Submitted to the Annals of Statistics
arXiv: arXiv:1807.02737

SUPPLEMENTARY MATERIAL FOR "A CAUSAL BOOTSTRAP"

By Guido Imbens[*] and Konrad Menzel[†]

Stanford GSB[*] and New York University[†]

This online supplement contains additional formal results and proofs for [7]. Section A gives formal conditions under which the pivotal version of the causal bootstrap achieves refinements over Gaussian inference. Section B contains the proofs for formal results in the paper.

APPENDIX A: REFINEMENTS

In this appendix, we state conditions under which the pivotal version of the causal bootstrap can achieve refinements over the Gaussian asymptotic approximation. For expositional ease, we only state a result regarding refinements for the case of completely randomized treatment assignment, where we assume

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i$$

In this section, we let $F_0(y_0), F_1(y_1)$ denote the respective marginal distributions of potential outcomes in the infinite superpopulation, and their (finite) population and sample counterparts are marked with p- and s-superscripts. We also assume that the bootstrap uses the respective empirical counterparts

$$\begin{aligned}\hat{F}_{0n}(y_0) &:= \frac{1}{n(1-p)}\sum_{i=1}^{N} R_i(1-W_i)\mathbb{1}\{Y_i \le y_0\}\\ \hat{F}_{1n}(y_1) &:= \frac{1}{np}\sum_{i=1}^{N} R_i W_i \mathbb{1}\{Y_i \le y_1\}\end{aligned}$$

as estimators for these marginal distributions.

We state sufficient conditions for refinements in terms of a general smooth functional $\tau(F_0, F_1)$ with variance bound $\sigma(F_0, F_1)$, where for the case of the average treatment effect,

$$\tau(F_0, F_1) := \mathbb{E}_{F_1}[Y_i(1)] - \mathbb{E}_{F_0}[Y_i(0)], \quad \text{and } \sigma(F_0, F_1) := \sigma(C^{iso}(F_0, F_1))$$

We assume throughout that the functionals $\tau(\cdot)$ and $\sigma(\cdot)$ are Fréchet differentiable: For a Banach space $\mathcal{H}$ we say that a functional $P \mapsto T(P)$, $T : \mathcal{H} \to \mathbb{R}$ is Fréchet differentiable at $P \in \mathcal{H}$ if there exists a bounded linear functional $T' : h \mapsto T'(P)(h)$, $T' : \mathcal{H} \to \mathbb{R}$ such that

$$\lim_{h \to 0} \frac{|T(P+h) - T(P) - T'(P)(h)|}{\|h\|} = 0$$

In a similar fashion, we can also define higher-order Fréchet derivatives: the $(s+1)$th derivative of $T(P)$ is a mapping $T^{(s+1)}(P)(h_1, \dots, h_{s+1})$ that is multilinear in its $s+1$ arguments and that is defined recursively as

$$\lim_{h_{s+1} \to 0} \frac{\left|T^{(s)}(P+h_{s+1})(h_1, \dots, h_s) - T^{(s)}(P)(h_1, \dots, h_s) - T^{(s+1)}(P)(h_1, \dots, h_s, h_{s+1})\right|}{\|h_{s+1}\|} = 0$$

Denoting $V_i := (Y_i(0), Y_i(1))$, we let

$$\begin{aligned} g_{1p}(V_i) &:= \tau'(F_{01})(\delta_{V_i} - F_{01}) \\ g_{2p}(V_i, V_j) &:= \tau''(F_{01})(\delta_{V_i} - F_{01}, \delta_{V_j} - F_{01}) \end{aligned}$$

where δ_v is the point mass at $V_i = v$, i.e. a distribution with $P(V_i = v) = 1$.

Assumption A.1. ***(Smooth Functionals)*** *(a) The functionals $\tau(F_0, F_1), \tau(F_0, F_1)$, and $\sigma(F_0, F_1)$ are three times Fréchet-differentiable, with two bounded derivatives. Furthermore, (b) the random variable $g_1(V_i)$ satisfies Cramér's condition,*

$$\varrho := 1 - \sup_{t:\sqrt{n} \geq |t| \geq t_0} \mathbb{E}\left[\exp\left\{itg_{1p}(V_i)\right\}\right] > 0 \tag{A.1}$$

where $t_0 := b_1/\mathbb{E}\left|g_{1p}(V_i)\right|^3$ for some small constant $b_1 > 0$, and the expectation is taken over the distribution $C^{iso}(F_0, F_1)$, and the analogous condition holds for the lower bound.

Part (a) imposes smoothness conditions on the functionals characterizing the bounds, in particular for the functional $T(F_0, F_1) := (\tau(F_0, F_1) - \tau(F_{01}^p))/\sigma(F_0, F_1)$, we have that $\|T^{(2)}\|_\infty^2 < \infty$. Part (b) translates the formulation of Cramérs condition from [5] into the notation of this paper.

Theorem A.1. ***(Refinements)*** *Suppose that Assumptions 1.1, 1.2, and A.1 hold, and that the population consists of draws from a distribution of $Y_i(0), Y_i(1)$ whose marginals are non-lattice with six bounded moments. Then we have*

$$\left\| \mathbb{P}_n^* \left(\sqrt{n} \frac{\tau(\hat{F}_{0n}^*, \hat{F}_{1n}^*) - \tau(\hat{F}_{0n}, \hat{F}_{1n})}{\sigma(\hat{F}_{0n}^*, \hat{F}_{1n}^*)} \right) - \mathbb{P}\left(\sqrt{n} \frac{\tau(\hat{F}_{0n}, \hat{F}_{1n}) - \tau(F_0^s, F_1^s)}{\sigma(\hat{F}_{0n}, \hat{F}_{1n})} \right) \right\|_\infty = O_P(n^{-1})$$

Proof of Theorem A.1: Let $n_* := \min\{n, N-n\}$. If $n_* = N-n$ and its value remains bounded as $n \to \infty$, then we can ignore the contribution of sampling uncertainty and will focus on randomization error alone. Hence, without loss of generality we can focus on the second case in which both $n, n_* \to \infty$. In the following we let $V_i := (Y_i(0), Y_i(1), X_i)'$ denote the attributes and potential outcomes for the ith observation, and without loss of generality we assume that observations are ordered such that $R_i = 1$ for $i = 1, \dots, n$ and $R_i = 0$ for $i = n+1, \dots, N$.

Separate contributions of sampling and randomization error. We can write the estimation error in the upper bound for τ as

$$\sqrt{n}(\hat{\tau} - \tau) = \sqrt{n}(\tau(\hat{F}_{0n}, \hat{F}_{1n}) - \tau(F_{0n}^s, F_{1n}^s)) + \sqrt{n}(\tau(F_0^s, F_1^s) - \tau(F_0^p, F_1^p)) =: B_1 + B_2$$

In the following we suppress the L, U superscripts and take the expansions to apply to either bound. We first consider separate stochastic expansions for the terms B_1, B_2, noting that B_1 is conditionally independent of B_2 given F_{01}^s. We also let S_1^2, S_2^2 denote the respective variances of B_1, B_2 under the distribution $C(\hat{F}_{1n}, \hat{F}_{0n})$.

Orthogonal decomposition for sampling error. We denote $V_i := (Y_i(0), Y_i(1), X_i)$. Also let

$$\begin{aligned} g_{1s}(V_i) &:= \frac{N-1}{N-n}\left(\mathbb{E}[\tau(F_0^s, F_1^s)|V_i] - \mathbb{E}[\tau(F_0^s, F_1^s)]\right) \\ g_{2s}(V_i, V_j) &:= \frac{(N-2)(N-3)}{(N-n)(N-n-1)}\left(\mathbb{E}[\tau(F_0^s, F_1^s)|V_i, V_j] - \mathbb{E}[\tau(F_0^s, F_1^s)] - \frac{N-n}{N-2}\left(g_{1s}(V_i) + g_{1s}(V_j)\right)\right) \end{aligned}$$

where we also let $\sigma_{1s}^2 := \mathrm{Var}^p(g_{1s}(V_i))$.

Now consider $T_1 \equiv T_1(Z_1, \dots, Z_n) := B_1/S_1$. By Theorem 1 in [4], T_1 can be expanded according to

$$T_1 = \mathbb{E}_p T_1 + U_{11} + U_{21} + R_{31}$$

where $U_{11} := \sum_{i=1}^{N} g_{1s}(V_i)$ and $U_{21} := \sum_{i=1}^{N}\sum_{j=1}^{N} g_{2s}(V_i, V_j)$ and

$$\mathbb{E}_p R_{31}^2 \le n_*^{-1}\mathbb{E}\left[(n_* D_1 D_2 T_1)^2\right] =: n_*^{-1}\delta_s$$

where

$$D_1D_2T_1 := T_1(V_1, V_2, \dots, V_n) - T_1(V_{n+1}, V_2, \dots, V_n) - T_1(V_1, V_{n+2}, \dots, V_n) + T_1(V_{n+1}, V_{n+2}, \dots, V_n)$$

Adapting the formal argument in sections 2.3 and 2.6 of [2] we can bound

$$\delta_s := \mathbb{E}\left[(n_* D_1 D_2 T_1)^2\right] \le C n_*^{-2}\|T^{(2)}\|_\infty^2$$

for some constant C, noting that the relevant bounds are in terms of marginal distributions and therefore continue to apply for sampling without replacement. It follows from Assumption A.1 that $\delta_s = O(n_*^{-2})$.

Orthogonal decomposition for randomization error. From (5.2) and Section 2.6 in [2] we obtain a similar decomposition for the randomization error conditional on the sample. Specifically, let

$$\begin{aligned}
g_{1r}(V_i) &:= \frac{n-1}{n-n_1}\sum_{w=0}^{1}\xi(w)\left(\mathbb{E}[\tau(F_0^s, F_1^s)|Y_i(z)] - \mathbb{E}[\tau(F_0^s, F_1^s)]\right)\\
g_{2r}(V_i, V_j) &:= \frac{(n-2)(n-3)}{(n-n_1)(n-n_1-1)}\sum_{w,w'=0}^{1}\xi(w)\xi(w')\\
&\quad \times\left(\mathbb{E}[\tau(F_0^s, F_1^s)|Y_i(z), Y_j(z')] - \mathbb{E}[\tau(F_0^s, F_1^s)]\frac{n-n_1}{n-2}\left(g_{1r}(V_i) + g_{1r}(V_j)\right)\right)
\end{aligned}$$

where $\xi(w) := (w-p)/(1-p)$, and we also let $\sigma_{1r}^2 := \mathrm{Var}^p(g_{1r}(V_i))$. Then, applying Theorem in [4] again, we can expand $T_1 \equiv T_2(V_1, \dots, V_n) := B_2/S_2$ as

$$T_2 = \mathbb{E}_s T_2 + U_{12} + U_{22} + R_{32}$$

where $U_{12} := \sum_{i=1}^{N} g_{1r}(V_i)$ and $U_{22} := \sum_{i=1}^{N}\sum_{j=1}^{N} g_{2r}(V_i, V_j)$ and

$$\mathbb{E}_p R_{32}^2 \le C(n_* p)^{-2}\|T^{(2)}\|_\infty^2$$

Edgeworth expansion. By Theorem 1 in [4], these bounds on the remainders R_{1n}, R_{2n} establish the validity of separate orthogonal expansions

$$T_1 + T_2 = \mathbb{E}[T_1 + T_2] + (U_{11} + U_{21}) + (U_{12} + U_{22}) + (R_{1n} + R_{2n}) \tag{A.2}$$

where $\mathbb{E}[R_{1n}^2 + R_{2n}^2] \le C\|T^{(2)}\|_\infty^2(np)^{-1}$. Noting that U_{21}, U_{22} are conditionally independent of U_{11}, U_{12}, it follows from Theorem 1.1 in [5] and the law of iterated expectations that the sum $\sqrt{n}(\hat{\tau} - \tau)/\sigma = (B_1 + B_2)/(S_1 + S_2)$ also has an Edgeworth expansion

$$\mathbb{P}\left(\sqrt{n}\frac{\tau(\hat{F}_{0n}, \hat{F}_{1n}) - \tau(F_0^p, F_1^p)}{\sigma(\hat{F}_{1n}, \hat{F}_{0n})} \le t\right) = \Phi(t) - \frac{(1-2q)\alpha_s + (1-2p)\alpha_r + 3(\kappa_s + \kappa_r)}{6n(N-n)/N}\Phi'''(t) + \Delta_n$$

where $\Delta_n = O((np)^{-1})$, where the coefficient on the second term in the Edgeworth expansion, $n^{-1/2}(1-2q)\alpha_s + (1-2p)\alpha_r + 3(\kappa_s + \kappa_r)$ corresponds, up to order n^{-1}, to the third moment of the approximand $(U_{11} + U_{21}) + (U_{12} + U_{22})$. Specifically, we have

$$\begin{aligned}
\alpha_s &:= \frac{1}{N}\sum_{i=1}^{N}\sigma^{-3}g_{1s}(V_i)^3\\
\alpha_r &:= \frac{1}{N}\sum_{i=1}^{N}\sigma^{-3}g_{1r}(V_i)^3\\
\kappa_s &:= \frac{2Nq(1-q)}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\sigma^{-3}g_{2s}(V_i, V_j)g_{1s}(V_i)g_{1s}(V_j)\\
\kappa_r &:= \frac{2Nqp(1-p)}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\sigma^{-3}g_{2r}(V_i, V_j)g_{1r}(V_i)g_{1r}(V_j)
\end{aligned}$$

Under Assumption A.1 we can verify that these cumulants are bounded as N increases.

Edgeworth expansion for bootstrap distribution. From the basic consistency argument, the bootstrap estimate $\hat{F}_{01n} := C(\hat{F}_{0n}, \hat{F}_{1n})$ is consistent for the finite-population distribution $F_{01}^p := C(F_0^p, F_1^p)$, and inherits its main properties from the same distribution for the infinite meta-population, $F_{01} := C(F_0, F_1)$. Specifically, the bootstrap distribution satisfies an analogous orthogonal expansion to (A.2), and Cramér's condition (A.1) holds with the same constant ϱ.

Specifically, given $\hat{F}_{0n}, \hat{F}_{1n}$ we define the bootstrap influence functions $g_{1s}^*(w_1), g_{1r}^*(w_1), g_{2s}^*(w_1, w_2), g_{2r}^*(w_1, w_2)$, and σ^* in analogy to their sampling/randomzation counterparts. We then have that for the bootstrap distribution,

$$\mathbb{P}_n^* \left(\sqrt{n} \frac{\tau(\hat{F}_{0n}^*, \hat{F}_{1n}^*}{\sigma(\hat{F}_{1n}^*, \hat{F}_{0n}^*)} - \tau(\hat{F}_{0n}, \hat{F}_{1n})) \leq t \right) = \Phi(t) - \frac{(1-2q)\alpha_s^* + (1-2p)\alpha_r^* + 3(\kappa_s^* + \kappa_{r^*})}{6n(N-n)/N} \Phi'''(t) + \Delta_n^*$$

where $\Delta_n^* = O((np)^{-1})$, and

$$\begin{aligned}
\alpha_s^* &:= \frac{1}{N} \sum_{i=1}^N (\sigma^*)^{-3} g_{1s}^*(V_i^*)^3 \\
\alpha_r^* &:= \frac{1}{N} \sum_{i=1}^N (\sigma^*)^{-3} g_{1r}^*(V_i^*)^3 \\
\kappa_s^* &:= \frac{2Nq(1-q)}{N-1} \sum_{i=1}^N \sum_{j=1}^{i-1} (\sigma^*)^{-3} g_{2s}^*(V_i, V_j) g_{1s}^*(V_i) g_{1s}^*(V_j) \\
\kappa_r^* &:= \frac{2Nqp(1-p)}{N-1} \sum_{i=1}^N \sum_{j=1}^{i-1} (\sigma^*)^{-3} g_{2r}^*(V_i, V_j) g_{1r}^*(V_i) g_{1r}^*(V_j)
\end{aligned}$$

Convergence of bootstrap cumulants. It remains to be shown that σ^* and the cumulants $\alpha_1^*, \alpha_2^*, \kappa_1^*, \kappa_2^*$ converge in probability at a root-n rate to their population counterparts. Under the assumption of six bounded moments for $g_{1s}, g_{1r}, g_{12s}, g_{12r}$ and noting that $0 < p < 1$, this can be established using a finite-population CLT. The conclusion then follows from Theorem 1.1 in [5] □

APPENDIX B: DERIVATIONS AND PROOFS FOR RESULTS IN THE MAIN TEXT

B.1. Randomization Distribution for $\hat{F}_0(y_0), \hat{F}_1(y_1)$.

We first compute the randomization covariance $\text{Cov}_{W,R}^p(\hat{F}_0(y_0), \hat{F}_1(y_1))$ given the population distribution $F_{01}^p(y_0, y_1)$, where

$$\begin{aligned}
\hat{F}_0(y_0) &= \frac{1}{n(1-p)} \sum_{i=1}^N R_i(1-W_i) \mathbb{1}\{Y_i(0) \leq y_0\} \\
\hat{F}_1(y_1) &= \frac{1}{np} \sum_{i=1}^N R_i W_i \mathbb{1}\{Y_i(1) \leq y_1\}
\end{aligned}$$

In the following we write $A_{0i} := \mathbb{1}\{Y_i(0) \leq y_0\}$ and $A_{1i} := \mathbb{1}\{Y_i(1) \leq y_1\}$, and take any moments to be with respect to the distribution of R_i and W_i and conditional on the values of $(Y_i(0), Y_i(1))_{i=1}^N$ in the population. We then have

$$\begin{aligned}
\mathbb{E}[\hat{F}_0(y_0)\hat{F}_1(y_1)] &= \frac{1}{n^2 p(1-p)} \mathbb{E}\left[\sum_{i=1}^N \sum_{j=1}^N R_i R_j (1-W_i) W_j A_{0i} A_{1j} \right] \\
&= \frac{1}{n^2 p(1-p)} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}\left[R_i R_j (1-W_i) W_j\right] A_{0i} A_{1j} \\
&= \frac{1}{n^2 p(1-p)} \sum_{i=1}^N \sum_{j \neq i} \mathbb{E}\left[R_i R_j\right] \mathbb{E}\left[(1-W_i) W_j\right] A_{0i} A_{1j} \\
&= \frac{1}{n^2 p(1-p)} \sum_{i=1}^N \sum_{j \neq i} \frac{n(n-1)}{N(N-1)} \frac{n^2 p(1-p)}{n(n-1)} A_{0i} A_{1j} \\
&= \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i} A_{0i} A_{1j} = \frac{1}{N(N-1)} \left(\left[\sum_{i=1}^N A_{0i} \right] \left[\sum_{j=1}^N A_{1j} \right] - \sum_{i=1}^N A_{0i} A_{1i} \right)
\end{aligned}$$

so that

$$\mathrm{Cov}(\hat{F}_0(y_0), \hat{F}_1(y_1)) = -\frac{1}{N-1}\left(F_{01}^p(y_0, y_1) - F_0^p(y_0)F_1^p(y_1)\right)$$

To evaluate $\mathrm{Cov}\left(\hat{F}_0(y_0), \hat{F}_0(y_1)\right)$, let $B_{0i} := \mathbb{1}\{Y_i(0) \le y_0\} - F_0^p(y_0)$ and $B_{1i} := \mathbb{1}\{Y_i(0) \le y_1\} - F_0^p(y_1)$. We can then write

$$\begin{aligned}
\mathrm{Cov}\left(\hat{F}_0(y_0), \hat{F}_0(y_1)\right) &= \frac{1}{n^2(1-p)^2}\sum_{i=1}^N\sum_{j=1}^N \mathbb{E}\left[R_iR_j(1-W_i)(1-W_j)\right]B_{0i}B_{1j} \\
&= \frac{1}{n^2(1-p)^2}\left[\sum_{i=1}^n \frac{n}{N}\frac{n(1-p)}{n}B_{0i}B_{1i} + \sum_{i=1}^N\sum_{j\neq i}\frac{n(n-1)}{N^2}\frac{n(1-p)(n(1-p)-1)}{n^2}B_{0i}B_{1j}\right] \\
&= \frac{1}{n(1-p)N}\sum_{i=1}^N B_{0i}B_{1i} + \frac{(n-1)(n(1-p)-1)}{n^2(1-p)}\left[\frac{1}{N^2}\sum_{i=1}^N\sum_{j\neq i}B_{0i}B_{1j}\right] \\
&= \left[\frac{1}{N}\sum_{i=1}^N B_{0i}B_{1i}\right]\left(\frac{1}{n(1-p)} - \frac{(n-1)(n(1-p)-1)}{Nn^2(1-p)}\right) \\
&= \left(\min\{F_0^p(y_0), F_0^p(y_1)\} - F_0^p(y_0)F_0^p(y_1)\right)\left(\frac{1}{n(1-p)} - \frac{1}{N} + O\left(\frac{1}{nN}\right)\right)
\end{aligned}$$

Similarly,

$$\mathrm{Cov}\left(\hat{F}_1(y_0), \hat{F}_1(y_1)\right) = \left(\min\{F_1^p(y_0), F_1^p(y_1)\} - F_1^p(y_0)F_1^p(y_1)\right)\left(\frac{1}{np} - \frac{1}{N} + O\left(\frac{1}{nN}\right)\right)$$

Furthermore,

$$\begin{aligned}
\mathrm{Cov}\left(\hat{F}_0(y_0), \hat{F}_0(y_1)\right) &= \frac{1}{n}\min\{F_0^p(y_0), F_0^p(y_1)\} - F_0^p(y_0)F_0^p(y_1) \\
\mathrm{Cov}\left(\hat{F}_1(y_0), \hat{F}_1(y_1)\right) &= \frac{1}{n}\min\{F_1^p(y_0), F_1^p(y_1)\} - F_1^p(y_0)F_1^p(y_1)
\end{aligned}$$

We let $\mathbf{H}$ denote the covariance kernel of the randomization process with elements

$$\begin{aligned}
H_{00}(y_0, y_0') &= \lim_n n\mathrm{Cov}(\hat{F}_0(y_0), \hat{F}_0(y_0')) = \left(\frac{1}{1-p} - \frac{n}{N}\right)\left(\min\{F_0^p(y_0), F_0^p(y_0')\} - F_0^p(y_0)F_0^p(y_0')\right) \\
H_{01}(y_0, y_1) &= \lim_n n\mathrm{Cov}(\hat{F}_0(y_0), \hat{F}_1(y_1')) = \lim_n \frac{n}{N}\left(F_{01}^p(y_0, y_1) - F_0^p(y_0)F_1^p(y_1)\right) \quad \text{(B.1)} \\
H_{11}(y_1, y_1') &= \lim_n n\mathrm{Cov}(\hat{F}_1(y_1), \hat{F}_1(y_1')) = \left(\frac{1}{p} - \frac{n}{N}\right)\left(\min\{F_1^p(y_1), F_1^p(y_1')\} - F_1^p(y_1)F_0^p(y_1')\right)
\end{aligned}$$

Note also that $\frac{1}{1-p} \ge 1 \ge \frac{n}{N} \ge 0$ and $\frac{1}{p} \ge 1 \ge \frac{n}{N} \ge 0$, so that $H_{00}(\cdot,\cdot)$ and $H_{11}(\cdot,\cdot)$ are nonnegative.

B.2. Proofs for Section 2.5.

B.2.1. *Least Favorable Coupling for the Average Treatment Effect.* We first prove a more general result than Proposition 2.1 by showing that the isotone coupling of potential outcomes in fact results in a distribution for the ATE parameter which is dominates by that under any other coupling in the sense of second-order stochastic dominance (SOSD):

Lemma B.1. ***(Ordering of Distributions)*** *Let F_{01} be an arbitrary joint distribution with marginal distributions F_0 and F_1, and let $F_{01}^{iso} := C^{iso}(F_0, F_1)$ be the joint distribution under the isotone coupling. Then for any convex function, the randomization distribution for $\hat{\tau}_{ATE}$ satisfies*

$$\mathbb{E}_{F_{01}^{iso}}[v(\hat{\tau}_{ATE})] \ge \mathbb{E}_{F_{01}}[v(\hat{\tau}_{ATE})]$$

For any strictly convex function $v(\cdot)$ this inequality is strict whenever $F_{01} \neq F_{01}^{iso}$.

This result is a straightforward consequence of the familiar observation that the isotone (assortative) coupling of potential outcomes results in the distribution of $Y_i(0)+Y_i(1)$ which is dominated according to second-order stochastic dominance by the distribution resulting from any other coupling (see e.g. [1], [6], and [8]). For illustrative purposes, we give a complete proof here.

Proof: In order to establish second-order stochastic dominance with respect to the isotone assignment $Y_i(1) = F_1^{-1}(F_0(Y_i(0)))$, consider the expectation of $v(\hat{\tau}_{ATE})$ for any convex function $v(u)$. Note that for any pair of observations i, j we can write

$$\hat{\tau}_{ATE} = \frac{1}{n}\left(B_{-ij} + R_iW_i\left(Y_i(0)/(1-p) + Y_i(1)/p\right) + R_jW_j\left(Y_j(0)/(1-p) + Y_j(1)/p\right)\right)$$

where $B_{-ij} := \sum_{k\neq i,j} R_k\left(W_k(Y_k(0)/(1-p) + Y_k(1)/p) - Y_k(0)/(1-p)\right) - (R_iY_i(0) + R_jY_j(0))/(1-p)$.

We can now consider the change in $\mathbb{E}[v(\hat{\tau}_{ATE})]$ from pairwise substitutions of potential outcomes between units i and j. Specifically suppose that under the initial coupling, the potential outcomes for unit i are given by $Y_i(0), Y_i(1)$, and the potential outcomes for unit j are $Y_j(0), Y_j(1)$. We then consider the effect of switching the assignment to potential outcomes $Y_i(0), Y_j(1)$ for unit i, and potential outcomes $Y_j(0), Y_i(1)$ for unit j.

Since W_i, W_j are independent of W_k, that change leads to an increase in $\mathbb{E}[v(\hat{\tau}_{ATE})]$ if and only if

$$\begin{aligned} 0 \quad \leq \quad & P(W_i = 1, W_j = 0)\left\{v\left(\frac{1}{n}\left(B_{-ij} + Y_i(0)/(1-p) + Y_j(1)/p\right)\right) - v\left(\frac{1}{n}\left(B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p\right)\right)\right\} \\ & + P(W_i = 0, W_j = 1)\left\{v\left(\frac{1}{n}\left(B_{-ij} + Y_j(0)/(1-p) + Y_i(1)/p\right)\right) - v\left(\frac{1}{n}\left(B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p\right)\right)\right\} \\ = \quad & p(1-p)\left\{ v\left(\frac{1}{n}\left(B_{-ij} + Y_i(0)/(1-p) + Y_j(1)/p\right)\right) + v\left(\frac{1}{n}\left(B_{-ij} + Y_i(0)/(1-p) + Y_j(1)/p\right)\right)\right. \\ & \left. - v\left(\frac{1}{n}\left(B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p\right)\right) - v\left(\frac{1}{n}\left(B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p\right)\right)\right\} \end{aligned}$$

for any pair of observations with $R_i = R_j = 1$. Noting that for any convex function $v(\cdot)$, $v(b + x_0 + x_1)$ is supermodular in $x = (x_0, x_1)'$, this difference is nonnegative whenever $Y_i(0) - Y_j(0)$ and $Y_i(1) - Y_j(1)$ have the opposite sign. Furthermore, if in addition $v(\cdot)$ is strictly convex, the first inequality is strict.

Since any coupling of potential outcomes can be obtained from the isotone assignment by pairwise substitutions of this form, the isotone assignment maximizes the expectation

$$\mathbb{E}[v(\hat{\tau}_{ATE})] = \mathbb{E}\left[v\left(\frac{1}{n}\sum_{i=1}^{N} R_i\left\{W_iY_i(1)/p - (1-W_i)Y_i(0)/(1-p)\right\}\right)\right]$$

for all convex functions $v(\cdot)$. Therefore the distribution of $\hat{\tau}_{ATE}$ under the isotone assignment dominates that under any alternative coupling, as claimed above. □

Proof of Proposition 2.1:. The claim in the proposition follows immediately from Lemma B.1 and the observation that the function $v(y) = y^2$ is strictly convex □

B.3. Proofs for Section 5.

B.3.1. *Proof of Theorem 5.1.* From standard results (see e.g. Example 19.6 in [9]), the class $\mathcal{F} := \{(-\infty, y] : y \in \mathbb{R}\}$ is Glivenko-Cantelli, so that $(\hat{F}_0 - F_0^p, \hat{F}_1 - F_1^p)$ converges to zero almost surely as an element of the space of bounded functions on $\mathbb{R}$. Since Assumption 2.1 is sufficient to guarantee that the functionals $\tau(F_0, F_1)$ and $\sigma(F_0, F_1)$, are continuous in F_0, F_1, the claim of the Theorem follows immediately from the continuous mapping theorem (see e.g. Theorem 18.11 in [9]) □

For the proof of Theorem 5.2, we need to characterize functional convergence of the randomization process. To that end, we first introduce some standard notation from empirical process theory (see [10]). Let $\mathcal{F} := \{\mathbb{1}\{y \leq (-\infty, t]\} : t \in \mathbb{R}\}$ be the class of indicator functions for the left-open half-lines on $\mathbb{R}$ and let $\ell^\infty(\mathcal{F})$ be the space of bounded functions from $\mathcal{F}$ to $\mathbb{R}$ endowed with the norm $\|z\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |z(f)|$. Also, let BL_1 denote the set of all functions $h : \ell^\infty\mathcal{F} \mapsto [0, 1]$ with $|h(z_1) - h(z_2)| \leq \|z_1 - z_2\|_{\mathcal{F}}$.

Lemma B.2. *Suppose that $(Y_i(0), Y_i(1)) \overset{iid}{\sim} F_{01}$. Then the randomization process*

$$\hat{G}_n := \sqrt{n}\begin{pmatrix} \hat{F}_0 - F_0^p \\ \hat{F}_1 - F_1^p \end{pmatrix}$$

converges in outer probability to $\mathbb{G}$ under the bounded Lipschitz metric,

$$\sup_{h\in BL_1} |\mathbb{E}_W h(\hat{G}_n) - \mathbb{E}h(\mathbb{G})| \to 0$$

in outer probability, where $\mathbb{G}$ is a Gaussian process with covariance kernel $\mathbf{H}$.

Proof: As before, denote the joint c.d.f. of potential outcomes (observed and counterfactuals) for the n units included in the sample with

$$F_{01}^s(y_0, y_1) := \frac{1}{n}\sum_{i=1}^N R_i \mathbb{1}\{Y_i(0) \le y_0, Y_i(1) \le y_1\}$$

and the empirical c.d.f. among the units included in the sample for which $W_i = 1$,

$$F_{01}^t(y_0, y_1) := \frac{1}{np}\sum_{i=1}^N R_i W_i \mathbb{1}\{Y_i(0) \le y_0, Y_i(1) \le y_1\}$$

Using this notation we can write

$$\sqrt{n}(F_{01}^t(y_0, y_1) - F_{01}^p(y_0, y_1)) = \sqrt{n}(F_{01}^t(y_0, y_1) - F_{01}^s(y_0, y_1)) + \sqrt{n}(F_{01}^s(y_0, y_1) - F_{01}^p(y_0, y_1))$$

Since R_i, W_i are drawn at random and without replacement, it follows from Theorem 3.1 in [3] that

$$\begin{aligned} \sqrt{n}(F_{01}^t(y_0, y_1) - F_{01}^s(y_0, y_1)) &\rightsquigarrow \mathbb{G}_{F_{01}^s} \\ \sqrt{n}(F_{01}^s(y_0, y_1) - F_{01}^p(y_0, y_1)) &\rightsquigarrow \mathbb{G}_{F_{01}^p} \end{aligned}$$

for Brownian bridges $\mathbb{G}_{F_{01}^s}$ and $\mathbb{G}_{F_{01}^p}$. Since for any joint distribution $F_{01}(y_0, y_1)$ the marginals satisfy $\lim_{y_1 \to \infty} F_{01}(y_0, y_1) = F_0(y_0)$ for each y_0, weak convergence of the joint process implies weak convergence of the marginal empirical processes,

$$\begin{aligned} \sqrt{n}(F_0^t - F_0^p) &\rightsquigarrow \mathbb{G}_{F_0^s} + \mathbb{G}_{F_0^p} \\ \sqrt{n}(F_1^t - F_1^p) &\rightsquigarrow \mathbb{G}_{F_1^s} + \mathbb{G}_{F_1^p} \end{aligned}$$

Finally, $\hat{F}_1(y_1) \equiv F_1^t(y_1)$ and $\hat{F}_0(y_0) \equiv \frac{1}{(p-1)}(F_0^s(y_0) - pF_0^t(y_0))$, establishing the claim, where the structure of the covariance kernel follows from the point-wise calculations in the derivation of (B.1) □

B.3.2. *Proof of Theorem 5.2:.* From Assumption 2.1 it is immediate that $\tau(F_0, F_1)$ is Hadamard-differentiable. Lemma B.2 and the functional delta method, see e.g. Theorem 20.8 in [9], then imply asymptotic normality of $\sqrt{n}(\hat{\tau} - \tau)/\sigma(F_0, F_1)$. Theorem 5.2 then follows from Slutsky's theorem and consistency of $\hat{\sigma}$ from Theorem 5.1 □

We next turn to the bootstrap distribution: Consider the bootstrap replicates

$$\hat{F}_0^*(y_0) := \frac{1}{n(1-p)}\sum_{i=1}^n R_i^*(1 - W_i^*)\mathbb{1}\{Y_i^*(0) \le y_0\}, \qquad \hat{F}_1^*(y_1) := \frac{1}{np}\sum_{i=1}^n R_i^* W_i^* \mathbb{1}\{Y_i^*(1) \le y_1\}$$

by randomizing from $\hat{F}_{01}$. We also define the asymptotic covariance kernel $\mathbf{H}^{iso}$ corresponding to the coupling C^{iso} in analogy to (B.1) where F_{01} is replaced with $C^{iso}(F_0, F_1)$. We first show the two following Lemmas:

Lemma B.3. *Suppose that* $(Y_i(0), Y_i(1)) \overset{iid}{\sim} F_{01}$. *Then for any copula* $C : [0,1]^2 \to [0,1]$,

$$\sup_{y_0, y_1 \in \mathbb{R}} \left| C(\hat{F}_0, \hat{F}_1)(y_0, y_1) - C(F_0^p, F_1^p)(y_0, y_1) \right| \overset{a.s.}{\to} 0$$

Proof: From standard results, the class $\mathcal{F} := \{(-\infty, y] : y \in \mathbb{R}\}$ is Glivenko-Cantelli, so that $(\hat{F}_0 - F_0^p, \hat{F}_1 - F_1^p)$ converges to zero almost surely as an element of the space of bounded functions on $\mathbb{R}$. Noting that any copula $C : [0,1]^2 \to [0,1]$ is a bounded nondecreasing function in each of its arguments, it follows that

$$\sup_{y_0, y_1 \in \mathbb{R}} \left| C(\hat{F}_0, \hat{F}_1)(y_0, y_1) - C(F_0^p, F_1^p)(y_0, y_1) \right| \overset{a.s.}{\to} 0$$

establishing the claim □

Lemma B.4. *Suppose that* $(Y_i(0), Y_i(1)) \overset{iid}{\sim} F_{01}$. *Then the bootstrap process*

$$\hat{G}_n^* := \sqrt{n}\begin{pmatrix} \hat{F}_0^* - \hat{F}_0 \\ \hat{F}_1^* - \hat{F}_1 \end{pmatrix}$$

converges in outer probability to $\mathbb{G}$ *under the bounded Lipschitz metric, that is*

$$\sup_{h \in BL_1} \left| \mathbb{E}_W h(\hat{G}_n^*) - \mathbb{E}h(\mathbb{G}) \right| \quad \to \quad 0$$

in outer probability, where $\mathbb{G}$ *is a Gaussian processes with covariance kernel* $\mathbf{H}$.

PROOF: By construction of the coupling $(Y_i^*(0), Y_i^*(1))$, the marginal distributions of $Y_i^*(0)$ and $Y_i^*(1)$ are equal to $\hat{F}_0$ and $\hat{F}_1$, respectively. By construction of the bootstrap, the bootstrap replications $\hat{F}_0^*, \hat{F}_1^*$ are generated by randomization from the samples $(\tilde{Y}_i(1), \tilde{Y}_i(1))_{i=1}^n$ corresponding to the joint distribution $\hat{F}_{01} := C^{iso}(\hat{F}_0, \hat{F}_1)$.

Now let $\hat{\mathbf{H}}^{iso}$ the covariance kernel obtained from (B.1) replacing F_0 with $\hat{F}_0$, F_1 with $\hat{F}_1$, and F_{01} with $C^{iso}(\hat{F}_0, \hat{F}_1)$, respectively. By construction, the bootstrap distribution of $\hat{G}_n^*$ conditional on $\hat{F}_0, \hat{F}_1$ have covariance given by $\hat{\mathbf{H}}^{iso}$. Finally, $\hat{\mathbf{H}}^{iso}$ is a continuous function of $C^{iso}(\hat{F}_0, \hat{F}_1)$. Hence by Lemma B.3 and the continuous mapping theorem we have that

$$\|\hat{\mathbf{H}}^{iso} - \mathbf{H}^{iso}\| \overset{a.s.}{\to} 0$$

which completes the proof.

The claim of the Lemma then follows from the same arguments as in Lemma B.2 and the continuous mapping theorem □

B.3.3. *Proof of Theorem 5.3:.* It follows from Assumption 2.1 that $\tau(F_0, F_1), \sigma(F_0, F_1)$ are Hadamard differentiable, so that Theorem B.4 follows from Lemma B.4 and the functional Delta method (e.g. Theorem 20.8 in [9]) □

B.4. Proofs for Section 6.

B.4.1. *Proof of Proposition 6.1..* To derive the least favorable coupling given covariates X_i, notice that by the conditional variance (ANOVA) identity,

$$\mathrm{Var}(W_i(Y_i(0) + Y_i(1))) = \mathbb{E}[\mathrm{Var}(W_i(Y_i(0) + Y_i(1))|X_i)] + \mathrm{Var}(\mathbb{E}[W_i(Y_i(0) + Y_i(1))|X_i])$$

Since the conditional mean $\mathbb{E}[Y_i(0)+Y_i(1)|X_i, W_i]$ is invariant to the coupling, the unconditional variance of $W_i(Y_i(0)+Y_i(1))$ is maximized by the conditional copula $C_X(\cdot|x)$ which maximizes the conditional variance of $Y_i(0) + Y_i(1)$. Hence, applying Proposition 2.1 to the conditional distribution of $Y_i(0) + Y_i(1)$ given X_i, it follows that the least favorable coupling in $\mathcal{C}_X$ is the isotone coupling,

$$C_X(u, v|x) := \min\{u, v\}, \quad \text{for all } x$$

By the same line of reasoning, the least favorable coupling in the set $\mathcal{C}_B$ is the isotone coupling conditional on $B_i = b(X_i)$, which by Lemma 6.1 yields a randomization distribution with larger asymptotic variance than any joint distribution $F_{01}(y_0, y_1) := C_X[F_0, F_1]$, which establishes the claim □

B.4.2. *Proof of Theorem 6.1.* The proofs of Theorems 5.1-5.3 followed from Lemmas B.2-B.4 and the observation that for the completely randomized case, the functional $\sigma(F_0, F_1)$ is Hadamard-differentiable given the assumptions. By inspection, the remaining steps go through unchanged for the observational case as well.

For the extension of Lemma B.2 to the joint distribution of $Y_i(0)$ ($Y_i(1)$, respectively), $b(X_i)$, and $e(X_i)$, notice first that Theorem 3.1 in [3] holds for multivariate distributions of arbitrary (finite) dimension. Specifically, if we let $F_{01}^p(y_0, y_1, b, e)$ and $F_{01}^s(y_0, y_1, b, e)$ denote the joint empirical c.d.f. for $(Y_i(0), Y_i(1), b(X_i), e(X_i))$ in the population and the sample, respectively, we have that

$$\sqrt{n}(F_{01}^s(y_0, y_1, b, e) - F_{01}^p(y_0, y_1, b, e)) \rightsquigarrow \mathbb{G}_{F_{01}^p}$$

where $\mathbb{G}_{F_{01}^p}$ is a Brownian bridge with covariance kernel depending on F_{01}^p.

Next, let $F_{01}^s(y_0, y_1, b, e)$ be the empirical c.d.f. of $(Y_i(0), Y_i(1), b(X_i), e(X_i))$ among the units with $R_i = W_i = 1$. Since W_i are independent draws from a Bernoulli distribution with conditional success probabilities $e(X_i)$, we can use a conditional multiplier CLT, see e.g. Theorem 2.9.7. in [10] to obtain

$$\sqrt{n}(F_{01}^t(y_0, y_1, b, e) - F_{01}^s(y_0, y_1, b, e)) \rightsquigarrow \mathbb{G}_{F_{01}^s}$$

for a Brownian bridge $\mathbb{G}_{F_{01}^s}$. By Assumption 6.1, the components $(F_{01}^s(y_0, y_1, b, e) - F_{01}^p(y_0, y_1, b, e)$ and $F_{01}^t(y_0, y_1, b, e) - F_{01}^s(y_0, y_1, b, e)$ are uncorrelated, so that $\mathbb{G}_{F_{01}^s}$ and $\mathbb{G}_{F_{01}^p}$ are independent.

It then follows that the randomization process

$$\hat{G}_n := \sqrt{n}\begin{pmatrix} \hat{F}_0(y_0, b) - F_0^p(y_0, b) \\ \hat{F}_1(y_1, b) - F_1^p(y_1, b) \end{pmatrix}$$

converges in outer probability to $\mathbb{G}$ under the bounded Lipschitz metric,

$$\sup_{h \in BL_1} |\mathbb{E}_W h(\hat{G}_n) - \mathbb{E}h(\mathbb{G})| \to 0$$

in outer probability, where $\mathbb{G}$ is a Gaussian process with a covariance kernel $\mathbf{H}$ depending on the joint distribution $F_{01}(y_0, y_1, b)$ that is derived in analogy to the calculations in Appendix B.1.

We next show that we can extend Lemma B.3 to the observational case, that is

$$\sup_{y_0,y_1\in\mathbb{R}} \left| C_B[\hat{F}_0, \hat{F}_1](y_0, y_1) - C_B[F_0^p, F_1^p](y_0, y_1) \right| \overset{a.s.}{\to} 0$$

Here the main difference to the completely randomized case is that the arguments F_0, F_1 are the joint distributions of $b(X_i)$ with $Y_i(0)$ and $Y_i(1)$, respectively, whereas the coupling $C_B[F_0, F_1]$ is defined in terms of the conditional distributions given $b(X_i)$. Under the conditions in Assumption 6.4 it follows from standard arguments that $\hat{F}_0(y_0|b)$ and $\hat{F}_1(y_1|b)$ are uniformly consistent for the population distributions $F_0(y_0|b)$ and $F_1(y_1|b)$, respectively. In particular, under Assumption 6.4, the denominator of the conditional probability, $F_B(b)$, is bounded away from zero with probability approaching one everywhere on the support of $b(X_i)$. The conclusion of Lemma B.3 then follows from the dominated convergence theorem for the integral over $F_B(B_i)$, noticing that the isotone copula $C_B^{iso}(F_0, F_1)$ is continuous in its arguments.

Given Lemma B.3, the conclusion of Lemma B.4 then follow using the same steps. In particular, the bootstrap process

$$\hat{G}_n^* := \sqrt{n} \left(\begin{array}{c} \hat{F}_0^*(y_0, b) - \hat{F}_0(y_0, b) \\ \hat{F}_1^*(y_1, b) - \hat{F}_1(y_1, b) \end{array} \right)$$

converges in outer probability to $\mathbb{G}^*$, a Brownian bridge with covariance kernel $\mathbf{H}^*$, where $\mathbf{H}^*$ is the covariance kernel derived in Appendix B.1 corresponding to the joint distribution $F_{01} \equiv C_B[F_0, F_1]$. The conclusions of Theorem 6.1 then follows from Lemma 6.1 which establishes that the variance bound $\sigma_B^2(F_0, F_1)$ is conservative for $\sigma^2(F_{01})$ □

REFERENCES

[1] Becker, G. (1973): “A Theory of Marriage, Part I,” *Journal of Political Economy*, 81(4), 813–846.

[2] Bentkus, V., F. Götze, and W. van Zwet (1997): “An Edgeworth Expansion for Symmetric Statistics,” *Annals of Statistics*, 25(2), 851–896.

[3] Bickel, P. (1969): “A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case,” *Annals of Mathematical Statistics*, 40(1), 1–23.

[4] Bloznelis, M., and F. Götze (2001): “Orthogonal Decomposition of Finite Population Statistics and its Applications to Distributional Asymptotics,” *Annals of Statistics*, 29(3), 899–917.

[5] ——— (2002): “An Edgeworth Expansion for Symmetric Finite Population Statistics,” *Annals of Probability*, 30(3), 1238–1265.

[6] Fan, Y., and S. Park (2010): “Sharp Bounds on the Distribution of Treatment Effects and their Statistical Inference,” *Econometric Theory*, 26(3), 931–951.

[7] Imbens, G., and K. Menzel (2021): “A Causal Bootstrap,” *Annals of Statistics*, forthcoming.

[8] Stoye, J. (2010): “Partial Identification of Spread Parameters,” *Quantitative Economics*, 1, 323–357.

[9] van der Vaart, A. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[10] van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer, New York.

Stanford Graduate School of Business
and Department of Economics,
655 Knight Way
Stanford, California 94305, USA
imbens@stanford.edu

New York University, Department of Economics
19 W4 St, 6th Floor
New York, NY 10012, USA
km125@nyu.edu