Estimating Functionals of the Joint Distribution of Potential Outcomes with Optimal Transport

Daniel Ober-Reynolds

University of California, Los Angeles

doberreynolds@gmail.com

February, 2024

Introduction

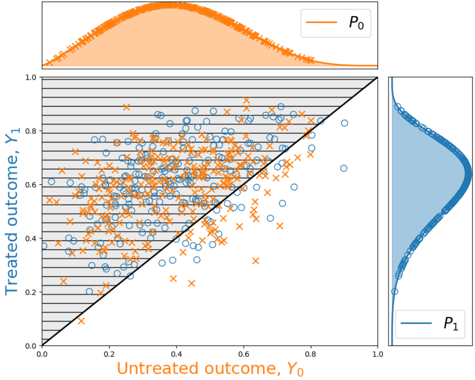
The fundamental problem of causal inference

It is impossible to observe the [treated outcome] and [untreated outcome] on the same unit and, therefore, it is impossible to observe the effect...

(Holland, 1986)

- Parameters of the joint distribution of potential outcomes are not point identified.
- **This paper**
  - shows optimal transport characterizes sharp bounds,
  - accomodates noncompliance through a standard IV model, and
  - provides simple, computationally convenient estimators.

The fundamental problem of causal inference

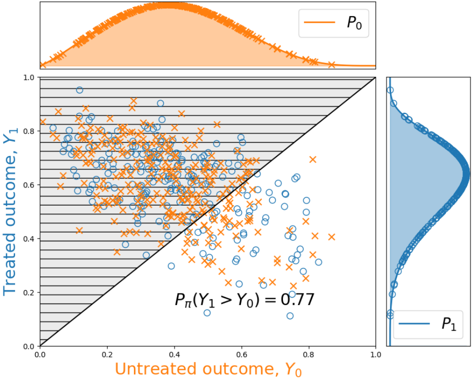


- Never observe (Y_1, Y_0), because each unit is treated $(D = 1)$ or untreated $(D = 0)$:

$$\text{Observed outcome } Y = DY_1 + (1 - D)Y_0$$

- The marginal distributions P_1 and P_0 are identified - but have less information.
- For example, what share of units benefit from treatment?

Example 1: the share benefiting from treatment



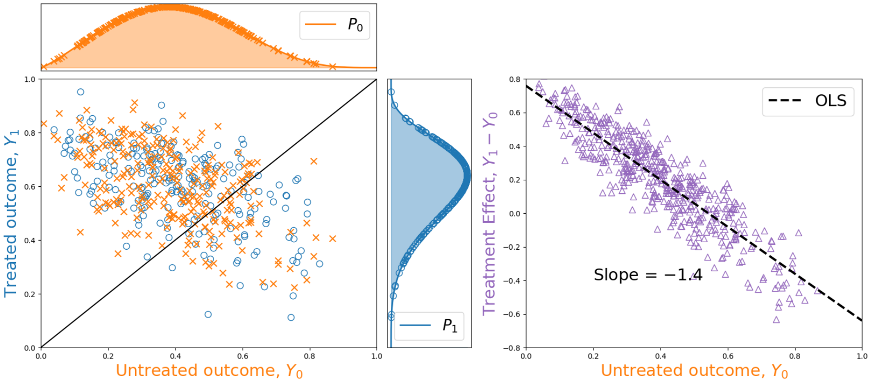
- Many joint distributions π share marginal distributions P_1, P_0:

$$\Pi(P_1, P_0) = \{\pi \;:\; \pi_1 = P_1,\; \pi_0 = P_0\}$$

- Optimizing $P(Y_1 > Y_0)$ over $\Pi(P_1, P_0)$ implies bounds:

$$\min_{\pi \in \Pi(P_1, P_0)} P_\pi(Y_1 > Y_0) \qquad\qquad \max_{\pi \in \Pi(P_1, P_0)} P_\pi(Y_1 > Y_0)$$

Example 2: who sees larger benefits from treatment?



- Do those with smaller Y_0 see larger $Y_1 - Y_0$?

$$\text{OLS slope} = \frac{\mathrm{Cov}(Y_1 - Y_0, Y_0)}{\mathrm{Var}(Y_0)} = \frac{E[(Y_1 - Y_0)Y_0] - (E[Y_1] - E[Y_0])E[Y_0]}{E[Y_0^2] - (E[Y_0])^2}$$

- Optimizing $E[(Y_1 - Y_0)Y_0]$ over $\Pi(P_1, P_0)$ implies bounds on OLS slope:

$$\min_{\pi \in \Pi(P_1, P_0)} E_\pi[(Y_1 - Y_0)Y_0] \qquad \max_{\pi \in \Pi(P_1, P_0)} E_\pi[(Y_1 - Y_0)Y_0]$$

This paper

- Parameter of interest:

$$\gamma = g(\theta, \eta) \in \mathbb{R},$$

where $\theta = E[c(Y_1, Y_0)] \in \mathbb{R}$ and $\eta = (E[\eta_1(Y_1)], E[\eta_0(Y_0)]) \in \mathbb{R}^{K_1+K_0}$.

  - Example 3: $\gamma = \text{Var}(Y_1 - Y_0) = E[(Y_1 - Y_0)^2] - (E[Y_1] - E[Y_0])^2$

- Characterize sharp identified set with optimal transport:

$$OT_c(P_1, P_0) = \min_{\pi \in \Pi(P_1, P_0)} E_\pi[c(Y_1, Y_0)]$$

- Propose and study sample analogue estimators of the bounds.

- Empirical application: who sees larger benefits from the NSW job training?

Related literature

- Joint distribution of potential outcomes
  - CDF or quantiles of $Y_1 - Y_0$: Manski (1997), Heckman et al. (1997), Firpo (2007), Fan and Park (2010), Fan and Park (2012), Firpo and Ridder (2019), Callaway (2021), Frandsen and Lefgren (2021).
  - General methods: Russell (2021) Fan et al. (2023), Ji et al. (2023), **this paper**.

- Optimal transport in econometrics
  - Partial identification: Galichon and Henry (2011), Ekeland et al. (2010)
  - Causal inference: Dunipace (2021), Gunsilius and Xu (2021), Torous et al. (2021)
  - Joint distribution of (Y_1, Y_0): Ji et al. (2023), **this paper**.

$\Rightarrow$ This paper contributes identification and estimators that

i. cover a large class of parameters while remaining tractable,
ii. allow for simple bootstrap inference, and
iii. accomodate noncompliance through a standard IV model.

Overview

Overview

1. Setting and parameter class
2. Identification
3. Estimators
4. Simulations
5. Application

Setting

- For this talk, focus on unconfoundedness.

Assumption 1 (Setting, simplified) $\{Y_i, D_i, X_i\}_{i=1}^n$ is an i.i.d. sample with

$$Y \in \mathcal{Y} \subseteq \mathbb{R}, \qquad D \in \{0,1\}, \qquad X \in \mathcal{X} = \{x_1, \ldots, x_M\}$$

generated from a distribution satisfying

(i) Potential outcomes: $Y = DY_1 + (1-D)Y_0$

(ii) Unconfoundedness: $(Y_1, Y_0) \perp D \mid X$

(iii) $P(D = d, X = x) > 0$ for each (d, x)

- In the paper, binary IV satisfying monotonicity condition (Imbens and Angrist, 1994).

Setting w/IV

Parameter class

- Parameter of interest:

$$\gamma = g(\theta, \eta) \in \mathbb{R}$$

where $\theta = E[c(Y_1, Y_0)] \in \mathbb{R}$ and $\eta = (E[\eta_1(Y_1)], E[\eta_0(Y_0)]) \in \mathbb{R}^{K_1+K_0}$

Assumption 2 (Cost function) Either

(i) $c(y_1, y_0)$ is Lipschitz continuous and $\mathcal{Y}$ is compact, or

(ii) $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$ and the CDFs $F_{d|x}(y) = P(Y_d \leq y \mid X = x)$ are continuous.

Remark: If $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$ but $F_{d|x}(\cdot)$ are not continuous, inference remains valid for an outer identified set.

Parameter class

- Parameter of interest:

$$\gamma = g(\theta, \eta) \in \mathbb{R}$$

where $\theta = E[c(Y_1, Y_0)] \in \mathbb{R}$ and $\eta = (E[\eta_1(Y_1)], E[\eta_0(Y_0)]) \in \mathbb{R}^{K_1+K_0}$.

Assumption 3 (Function of moments, simplified)

(i) $\eta_1(Y)$ and $\eta_0(Y)$ have finite second moments,

(ii) $g(\cdot, \cdot)$ is continuously differentiable, and

(iii) $g(\cdot, \eta)$ is monotonic.

Remark: Assumption 3 (iii) is relaxed in the paper.

Full assumption 3

Parameter class: motivating examples

- Share benefiting: $P(Y_1 > Y_0)$
  - Allcott et al. (2020): deactivating Facebook affects subjective well-being.
- Share benefiting above cost: $P(Y_1 - Y_0 > \text{cost})$
  - Friebel et al. (2023): employee referral programs increase grocery store profit.
- Who benefits more from treatment? $\text{Cov}(Y_1 - Y_0, Y_0)/\text{Var}(Y_0)$
  - Application: NSW job experience increases post-training annual income.
- Expected percent change: $E\left[\frac{Y_1 - Y_0}{Y_0}\right]$
  - This parameter is often approximated with $E[\log(Y_1) - \log(Y_0)]$.
- Quantiles of $Y_1 - Y_0$
  - Median is more representative than mean when distribution is skewed.

Overview

Optimal transport

$$OT_c(P_1, P_0) = \min_{\pi \in \Pi(P_1, P_0)} E_\pi[c(Y_1, Y_0)]$$

- Choose a joint distribution with given marginals to minimize costs.
  - Feasible set: $\Pi(P_1, P_0) = \{\pi : \pi_1 = P_1,\ \pi_0 = P_0\}$
  - Cost function: $c(y_1, y_0)$
- Often interpreted in other contexts, but here intended literally.
- Attained under mild conditions.

Identification without covariates

- $\{Y_i, D_i\}_{i=1}^n$ identifies marginal distributions P_1 and P_0.
- Identified set for $P_{1,0}$ is set of joint distributions with marginals P_1, P_0:

$$\Pi(P_1, P_0) = \{\pi : \pi_1 = P_1, \pi_0 = P_0\}$$

- Bounds on $\theta = E_{P_{1,0}}[c(Y_1, Y_0)]$ for continuous c:

$$\theta^L = \min_{\pi \in \Pi(P_1,P_0)} E_\pi[c(Y_1, Y_0)], \qquad \theta^H = \max_{\pi \in \Pi(P_1,P_0)} E_\pi[c(Y_1, Y_0)]$$
$$= OT_c(P_1, P_0), \qquad = -OT_{-c}(P_1, P_0)$$

- Bounds on $\gamma = g(\theta, \eta)$:

$$\gamma^L = \min_{t \in [\theta^L, \theta^H]} g(t, \eta), \qquad \gamma^H = \max_{t \in [\theta^L, \theta^H]} g(t, \eta)$$

CDF?

Identification with covariates

- $\{Y_i, D_i, X_i\}_{i=1}^n$ identifies marginal *conditional* distributions $P_{1|x}$ and $P_{0|x}$.

$$Y_d \mid X = x \sim P_{d|x}$$

- Identified set for $P_{1,0|x}$ is set of joint distributions with marginals $P_{1|x}$, $P_{0|x}$:

$$\Pi(P_{1|x}, P_{0|x}) = \left\{\pi_{1,0|x} : \pi_{1|x} = P_{1|x},\ \pi_{0|x} = P_{0|x}\right\}$$

- Bounds on $\theta = E_{P_{1,0}}[c(Y_1, Y_0)] = E[\overbrace{E_{P_{1,0|X}}[c(Y_1, Y_0) \mid X]}^{:=\theta_X}]$ for continuous c:

$$\theta_x^L = OT_c(P_{1|x}, P_{0|x}), \qquad \theta_x^H = -OT_{-c}(P_{1|x}, P_{0|x})$$
$$\theta^L = E[\theta_X^L], \qquad \theta^H = E[\theta_X^H]$$

- Bounds on $\gamma = g(\theta, \eta)$:

$$\gamma^L = \min_{t \in [\theta^L, \theta^H]} g(t, \eta), \qquad \gamma^H = \max_{t \in [\theta^L, \theta^H]} g(t, \eta)$$

CDF?

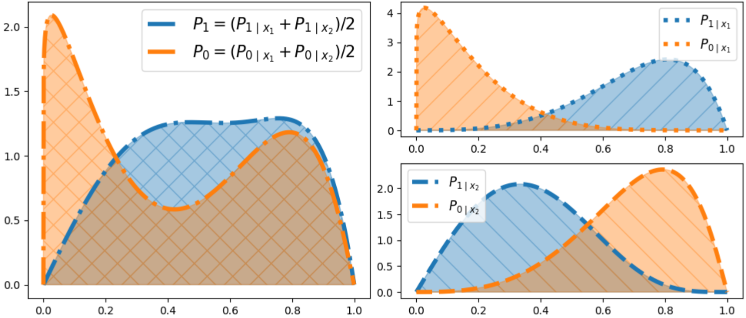
Covariates tighten identified bounds

- Covariates tighten bounds,

$$OT_c(P_1, P_0) \leq \theta^L, \qquad \theta^H \leq -OT_{-c}(P_1, P_0).$$

  - Why? The optimization has additional constraints.
  - $\theta^L = E[OT_c(P_{1|X}, P_{0|X})]$ looks for $\pi \in \Pi(P_1, P_0)$ also matching $(P_{1|x}, P_{0|x})$.

- Bounds on $P(Y_1 > Y_0)$: not sharp $[0.25, 1]$, sharp: $[0.44, 0.68]$.



Theorem: identification

- For continuous c,

$$\begin{array}{lll} \text{Bounds on } \theta_x: & \theta_x^L = OT_c(P_{1|x}, P_{0|x}), & \theta_x^H = -OT_{-c}(P_{1|x}, P_{0|x}) \\ \text{Bounds on } \theta: & \theta^L = E[\theta_X^L] & \theta^H = E[\theta_X^H] \\ \text{Bounds on } \gamma: & \gamma^L = \min_{t \in [\theta^L, \theta^H]} g(t, \eta), & \gamma^H = \max_{t \in [\theta^L, \theta^H]} g(t, \eta) \end{array}$$

Theorem (identification)

Suppose assumptions 1, 2, and 3 are satisfied. Then the sharp identified set for $\gamma = g(\theta, \eta)$ is $[\gamma^L, \gamma^H]$.

CDF? IV Aside Quantile details

Overview

Optimal transport

$$OT_c(P_1, P_0) = \underbrace{\min_{\pi \in \Pi(P_1,P_0)} E_\pi[c(Y_1, Y_0)]}_{\text{Primal Problem}} \underset{\substack{\uparrow \\ \text{Strong} \\ \text{Duality}}}{=} \underbrace{\max_{(\varphi,\psi)\in\Phi_c} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)]}_{\text{Dual Problem}}$$

$$\Pi(P_1, P_0) = \{\pi : \pi_1 = P_1,\ \pi_0 = P_0\} \qquad \Phi_c = \{(\varphi, \psi) : \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)\}$$

- The primal problem is used in identification.
- The dual problem is used for estimation.
- Strong duality holds under the cost function assumptions. Each problem is attained, too.

Estimators: recall identification

- Distributions of $Y_d \mid X = x \sim P_{d|x}$:

$$E_{P_{d|x}}[f(Y_d)] = \frac{E[f(Y)\mathbb{1}\{D = d, X = x\}]}{P(D = d, X = x)}$$

- Using strong duality,

$$OT_c(P_{1|x}, P_{0|x}) = \max_{(\varphi,\psi)\in\Phi_c} E_{P_{1|x}}[\varphi(Y_1)] + E_{P_{0|x}}[\psi(Y_0)].$$

- The identified set for γ is $[\gamma^L, \gamma^H]$, where for c continuous,

$$\begin{aligned} \theta_x^L &= OT_c(P_{1|x}, P_{0|x}), & \theta_x^H &= -OT_{-c}(P_{1|x}, P_{0|x}) \\ \theta^L &= E[\theta_X^L], & \theta^H &= E[\theta_X^H] \\ \gamma^L &= \min_{t\in[\theta^L,\theta^H]} g(t,\eta), & \gamma^H &= \max_{t\in[\theta^L,\theta^H]} g(t,\eta) \end{aligned}$$

CDF?

Estimators: sample analogues

- Estimate $P_{d|x}$ with sample analogues $\hat{P}_{d|x}$:

$$E_{\hat{P}_{d|x}}[f(Y_d)] = \frac{\frac{1}{n}\sum_{i=1}^{n} f(Y_i)\mathbb{1}\{D_i = d, X_i = x\}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{D_i = d, X_i = x\}}$$

- Using strong duality,

$$OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}) = \max_{(\varphi,\psi)\in\Phi_c} E_{\hat{P}_{1|x}}[\varphi(Y_1)] + E_{\hat{P}_{0|x}}[\psi(Y_0)].$$

- Estimate the endpoints of $[\gamma^L, \gamma^H]$ with plug-in estimators. For c continuous,

$$\hat{\theta}_x^L = OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}), \qquad \hat{\theta}_x^H = -OT_{-c}(\hat{P}_{1|x}, \hat{P}_{0|x})$$

$$\hat{\theta}^L = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{X_i}^L, \qquad \hat{\theta}^H = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{X_i}^H$$

$$\hat{\gamma}^L = \min_{t\in[\hat{\theta}^L, \hat{\theta}^H]} g(t, \hat{\eta}), \qquad \hat{\gamma}^H = \max_{t\in[\hat{\theta}^L, \hat{\theta}^H]} g(t, \hat{\eta})$$

CDF?

Estimators: computing $OT_c(\hat{P}_{1|x}, \hat{P}_{0|x})$

$$OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}) = \max_{(\varphi,\psi)\in\Phi_c} E_{\hat{P}_{1|x}}[\varphi(Y_1)] + E_{\hat{P}_{0|x}}[\psi(Y_0)].$$

- To evaluate $E_{\hat{P}_{d|x}}[f(Y_d)]$ for any function f, only the values $f_i = f(Y_i)$ matter.

$$E_{\hat{P}_{d|x}}[f(Y_d)] = \sum_{i=1}^{n} \omega_{d,x,i} \times f_i, \qquad \omega_{d,x,i} = \frac{\mathbb{1}\{D_i = d, X_i = x\}/n}{\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{D_j = d, X_j = x\}}.$$

- Computating $OT_c(\hat{P}_{1|x}, \hat{P}_{0|x})$ is straightforward linear programming:

$$OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}) = \max_{\{\varphi_i,\psi_i\}_{i=1}^n} \sum_{i=1}^{n} \omega_{1,x,i} \times \varphi_i + \sum_{i=1}^{n} \omega_{0,x,i} \times \psi_i$$
$$\text{s.t. } \varphi_i + \psi_j \leq c(Y_i, Y_j) \text{ for all } 1 \leq i,j \leq n,$$

- Dimension is reduced by ignoring φ_i, ψ_i, and constraints where $\omega_{d,x,i} = 0$.

Convergence in distribution: theorem

- Let P be the distribution of an observation, and $\mathbb{P}_n$ the empirical distribution.

$$(\hat{\gamma}^L, \hat{\gamma}^H) = T(\mathbb{P}_n), \qquad\qquad (\gamma^L, \gamma^H) = T(P)$$

Theorem (Weak convergence)

Suppose assumptions 1, 2, and 3 hold. Then

$$\sqrt{n}((\hat{\gamma}^L, \hat{\gamma}^H) - (\gamma^L, \gamma^H)) \xrightarrow{L} T'_P(\mathbb{G})$$

where $\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ *and* $T'_P(\cdot)$ *is the Hadamard directional derivative of* $T(\cdot)$ *at* P.

$T(\cdot)$ details | Proof sketch

Inference: bootstrap

- Estimating the asymptotic distribution is necessary for inference.
- The bootstrap provides an attractive procedure.
  - Bootstrap draw: $\{Y_i^*, D_i^*, X_i^*\}_{i=1}^n$
  - Bootstrap empirical distribution: $\mathbb{P}_n^*$
- Compute $T(\mathbb{P}_n^*)$ the same way as $T(\mathbb{P}_n)$: let $\omega_{d,x,i}^* = \frac{\mathbb{1}\{D_i^*=d, X_i^*=x\}/n}{\frac{1}{n}\sum_{j=1}^n \mathbb{1}\{D_j^*=d, X_j^*=x\}}$,

$$OT_c(\hat{P}_{1|x}^*, \hat{P}_{0|x}^*) = \max_{\{\varphi_i, \psi_i\}_{i=1}^n} \sum_{i=1}^n \omega_{1,x,i}^* \varphi_i + \sum_{i=1}^n \omega_{0,x,i}^* \psi_i$$

$$\text{s.t. } \varphi_i + \psi_j \leq c(Y_i, Y_j) \text{ for all } 1 \leq i, j \leq n$$

Inference: bootstrap

Assumption 4 (Unique solutions, informal) For each instance of optimal transport in $T(P)$, the solution to the dual problem is suitably unique.

Theorem (Bootstrap consistency)

Suppose assumptions 1, 2, 3, and 4 hold. Then $T'_P(\mathbb{G})$ is bivariate normal, and conditional on $\{Y_i, D_i, X_i\}_{i=1}^n$,

$$\sqrt{n}(T(\mathbb{P}_n^*) - T(\mathbb{P}_n)) \xrightarrow{L} T'_P(\mathbb{G})$$

in outer probability.

Precise assumption 4

Inference: bootstrap

- Bootstrap works with assumption 4 (unique solutions)...when does that happen?

Lemma (Unique solutions) Suppose that

(i) $c(y_1, y_0)$ is continuously differentiable, and

(ii) for each x, $\text{Supp}(Y_d \mid X = x) = [y^{\ell}_{d,x}, y^{u}_{d,x}]$ is bounded.

then assumption 4 holds.

- Assumption 4 may hold without this lemma's conditions.

Inference: bootstrap alternative

- Only require assumptions 1, 2, and 3 to claim

$$\sqrt{n}((\hat{\gamma}^L, \hat{\gamma}^H) - (\gamma^L, \gamma^H)) \xrightarrow{L} T'_P(\mathbb{G}).$$

- But without assumption 4, $T'_P(\mathbb{G})$ may not be bivariate Normal,

$\implies$ The bootstrap is not consistent.

- The paper shows a consistent alternative.
  - Follows Fang and Santos (2019): estimating the derivative $T'_P(\cdot)$.
  - Implementation is more involved, but still computationally tractable.

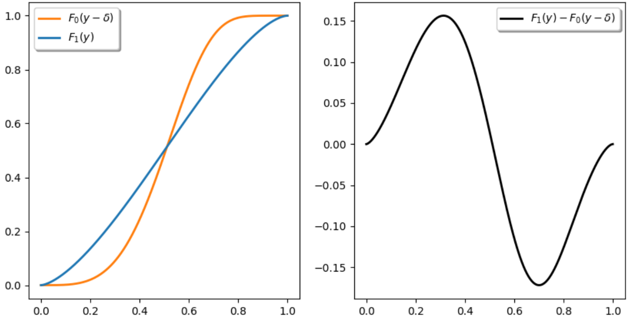
Overview

Simulations: parameter and DGP

- Parameter $\gamma = \theta = P(Y_1 - Y_0 \leq \delta)$ has simple bounds:

$$\gamma^L = \sup_y \{F_1(y) - F_0(y-\delta)\}, \qquad \gamma^H = 1 + \inf_y \{F_1(y) - F_0(y-\delta)\}$$

- For simplicity: no X, $P(D=1) = 1/2$, distributions of Y_1, Y_0:



- Unique solutions $\implies$ bootstrap is valid.

Simulations: confidence set

- Asymptotic $1-\alpha$ confidence set for $[\gamma^L, \gamma^H]$:

  (i) Using $\{Y_i, D_i, X_i\}_{i=1}^n$, compute estimators:

  $$(\hat{\gamma}^L, \hat{\gamma}^H) = T(\mathbb{P}_n)$$

  (ii) For each $b = 1, \ldots, B$, draw $\{Y_{i,b}^*, D_{i,b}^*, X_{i,b}^*\}_{i=1}^n$ to define $\mathbb{P}_{n,b}^*$ and compute:

  $$(\hat{\gamma}_b^{L*}, \hat{\gamma}_b^{H*}) = T(\mathbb{P}_{n,b}^*)$$

  (iii) Let $\hat{c}_{1-\alpha}$ be the $1-\alpha$ quantile of $\{\max\{\sqrt{n}(\hat{\gamma}_b^{L*} - \hat{\gamma}), -\sqrt{n}(\hat{\gamma}_b^{H*} - \hat{\gamma}^H)\}\}_{b=1}^B$, and

  $$CI = [\hat{\gamma}^L - \hat{c}_{1-\alpha}/\sqrt{n}, \hat{\gamma}^H + \hat{c}_{1-\alpha}/\sqrt{n}]$$

Simulations: finite sample bias and correction

- *CI* has exact *asymptotic* coverage. What about small samples?
  - max over sample averages is biased upward (Haile and Tamer, 2003).
  - Leads to $[\hat{\gamma}^L, \hat{\gamma}^H]$ that tend to be "too narrow" in small samples.

- Bootstrap bias correction (Efron and Tibshirani, 1994; Horowitz, 2001):

$$(\widehat{bias}^L, \widehat{bias}^H) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\gamma}^{L*}, \hat{\gamma}^{H*}) - (\hat{\gamma}^L, \hat{\gamma}^H),$$

$$\hat{\gamma}^L_{BC} = \hat{\gamma}^L - \widehat{bias}^L, \qquad \hat{\gamma}^H_{BC} = \hat{\gamma}^H - \widehat{bias}^H$$

- Bootstrap bias corrected confidence interval:

$$CI_{BC} = [\hat{\gamma}^L_{BC} - \hat{c}_{1-\alpha}/\sqrt{n}, \hat{\gamma}^H_{BC} + \hat{c}_{1-\alpha}/\sqrt{n}]$$

Simulations: results

- 300 simulations, 3,000 bootstrap draws, targeting 95% coverage.

Table: Simulations, $P(Y_1 - Y_0 \leq \delta)$

n	Bias $\hat{\gamma}^L$	Bias $\hat{\gamma}^H$	St. Dev. $\hat{\gamma}^L$	St. Dev. $\hat{\gamma}^H$	Emp. Coverage CI
100	0.047	-0.051	0.065	0.066	0.900
200	0.031	-0.031	0.049	0.049	0.917
300	0.030	-0.021	0.040	0.040	0.893

Table: Simulations, $P(Y_1 - Y_0 \leq \delta)$, w/Bias Correction

n	Bias $\hat{\gamma}^L_{BC}$	Bias $\hat{\gamma}^H_{BC}$	St. Dev. $\hat{\gamma}^L_{BC}$	St. Dev. $\hat{\gamma}^H_{BC}$	Emp. Coverage CI_{BC}
100	0.021	-0.026	0.071	0.071	0.927
200	0.013	-0.015	0.052	0.051	0.953
300	0.015	-0.007	0.042	0.042	0.957

Overview

A randomized job training experiment

- The National Supported Work Demonstration Program (NSW)
  - Disadvantaged workers randomized to treatment (guaranteed job, meeting w/counselor) or control.
  - Diamond and Sekhon (2013) subsample: men, 297 treated and 425 control
  - Outcome Y is 1978 real earnings, one year after treatment ended.

Table: Balance table

	base inc.	age	yrs. educ.	HS dropout	black	hispanic	married
control	3672.49	24.45	10.19	0.81	0.80	0.11	0.16
	(6521.53)	(6.59)	(1.62)	(0.39)	(0.40)	(0.32)	(0.36)
treated	3571.00	24.63	10.38	0.73	0.80	0.09	0.17
	(5773.13)	(6.69)	(1.82)	(0.44)	(0.40)	(0.29)	(0.37)

Note: Standard deviations in parentheses.

Who saw larger benefits from treatment?

- *Question:* Who saw larger benefits from the NSW treatment?
- *Parameter:* The OLS slope coefficient $Y_1 - Y_0 = \alpha + \gamma Y_0 + \varepsilon$

$$\gamma = \frac{\text{Cov}(Y_1 - Y_0, Y_0)}{\text{Var}(Y_0)} = \frac{\overbrace{E[(Y_1 - Y_0)Y_0]}^{\theta} - (E[Y_1] - E[Y_0])E[Y_0]}{E[Y_0^2] - (E[Y_0])^2}$$

- *Interpretation:* $\gamma < 0$ implies workers with below average Y_0 tend to see above average $Y_1 - Y_0$

NSW results

- Discretized age and baseline income are informative covariates.
  - age bins: $[16, 23]$, $(23, \infty)$
  - baseline income bins: $[0, 0]$, $(0, 4000]$, $(4000, \infty)$

Table: Estimates of bounds for γ, the OLS Slope

	Lower Bound	Upper Bound	95% *CI*
No Covariates	-1.78	0.19	[-2.01, 0.42]
Disc. Age and Inc.	-1.72	0.00	[-1.95, 0.22]
With Bias Corr.	-1.73	0.04	[-1.96, 0.27]

NSW results: conditional on covariate values

Table: Estimates conditional on covariate values

age	base inc.	$\hat{\gamma}^L_{BC}$	$\hat{\gamma}^H_{BC}$	95% CI_{BC}	n
(16, 23]	0	-1.97	0.28	[-2.26, 0.56]	140
	(0, 4000]	-1.74	-0.15	[-1.9, 0.01]	141
	(4000, ∞)	-1.45	-0.44	[-1.63, -0.27]	90
(23, ∞)	0	-2.13	0.81	[-2.65, 1.33]	187
	(0, 4000]	-1.39	-0.16	[-1.93, 0.38]	56
	(4000, ∞)	-1.66	0.03	[-2.08, 0.45]	108

- Among young men with + base income, low Y_0 is associated with high $Y_1 - Y_0$.
- This subset's vulnerable individuals see larger benefits from treatment.

Conclusion

- Summary:
  - Parameters of the joint distribution of potential outcomes are not point identified.
  - Sharp bounds are characterized with optimal transport.
  - Sample analogue estimators are computationally and analytically attractive.
- Ongoing and future work:
  - Accomodate plausible support restrictions, such as $Y_1 \geq Y_0$.
  - Support function approach to consider parameters depending on more than one joint moment.

References I

Abadie, Alberto (2003). "Semiparametric instrumental variable estimation of treatment response models". In: *Journal of econometrics* 113(2), pp. 231–263.

Allcott, Hunt et al. (2020). "The welfare effects of social media". In: *American Economic Review* 110(3), pp. 629–676.

Beaman, Lori et al. (2013). "Profitability of fertilizer: Experimental evidence from female rice farmers in Mali". In: *American Economic Review* 103(3), pp. 381–386.

Callaway, Brantly (2021). "Bounds on distributional treatment effect parameters using panel data with an application on job displacement". In: *Journal of Econometrics* 222(2), pp. 861–881.

Diamond, Alexis and Jasjeet S Sekhon (2013). "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies". In: *Review of Economics and Statistics* 95(3), pp. 932–945.

Dunipace, Eric (2021). "Optimal transport weights for causal inference". In: *arXiv preprint arXiv:2109.01991*.

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Ekeland, Ivar, Alfred Galichon, and Marc Henry (2010). "Optimal transportation and the falsifiability of incompletely specified economic models". In: *Economic Theory* 42, pp. 355–374.

Fan, Yanqin and Sang Soo Park (2010). "Sharp bounds on the distribution of treatment effects and their statistical inference". In: *Econometric Theory* 26(3), pp. 931–951.

Fan, Yanqin and Sang Soo Park (2012). "Confidence intervals for the quantile of treatment effects in randomized experiments". In: *Journal of Econometrics* 167(2), pp. 330–344.

Fan, Yanqin, Xuetao Shi, and Jing Tao (2023). "Partial identification and inference in moment models with incomplete data". In: *Journal of Econometrics* 235(2), pp. 418–443.

Fang, Zheng and Andres Santos (2019). "Inference on directionally differentiable functions". In: *The Review of Economic Studies* 86(1), pp. 377–412.

Firpo, Sergio (2007). "Efficient semiparametric estimation of quantile treatment effects". In: *Econometrica* 75(1), pp. 259–276.

Firpo, Sergio and Geert Ridder (2019). "Partial identification of the treatment effect distribution and its functionals". In: *Journal of Econometrics* 213(1), pp. 210–234.

References II

Frandsen, Brigham R and Lars J Lefgren (2021). "Partial identification of the distribution of treatment effects with an application to the Knowledge is Power Program (KIPP)". In: *Quantitative Economics* 12(1), pp. 143–171.

Friebel, Guido et al. (2023). "What do employee referral programs do? Measuring the direct and overall effects of a management practice". In: *Journal of Political Economy* 131(3), pp. 633–686.

Galichon, Alfred and Marc Henry (2011). "Set identification in models with multiple equilibria". In: *The Review of Economic Studies* 78(4), pp. 1264–1298.

Gunsilius, Florian and Yuliang Xu (2021). "Matching for causal effects via multimarginal unbalanced optimal transport". In: *arXiv preprint arXiv:2112.04398*.

Haile, Philip A and Elie Tamer (2003). "Inference with an incomplete model of English auctions". In: *Journal of Political Economy* 111(1), pp. 1–51.

Heckman, James J, Jeffrey Smith, and Nancy Clements (1997). "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts". In: *The Review of Economic Studies* 64(4), pp. 487–535.

Holland, Paul W (1986). "Statistics and causal inference". In: *Journal of the American statistical Association* 81(396), pp. 945–960.

Horowitz, Joel L (2001). "The bootstrap". In: *Handbook of econometrics*. Vol. 5. Elsevier, pp. 3159–3228.

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62(2), pp. 467–475.

Ji, Wenlong, Lihua Lei, and Asher Spector (2023). "Model-Agnostic Covariate-Assisted Inference on Partially Identified Causal Effects". In: *arXiv preprint arXiv:2310.08115*.

Manski, Charles F (1997). "Monotone treatment response". In: *Econometrica: Journal of the Econometric Society*, pp. 1311–1334.

Russell, Thomas M (2021). "Sharp bounds on functionals of the joint distribution in the analysis of treatment effects". In: *Journal of Business & Economic Statistics* 39(2), pp. 532–546.

Torous, William, Florian Gunsilius, and Philippe Rigollet (2021). "An optimal transport approach to causal inference". In: *arXiv preprint arXiv:2108.05858*.

Appendix: full setting

Assumption 1 (Setting). $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$ is an i.i.d. sample, with

$$Y \in \mathcal{Y} \subseteq \mathbb{R}, \qquad D \in \{0,1\}, \qquad Z \in \{0,1\}, \qquad X \in \mathcal{X} = \{x_1, \ldots, x_M\}$$

generated from a distribution satisfying

(i) Potential outcomes: $Y = DY_1 + (1-D)Y_0$,

(ii) Potential treatment statuses: $D = ZD_1 + (1-Z)D_0$, with $D_z \in \{0,1\}$,

(iii) Instrument exogeneity: $(Y_1, Y_0, D_1, D_0) \perp Z \mid X$,

(iv) Monotonicity: $D_1 \geq D_0$ almost surely,

(v) Existence of compliers: $P(D_1 > D_0, X = x) > 0$ for each x, and

(vi) $P(X = x, Z = z) > 0$ for each (x, z)

- Terminology: always-taker, complier, defier, never-taker.

	$D_0 = 1$	$D_0 = 0$
$D_1 = 1$	Always-takers	Compliers
$D_1 = 0$	~~Defiers~~	Never-takers

- Monotonicity rules out defiers. Focus on distribution of compliers.

Setting Aside, IV

Appendix: identification of $P(Y_1 - Y_0 \leq \delta)$

- $OT_c(P_1, P_0)$ is well behaved (attained, strong duality holds, etc) when $c(y_1, y_0)$ is bounded and lower semicontinuous

- If $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$, let

$$c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}, \qquad c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\}$$
$$\theta_x^L = OT_{c_L}(P_{1|x}, P_{0|x}), \qquad \theta_x^H = 1 - OT_{c_H}(P_{1|x}, P_{0|x})$$

- The form of the bounds remains the same:

$$\theta^L = E[\theta_X^L], \qquad \theta^H = E[\theta_X^H]$$
$$\gamma^L = \min_{t \in [\theta^L, \theta^H]} g(t, \eta), \qquad \gamma^H = \max_{t \in [\theta^L, \theta^H]} g(t, \eta)$$

- Identified sets are still sharp when CDFs are continuous:

$$F_{d|x}(y) = P(Y_d \leq y \mid X = x)$$

Appendix: aside, CDF results are conservative when continuity fails

$$OT_c(P_1, P_0) = \inf_{\pi \in \Pi(P_1, P_0)} E_\pi[c(Y_1, Y_0)]$$

- Bounds on $\theta = P(Y_1 - Y_0 \leq \delta)$ are found with

$$c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}, \qquad c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\},$$
$$\theta^L = OT_{c_L}(P_1, P_0), \qquad \theta^H = 1 - OT_{c_H}(P_1, P_0)$$

  Using OT results, show that if marginal CDFs F_d are continuous then $\Theta_{ID} = [\theta^L, \theta^H]$.

- As a byproduct, recover the famed **Makarov bounds** studied by Fan and Park (2010)

$$\theta^L = \sup_y \{F_1(y) - F_0(y - \delta)\}, \qquad \theta^H = 1 + \inf_y \{F_1(y) - F_0(y - \delta)\}$$

- **Furthermore,** $\mathbb{1}\{y_1 - y_0 < \delta\} \leq \mathbb{1}\{y_1 - y_0 \leq \delta\}$ implies the bounds are conservative: $\Theta_{ID} \subseteq [\theta^L, \theta^H]$ whether or not F_d are continuous.

Identification | Identification Thm | Estimators

Appendix: full assumption 3

- Parameter of interest:
$$\gamma = g(\theta, \eta) \in \mathbb{R}$$
where $\theta = E[c(Y_1, Y_0)] \in \mathbb{R}$ and $\eta = (E[\eta_1(Y_1)], E[\eta_0(Y_0)]) \in \mathbb{R}^{K_1+K_0}$.

Assumption 3 (Function of moments)

(i) $E[\|\eta_d(Y)\|^2] < \infty$ for $d = 1, 0$,

(ii) $g(\cdot, \eta)$ is continuous, and

(iii) the functions
$$g^L(t^L, t^H, e) = \min_{t \in [t^L, t^H]} g(t, e), \qquad g^H(t^L, t^H, e) = \max_{t \in [t^L, t^H]} g(t, e)$$
are continuously differentiable at $(t^L, t^H, e) = (\theta^L, \theta^H, \eta)$.

Remark: A3 (ii), (iii) implied by g continuously differentiable and $g(\cdot, \eta)$ monotonic

Back

Appendix: quantiles

- Suppose the parameter of interest is q_τ solving

$$P(Y_1 - Y_0 \leq q_\tau) = \tau$$

- View CDF bounds as a function: $\theta(\delta) = P(Y_1 - Y_0 \leq \delta)$

$$\begin{aligned} c_{L,\delta}(y_1, y_0) &= \mathbb{1}\{y_1 - y_0 < \delta\}, & c_{H,\delta}(y_1, y_0) &= \mathbb{1}\{y_1 - y_0 > \delta\}, \\ \theta_x^L(\delta) &= OT_{c_L}(P_{1|x}, P_{0|x}), & \theta_x^H(\delta) &= 1 - OT_{c_H}(P_{1|x}, P_{0|x}) \\ \theta^L(\delta) &= E[\theta_X^L(\delta)] & \theta^H(\delta) &= E[\theta_X^H(\delta)] \end{aligned}$$

and let $Q_{I,\tau}$ be the sharp identified set for q_τ.

Lemma (Identification of q_τ). Suppose assumptions 1 and 2(ii) hold. Then $q \in Q_{I,\tau}$ if and only if $\theta^L(q) \leq \tau \leq \theta^H(q)$.

Examples

Appendix: aside, IV

- Identification extends easily to IV.
- Consider the binary IV potential outcomes framework of Abadie (2003): $Z \in \{0, 1\}$,

$$D = ZD_1 + (1 - Z)D_0, \qquad (Y_1, Y_0, D_1, D_0) \perp Z \mid X, \qquad D_1 \geq D_0$$

  units with $D_1 > D_0$ are known as *compliers*.
- This model identifies marginal distributions of potential outcomes of compliers:

$$Y_d \mid D_1 > D_0, X = x$$

- Same identification applies to parameters conditional on compliance. E.g.,

$$P(Y_1 > Y_0 \mid D_1 > D_0)$$

Ident. Thm. IV Setting

Appendix: definition of T

- Proof defines a set of universally bounded functions

$$\mathcal{F} \subseteq \{f : \mathcal{Y} \times \{0,1\} \times \mathcal{X} \to \mathbb{R}\}$$

- View $\mathbb{P}_n$, P as bounded functions on $\mathcal{F}$:

$$\ell^\infty(\mathcal{F}) = \left\{ g : \mathcal{F} \to \mathbb{R} \, ; \, \|g\|_\infty = \sup_{f \in \mathcal{F}} |g(f)| < \infty \right\}$$

- The map $T : \ell^\infty(\mathcal{F}) \to \mathbb{R}^2$ is described by $P \mapsto (P_{1|x}, P_{0|x}, \eta)$ and

$$\begin{aligned} \theta_x^L &= OT_c(P_{1|x}, P_{0|x}), & \theta_x^H &= -OT_{-c}(P_{1|x}, P_{0|x}) \\ \theta^L &= E[\theta_X^L], & \theta^H &= E[\theta_X^H] \\ \gamma^L &= \min_{t \in [\theta^L, \theta^H]} g(t, \eta), & \gamma^H &= \max_{t \in [\theta^L, \theta^H]} g(t, \eta) \end{aligned}$$

Weak convergence theorem

Appendix: proof sketch (1/3)

1. Will view P, $\mathbb{P}$ as maps in $\ell^\infty(\mathcal{F})$ for Donsker set $\mathcal{F}$ (defined later), and $T : \ell^\infty(\mathcal{F}) \to \mathbb{R}^2$.

2. To show $T(\cdot)$ is (Hadamard) directionally differentiable, suffices to show OT_c is directionally differentiable.

3. By strong duality,

$$OT_c(P_{1|x}, P_{0|x}) = \sup_{(\varphi,\psi)\in\Phi_c} E_{P_{1|x}}[\varphi(Y_1)] + E_{P_{0|x}}[\psi(Y_0)]$$

$$\Phi_c = \{(\varphi, \psi) : \varphi(y_1) + \psi(y_0) \le c(y_1, y_0)\}$$

Weak convergence theorem

Appendix: proof sketch (2/3)

$$OT_c(P_{1|x}, P_{0|x}) = \sup_{(\varphi,\psi)\in\Phi_c} E_{P_{1|x}}[\varphi(Y_1)] + E_{P_{0|x}}[\psi(Y_0)]$$

$$\Phi_c = \{(\varphi,\psi) : \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)\}$$

4. Φ_c is a large set, but much of it can be ignored:
   - If $\varphi(y_1) \leq \tilde{\varphi}(y_1)$, then $E_{P_{1|x}}[\varphi(Y_1)] \leq E_{P_{1|x}}[\tilde{\varphi}(Y_1)]$
   - Any pair (φ,ψ) where $\varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)$ is "slack" can be ignored

5. This observation leads to

$$\sup_{(\varphi,\psi)\in\Phi_c} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)] = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)] \quad (1)$$

   (i) if $c(y_1, y_0)$ is L-Lip. and $\mathcal{Y}$ is compact, $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are L-Lip. and universally bounded.
   (ii) if $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$, $\mathcal{F}_c$ is the set of intervals, $\mathcal{F}_c^c$ the complements of intervals.

6. Finally, $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ is compact and $E_{P_{1|x}}[\varphi(Y_1)] + E_{P_{0|x}}[\psi(Y_0)]$ is continuous

$\implies$ OT_c, and therefore $T(\cdot)$, are Hadamard directionally differentiable.

Weak convergence theorem | c-concavity

Appendix: proof sketch (3/3)

7. Define $\mathcal{F}$ to be union of $\mathcal{F}_c$ and $\mathcal{F}_c^c$ (and nuisance moments, all $\times$ indicators).

8. $\mathcal{F}$ is Donsker $\implies \sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$.

9. Functional delta method implies the result,

$$\sqrt{n}((\hat{\gamma}^L, \hat{\gamma}^H) - (\gamma^L, \gamma^H)) = \sqrt{n}(T(\mathbb{P}_n) - T(P)) \xrightarrow{L} T_P'(\mathbb{G}).$$

Weak convergence theorem

Appendix: c-concavity

$$OT_c(P_1, P_0) = \sup_{(\varphi,\psi)\in\Phi_c} \underbrace{E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)]}_{J(\varphi,\psi)},$$

- Define the c-transforms:

$$\varphi^c(y_0) = \inf_{y_1}\{c(y_1, y_0) - \varphi(y_1)\}, \qquad \psi^c(y_1) = \inf_{y_0}\{c(y_1, y_0) - \psi(y_0)\}$$

  call φ^c (and ψ^c) c-concave functions.

- For any $(\varphi, \psi) \in \Phi_c = \{(\varphi, \psi)\ ;\ \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)\}$,
  (i) $(\varphi, \varphi^c), \in \Phi_c$
  (ii) If $(\varphi, \psi) \in \Phi_c$, then $\psi(y_0) \leq \varphi^c(y_0)$ for all y_0, so
  (iii) $J(\varphi, \psi) \leq J(\varphi, \varphi^c)$ by monotonicity of $E_{P_d}[\cdot]$.
  $\implies$ The dual problem can be restricted to c-concave functions.

- c-concave functions often inherit properties of c:
  - Lipschitz continuity, boundedness, etc.
  - These properties are used to define $\mathcal{F}_c$ and $\mathcal{F}_c^c$

Proof sketch | Weak convergence theorem

Appendix: formal assumption 4

- Let P be the distribution of an observation: $(Y, D, Z, X) \sim P$.
- Let $\mathcal{Y}_{d,x}$ be the support of $Y \mid D = d, X = x$, and $\mathbb{1}_{\mathcal{Y}_{d,x}}(y) = \mathbb{1}\{y \in \mathcal{Y}_{d,x}\}$
- Define c_L, c_H:
  (i) If assumption 2 (i) holds, let $c_L = c(y_1, y_0)$ and $c_H(y_1, y_0) = -c(y_1, y_0)$.
  (ii) If assumption 2 (ii) holds, let $c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}$ and $c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\}$.

Assumption 4 (Unique solutions) For each $x \in \mathcal{X}$, each $c \in \{c_L, c_H\}$, and any

$$(\varphi_1, \psi_1), (\varphi_2, \psi_2) \in \underset{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)}{\arg\max} E_{P_{1|x}}[\varphi(Y_1)] + E_{P_{0|x}}[\psi(Y_0)],$$

there exists $s \in \mathbb{R}$ such that

$$\mathbb{1}_{\mathcal{Y}_{1,x}} \times \varphi_1 = \mathbb{1}_{\mathcal{Y}_{1,x}} \times (\varphi_2 + s),\ P - a.s., \quad \mathbb{1}_{\mathcal{Y}_{0,x}} \times \psi_1 = \mathbb{1}_{\mathcal{Y}_{0,x}} \times (\psi_2 - s),\ P - a.s.$$

Assumption 4 Why c_L, c_H?