 1 (wID) and the learning bounds (and learning guarantees) of the learning objective $\mathscr{L}(\mathscr{W}, h)$.

Algo. 2 WERM-ID-R$(\mathscr{D}, G, \mathbf{x}, y)$

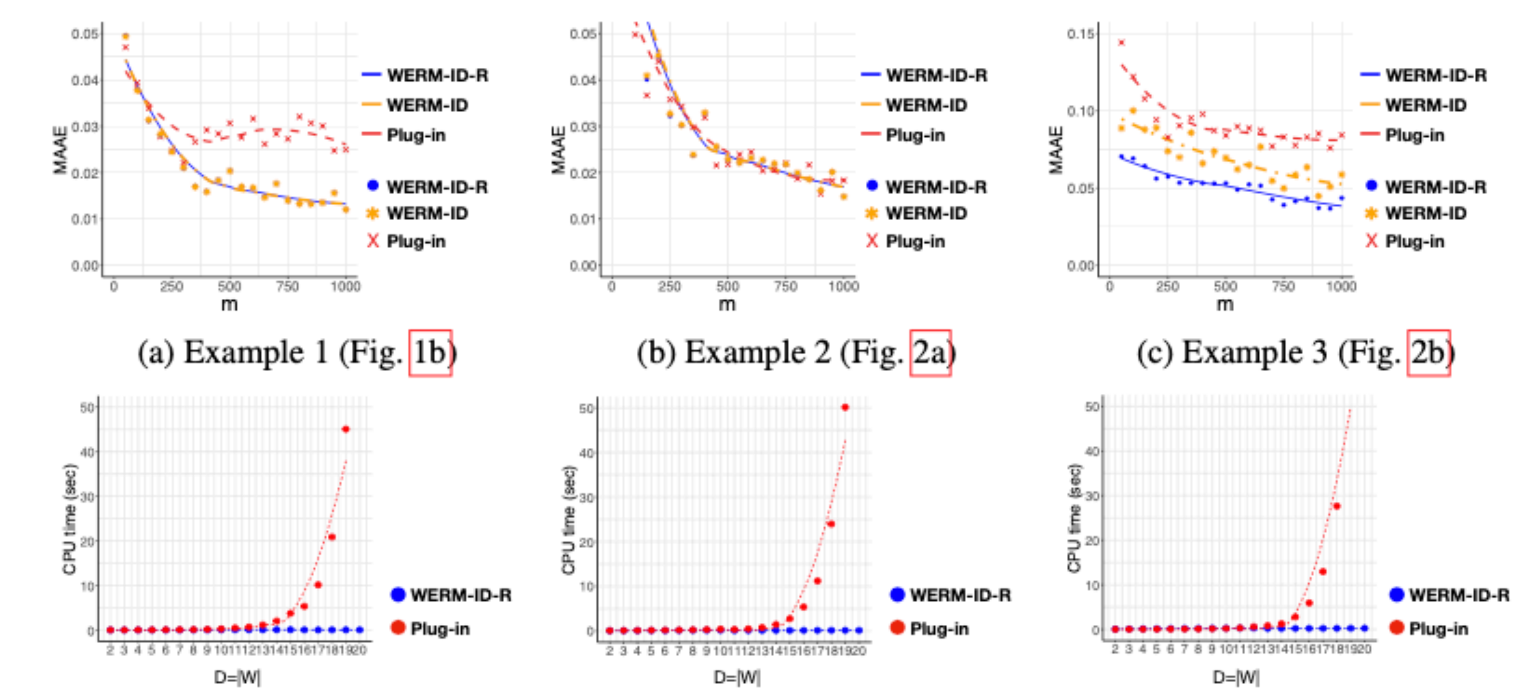
1. Run wID$(\mathbf{x}, y, G, P)$ and derive $(\mathscr{W}^*, \mathbf{R})$ s.t. $P(y \mid do(\mathbf{x})) = P^{\mathscr{W}^*}(y \mid \mathbf{r})$.
2. Evaluate $\widehat{\mathscr{W}^*}$ from samples $\mathscr{D}$.
3. Learn $\mathscr{W} \equiv \arg\min_{\mathscr{W}' \in \mathscr{H}_{\mathscr{W}}} \mathscr{L}_{\mathscr{W}}(\mathscr{W}', \lambda_{\mathscr{W}}, \widehat{\mathscr{W}^*})$.
4. Learn $h \equiv \arg\min_{h' \in \mathscr{H}} \mathscr{L}_h(h', \mathscr{W}, \lambda_h)$.

- Algo. 2 is time-efficient (i.e., Algo. 2 runs in polynomial w.r.t. sample sizes and the number of variables in G).

(Informal) **Theorem 3:** Time complexity of Algo. 2.

Let $n \equiv |\mathbf{V}|$ and $m \equiv |\mathscr{D}|$. Let $T_1(m)$ denote the time complexity for estimating conditional distribution; $T_2(m)$ denote the time complexity for optimizing $\mathscr{L}_h$ and $\mathscr{L}_{\mathscr{W}}$. Then, Algo. 2 runs in $O\left(\text{poly}(n) + n(m + nT_1(m)) + T_2(m)\right)$.

Simulation



(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)

Simulation results — **(Top)** Comparing accuracies of the proposed estimators with plug-in estimator. (**Bottom**) Comparing the running time between the proposed vs. plug-in estimator.

- The simulation results for various causal instances implies that the proposed estimator is sample and time-efficient compared to the plug-in estimator, the only viable for arbitrary causal functional.

Summary & Contribution

- We develop a sound and complete algorithm (Algo. 1) that generates any identifiable causal functionals as weighted distributions, amenable to WERM method.
- We formulate the causal estimation problem as an WERM optimization. We introduce a learning objective, inspired by generalization error bound, and provide theoretical learning guarantee to the solution (Thm. 2).
- We develop a practical and systematic algorithm (Algo. 2, Thm. 3) for learning target causal effects from finite samples given a causal graph, based on the proposed framework. The practical effectiveness of this approach is demonstrated through simulated studies.

References

[1] Swaminathan, Adith, and Thorsten Joachims. "Counterfactual risk minimization: Learning from logged bandit feedback." *International Conference on Machine Learning.* 2015.

[2] Johansson, Fredrik D., et al. "Learning weighted representations for generalization across designs." *arXiv preprint arXiv:1802.08598* (2018).

[3] Cortes, Corinna, Yishay Mansour, and Mehryar Mohri. "Learning bounds for importance weighting." *Advances in neural information processing systems.* 2010.