A CAUSAL BOOTSTRAP

BY GUIDO IMBENS† AND KONRAD MENZEL‡

Stanford GSB† and New York University‡

The bootstrap, introduced by Efron [14], has become a very popular method for estimating variances and constructing confidence intervals. A key insight is that one can approximate the properties of estimators by using the empirical distribution function of the sample as an approximation for the true distribution function. This approach views the uncertainty in the estimator as coming exclusively from sampling uncertainty. We argue that for causal estimands the uncertainty arises entirely, or partially, from a different source, corresponding to the stochastic nature of the treatment received. We develop a bootstrap procedure for inference regarding the average treatment effect that accounts for this uncertainty, and compare its properties to that of the classical bootstrap. We consider completely randomized and observational designs as well as designs with imperfect compliance.

1. Introduction. The bootstrap, introduced by Efron [14], has become a very popular method for constructing hypothesis tests or confidence intervals. This popularity stems in part from the fact that it provides approximations to the distribution of an estimator or statistic that are in certain cases superior to those obtained from using a Gaussian asymptotic approximation together with estimated standard errors (asymptotic refinement). While the classical bootstrap is designed to approximate distributions that result from repeated sampling from a large population, this paper shows how to adapt the bootstrap principle when the estimand of interest is a causal parameter, and the data is generated by a randomized experiment. We also consider observational studies and designs with imperfect compliance when the population of interest may be finite.

Permutation tests, such as Fisher’s exact test (see e.g. [28]), can yield exact p-values under the auxiliary hypothesis that treatment effects are constant across units, however we argue below that those methods are not suitable for forming confidence intervals for parameters describing the distribution of causal effects in a given population. For the average treatment effect, causal standard errors have been proposed by [3], as well as [1]. These methods impose no restrictions on treatment effect heterogeneity but their use generally relies on a Gaussian limiting approximation. We propose a bootstrap approach to causal inference which also does not restrict treatment heterogeneity, but improves on the Gaussian asymptotic approximation in samples of small or moderate size.

Using the potential outcome framework, e.g., [28], we are interested in the average causal effect of a binary variable $W_i \in \{0, 1\}$ (the “treatment”) on an outcome variable whose potential outcomes we denote with $Y_i(0), Y_i(1)$, for a population of $N$ units $i = 1, \ldots, N$. Implicitly we assume that the potential outcomes $Y_i(w)$ for unit $i$ do not vary with the treatment status assigned to other units, known as the Stable Unit Treatment Value Assumption (SUTVA, Rubin [39]).

The researcher has a random sample from that population, where $R_i \in \{0, 1\}$ denotes an indicator whether the $i$th unit is included in the sample. For all units in the sample we observe the treatment $W_i$ and the realized outcome $Y_i := Y_i(W_i)$. One common estimand is the average effect for the $N$
units in the population:

\[ \tau_{ATE} := \frac{1}{N} \sum_{i=1}^{N} \left( Y_i(1) - Y_i(0) \right). \]

To illustrate the main idea, we first assume that the data arise from a completely randomized experiment, where \( n \leq N \) units are selected at random from the population as experimental subjects, of which \( n_1 \) units are then randomly assigned to receive the treatment, and the remaining \( n_0 = n - n_1 \) units are assigned to the control group. We later provide versions of the bootstrap for observational designs in Section 6 and designs with imperfect compliance in Appendix A.

Specifically, we assume that the population of interest in turn consists of \( N \) i.i.d. draws from an encompassing distribution \( F_{01} \).

**Assumption 1.1.** (Sampling Experiment) The population consists of \( N \) units with potential values \((Y_i(0), Y_i(1))_{i=1}^{N}\) which are i.i.d. draws from the distribution \( F_{01}(y_0, y_1) \). The \( n \) observed units are sampled at random and without replacement from the population,

\[ Y_i(0), Y_i(1) \perp \perp R_i \]

where we denote \( q := \frac{n}{N} \in (0, 1] \).

The main role of the superpopulation distribution \( F_{01}(y_0, y_1) \) in our argument is to impose certain regularity conditions on the asymptotic sequence in \( N \) which are analogous to those assumed in [21] and [30].

We assume throughout that the treatment \( W_i \in \{0, 1\} \) is binary, and that the outcome \( Y_i(W_i) \) for unit \( i \) does not vary with the treatment status assigned to other units. The latter requirement is also known as individualistic treatment response, or Stable Unit Treatment Value Assumption (SUTVA). Furthermore, we initially assume that the experiment is completely randomized (see Section 6 for extensions to settings with unconfounded treatment assignment):

**Assumption 1.2.** (Complete Randomization) Treatment assignment is completely randomized, that is for each unit with \( R_i = 1 \) we have

\[ (Y_i(0), Y_i(1)) \perp \perp W_i \]

where \( W_i = 1 \) for \( n_1 \) units selected at random and without replacement from the \( n \) observations with \( R_i = 1 \), and the propensity score \( p := \frac{n_1}{n} \) satisfies \( 0 < p < 1 \).

For greater clarity of exposition we also assume at first that the researcher observes no further covariate information. The approach of this paper can be generalized to observational studies under unconfoundedness, and experiments with imperfect compliance for which unconfoundedness fails, but intention to treat is (conditionally) independent of potential outcomes and can serve as an instrumental variable to identify causal effects on a population of compliers, see Section 6 and Appendix A.

1.1. Sampling Uncertainty and Design Uncertainty. As in [1] we wish to distinguish explicitly between two types of uncertainty in estimators, sampling uncertainty arising from the stochastic nature of \( R := (R_1, \ldots, R_N)' \), and design uncertainty arising from the stochastic nature of \( W := (W_1, \ldots, W_n)' \).
To characterize sampling uncertainty we postulate the existence of a large, possibly infinite, population. We draw a random sample from this population, and observe for each unit in this sample a set of values, say, a pair \((Y_i, W_i)\). We may be interested in the difference between the population averages of \(Y_i\) for the subpopulations with \(W_i = 0\) and \(W_i = 1\). We can estimate this object using the difference in average outcomes by \(W_i\) values in the sample. This estimator differs from the target because we do not observe all units in the population. Had we drawn a different random sample, with different units, the value of the estimator would have been different. See Table 1, where \(R_i\) is the sampling indicator, equal to 1 for sampled units and 0 otherwise. The uncertainty arising from the randomness in \(R\) is captured by the conventional standard error.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Actual Sample</th>
<th>Alternative Sample I</th>
<th>Alternative Sample II</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y_i)</td>
<td>(W_i)</td>
<td>(R_i)</td>
<td>(Y_i)</td>
</tr>
<tr>
<td>1</td>
<td>9.1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>-3.6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>(N)</td>
<td>?</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>

In a randomized experiment the uncertainty is not necessarily only of this sampling variety. Instead we can think of the uncertainty arising from the stochastic nature of the assignment, \(W\). For units with \(W_i = 0\) we observe \(Y_i(0)\), and for units with \(W_i = 1\) we observe \(Y_i(1)\). In our sample units have a particular set of assignments. In a repeated sampling thought experiment the units in the sample would have remained the same, but their assignments could have been different, leading to a different value for the estimator. See Table 2. In general we can have both the types of uncertainty present.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Actual Sample</th>
<th>Alternative Sample I</th>
<th>Alternative Sample II</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y_i(1))</td>
<td>(Y_i(0))</td>
<td>(W_i)</td>
<td>(Y_i(1))</td>
</tr>
<tr>
<td>1</td>
<td>3.2</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-1.6</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>2.3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td>-3.1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>(N)</td>
<td>-5.7</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>

1.2. Notation. In the following, we denote the distribution of potential outcomes in the population with \(F_{01}^P(y_0, y_1) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}\) and the size of that population with
The distribution in the sample of size \( n \) is denoted with \( F_{01}^s(y_0, y_1) := \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\} \). \( F_{01}^s(\cdot, \cdot) \) is not quite the empirical distribution because we only observe one of the values in each pair \((Y_i(0), Y_i(1))\). We use \( p \) superscripts throughout to indicate population quantities, and \( s \) superscripts to denote their sample analogs. The number of treated units in the sample is denoted with \( n_1 \), the number of control units with \( n_0 = n - n_1 \), and the respective shares of treated and control units with \( p := n_1/n \), so that \( 1 - p = n_0/n \). We also define the empirical c.d.f. for either potential outcome given the randomized treatment as \( \hat{F}_0(y_0) := \frac{1}{n_0} \sum_{i=1}^{N} R_i (1 - W_i) \mathbb{1}\{Y_i(0) \leq y_0\} \) and \( \hat{F}_1(y_1) := \frac{1}{n_1} \sum_{i=1}^{N} R_i W_i \mathbb{1}\{Y_i(1) \leq y_1\} \).

**2. The Causal Bootstrap for Average Treatment Effects.** In this section we consider causal bootstrap inference for the population average treatment effect \( \tau_{ATE} \) defined in (1.1). The estimator we use is the difference in sample averages by treatment status:

\[
\hat{\tau}_{ATE} := \bar{Y}_1 - \bar{Y}_0,
\]

where

\[
\bar{Y}_1 := \frac{1}{n_1} \sum_{i=1}^{N} R_i W_i Y_i, \quad \text{and} \quad \bar{Y}_0 := \frac{1}{n_0} \sum_{i=1}^{N} R_i (1 - W_i) Y_i.
\]

The repeated sampling perspective we take is one where the potential outcomes \((Y_i(0), Y_i(1))\) are fixed for all \( N \) units in the population. The stochastic properties of the estimator arise from the stochastic nature of the assignment and sampling, which are both sources of randomness in the average of realized outcomes by treatment status, where we regard \( n, n_0, n_1 \) as fixed.

**2.1. The True Variance of the Estimator for the Average Treatment Effect.** Here we present the true variance of the estimator \( \hat{\tau}_{ATE} \) under random assignment of the treatment. From the \( n \) experimental subjects, \( n_1 \) are selected at random to receive the active treatment, and the remainder are assigned to the control group. Define

\[
\bar{Y}(0) = \frac{1}{N} \sum_{i=1}^{N} Y_i(0), \quad \bar{Y}(1) = \frac{1}{N} \sum_{i=1}^{N} Y_i(1),
\]

\[
S_0^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(0) - \bar{Y}(0))^2, \quad S_1^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(1) - \bar{Y}(1))^2,
\]

and

\[
S_{01}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(1) - Y_i(0) - \tau_{ATE})^2.
\]

Then the exact variance of \( \hat{\tau} \), over the randomization distribution, is

\[
\mathbb{V}(\hat{\tau}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{N}.
\]

See, for example, [36], [3], [13], and [1].
2.2. An Analytical Variance Estimator. Define

\[
\hat{S}_0^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{N} R_i (1 - W_i) (Y_i - \overline{Y}_0)^2, \quad \text{and} \quad \hat{S}_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{N} R_i W_i (Y_i - \overline{Y}_1)^2.
\]

Then the standard variance estimator is

\[
\hat{V}_{\text{Neyman}} := \frac{\hat{S}_0^2}{n_0} + \frac{\hat{S}_1^2}{n_1}.
\]

This estimator ignores the third term in the variance, which is negative, so \(\hat{V}_{\text{Neyman}}\) in general overestimates the true variance. It is possible to give sharp bounds for \(S_{01}^2\) given the respective marginal distributions of \(Y_i(0)\) and \(Y_i(1)\). [3] proposed a consistent estimator for the resulting bounds on \(V(\hat{\tau})\) that can be expressed as

\[
\hat{V}_{\text{AGL}} := \frac{\hat{S}_0^2}{n_0} + \frac{\hat{S}_1^2}{n_1} - \frac{\hat{S}_{01}^2}{N}
\]

where \(\hat{S}_{01}^2\) is an estimator of the sharp lower bound for \(S_{01}^2\).\(^1\)

2.3. The Classical Bootstrap. The classical bootstrap corresponds to the case where the uncertainty is purely sampling uncertainty. The bootstrap approximates the cumulative distribution function of the pairs \((Y_i, W_i)\), \(F_{YW}(\cdot, \cdot)\) in the population by the empirical distribution function \(\hat{F}_{YW}(\cdot, \cdot)\), where

\[
\hat{F}_{YW}(w, y) := \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i \leq y, W_i \leq w\}.
\]

It then calculates properties of the estimator given that approximate distribution \(\hat{F}_{YW}(\cdot, \cdot)\). One can interpret the standard bootstrap as imputing all the missing values of \((Y_i, W_i)\) in the population by replicates of the observed values, and thus constructing an artificial population from which we then draw random samples. This perspective is helpful to contrast the different approach underlying the causal bootstrap.

2.4. The Causal Bootstrap. Here we initially take the perspective that the uncertainty is solely arising from the stochastic nature of the assignment, as in Table 2. In the spirit of the above interpretation of the standard bootstrap, we use the observed data to impute all the missing values in the population. Then we simulate the estimator using this partly imputed population.

The difference with the standard bootstrap is in the nature of the missing data process, and how we impute them. Consider unit 1 in Table 2. In the actual sample this unit receives the active treatment, and so we observe \(Y_1(1)\), but we do not know the value of the control outcome for this unit, \(Y_1(0)\).

A natural approach is to impute the missing value of \(Y_1(0)\) using one of the observed values for \(Y_i(0)\), that is, one of the realized values of \(Y_i\) for control units. The question is which one to use. It turns out that it matters how we choose to impute the missing values from the observed values.

This issue is related to the term \(S_{01}^2\) in the true variance of the estimator \(\hat{\tau}\) for the average treatment effect, the term that is not consistently estimable, and which we typically ignore in practice.

\(^1\)Such an estimator is

\[
\hat{S}_{01}^2 := \hat{S}_0^2 + \hat{S}_1^2 - 2\hat{\sigma}_{HN}^2(y_0, y_1)
\]

where \(\hat{\sigma}_{HN}^2(y_0, y_1)\) is a consistent estimator for the upper bound for \(\text{Cov}(Y_i(0), Y_i(1))\), see equation (8) of their paper.
To frame this question, it is useful to start with the joint distribution function of the pairs of potential outcomes in the population,

\[ F_{01}^p(y_0, y_1) := \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}. \]

The average treatment effect, and any other causal parameters of interest, can be written as a functional of this distribution,

\[ \tau := \tau(F_{01}^p). \]

Given \( F_{01}^p \), the assignment mechanism completely determines the distribution of any estimator, for example the difference in averages by treatment status, \( \hat{\tau} \). This is similar to the way in which in the sampling case knowledge of the joint population distribution allows us to deduce the properties of any estimator.

The problem, and the main difference with the sampling case is that for each unit in the population, at most one of the two potential outcomes \( Y_i(0) \) and \( Y_i(1) \) is observed so that there is no consistent estimator for \( F_{01}^p(\cdot, \cdot) \). In general, the joint distribution of potential values can be written as

\[ F_{01}^p(y_0, y_1) = C(F_0^p(y_0), F_1^p(y_1)) \]

where the copula \( C : [0,1]^2 \mapsto [0,1] \) is a function that is nondecreasing in either argument. By Sklar’s theorem (e.g. stated as Theorem 2.3.3 in [35]), such a copula exists even though it need not be unique unless the marginal distributions \( F_0^p, F_1^p \) are continuous. In the following, we let

\[ C := \{ C : [0,1]^2 \mapsto [0,1], C(u,v) \text{ nondecreasing in } u \text{ and } v \} \]

denote the set of all possible copulae.

It is important to note that although the marginal distributions \( F_0^p, F_1^p \) can be estimated consistently from a completely randomized experiment as sample size grows, the data on realized treatments and outcomes impose no empirical restrictions on the copula \( C(u,v) \) for the joint distribution of \( (Y_i(0), Y_i(1)) \). Hence, neither the parameter \( \tau(F_{01}^p) = \tau(C(F_0^p, F_1^p)) \) nor the distribution of an estimator \( \hat{\tau} \) need in general not be point-identified.

In the spirit of the variance estimator in [3], we address this challenge by simulating the distribution of \( \hat{\tau} \) using an estimator for the population distribution \( F_{01}^p \) that is conservative with respect to the copula in a sense to be made more precise below. To illustrate the broader conceptual idea, consider an estimator

\[ \hat{\tau} := \tau(\hat{F}_0, \hat{F}_1) \]

for a general functional \( \tau(F_{01}) \) of the distribution of potential values. Under regularity conditions, such an estimator admits a stochastic expansion of the form

\[ \hat{\tau} - \tau(F_{01}^p) = \mu(F_{01}^p) + n^{-1/2} \sigma(F_{01}^p) Z + n^{-1} \kappa(F_{01}^p) + o_P(n^{-1}) \]

where \( Z \sim N(0,1) \). The first-order “bias” term

\[ \mu(F_{01}^p) := \mathbb{E}_{F_{01}^p}[\hat{\tau}] - \tau(F_{01}^p) \]

and the scale parameter

\[ \sigma^2(F_{01}^p) := \lim_N n \text{Var}_{F_{01}}(\hat{\tau}) \]

See e.g. [8] for regularity conditions for finite-population expansions of this type.
are deterministic functions of the unknown distribution $F_{01}^p = C(F_{00}^p, F_{11}^p)$, and the limit for the asymptotic variance is taken as $N$ and $n := n_N$ grow large. The second-order approximation error $\kappa(F_{01}^p)$ is a tight random variable whose distribution also depends on $F_{01}^p$.

If the functional $\tau(F_{01}^p)$ is not point-identified, then $\mu(F_{01}^p)$ may take values in a set whose bounds may be characterized in terms of the marginal distributions $F_{00}^p$, $F_{11}^p$. Specifically, given the marginal distributions $F_{00}^p$, $F_{11}^p$ we have sharp bounds of the form

\[ (2.1) \quad \mu_L(F_{00}^p, F_{11}^p) := \inf_{\tilde{\mathcal{C}} \subseteq \mathcal{C}} \mu(\tilde{\mathcal{C}}(F_{00}^p, F_{11}^p)) \leq \mu(F_{01}^p) \leq \sup_{\tilde{\mathcal{C}} \subseteq \mathcal{C}} \mu(\tilde{\mathcal{C}}(F_{00}^p, F_{11}^p)) =: \mu_U(F_{00}^p, F_{11}^p), \]

that are generally available, see e.g. [23] and [34]. Similarly we can form bounds for the variance,

\[ (2.2) \quad \sigma_L(F_{00}^p, F_{11}^p) := \inf_{\tilde{\mathcal{C}} \subseteq \mathcal{C}} \sigma(\tilde{\mathcal{C}}(F_{00}^p, F_{11}^p)) \leq \sigma(F_{01}^p) \leq \sup_{\tilde{\mathcal{C}} \subseteq \mathcal{C}} \sigma(\tilde{\mathcal{C}}(F_{00}^p, F_{11}^p)) =: \sigma_U(F_{00}^p, F_{11}^p). \]

For a given inference problem, the bootstrap has to estimate these quantities conservatively with respect to the unknown copula $C(\cdot)$, which can be done iteratively as follows: we first need to determine which couplings $C_0^*$ attain the value of $\mu(C_0^*(F_0, F_1))$ which is least favorable for the inference problem at hand. Within the (not necessarily singleton) set $C_0^*$ of such couplings, we then determine the least-favorable value of $\sigma(C_1^*(F_0, F_1))$ for $C_1^* \in C_0^*$. We can apply this principle recursively either until the resulting set $C_k^*$ contains a unique copula, or until we reach the order of approximation desired for formal results regarding the bootstrap procedure. This results in an estimate $\hat{\theta}_{01}^* := C_k^*(\hat{F}_0, \hat{F}_1)$ for the population distribution $F_{01}^p$ that is conservative regarding the inference task at hand. The causal bootstrap then approximates the distribution of $\hat{\tau}$ by sampling and randomization from a population $\hat{F}_{01}^*$ using the known sampling and assignment mechanism.

2.5. Least Favorable Coupling for the Average Treatment Effect. In this paper, we consider the special case of two-sided confidence intervals based on a t-ratio for the sample average treatment effect. The case of the average treatment effect has been the main focus of the previous literature. It is a special case for our problem in that the copula does not matter for estimation - by inspection, the functional

\[ \tau_{ATE}(F_{01}) = \mathbb{E}_{F_{01}}[Y_i(1)] - \mathbb{E}_{F_{01}}[Y_i(0)] = \mathbb{E}_{F_{11}}[Y_i(1)] - \mathbb{E}_{F_{00}}[Y_i(0)] =: \tau(F_0, F_1) \]

does not depend on the copula, and the default estimator

\[ \hat{\tau}_{ATE} := \tau(\hat{F}_0, \hat{F}_1) \equiv \frac{1}{n_1} \sum_{i=1}^N R_i W_i Y_i - \frac{1}{n_0} \sum_{i=1}^N R_i (1 - W_i) Y_i \]

is known to be unbiased for $\tau(C(F_0, F_1))$ under any coupling so that $\mu(C(F_0, F_1)) \equiv 0$ for each $C$.

In order to ensure that the estimand is well-defined and satisfies other regularity conditions for the bootstrap, we make the following assumption:

**Assumption 2.1.** The first four moments of the respective marginal distributions of $Y_i(0)$ and $Y_i(1)$ are bounded.

Note that this is a condition on the distribution of the infinite superpopulation from which the $N$ units in the population of interest are drawn according to Assumption 1.1, and it imposes a regularity condition on the asymptotic sequence that is analogous to e.g. Condition 1 in [30].
Also for a two-sided confidence interval constructed from inverting a t-test based on \( \hat{\tau}_{ATE} \), the least favorable coupling must attain the upper bound for the asymptotic variance,

\[
\sigma^2_C(F_0, F_1) = \sup_{C \in \mathcal{C}} \sigma^2(C(F_0, F_1)) =: \sigma^2(F_0, F_1)
\]

We next show that \( \sigma^2(C(F_0, F_1)) \) is uniquely maximized at the joint distribution corresponding to the isotone assignment which matches values of \( Y_i(0) \) to values of \( Y_i(1) \) while preserving their respective marginal distributions. More formally, the joint distribution of the potential outcomes under the isotone coupling is characterized by the copula

\[
C^{iso}(u, v) := \min\{u, v\}
\]

We find that the upper bound on the variance is in fact uniquely attained at the isotone coupling. Therefore an estimator for the distribution of \( \hat{\tau}_{ATE} \) which assumes the isotone coupling is asymptotically conservative at any order of approximation.

**Proposition 2.1. (Least Favorable Coupling for the ATE).** Suppose that Assumption 2.1 holds. Then, given the marginal distributions \( F_0, F_1 \), the variance bound is uniquely attained at

\[
\sigma^2(F_0, F_1) := \lim_n \text{Var}_{F_{01}^{iso}}(\hat{\tau})
\]

where \( F_{01}^{iso} := C^{iso}(F_0, F_1) \) is the joint distribution corresponding to the isotone coupling.

The fact that the variance bound is attained at the isotone coupling is widely known (see e.g. [4], [15], [41], and [3]), for expositional purposes we provide a proof in the appendix. In fact, we establish the slightly stronger conclusion that the distribution under the isotone coupling is in fact minimal with respect to second-order stochastic dominance. For our approach it is also important to establish that this maximum is unique in the sense that the joint distribution resulting from any other coupling yields a variance that is strictly lower than \( \sigma^2_C(F_0, F_1) \). In particular, for confidence intervals based on the Gaussian asymptotic distribution, the isotone coupling does indeed constitute the least favorable coupling.

The general idea behind this procedure could in principle also be applied to other causal estimands of interest, noting that in general the functional \( \tau(F_{01}^{iso}) \) need not be fully determined by the marginal distributions \( F_0, F_1 \). Bounds for the distribution function of treatment effects \( Y_i(1) - Y_i(0) \) and its quantiles were given by [23], [17], [15], [32]. In each of those cases, the bounds for the functional in (2.1) do not coincide, and are not attained at the isotone coupling (see e.g. [35] for the least-favorable couplings for quantiles).\(^3\) [11] give conditions under which the isotone assignment does in fact constitute the least favorable bound for a functional of the joint distribution.

**2.6. Related Literature.** Worst case bounds on the distributions of potential outcomes and treatment effects and their quantiles have been analyzed by [23], [34], [19], [15], [16], [17], and [32]. This literature uses theoretical results on dependency bounds for functions of several random variables which were developed among others by [11], [33], [20], and [44]. [41] establishes that a class of spread parameters is monotone with respect to conventional stochastic orders of distribution, and shows how to derive parameter bounds for causal inference. Several of these studies also propose

\(^3\)When only non-trivial bounds for the causal estimand are available, there are additional technical challenges in construction of confidence sets (see e.g. [27] and [40]) and the use of the bootstrap ([18] and [25]). These issues have been addressed in those papers for sampling-based inference, adapting those insights to finite-population causal inference is beyond the scope of this paper.
inference procedures that account for sampling uncertainty rather than randomization error. In contrast, for our problem we need to explicitly construct the respective couplings that achieve the lower and upper bounds to the parameter, and in addition the largest randomization variance for an estimator of either bound.

[37] proposes a confidence interval for a causal parameter based on the least-favorable coupling for a binary outcome variable. [3] propose an estimator of the sharp upper bound for the randomization variance of the average treatment effect in completely randomized experiments. Our approach of embedding the finite-population randomization distribution into an asymptotic sequence of sampling experiments closely follows [1]. Our results make use of a finite-population CLT for the empirical process developed by [6] for the two-sample problem. Finite-sample central limit theorems for randomization inference were also provided by [29]. Bootstrap methods for sampling from finite populations (without replacement) have been proposed by [7] and [10]. For this problem the main challenge in generating the finite bootstrap population is that the size of the super-population \( N \) may be a non-integer multiple of \( n \). We propose a new alternative for estimating the potential outcome distribution for a super-population of exact size \( N \) and for which the marginal distributions coincide with their empirical analogs up to rounding error. Our results for observational studies can also be adapted to nest [21] and [30]'s analysis of regression adjustments in experimental data.

2.7. Relation to Fisher’s Exact Test. Bootstrap inference on the average treatment effect as proposed in this paper bears some conceptual similarities with Fisher’s exact test of the sharp null of no unit-level treatment effect, \( H_0 : Y_i(0) = Y_i(1) \) with probability 1 (see e.g. [38], [28], [13], [31]). One important distinction is that the justification for our procedure is only asymptotic, whereas the Fisher exact test is valid in finite samples.

Furthermore, the Fisher exact test evaluates the randomization distribution of the estimated ATE under the sharp null of no or a constant unit-level treatment effect. The sharp null not only implies that the joint distribution of \( Y_i(0) \) and \( Y_i(1) \) corresponds to the isotone assignment, but also equality of the marginal distributions \( F_0(y) = F_1(y) \), which may in fact be rejected by the data under the null of a zero average treatment effect. In that case even a conservative estimator of the randomization variance may in fact be smaller than that implied by zero, or constant, unit-level effects. More generally, when Fisher’s sharp null fails and \( F_0(y) \neq F_1(y) \), the bootstrap estimate of the randomization variance can in several important scenarios be smaller than that implicit in Fisher’s exact test, in which case our procedure is asymptotically more powerful.

Specifically, standard variance calculations (see e.g. [13]) imply that the implicit variance estimate for Fisher’s exact test under the null of no average effect is

\[
\text{Var}(\hat{\tau}_{\text{ATE}}) = \text{Var}(Y_i) \left( \frac{1}{n_1} + \frac{1}{n_0} \right) = \frac{n_0 S_0^2 + n_1 S_1^2}{n} \left( \frac{1}{n_1} + \frac{1}{n_0} \right).
\]

We can compare this to the actual variance stated in Section 2.1,

\[
\text{Var}(\hat{\tau}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{N}.
\]

Our bootstrap procedure implies a conservative estimate, i.e. a sharp lower bound for \( S_{01}^2 \) from the isotone coupling of the potential outcomes, which is strictly positive whenever the marginal distributions of \( Y_i(0) \) and \( Y_i(1) \) are not the same. The comparison between the terms \( \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} \) and \( \text{Var}(Y_i) \left( \frac{1}{n_1} + \frac{1}{n_0} \right) \) is generally ambiguous - [13] describes several cases in which the randomization variance implied by Fisher’s test is strictly larger, and his conclusions carry over to
the bootstrap procedure in this paper. On the other hand it is important to note that when 
\[ \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} > \text{Var}(Y_i) \left( \frac{1}{n_1} + \frac{1}{n_0} \right), \]
Fisher’s exact test over-rejects under the null hypothesis of no average treatment effect, so that this potential power advantage for the Fisher test only arises in situations in which the exact test does not provide a valid test of that null. For inference regarding functionals of a distribution, [12] showed that this problem can be addressed by studentization of the test statistic, where [31] proposed the use of [3]'s variance estimator for causal inference. Section 4 presents Monte Carlo evidence to illustrate these theoretical findings.

3. Bootstrap Procedure. This section describes the bootstrap procedure for confidence intervals for the average treatment effect, in which case the least favorable coupling is the isotone (rank-preserving) assignment by Proposition 2.1. The method allows for sampling and randomization uncertainty, where we consider a sampling experiment under which the researcher observes \( n \) units that are selected at random out of a population of \( N \) units. This section considers the leading case of completely randomized designs. We later show that the main procedure can be adapted to observational designs and designs with imperfect compliance, see Section 6 and Appendix A.

Given a sample generated according to Assumptions 1.1 and 1.2, we denote the point estimate for the average treatment effect
\[ \hat{\tau} := \tau(\hat{F}_0, \hat{F}_1) \]
and the upper variance bound
\[ \hat{\sigma} := \sigma(\hat{F}_0, \hat{F}_1) \]
which can be evaluated using the formula for \( \hat{V}_{AGL} = \sigma(\hat{F}_0, \hat{F}_1) \) given by [3]. We present the algorithm for the (recommended) pivotal version of the causal bootstrap based on the t-ratio
\[ T := \sqrt{n} \frac{\hat{\tau} - \tau}{\hat{\sigma}}. \]

The bootstrap procedure begins by expanding the sample of size \( n \) into a (potentially larger) empirical population of size \( N \). For each of the resulting \( N \) units, we then impute a value for the missing counterfactual according to the least favorable coupling. For each bootstrap replication, we then sample \( n \) units at random without replacement and assign treatment to \( n_1 \) of those units. Given the resulting sample of potential outcomes that would be observed under that assignment, we then compute the estimator for the average treatment effect and the corresponding t-ratio or other statistic. We can then obtain bootstrap estimates for percentiles or other features of the randomization distribution of that statistic by simulating a large number of bootstrap replications from that empirical population. We next describe each of these steps in greater detail.

3.1. Generating the Empirical Population. If we only observe a random sample of size \( n < N \) from the population of interest, we first need to construct an empirical population \( \left( \hat{Y}_i, \hat{W}_i \right)_{i=1}^N \) of the appropriate size. To this end, we generate replicates of the \( n \) observed units, however not necessarily of the same number for each observation when \( N \) is not an integer multiple of \( n \). When the full population is observed, \( n = N \), the empirical population coincides with the original sample, so that without loss of generality we can skip this step and set \( \hat{Y}_i = Y_i \) and \( \hat{W}_i = W_i \) for \( i = 1, \ldots, N \).

We propose the following procedure for generating the empirical population:

- We let \( Y_0^0 \leq \cdots \leq Y_0^{n_0} \) denote the ordered sample of values of \( Y_i \) for the \( n_0 \) units with \( W_i = 0 \), and \( Y_1^1 \leq \cdots \leq Y_1^{n_1} \) the ordered sample of values \( Y_i \) for the \( n_1 \) units with \( W_i = 1 \).
• Let \(N_0 = \left\lceil \frac{n_0 N}{n} \right\rceil\) and \(N_1 = N - N_0\) and define

\[
M_j^0 := \frac{j}{n_0 N_0} - \frac{j - 1}{n_0 N} \quad M_j^1 := \frac{j}{n_1 N_1} - \frac{j - 1}{n_1 N}
\]

\[(3.3)\]

We then generate the empirical population \((\tilde{Y}_i, \tilde{W}_i)^N_{i=1}\) by including \(M_j^0\) copies of \(Y_j^0\) with \(\tilde{W}_j = 0\) and \(M_j^1\) copies of \(Y_j^1\) with \(\tilde{W}_j = 1\).

By inspection, \(\sum_{j=1}^{n_0} M_j^0 + \sum_{j=1}^{n_1} M_j^1 = [N_0] + [N_1] = N\) so that this procedure ensures that the empirical population has size equal to \(N\). Also, for \(n\) and \(N\) sufficiently large, the respective empirical distributions of \(\tilde{Y}_i\) among units with \(\tilde{W}_i = 0\) and \(\tilde{Y}_i\) among units with \(\tilde{W}_i = 1\) are, equal to \(\hat{F}_0\) and \(\hat{F}_1\), respectively, up to an approximation error of the order \(n^{-1}\).

3.2. Imputing Missing Counterfactuals. For the specific case of two-sided inference for the average treatment effect, Proposition 2.1 shows that the least favorable coupling corresponds to the isotone assignment \(C^{iso}(u, v) := \min\{u, v\}\). For other inference problems, the missing counterfactuals would have to be imputed by drawing from the appropriate least-favorable coupling, following the strategy outlined in Section 2.4.

In order to generate an empirical population with joint distribution \(\hat{F}_{01} := C^{iso}(\hat{F}_0, \hat{F}_1)\), we can simply impute the missing counterfactuals according to:

\[
\tilde{Y}_i(0) := \begin{cases} \tilde{Y}_i & \text{if } \tilde{W}_i = 0 \\ \hat{F}_0^{-1}(\hat{F}_1(\tilde{Y}_i)) & \text{otherwise} \end{cases}
\]

\[
\tilde{Y}_i(1) := \begin{cases} \tilde{Y}_i & \text{if } \tilde{W}_i = 1 \\ \hat{F}_1^{-1}(\hat{F}_0(\tilde{Y}_i)) & \text{otherwise} \end{cases}
\]

\[(3.4)\]

For functionals \(\tau(F_{01})\) of the potential outcome distribution other than the average treatment effect, or inference problems other than two-sided confidence intervals, the least favorable coupling will be of a different form, so this step would have to be replaced by a procedure imputing the missing counterfactuals from a different coupling.

3.3. Resampling Algorithm. For the \(b\)th bootstrap replication, we initially draw \(n\) units \((Y_{ib}^*(0), Y_{ib}^*(1))\) from the empirical population at random and without replacement. We then generate \(W_{ib}^*, \ldots, W_{nb}^*\) by selecting \(n_1\) units from that sample without replacement and setting \(W_{ib}^* = 1\) for those selected units, and \(W_{ib}^* = 0\) for the remaining \(n - n_1\) units. We then obtain the bootstrap sample \(Y_{1b}^*, \ldots, Y_{nb}^*\), where \(Y_{ib}^* := Y_{ib}(W_{ib}^*)\).

We can then compute the bootstrap analogs of the estimated c.d.f.s \(\hat{F}_{0b}^*(y_0) := \frac{1}{n_0} \sum_{i=1}^{N} R_{ib}^*(1 - W_{ib}^*)1\{Y_{ib}^* \leq y_0\}\) and \(\hat{F}_{1b}^* := \frac{1}{n_1} \sum_{i=1}^{n} R_{ib} W_{ib}^*1\{Y_{ib}^* \leq y_1\}\), the corresponding estimates of the average treatment effect, and the variance bound,

\[
\hat{\tau}_b^* := \tau(\hat{F}_{0b}^*, \hat{F}_{1b}^*) = \frac{1}{n_1} \sum_{i=1}^{n} W_{ib}^* Y_{ib}^* - \frac{1}{n_0} \sum_{i=1}^{n} (1 - W_{ib}^*) Y_{ib}^*
\]

\[
\hat{\sigma}_b^* := \sigma(\hat{F}_{0b}^*, \hat{F}_{1b}^*)
\]
We then record the studentized values of the bootstrap estimates,

\[ T_b^* := \sqrt{n} \frac{\hat{\tau}_b^* - \hat{\tau}}{\hat{\sigma}_b^*} \]  

Repeating the resampling step \( B \) times, we obtain a sample \( (T_1^*, \ldots, T_B^*) \) that constitutes independent draws from the bootstrap estimator of the randomization distribution and can be used to construct critical values for tests or confidence intervals.

3.4. **Bootstrap Algorithm.** In summary, the proposed bootstrap algorithm proceeds in four main steps:

1. We create an empirical population of size \( N \), \( \left( \hat{Y}_i, \hat{W}_i \right)_{i=1}^N \) by selecting \( M_j^0 \) copies of \( Y_j \) with \( W_j = 0 \) and \( M_j^1 \) copies of \( Y_j^1 \) with \( W_j = 1 \), where \( M_j^0 \) and \( M_j^1 \) are defined in (3.3). If the sample is the population, \( n = N \), this step is redundant.
2. We then impute potential values \( \tilde{Y}_i(0), \tilde{Y}_i(1) \) for each unit \( i = 1, \ldots, N \), where \( \tilde{Y}_i(\hat{W}_i) = \hat{Y}_i \) and \( \tilde{Y}_i(1 - \hat{W}_i) \) is obtained according to (3.4).
3. Finally, we simulate the randomization distribution by repeatedly drawing \( n \) units \( Y_i^*(0), Y_i^*(1) \) out of that empirical population without replacement and generating randomization draws \( W_1^*, \ldots, W_n^* \) by setting \( W_{ib}^* = 1 \) for \( n_1 \) units sampled from \( \{1, \ldots, n\} \) without replacement, and \( W_{ib}^* = 0 \) for the remaining \( n - n_1 \) units. We then set \( Y_{ib}^* := Y_i^*(W_{ib}^*) \).
4. Given the sample \( (Y_{ib}^*, W_{ib}^*) \) we then compute the bootstrap version of the statistic \( T_b^* \) according to (3.5).

We can then approximate the bootstrap distribution of \( T_b^* \) by simulation, repeating steps 3 and 4 for the same empirical distribution.

3.5. **Confidence Intervals.** We consider confidence intervals constructed by inverting a t-test based on the point estimate \( \hat{\tau} := \tau(\hat{F}_0, \hat{F}_1) \) and given the variance bound \( \hat{\sigma} := \sigma(\hat{F}_0, \hat{F}_1) \) given in equation (3.2). The proposed confidence intervals for \( \tau \) are then of the form

\[ CI(1-\alpha) := \left[ \hat{\tau} - n^{-1/2} \hat{\sigma} \hat{c}(1-\alpha), \hat{\tau} - n^{-1/2} \hat{\sigma} \hat{c}(\alpha) \right] \]

We use bootstrap approximations to the randomization distribution of the t-ratio \( n^{1/2}(\hat{\tau} - \tau)/\hat{\sigma} \) under the least favorable coupling in order to determine the critical values. Specifically, let \( \hat{G}(z) := \frac{1}{B} \sum_{b=1}^B \mathbb{I}\{T_b^* \leq z\} \) denote the empirical distribution for the bootstrap samples obtained from the previous step. We then estimate the critical values using \( \hat{c}(\alpha) := \hat{G}^{-1}(\alpha) \) and \( \hat{c}(1-\alpha) := \hat{G}^{-1}(1-\alpha) \).

4. **Monte Carlo Simulations.** We next compare the performance of this causal bootstrap with the standard bootstrap and other alternative methods based on sampling or randomization designs. Specifically, we consider confidence intervals using Gaussian critical values with the respective analytic estimators of the sampling variance \( \hat{V}_{\text{Neyman}} \) and the causal variance \( \hat{V}_{\text{AGL}} \) given in Section 2.2. We also consider Gaussian inference using the variance estimators \( \hat{V}_{\text{s-boot}} \) and \( \hat{V}_{\text{c-boot}} \) obtained from the classical (sampling) bootstrap, and the causal bootstrap proposed in this paper. We compare these to confidence intervals from inverting Fisher’s exact test, and confidence intervals from the standard and the causal bootstrap for the t-statistic based on either sampling or causal variance estimate. Throughout we will restrict our attention to the case \( n = N \), i.e. when the full population of interest is observed.

We first consider three different simulation designs to illustrate the main points of comparison between the causal bootstrap and the main alternatives for causal inference.
• Design I sets $n_0 = n_1 = 100$ and draws potential outcomes according to $Y_i(0) \sim \mathcal{N}(0, 1), Y_i(1) = Y_i(0)$. In this setting, treatment effects are constant at $Y_i(1) - Y_i(0) \equiv 0$ and the marginal distributions $F_0^Y(y) \equiv F_1^Y(y)$, so that all procedures should be expected to do well.

• For Design II we again have $n_0 = n_1 = 100$, but generate potential outcomes as $Y_i(0) \sim \mathcal{N}(0, 1)$, and $Y_i(1) = 0$. In that case, the marginal distributions $F_0^Y(y)$ and $F_1^Y(y)$ are different, so that causal standard errors and the causal bootstrap should do better than their sampling analogs.

• Design III replicates Design II at a smaller sample size, where $n_0 = n_1 = 20$, and $Y_i(0) \sim \mathcal{N}(0, 1), Y_i(1) = 0$.

• For Design IV, $n_0 = n_1 = 20$, and we generate non-Gaussian potential outcomes where $Y_i(1) = 0$ and $Y_i(0)$ is a mixture that is drawn from $\mathcal{N}(0, 1)$ with probability 0.9, and from $\mathcal{N}(0, 16)$ with probability 0.1. This design highlights the difference between the bootstrap and Gaussian inference, which is no longer exact for this design.

Simulation results are shown in Table 3, where we compare coverage rates of nominal 95% confidence intervals, and the corresponding standard errors for each design. If a particular method does not directly calculate standard errors, we calculate the standard errors by taking the ratio of the difference between the upper and lower limit of the confidence interval and dividing by 2 times 1.96. We also report the nominal confidence level and theoretical causal standard error $\sqrt{\operatorname{Var}(\hat{\tau})}$ for each design in the last row (“Target”). With the exception of Fisher’s exact test, all methods rely on asymptotics, and should therefore not be expected to achieve exact coverage at the nominal level.

Under the first design, causal and sampling standard errors coincide, and the exact distribution of $\hat{\tau}_{ATE}$ is Gaussian. Furthermore, Fisher’s exact null holds true, so all methods should work well. Under the second design, the exact distribution of $\hat{\tau}_{ATE}$ is again Gaussian but a conservative estimate for the variance $\operatorname{Var}(Y_i(1) - Y_i(0))$ is strictly positive, so the causal standard error is strictly smaller than the sampling-based standard error. Hence for Design II, inference based on sampling based standard errors or the standard bootstrap should be expected to be conservative, whereas inference using causal standard errors or the causal bootstrap may still be conservative but coverage should be closer to the nominal level than for sampling-based methods.

Design III repeats Design II for a smaller sample size, which reveals a modest downward bias in the (bootstrap or analytical) causal standard errors, resulting in rejection rates exceeding the nominal level. Such a bias should be expected since the causal standard error corresponds to the plug-in estimator $\sigma(\hat{F}_0, \hat{F}_1)$, where under regularity conditions $\sigma(F_0, F_1)$ is a smooth but nonlinear functional of the marginal distributions $F_0, F_1$. While the bootstrap should not be expected to provide refinements for estimating the standard error (a non-pivotal quantity), a refinement for inference based on the studentized estimator corrects for the leading term of that bias and therefore result in rejection rates closer to the desired level. See also Appendix B for formal results.

Design IV has heterogeneous treatment effects and non-Gaussian marginal distributions for potential outcomes, where the tails of the marginal distribution of $Y_i(0)$ are thicker than for the Gaussian distribution. In this setting sampling-based methods should be more conservative than their causal analogs, and the pivotal bootstrap may provide refinements over Gaussian inference with causal standard errors. Since sample size for the third design is fairly small and all methods except for Fisher’s exact test rely on asymptotics, all inference methods exhibit modest size distortions. However simulated coverage rates for the pivotal bootstrap are close to the nominal level, and simulation results in Tables 5 and 6 confirm that coverage rates approach the desired nominal level when sample sizes get sufficiently large.\footnote{As for the classical bootstrap and other inference methods based on Gaussian asymptotic approximations, the...}
Table 3
95% Confidence Intervals And Standard Errors

<table>
<thead>
<tr>
<th>Variance Estimator</th>
<th>Critical Values</th>
<th>Pivotal Statistic</th>
<th>Design I (Rate, s.e.)</th>
<th>Design II (Rate, s.e.)</th>
<th>Design III (Rate, s.e.)</th>
<th>Design IV (Rate, s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9536, 0.1412</td>
<td>0.9950, 0.0999</td>
<td>0.9870, 0.2218</td>
<td>0.9776, 0.3330</td>
</tr>
<tr>
<td>( V_{\text{AGL}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9528, 0.1404</td>
<td>0.9524, 0.0706</td>
<td>0.9334, 0.1568</td>
<td>0.9116, 0.2354</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9518, 0.1405</td>
<td>0.9944, 0.0994</td>
<td>0.9850, 0.2162</td>
<td>0.9744, 0.3245</td>
</tr>
<tr>
<td>( \hat{V}_{\text{c-boot}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9512, 0.1400</td>
<td>0.9494, 0.0704</td>
<td>0.9392, 0.1548</td>
<td>0.9084, 0.2325</td>
</tr>
<tr>
<td>( \hat{V}_{\text{AGL}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9512, 0.1400</td>
<td>0.9494, 0.0704</td>
<td>0.9392, 0.1548</td>
<td>0.9084, 0.2325</td>
</tr>
<tr>
<td>N/A</td>
<td>Fisher Exact</td>
<td>No</td>
<td>0.9516, 0.1411</td>
<td>0.9630, 0.0999</td>
<td>0.9626, 0.2219</td>
<td>0.9698, 0.3332</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>0.9534, 0.1421</td>
<td>0.9954, 0.1012</td>
<td>0.9900, 0.2404</td>
<td>0.9838, 0.3865</td>
</tr>
<tr>
<td>( V_{\text{AGL}} )</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>0.9528, 0.1433</td>
<td>0.9954, 0.1012</td>
<td>0.9900, 0.2404</td>
<td>0.9838, 0.3865</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>0.9526, 0.1414</td>
<td>0.9510, 0.0715</td>
<td>0.9446, 0.1681</td>
<td>0.9434, 0.2802</td>
</tr>
<tr>
<td>( \hat{V}_{\text{AGL}} )</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>0.9530, 0.1419</td>
<td>0.9510, 0.0715</td>
<td>0.9446, 0.1681</td>
<td>0.9434, 0.2802</td>
</tr>
</tbody>
</table>

We can also compare causal and sampling-based inference in terms of local power. To this end we modify Designs I and II by shifting \( Y_i(1) \) by \( \tau_n = 2/\sqrt{n} \), which is also the average treatment effect under that local alternative. Table 4 reports rejection rates for tests of the null hypothesis of a zero average treatment effect. To interpret the results, note that under Design I marginal distributions are equal so that neither method is conservative, whereas under Design II marginal distributions are different so sampling-based methods (and Fisher’s exact test) are. We can see that for cases in which standard, sampling-based methods are conservative, causal methods increase rejection rates under the local alternative by around 25-30 percentage points throughout.

Table 4
Power against local Alternatives, \( \tau = 2/\sqrt{n} \).

<table>
<thead>
<tr>
<th>Variance Estimator</th>
<th>Critical Values</th>
<th>Pivotal Statistic</th>
<th>Design I (Rate, s.e.)</th>
<th>Design II (Rate, s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.1828, 0.1676</td>
<td>0.1748, 0.1656</td>
</tr>
<tr>
<td>( V_{\text{AGL}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.1934, 0.1750</td>
<td>0.1766, 0.1676</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.1966, 0.1742</td>
<td>0.1770, 0.1672</td>
</tr>
<tr>
<td>( \hat{V}_{\text{c-boot}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.1942, 0.1790</td>
<td>0.1776, 0.1682</td>
</tr>
<tr>
<td>N/A</td>
<td>Fisher Exact</td>
<td>No</td>
<td>0.0274, 0.0714</td>
<td>0.1018, 0.1200</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>0.1650, 0.1630</td>
<td>0.1712, 0.1668</td>
</tr>
<tr>
<td>( V_{\text{AGL}} )</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>0.1760, 0.1662</td>
<td>0.1720, 0.1672</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>0.1714, 0.1644</td>
<td>0.1712, 0.1638</td>
</tr>
<tr>
<td>( \hat{V}_{\text{AGL}} )</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>0.1652, 0.1624</td>
<td>0.1698, 0.1628</td>
</tr>
</tbody>
</table>

We next illustrate the role of the coupling of the potential values where we draw \((Y_i(0), Y_i(1))\) from a bivariate Gaussian distribution with variances \(\text{Var}(Y_i(0)) = 0.5\) and \(\text{Var}(Y_i(1)) = 2\) and correlation coefficient of the two potential values, \(\rho_{01} \in \{-1, 0, 1\}\). From our theoretical results, we

asymptotic justification for causal bootstrap generally requires bounded moments for potential outcomes up to the order given in Assumption 2.1. This can be confirmed by simulation experiments varying the thickness of the tails of \(Y_i(0), Y_i(1)\), the results of which are available from the authors on request.
Table 5
Coverage of nominal 95% Confidence Intervals, Gaussian Potential Outcomes with Different Couplings

<table>
<thead>
<tr>
<th>Variance Estimator</th>
<th>Critical Values</th>
<th>Pivotal Statistic</th>
<th>((n_0, n_1) = (50, 20))</th>
<th>((n_0, n_1) = (200, 80))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{V}_{\text{Neyman}})</td>
<td>Gaussian</td>
<td>No</td>
<td>(\hat{\rho}_{01} = 1)</td>
<td>(0.9542)</td>
</tr>
<tr>
<td>(\hat{V}_{\text{AGL}})</td>
<td>Gaussian</td>
<td>No</td>
<td>(\hat{\rho}_{01} = 0)</td>
<td>(0.9376)</td>
</tr>
<tr>
<td>(\hat{V}<em>{t</em>{\text{boot}}})</td>
<td>Gaussian</td>
<td>No</td>
<td>(\hat{\rho}_{01} = -1)</td>
<td>(0.9486)</td>
</tr>
<tr>
<td>(\hat{V}<em>{c</em>{\text{boot}}})</td>
<td>Gaussian</td>
<td>No</td>
<td>(0.9340)</td>
<td>(0.9482)</td>
</tr>
<tr>
<td>N/A</td>
<td>Fisher Exact</td>
<td>No</td>
<td>(0.9306)</td>
<td>(0.8742)</td>
</tr>
<tr>
<td>(\hat{V}_{\text{AGL}})</td>
<td>Fisher Exact</td>
<td>Yes</td>
<td>(0.9988)</td>
<td>(0.9964)</td>
</tr>
<tr>
<td>(\hat{V}_{\text{Neyman}})</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>(0.9646)</td>
<td>(0.9668)</td>
</tr>
<tr>
<td>(\hat{V}_{\text{AGL}})</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>(0.9632)</td>
<td>(0.9666)</td>
</tr>
<tr>
<td>(\hat{V}_{\text{Neyman}})</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>(0.9498)</td>
<td>(0.9520)</td>
</tr>
<tr>
<td>(\hat{V}_{\text{AGL}})</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>(0.9488)</td>
<td>(0.9488)</td>
</tr>
</tbody>
</table>

should expect Gaussian inference using causal standard errors and the causal bootstrap to have asymptotically exact coverage under the isotonic coupling \(\hat{\rho}_{01} = 1\) and be conservative when \(\hat{\rho}_{01} < 1\). Furthermore, for any coupling this design implies heterogeneous treatment effects, so that Fisher’s exact test does not in general control nominal confidence size for the average treatment effect. Given the calculations in Section 2.7 we designed the experiment deliberately to illustrate the potential of Fisher’s exact procedure to underestimate the spread of the randomization distribution, where \(n_0 > n_1\) and \(\text{Var}(Y_i(1)) > \text{Var}(Y_i(0))\). We also implement a version of the permutation test based on the studentized treatment contrast, as suggested by [12], to remedy this problem of Fisher’s procedure for tests of the null of no average treatment effect (following the suggestion by [31] we use the causal standard error derived from \(\hat{V}_{\text{AGL}}\)). Since the potential outcomes follow a Gaussian distribution, we should not expect refinements for the bootstrap relative to Gaussian inference.

In Table 5 we report simulated coverage rates for nominal 95% confidence intervals for the average treatment effect, where for either sample size we report the lowest coverage rate across the three different couplings in a separate column. The simulation results broadly confirm the theoretical predictions. Overcoverage from using sampling-based, rather than causal estimators for the variance or the bootstrap is not evident from the design with smaller sample sizes \((n_0 = 50, n_1 = 20)\), but becomes clearly visible once we move to the design with a larger number of units \((n_0 = 200, n_1 = 80)\). The confidence interval based on the non-pivotal version Fisher’s exact test has coverage that is consistently below the nominal 95% level, however the pivotal version of the procedure suggested by [12] appears to result in valid asymptotic inference.

Next we compare coverage rates of these confidence intervals as the size of the sample increases, where we choose a design with non-Gaussian distributions for the potential outcomes. Specifically, we let \(Y_i(0) \equiv 0\) and \(Y_i(1)|S_i(1) \sim \mathcal{N}(0, S_i^2)\), where \(S_i = 1\) with probability 0.9, and \(S_i = 4\) with probability 0.1. Since the marginal distributions for \(Y_i(0)\) and \(Y_i(1)\) are different, the difference between sampling variance and the upper bound for the causal variance is nontrivial. Furthermore, while we do not give formal results, under certain regularity conditions the pivotal causal bootstrap should be expected to provide refinements over the Gaussian limiting approximation to the randomization distribution.

Table 6 shows simulated coverage rates for the different confidence intervals at the nominal 95% significance level under this design. The results show that coverage rates for both the sampling-based variance estimators and bootstrap are higher throughout than for their causal analogs. The comparison between Gaussian confidence intervals using the causal variance estimators, \(\hat{V}_{\text{AGL}}\) and
Table 6

Coverage of nominal 95% Confidence Intervals, non-Gaussian Potential Values with Isotone Coupling

<table>
<thead>
<tr>
<th>Variance Estimator</th>
<th>Critical Values</th>
<th>Pivotal Statistic</th>
<th>(20, 20)</th>
<th>(50, 50)</th>
<th>(100, 100)</th>
<th>(200, 200)</th>
<th>(500, 500)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9768</td>
<td>0.9866</td>
<td>0.9914</td>
<td>0.9932</td>
<td>0.9924</td>
</tr>
<tr>
<td>( \hat{V}_{\text{AGL}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9186</td>
<td>0.9358</td>
<td>0.9396</td>
<td>0.9450</td>
<td>0.9436</td>
</tr>
<tr>
<td>( \hat{V}_{\text{c-boot}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9752</td>
<td>0.9864</td>
<td>0.9912</td>
<td>0.9928</td>
<td>0.9924</td>
</tr>
<tr>
<td>( \hat{V}_{\text{c-boot}} )</td>
<td>Gaussian</td>
<td>No</td>
<td>0.9144</td>
<td>0.9336</td>
<td>0.9378</td>
<td>0.9436</td>
<td>0.9436</td>
</tr>
<tr>
<td>N/A</td>
<td>Fisher Exact</td>
<td>No</td>
<td>0.9752</td>
<td>0.9864</td>
<td>0.9912</td>
<td>0.9928</td>
<td>0.9924</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>0.9870</td>
<td>0.9912</td>
<td>0.9940</td>
<td>0.9942</td>
<td>0.9934</td>
</tr>
<tr>
<td>( \hat{V}_{\text{AGL}} )</td>
<td>Std. Bootstrap</td>
<td>Yes</td>
<td>0.9870</td>
<td>0.9912</td>
<td>0.9940</td>
<td>0.9942</td>
<td>0.9934</td>
</tr>
<tr>
<td>( \hat{V}_{\text{Neyman}} )</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>0.9470</td>
<td>0.9532</td>
<td>0.9582</td>
<td>0.9548</td>
<td>0.9482</td>
</tr>
<tr>
<td>( \hat{V}_{\text{AGL}} )</td>
<td>Causal Bootstrap</td>
<td>Yes</td>
<td>0.9470</td>
<td>0.9532</td>
<td>0.9582</td>
<td>0.9548</td>
<td>0.9482</td>
</tr>
</tbody>
</table>

\( \hat{V}_{\text{c-boot}} \), respectively, to the pivotal causal bootstrap is also indicative of refinements, where the confidence interval based on the pivotal causal bootstrap has coverage rates much closer to the nominal level for small sample sizes, but that advantage vanishes as \( n_0, n_1 \) grow large.

5. Large Sample Theory. To characterize the asymptotic properties of the bootstrap procedure, we can cast the statistical experiment of sampling from a finite population with subsequent randomization of treatment among the sampled units as a two-stage scheme of sampling without replacement from nested finite populations. Specifically, in a first step we draw \( n \) units without replacement from the population of \( N \) units. In a second step, we draw \( n_1 \) units at random and without replacement from that sample to receive the treatment \( W_i = 1 \), whereas the remaining \( n_0 = n - n_1 \) units are assigned \( W_i = 0 \). This second step is conditionally independent of the first.

To characterize the contribution of sampling uncertainty to the distribution of the functional we define

\[
F_{01}^p(y_0, y_1) := \frac{1}{N} \sum_{i=1}^{N} I\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}
\]

\[
F_{01}^s(y_0, y_1) := \frac{1}{n} \sum_{i=1}^{N} R_i I\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}
\]

with corresponding marginals \( F_{01}^p, F_{01}^s, F_{01}^p, F_{01}^s \). In particular,

\[
(5.1) \quad F_{01}^s(y_0, y_1) - F_{01}^p(y_0, y_1) = \frac{1}{n} \sum_{i=1}^{n} (R_i - q) I\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}
\]

Turning to the contribution of design uncertainty, we define

\[
\hat{F}_0(y_0) := \frac{1}{n(1-p)} \sum_{i=1}^{n} R_i (1 - W_i) I\{Y_i(0) \leq y_0\}
\]

\[
\hat{F}_1(y_1) := \frac{1}{np} \sum_{i=1}^{n} R_i W_i I\{Y_i(1) \leq y_1\}
\]

We then have

\[
(5.2) \quad \left( \frac{\hat{F}_0(y_0) - F_{01}^s(y_0)}{\hat{F}_1(y_1) - F_{01}^s(y_1)} \right) = \frac{1}{np} \sum_{i=1}^{N} R_i (W_i - p) \left( -\frac{p}{1-p} I\{Y_i(0) \leq y_0\} \right)
\]
Taken together, (5.1) and (5.2) characterize the uncertainty from sampling and randomization in estimating the respective marginal distributions of $Y_i(0)$ and $Y_i(1)$ as a two-stage process of drawing without replacement from nested finite populations. An asymptotic Donsker Theorem for empirical processes based on sampling without replacement from a finite population is available from [6].

We now state the limiting properties of the bootstrap as $N$ and $n$ grow large. Specifically, we derive the limits of the randomization and bootstrap distributions. We then show that the latter is an asymptotically conservative estimator of the former for the purposes of forming confidence intervals.

5.1. Consistency and Randomization CLT. Consistency of the estimated bounds follows from consistency of $\hat{F}_0(y_0)$ and $\hat{F}_1(y_1)$ for $F_0(y_0)$ and $F_1(y_1)$, respectively, and the continuous mapping theorem, noting that the conditions in Assumption 2.1 are sufficient for the parameter bounds to be continuous functions of $F_0(y_0)$ and $F_1(y_1)$.

**Theorem 5.1.** *(Consistency)* Suppose Assumptions 2.1, 1.1, and 1.2 hold. Then $\hat{\tau}$ and $\hat{\sigma}$ are consistent for $\tau(\hat{F}_0, \hat{F}_1)$ and $\sigma(\hat{F}_0, \hat{F}_1)$, respectively.

For a randomization CLT for the estimated bounds we first establish a functional CLT for the randomization processes

$$
\hat{G}_0 := \sqrt{n}(\hat{F}_0 - F_0^p) \\
\hat{G}_1 := \sqrt{n}(\hat{F}_1 - F_1^p)
$$

for conditional distributions of potential outcomes. We argue that Assumption 2.1 is sufficient to establish Hadamard differentiability of the functionals $\tau(\hat{F}_0, \hat{F}_1)$, $\sigma(\hat{F}_0, \hat{F}_1)$ so that asymptotic normality of $\sqrt{n} \frac{\hat{\tau} - \tau}{\hat{\sigma}}$ follows from the functional Delta rule and Slutsky’s theorem.

**Theorem 5.2.** *(Randomization CLT)* Suppose Assumptions 2.1, 1.1, and 1.2 hold. Then the asymptotic distribution of the t-ratio for $\hat{\tau}_{ATE}$ is given by

$$
\sqrt{n} \frac{\hat{\tau} - \tau}{\hat{\sigma}} \xrightarrow{d} N \left( 0, \frac{\sigma^2(F_{01}^p)}{\sigma^2(F_{01}^p, F_{11}^p)} \right)
$$

where $\sigma^2(F_{01}) := \lim_n n \text{Var}_{F_{01}}(\hat{\tau})$.

The proof of this result is given in the appendix. The formal argument adapts a finite-population CLT for the empirical process developed by [6] for the two-sample problem to the case of sampling and randomization in a finite population.

5.2. Bootstrap CLT. For a bootstrap replication, denote the empirical distributions of $Y_i^*|W_i^* = 0$ and $Y_i^*|W_i^* = 1$ with $\hat{F}_0^*$ and $\hat{F}_1^*$, respectively. Also, let $\hat{\tau}^* = \tau(\hat{F}_0^*, \hat{F}_1^*)$ and $\hat{\sigma}^* = \sigma(\hat{F}_0^*, \hat{F}_1^*)$.

We then establish a CLT for the bootstrap analogs

$$
\hat{G}_0^* := \sqrt{n}(\hat{F}_0^* - \hat{F}_0) \\
\hat{G}_1^* := \sqrt{n}(\hat{F}_1^* - \hat{F}_1).
$$

A CLT for the bootstrapped bounds $\sqrt{n} \frac{\hat{\tau}^* - \tau}{\hat{\sigma}^*}$ then relies again on Hadamard differentiability of the variance bounds and the Delta rule for the bootstrap.

A bootstrap CLT can be shown using analogous steps as in a proof for Theorem 5.2, where the randomization distribution is generated based on an estimator for $F_{01}^p$ based on the estimated distributions $\hat{F}_0^p(y_0)$, $\hat{F}_1^p(y_1)$ and the respective least-favorable coupling $C^{iso}(\cdot)$. 

17
Theorem 5.3. **Bootstrap CLT** Suppose Assumptions 2.1, 1.1, and 1.2 hold. Then the asymptotic distribution of the bootstrapped t-ratio for \( \hat{\tau}_{ATE} \) is given by
\[
\sqrt{n} \frac{\hat{\tau}^* - \hat{\tau}}{\sigma^*} \xrightarrow{d} N(0,1)
\]

Most importantly, by Theorem 5.1 the bootstrap estimator for the randomization distribution for \( \hat{\tau} - \tau \hat{\sigma} \) has asymptotic variance equal to 1, whereas the asymptotic variance of the randomization distribution is \( \sigma^2(F_{01}) \) which is less than 1 by construction. That is, the bootstrap algorithm in section 3 converges to a “least-favorable” limiting experiment in an appropriate sense. Note also that the formal argument in the proofs of Theorems 5.1-5.3 immediately apply to any other functional \( \tau(F_0, F_1) \) that is Hadamard-differentiable in \( F_0, F_1 \), and for which the variance bound \( \sigma^2(F_0, F_1) \) is continuous in \( F_0, F_1 \). Furthermore, under slightly stronger conditions the causal bootstrap of the t-ratio achieves refinements, see Appendix B for a formal result.

5.3. **Asymptotic Validity of Confidence Intervals.** It remains to show that confidence intervals of the form (3.6) that are constructed under the “least-favorable” limiting experiment are indeed conservative given the CLT under the true randomization distribution in Theorem 5.2. This can be proven by combining the randomization and bootstrap CLTs, replacing the unidentified randomization variance with an estimate of the bound \( \sigma(F_0, F_1) \geq \sigma(F_{01}) \).

**Corollary 5.1. (Asymptotic Validity of Confidence Intervals)** Under Assumptions 2.1, 1.1, and 1.2, the \( 1 - \alpha \) confidence interval (3.6) using bootstrap critical values is asymptotically valid,
\[
\liminf_n \inf_{F_{01}^p} \mathbb{P}_{F_{01}^p} (\tau(F_{01}^p) \in CI(1 - \alpha)) \geq 1 - \alpha
\]
Given Theorems 5.2 and 5.3 this result follows immediately from the definition of the variance bound \( \sigma(F_0, F_1) \).

6. **Observational Studies.** In this section we show how to adapt our procedure to designs in which treatment assignment is not completely randomized but in which the assignment probability may vary with observable attributes of a unit, which we denote by a vector \( X_i \) for unit \( i \). As before, we assume that the population of interest consists of \( N \) units, and the researcher observes a random sample of size \( n \leq N \) from that population, where we let \( R_i \) denote an indicator variable that equals one if unit \( i \) is included in that sample. We continue to consider the problem of inference regarding the average treatment effect in the population,
\[
\tau_{ATE} := \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))
\]
Our main results focus on the case in which units are assigned to treatment based on observable attributes. Specifically, we assume that after conditioning on observable covariates \( X_i \), treatment assignment \( W_i \) is independent of the potential outcomes for that unit:

**Assumption 6.1. (Unconfoundedness)** Treatment assignment is independent across units \( i = 1, \ldots, n \) and strongly ignorable given \( X_i \),
\[
(Y_i(0), Y_i(1)) \perp \!
\!
\!
\perp W_i | X_i, R_i
\]
This assumption on the assignment mechanism differs from our treatment of the completely randomized case in two important ways: for one, the probability of assignment may vary across units, possibly depending on unit attributes which are also predictive of potential outcomes. Moreover, rather than assuming a fixed number of treated units, treatment assignment in the observable case is assumed to be independent across units, conditional on \( X_i \). In our view, this is a more plausible assumption for observational studies for which treatment is not assigned following an experimental protocol but determined by factors outside the researcher’s control.

Assumption 6.1 is usually referred to as unconfoundedness or ignorability of treatment assignment, see [28]. Under unconfoundedness and independent assignment, the assignment mechanism for a binary treatment is fully described by the propensity score \( e(x) \),

\[
    e(x) := P(W_i = 1 | X_i = x)
\]

For greater clarity of exposition we will focus on the case in which \( e(x) \) is known, but it is possible to extend our approach to the case in which the propensity score is unknown and has to be estimated (see [22] and [24]).

We also assume that the respective supports of the distributions of \( X_i \) of the treated and untreated units coincide in the sense that assignment is non-deterministic for any value of \( X_i \). In addition, our formal argument assumes that these conditional assignment probabilities are in fact bounded away from zero and one.

**Assumption 6.2.** *(Overlap)* The propensity score satisfies \( 0 < \underline{e} \leq e(x) \leq \overline{e} < 1 \) for all values of \( x \) in the support of \( X_i \).

Given these assumptions, a natural estimator for the average treatment effect (ATE) is the weighting estimator,

\[
    \hat{\tau}_{ATE} := \frac{1}{n} \sum_{i=1}^{n} \left( \frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i)Y_i}{1 - e(X_i)} \right)
\]

In the following, we show how to adapt the causal bootstrap to approximate the distribution of \( \hat{\tau} \) to account for uncertainty both from sampling from the finite population of size \( N \), as well as the randomness in the mechanism for assigning treatment status \( W_i \).

We characterize asymptotic properties for a sampling experiment under which the finite population \( F_{01}^{p} \) consists of \( N \) i.i.d. draws from a distribution (or infinite meta-population with distribution) \( F_{01} \). While causal inference takes that finite population as given, the quality of asymptotic approximations depends on properties of that underlying meta-population. In addition, estimation of the conditional distributions of potential outcomes given \( X_i \) require either that attributes are discrete or that the underlying distribution is sufficiently smooth to permit consistent estimation of conditional distributions as \( N \) and \( n \) grow.

**Assumption 6.3.** *(Superpopulation)* For each unit \( i = 1, \ldots, N \) in the population, attributes \( X_i \) are i.i.d. draws from the distribution \( F_X(x) \), and potential values of \( Y_i(0), Y_i(1) \) are independent draws from the distribution \( F_{01}(y_0, y_1 | x) \). \( F_X \) and \( F_{01} \) have bounded p.d.f.s \( f_X(x) \) and \( f_{01}(y_0, y_1 | x) \), respectively, that are twice continuously differentiable in the continuously distributed components of \( x \).

---

5 [26], see also chapter 12 in [28]
Under Assumptions 6.1-6.3, the marginal distribution of either potential outcome given $X_i$,

$$
F_0(y_0|x) := \mathbb{P}_{F_{01}}(Y_i(0) \leq y_0|X_i = x) = \frac{\mathbb{P}_{F_{00}}(Y_i \leq y_0, W_i = 0|X_i = x)}{1 - e(x)}
$$

$$
F_1(y_1|x) := \mathbb{P}_{F_{01}}(Y_i(1) \leq y_1|X_i = x) = \frac{\mathbb{P}_{F_{00}}(Y_i \leq y_1, W_i = 1|X_i = x)}{e(x)}
$$

is identified from the subsamples $\{(X_i, Y_i)|W_i = 0\}$ and $\{(X_i, Y_i)|W_i = 1\}$ and can be estimated consistently. However, the joint distribution conditional on $X_i$ is not identified. Specifically, the joint distribution of potential values conditional on $X_i$ can be written as

$$
F_{01}(y_0, y_1|x) = C_X(F_0(y_0|x), F_1(y_1|x)|x)
$$

where the conditional copula $C_X : [0, 1]^2 \times \mathcal{X} \rightarrow [0, 1]$ is a function that is nondecreasing in either argument for each value of $x$. By Sklar’s theorem, such a copula exists for each $x$ in the support of $X_i$ but need not be unique unless the marginal distributions $F_0, F_1$ are continuous. In the following, we let

$$
C_X := \{C_X : [0, 1]^2 \times \mathcal{X} \rightarrow [0, 1], C_X(u, v|x) \text{ nondecreasing in } u \text{ and } v\}
$$

denote the set of all possible conditional copulae. The (unconditional) joint distribution of potential values is then given by

$$
F_{01}(y_0, y_1) = \int C_X(F_0(y_0|x), F_1(y_1|x)|x)dF_X(x) =: C_X[0, F_1](y_0, y_1)
$$

where the arguments $F_0 := F_0(y_0, x)$ and $F_1 := F_1(y_1, x)$ are taken to be the joint distributions of either potential value and $X_i$.

6.1. Balancing Score. Nonparametric estimation of the conditional distributions of potential outcomes given $X_i$ may suffer from a curse of dimensionality in the number of attributes $X_i$, or may require smoothing in continuously distributed components of $X_i$. We therefore consider an implementation of our approach that conditions on a balancing score $b(X_i)$ given a mapping $b : \mathcal{X} \rightarrow \mathcal{B} \subset \mathbb{R}^d_B$ chosen by the researcher, where $\mathcal{B}$ is a subset of a $d_B$-dimensional Euclidean space. The balancing score $b(x)$ does not have to include the propensity score, but only serves as a basis for incorporating covariate information into the construction of worst-case bounds for the functionals of interest.

We then consider conditional couplings given $B_i := b(X_i)$ which result in joint distributions of the form

$$
\hat{F}_{01}(y_0, y_1) = \int C_B(F_0(y_0|b), F_1(y_1|b)|b)dF_B(b) =: C_B[0, F_1](y_0, y_1)
$$

where he arguments $F_0, F_1$ are taken to be the joint c.d.f.s $F_0(y_0, b)$ and $F_1(y_1, b)$ of either potential value with $b(X_i)$, and $F_B(b)$ denotes the marginal distribution of $B_i := b(X_i)$. We also denote the set of conditional couplings given $B_i$ with

$$
C_B := \{C_B : [0, 1] \times \mathcal{B} \rightarrow [0, 1], C(u, v|b) \text{ is nondecreasing in } u \text{ and } v\}
$$

Given the choice of a balancing score $b(X_i)$, we use the following nonparametric estimators for the joint c.d.f.s $F_0, F_1$ of $b(X_i)$ with $Y_i(0)$ and $Y_i(1)$, respectively:

$$
\hat{F}_{0n}(y_0, b) := \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - W_i)\mathbb{I}\{Y_i \leq y_0, b(X_i) \leq b\}}{1 - e(X_i)}
$$

$$\hat{F}_{1n}(y_1, b) := \frac{1}{n} \sum_{i=1}^{n} \frac{W_i\mathbb{I}\{Y_i \leq y_0, b(X_i) \leq b\}}{e(X_i)}$$


Note that the respective sums of the weights \( \frac{W_1}{\epsilon(X_1)}, \ldots, \frac{W_n}{\epsilon(X_n)} \) and \( \frac{1-W_1}{1-\epsilon(X_1)}, \ldots, \frac{1-W_n}{1-\epsilon(X_n)} \) are not guaranteed to be equal to one for a given realization of \( W_1, \ldots, W_n \), which can in principle be addressed using re-weighting (see e.g. \cite{24} for a discussion). This does not pose a substantial challenge to our approach, and we therefore develop our approach in terms of estimators with unadjusted inverse propensity score weights.

To obtain conditional distributions we consider two scenarios: (a) the case in which \( b(X_i) \) is discrete with finitely many support points, and (b) the case in which it is continuously distributed on a subset of \( \mathbb{R}^{dB} \). In the discrete case, we have \( \hat{F}_{wn}(y_w|b) = \frac{\hat{F}_{wn}(y_w, b)}{\hat{F}_n(b)} \) for each \( w = 0, 1 \) and value of \( b \) in the support of \( b(X_i) \). For the continuous case, we can form

\[
\begin{align*}
\hat{F}_{0n}(y_0|b) &= \frac{1}{nh^{-dB}} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y_0\} \frac{1-W_i}{1-\epsilon(X_i)} K\left(\frac{b(X_i)-b}{h}\right) \\
\hat{F}_{1n}(y_1|b) &= \frac{1}{nh^{-dB}} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y_1\} \frac{W_i}{\epsilon(X_i)} K\left(\frac{b(X_i)-b}{h}\right)
\end{align*}
\]

for a bandwidth \( h \equiv h_n > 0 \) which may shrink to zero at an appropriate rate, and a function \( K : \mathbb{R}^{dB} \to \mathbb{R}_+ \) satisfying the usual assumptions for the kernel function in nonparametric estimation.\(^{6}\)

**Assumption 6.4.** **(Conditioning on \( b(X_i) \))** (a) We either have that \( B \) is a finite set \( \{b_1, \ldots, b_K\} \) with \( \mathbb{P}(b(X_i) = b_k) \geq \underline{p}_B \) for each \( k = 1, \ldots, K \) and some \( \underline{p}_B > 0 \), or that \( b(X_i) \) is continuously distributed on a set \( B \subseteq \mathbb{R}^{dB} \) with density bounded away from zero. (b) If \( b(X_i) \) is continuously distributed, the smoothing estimator uses a bandwidth \( h = h_n \) satisfying \( h_n \to 0 \) and \( nh^{-dB} \to \infty \). (c) The first four moments of the conditional distribution of \( Y_i(0), Y_i(1) \) given \( b(X_i) \) are bounded almost surely. (d) The conditional distribution given the balancing score \( F_{01}(y_0, y_1|b) \) has a bounded p.d.f. \( f_{01}(y_0, y_1|b) \) that is twice continuously differentiable in the continuously distributed components of \( b(X_i) \).

The rate condition for the bandwidth is sufficient for asymptotic validity, in practice the bandwidth should be chosen at the optimal rate for estimating a conditional c.d.f. given the dimension of \( b(X_i) \) and the degree of smoothness assumed for \( F_0(y|b), F_1(y|b) \) as a function of \( b \). A more rigorous discussion of optimal bandwidth selection is beyond the scope of this paper.

We argue that if implemented correctly, smoothing or binning of covariates \( X_i \) using a balancing score leads to inference that is unambiguously conservative. Specifically, the following Lemma states that the least favorable coupling given a coarser conditioning set or oversmoothing produces bounds that are always wider than the sharp bounds given the full attribute vector \( X_i \).

**Lemma 6.1.** For a functional \( T(F_{01}) \) of the distribution \( F_{01} = F_{01}(y_0, y_1) \) and arbitrary function \( b : \mathcal{X} \to B \) we have that

\[
\sup_{C_X \in \mathcal{C}_X} T \left( \int C_X(F_0(y_0|x), F_1(y_1|x)|x)dF_X(x) \right) \leq \sup_{C_B \in \mathcal{C}_B} T \left( \int C_B(F_0(y_0|b(x)), F_1(y_1|b(x))|b(x))dF_X(x) \right)
\]

\(^{6}\)To be specific, we assume that the kernel function is the p.d.f. of a continuously distributed \( dB \)-variate random variable with \( \int K(u)du = 1 \), and moments \( \int uK(u)du = 0, \int u^2K(u)du < \infty \), and \( \int K(u)^2du < \infty \).
that confidence intervals achieve at least nominal asymptotic coverage. When estimating the marginal distributions in practical terms, this also suggests that the researcher may want to err on the side of over-smoothing values, so that making only partial use of that information produces bounds that are not sharp. In unconfoundedness, covariates are generally informative about the joint distribution of potential outcomes, which assumes the isotone coupling conditional on \( b(X_i) \) is (uniquely) attained at the isotone conditional coupling, \( C_{\text{iso}}(u,v|b) := \min\{u,v\} \).

Moreover, from Lemma 6.1, we also have that
\[
\sigma^2(F_{01}) \leq \sup_{C_X \in C_X} \sigma^2(C_X[F_0, F_1]) \leq \sup_{C_B \in C_B} \sigma^2(C_B[F_0, F_1]) = \sigma^2_B(F_0, F_1)
\]
for any distribution \( F_{01}(y_0,y_1|x) \) with the conditional marginals \( F_0(y_0|x), F_1(y_1|x) \). Therefore an estimator for the distribution of \( \hat{\tau}_{ATE} \) which assumes the isotone coupling conditional on \( b(X_i) \) is asymptotically conservative at any order of approximation.

We formalize these claims in the following proposition:

**Proposition 6.1. (Conservative Coupling for the ATE).** Suppose that Assumption 2.1 holds. Then, given the marginal distributions \( F_0, F_1 \), the variance bound
\[
\sigma^2_B(F_0, F_1) := \lim_{N} \text{Var}_{C_{\text{iso}}[F_0,F_1]}(\hat{\tau}) \geq \sigma^2(F_{01})
\]
for any joint distribution \( F_{01}(y_0,y_1,x) \) with marginals \( F_0(y_0,x), F_1(y_1,x) \).

A proof for this proposition is given in the appendix.
6.2. Bootstrap Algorithm. Given that least favorable coupling for the balancing score $b(X_i)$, we can now impute missing counterfactuals as follows: We first obtain the empirical conditional rank of observation $i$ as

$$
\hat{V}_i := \hat{F}_{W_i}(Y_i | b(X_i))
$$

Then the isotone assignment conditional on $b(X_i)$ can be constructed by imputing the values

$$
(6.1) \quad \hat{Y}_{W_i} := Y_i \quad \text{and} \quad \hat{Y}_{(1-W_i)i} := \hat{F}_{(1-W_i)}(\hat{V}_i | b(X_i))
$$

Notice that this coupling preserves the estimated conditional distribution of either potential outcome given $b(X_i)$, possibly up to a discretization error when for the propensity scores neither $e(X_i)/(1-e(X_i))$ nor $(1-e(X_i))/e(X_i)$ is integer.

We then estimate the randomization distribution of $\hat{\tau}_{ATE}$ using the following algorithm:

1. For the $b$th bootstrap sample, draw $W_{1b}^*, \ldots, W_{nb}^*$ independently from Bernoulli distributions with $P(W_{ib}^* = 1 | X_i) := e(X_i)$.
2. Compute the treatment contrast

$$
\hat{\tau}_b^* := \frac{1}{n} \sum_{i=1}^{n} \left( \frac{W_{ib}^* \hat{Y}_{1i}}{e(X_i)} - \frac{(1 - W_{ib}^*) \hat{Y}_{0i}}{1 - e(X_i)} \right)
$$

3. For $B$ independent replications use the empirical distribution of $\hat{\tau}_1^*, \ldots, \hat{\tau}_B^*$ as the bootstrap estimator for the randomization distribution of $\hat{\tau}$.

We can then prove that the adaptation of the bootstrap procedure to the observational case is also asymptotically valid:

**THEOREM 6.1. (Causal Bootstrap for Observational Designs)** Suppose Assumptions 6.1-6.4 hold. Then the conclusions of Theorems 5.1-5.3 hold. In particular, the $1-\alpha$ confidence interval (3.6) using bootstrap critical values is asymptotically valid,

$$
\lim \inf_n \inf_{p_{01}} P_{p_{01}} (\tau(F_{01}^p) \in CI(1-\alpha)) \geq 1 - \alpha
$$

See the appendix for a proof.

6.3. Simulation Experiments for Observational Designs. We next compare our theoretical results to Monte Carlo simulations for observational studies. In all simulation designs, the covariate $X_i$ is scalar with $X_i \sim U[0,1]$, the uniform distribution. The propensity score is non-constant, where $e(x) = 0.4$ if $x \leq 0.5$, and $e(x) = 0.6$ if $x > 0.5$. In all designs, the conditional moments of $Y_i(0), Y_i(1)$ given $X_i = x$ are at least twice continuously differentiable in $x$.

To construct the conditionally isotone coupling given $X_i$, we smooth estimates using a standard Gaussian kernel with bandwidth $h := (n/2)^{-0.15}$, where $n/2$ is the expected number of treated or control units under this design. Since Lemma 6.1 implies that oversmoothing guarantees conservative inference, the rate for the bandwidth was chosen to be slower than the MSE-optimal rate that balances bias and variance, $h_{opt} \asymp (n/2)^{-0.2}$. We chose a kernel function with unbounded support to ensure that the estimated conditional distributions for $Y_i(0)$ and $Y_i(1)$ are well-defined in all simulation and bootstrap samples even for narrower bandwidths.

We consider three different designs,
The first example generates potential outcomes according to $Y_i(0) = Y_i(1) \sim N(0, 1)$. This is a baseline case which also satisfies the assumptions of a completely randomized experiment and Fisher’s sharp null. Furthermore, the distribution of the data is Gaussian, so that in finite samples departures from exact coverage at the nominal level can be attributed to smoothing bias or estimation of the (sampling or causal) variance.

For the second design, $Y_i(0) = \frac{1}{2} (X_i^2 - \frac{1}{3})$ and $Y_i(1) = Y_i(0) + \frac{1}{2} N(0, 1)$, so that both potential outcomes vary with $X_i$ but the conditional average treatment has zero mean but it not constant. Since the propensity score is not constant, this design does not satisfy the (unconditional) independence assumption for completely randomized experiments.

Finally in Design 3, we let $Y_i(0) = \frac{1}{2} (\log U[0,1]) + 1$ and $Y_i(1) = (X_i - \frac{1}{2})$. For this simulation design the conditional mean of the unit-level treatment effect $Y_i(1) - Y_i(0)$ varies with $X_i$. Furthermore, the conditional distribution of either potential outcome given $X_i$ is not Gaussian.

For each design, we simulate coverage rates and estimated standard errors for sample sizes $n = 40$ and $n = 200$ to be able to assess which discrepancies in coverage rates from nominal levels are due to asymptotically negligible approximation errors, and which persist for arbitrarily large sample sizes. For each design, we simulated the model 5,000 times and all resampling methods (standard bootstrap, causal bootstrap, permutation test) were implemented using 1,000 replications.

We report our main findings in Table 6.3. The simulation evidence is mostly in line with theoretical predictions. Under the first design all methods do well, and the main problem is downward (finite-sample) bias in estimation of the causal standard error. This bias diminishes with sample size and primarily affects Gaussian plug-in inference, whereas the causal bootstrap is less affected.

For the second design, the causal variance bound is lower than the sampling variance estimator, so that causal methods have lower coverage rates. While the conditional distribution of both potential outcomes varies with $X_i$, performance of both procedures is comparable to their analogs for the completely randomized case. Furthermore, while substantially less conservative than their observational counterparts, the simulated confidence size for the causal inference procedures remains above the nominal 95 percent level. In light of Lemma 6.1, this is likely due to smoothing in the estimation of the conditional distributions of $Y_i(0)$ and $Y_i(1)$ given $X_i$. The bandwidth sequence $h$ was deliberately chosen to ensure conservative coverage but still converge to zero for very large $n$, so these results confirm that the strategy of oversmoothing estimates works as intended. For the third design, the differences between causal and sampling-based inference are slightly less pronounced but still visible, and results are qualitatively similar to those for Design 2.

In sum, this extension of the causal bootstrap to observational designs shares the main qualitative features of the, conceptually more straightforward, baseline case of completely randomized experiments.

7. Conclusion. The classical bootstrap is a popular method to approximate the asymptotic distribution of a statistic when the only source of statistical error is sampling variation. In this paper, we propose a bootstrap procedure for inference on causal parameters that accounts not only for sampling, but also design uncertainty. Since the joint distribution of potential values is fundamentally underidentified, inference has to be based on a conservative estimate of that distribution. We show how to construct the least favorable randomization distribution for the estimated average treatment effect, both for completely randomized or observational designs, and show that the resampling procedure can be used to construct asymptotically valid confidence intervals.
APPENDIX A: IMPERFECT COMPLIANCE, INSTRUMENTAL VARIABLES, AND SAMPLE LATE

In this appendix we give a brief outline how to apply the causal bootstrap to designs with imperfect compliance. In such a design, the researcher controls the initial assignment of a binary instrumental variable (“intention to treat”) $W_i \in \{0, 1\}$. For $W_i = w$, unit $i$ then chooses a binary treatment status $D_i(w) = d \in \{0, 1\}$, and the potential outcome associated with $D_i = d$ is $Y_i(d)$. We assume that for each unit included in the sample, the researcher observes the instrumental variable $W_i$, the realized treatment status $D_i := D_i(W_i)$, and the realized potential outcome $Y_i := Y_i(D_i)$. For simplicity, we assume that these variables are observed for each unit $i = 1, \ldots, N$ in the population of interest - for a random sample of size $n < N$, the procedure can be adapted in a manner analogous to the completely randomized case. We also restrict our attention to the case in which no additional unit-level attributes are observed.

We furthermore assume that intention to treat is independent of potential outcomes, both with respect to treatment status and the outcome of interest:

**Assumption A.1.** (Independence) The instrumental variable $W_i$ is independent of potential outcomes, 

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

Independence may result either from the design of randomized controlled trial (RCT) as common in social experiments, or quasi-experimental variation of a variable that affects the likelihood of treatment.

This problem has been analyzed in a potential outcomes framework by [2] under the additional assumption that for each unit, treatment status is nondecreasing with the initial assignment,

**Assumption A.2.** (Monotonicity) Treatment status is monotone in $w$,

$$D_i(1) \geq D_i(0) \text{ for all } i = 1, \ldots, N$$

Following conventional terminology, we categorize units in terms of how their treatment status varies with the instrumental variable: If $D_i(0) = D_i(1) = 0$, unit $i$ is a never taker, if $D_i(0) = D_i(1) = 1$, $i$ is an always taker; if $D_i(0) = 0, D_i(1) = 1$, $i$ is a complier, and if $D_i(0) = 1, D_i(1) = 0$, $i$ is a defier. The monotonicity assumption rules out the existence of defiers in the population of interest.

Under these two assumptions, [2] show that it is possible to identify the Local Average Treatment Effect (LATE),

$$\tau_{LATE} := \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)) \mathbb{1} \{D_i(0) = 0, D_i(1) = 1\}$$

nonparametrically from the observable data, which is the average treatment effect among the compliers in $\{1, \ldots, N\}$ with respect to the initial assignment $W_i$. A natural estimator for this causal parameter is the Wald (instrumental variables) estimator

$$\hat{\tau}_{Wald} := \frac{1}{N} \sum_{i=1}^{N} W_i(Y_i - \bar{Y}_N)$$

$$\frac{1}{N} \sum_{i=1}^{N} W_i(D_i - \bar{D}_N)$$
where \( \bar{Y}_N := \frac{1}{N} \sum_{i=1}^{N} Y_i \) and \( \bar{D}_N := \frac{1}{N} \sum_{i=1}^{N} D_i \) are the sample means of \( Y_i \) and \( D_i \), respectively.

We next show how to identify the marginal distributions of potential values \( Y_i(0) \) and \( Y_i(1) \) for each of these subpopulations: Under Assumptions A.1 and A.2, we can also identify the share of either complier population from the observable data,

\[
\pi_N := P(D_i(0) = 0, D_i(1) = 0) = \frac{P(W_i = 1, D_i = 0)}{P(W_i = 1)}
\]

\[
\pi_A := P(D_i(0) = 1, D_i(1) = 1) = \frac{P(W_i = 0, D_i = 1)}{P(W_i = 0)}
\]

\[
\pi_C := P(D_i(0) = 0, D_i(1) = 1) = 1 - \pi_A - \pi_N
\]

Furthermore, we can identify the respective marginal distributions for potential outcomes \( Y_i(0), Y_i(1) \) in either complier group: define the conditional c.d.f.

\[F^p(y|w,d) := P(Y_i \leq y|W_i = w, D_i = d)\]

From Assumptions A.1 and A.2 and the law of iterated expectations, it follows that

\[F^p(y|0,0) := P(Y_i(0) \leq y|D_i(0) = 0, D_i(1) = 1) = P(Y_i(0) \leq y|D_i(0) = 0, D_i(1) = 0)\pi_N\]

\[F^p(y|0,1) := P(Y_i(1) \leq y|D_i(0) = 1, D_i(1) = 1) = P(Y_i(1) \leq y|D_i(0) = 0, D_i(1) = 0)\pi_A\]

\[F^p(y|1,0) := P(Y_i(0) \leq y|D_i(0) = 0, D_i(1) = 0) = P(Y_i(0) \leq y|D_i(0) = 1, D_i(1) = 0)\pi_N\]

\[F^p(y|1,1) := P(Y_i(1) \leq y|D_i(0) = 0, D_i(1) = 1) = P(Y_i(1) \leq y|D_i(0) = 1, D_i(1) = 1)\pi_A\]

so that we can express the c.d.f. of potential outcomes for either complier population as

\[F^p_0(y_0|N) := P(Y_i(0) \leq y_0|D_i(0) = 0, D_i(1) = 0) = \frac{F^p(y_0,0)}{\pi_N}\]

\[F^p_0(y_0|C) := P(Y_i(0) \leq y_0|D_i(0) = 0, D_i(1) = 1) = \frac{F^p(y_0,1,0) - F^p(y_0,0)\pi_N}{\pi_C}\]

\[F^p_1(y_1|C) := P(Y_i(1) \leq y_1|D_i(0) = 0, D_i(1) = 1) = \frac{F^p(y_1,1) - F^p(y_1,0,1)\pi_A}{\pi_C}\]

\[F^p_1(y_1|A) := P(Y_i(1) \leq y_1|D_i(0) = 1, D_i(1) = 1) = \frac{F^p(y_1,1,0)}{\pi_A}\]

and estimate the distributions using their sample analogs. Note that \( F^p_0(y_0|A) \) and \( F^p_0(y_0|N) \) are not identified.

Now we can rewrite \( \tau_{LATE} \) in terms of \( F^p_0(\cdot|C) \) and \( F^p_1(\cdot|C) \),

\[\tau_{LATE} := E_{F^p_0(\cdot|C)}[Y_i(1)] - E_{F^p_0(\cdot|C)}[Y_i(0)]\]

so that the parameter is identified from \( F^p_0(\cdot|C) \) and \( F^p_1(\cdot|C) \) alone. Moreover, we can see that we only need to impute missing counterfactuals for the compliers in the population, whereas by definition the always and never takers do not change their treatment status in response to a change in \( W_i \). Applying Proposition 2.1 on the conditional distribution given \( C \), we can furthermore conclude that the variance of \( \hat{\tau}_{Wald} \) is maximized at the isotone coupling between \( F^p_0(\cdot|C) \) and \( F^p_1(\cdot|C) \).

These considerations suggest the following bootstrap procedure:

- Create an empirical population of \( N \) units, where for the first \( N_N := \hat{\pi}_N N \) units we set \( \bar{D}_i(0) = \bar{D}_i(1) = 0 \), for the next \( N_C := \hat{\pi}_C N \) units we let \( \bar{D}_i(0) = 0, \bar{D}_i(1) = 1 \), and the remaining \( N_A \hat{\pi}_A N \) units we set \( \bar{D}_i(0) = \bar{D}_i(1) = 1 \).

- Assign potential outcomes \( \bar{Y}_i(0) = \bar{F}_0^{-1}\left( \frac{i}{N_N} \right) \) to the first \( N_N \) units, \( \bar{Y}_i(0) = \bar{F}_0^{-1}\left( \frac{i}{N_C} \right) \) to the next \( N_C \) units, and \( \bar{Y}_i(1) = \bar{F}_1^{-1}\left( \frac{i}{N_A} \right) \) to the remaining \( N_A \) units.

- Impute potential outcomes \( \bar{Y}_i(1) \) to the units \( N_N + 1, \ldots, N_N + N_C \) according to the isotone coupling of \( \bar{F}_1(n_1|C) \) to \( \bar{F}_0(n_0|C) \). The counterfactuals for the never and always takers will not be relevant for the remaining steps, so we do not impute any potential values \( \bar{Y}_i(1) \) for never takers or \( \bar{Y}_i(0) \) for always takers.

- For each bootstrap replication, draw \( R^*_i \) and \( W^*_i \) at random and include observations with \( R^*_i = 1 \), treatment status \( \bar{D}^*_i := \bar{D}_i(W^*_i) \), and outcome \( Y^*_i := Y_i(\bar{D}^*_i) \).

The structure of this procedure is completely analogous to the completely randomized case with perfect compliance, except that we need to impute the missing counterfactuals for the presumed compliers, whose share \( \pi_C \) is also estimated from the data. Establishing formal properties of this procedure is beyond the scope of this paper.
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

\[(Y_i(0), Y_i(1)) \perp W_i\]

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

\[
\hat{F}_{0n}(y_0) := \frac{1}{n(1 - p)} \sum_{i=1}^{N} R_i (1 - W_i) 1\{Y_i \leq y_0\}
\]
\[
\hat{F}_{1n}(y_1) := \frac{1}{np} \sum_{i=1}^{N} R_i W_i 1\{Y_i \leq y_1\}
\]

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,

\[
\tau(F_0, F_1) := E_{F_1}[Y(1)] - E_{F_0}[Y(0)], \quad \text{and} \quad \sigma(F_0, F_1) := \sigma(C^{iso}(F_0, F_1))
\]

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 \(P : \mathcal{H} \to \mathbb{R}\) 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 \(T^{(s+1)}(P)(h_1, \ldots, h_{s+1})\) that is multilinear in its \(s+1\) arguments and is defined recursively as

\[
\lim_{h_{s+1} \to 0} \frac{\|T^{(s)}(P + h_1, \ldots, h_s) - T^{(s)}(P)(h_1, \ldots, h_s) - T^{(s+1)}(P)(h_1, \ldots, h_s, h_{s+1})\|}{\|h_{s+1}\|} = 0
\]

Denoting \(V_t := (Y_t(0), Y_t(1))\), we let

\[
g_{1p}(V_t) := \tau'(F_01)(\delta_{V_t} - F_{01})
\]
\[
g_{2p}(V_t, V_j) := \tau''(F_01)(\delta_{V_t} - F_{01}, \delta_{V_j} - F_{01})
\]

where \(\delta_v\) is the point mass at \(V_t = v\), i.e. a distribution with \(P(V_t = v) = 1\).

**Assumption B.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 \(g(V_t)\) satisfies Cramér’s condition,

\[(B.1) \quad \varrho := 1 - \sup_{t : \sqrt{n} \geq |h| \geq t_0} \mathbb{E} \left[ \exp \left\{ i t g_{1p}(V_t) \right\} \right] > 0\]

where \(t_0 := b_1 / \mathbb{E}[g_{1p}(V_t)]^3\) 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 \(T(F_0, F_1) := (\tau(F_0, F_1) - \tau(F_{01}^{0})) / \sigma(F_0, F_1)\), we have that \(\|T^{(2)}\|^2_{\infty} < \infty\). Part \(b\) translates the formulation of Cramér’s condition from [9] into the notation of this paper.

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

\[
\left\| F_n \left( \frac{\sqrt{n} \left( \hat{F}_{0n}, \hat{F}_{1n} \right) - \tau(\hat{F}_{0n}, \hat{F}_{1n})}{\sigma(\hat{F}_{0n}, \hat{F}_{1n})} \right) - P \left( \frac{\sqrt{n} \left( \hat{F}_{0n}, \hat{F}_{1n} - \tau(F_{0n}, F_{1n}) \right)}{\sigma(F_{0n}, F_{1n})} \right) \right\|_{\infty} = O_p(n^{-1})
\]
Proof of Theorem B.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 \( i \)th observation, and without loss of generality we assume that observations are ordered such that \( R_i = 1 \) for \( i = 1, \ldots, n \) and \( R_i = 0 \) for \( i = n + 1, \ldots, N \).

**Separate contributions of sampling and randomization error.** We can write the estimation error in the upper bound for \( \tau \) as
\[
\sqrt{n}(\hat{\tau} - \tau) = \sqrt{n}(\hat{\tau}(\hat{F}_{0n}, \hat{F}_{1n}) - \tau(F_0^*, F_1^*)) + \sqrt{n}(\tau(F_0^*, F_1^*) - \tau(F_0^*, F_1^*)) =: B_1 + B_2
\]
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 \( F_0^* \). 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
\[
g_{1s}(V_i) := \frac{N-1}{N-n} (\mathbb{E}[\tau(F_0^*, F_1^*)|V_i] - \mathbb{E}[\tau(F_0^*, F_1^*)])
\]
\[
g_{2s}(V_i, V_j) := \frac{(N-2)(N-3)}{(N-n)(N-n-1)} \left[ \mathbb{E}[\tau(F_0^*, F_1^*)|V_i, V_j] - \mathbb{E}[\tau(F_0^*, F_1^*)] - \frac{N-n}{N-2} (g_{1s}(V_i) + g_{1s}(V_j)) \right]
\]
where we also let \( \sigma_{1s}^2 := \text{Var}(g_{1s}(V_i)) \).

Now consider \( T_1 \equiv T_1(Z_1, \ldots, Z_n) := B_1/S_1 \). By Theorem 1 in [8], \( T_1 \) 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
\[
\mathbb{E}_p R_{31}^2 \leq n_*^{-1} \mathbb{E} \left[(n_*D_1D_2T_1)^2\right] := n_*^{-1} \delta_s
\]
where
\[
D_1D_2T_1 := T_1(V_1, V_2, \ldots, V_n) - T_1(V_{n+1}, V_2, \ldots, V_n) - T_1(V_1, V_{n+2}, \ldots, V_n) + T_1(V_{n+1}, V_{n+2}, \ldots, V_n)
\]
Adapting the formal argument in sections 2.3 and 2.6 of [5] we can bound
\[
\delta_s := \mathbb{E} \left[(n_*D_1D_2T_1)^2\right] \leq C n_*^{-2} \|T(2)\|^2_{\infty}
\]
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 B.1 that \( \delta_s = O(n_*^{-2}) \).

**Orthogonal decomposition for randomization error.** From (5.2) and Section 2.6 in [5] we obtain a similar decomposition for the randomization error conditional on the sample. Specifically, let
\[
g_{1r}(V_i) := \frac{n-1}{n-n_1} \sum_{w=0}^{n-n_1-1} \xi(w) (\mathbb{E}[\tau(F_0^*, F_1^*)|Y_i(z)] - \mathbb{E}[\tau(F_0^*, F_1^*)])
\]
\[
g_{2r}(V_i, V_j) := \frac{(n-2)(n-3)}{(n-n_1)(n-n_1-1)} \sum_{w, w'=0}^{n-n_1-1} \xi(w) \xi(w') \times \left( \mathbb{E}[\tau(F_0^*, F_1^*)|Y_i(z), Y_j(z')] - \mathbb{E}[\tau(F_0^*, F_1^*)] \right) \frac{n-n_1}{n-2} (g_{1r}(V_i) + g_{1r}(V_j))
\]
where \( \xi(w) := (w-p)/(1-p) \), and we also let \( \sigma_{1r}^2 := \text{Var}(g_{1r}(V_i)) \). Then, applying Theorem in [8] again, we can expand \( T_2 \equiv T_2(V_1, \ldots, V_n) := B_2/S_2 \) as
\[
T_2 = \mathbb{E}_p T_2 + U_{12} + U_{22} + R_{32}
\]
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
\[
\mathbb{E}_p R_{32}^2 \leq C(n_*p)^{-2} \|T(2)\|^2_{\infty}
\]
**Edgeworth expansion.** By Theorem 1 in [8], these bounds on the remainders $R_{1n}, R_{2n}$ establish the validity of separate orthogonal expansions

\[
(B.2) \quad T_1 + T_2 = E[T_1 + T_2] + (U_{11} + U_{21}) + (U_{12} + U_{22}) + (R_{1n} + R_{2n})
\]

where $E[R_{1n} + R_{2n}^2] \leq C\|T(2))\|_{\infty}^2(n^p)^{-1}$. Noting that $U_{21}, U_{22}$ are conditionally independent of $U_{11}, U_{12},$ it follows from Theorem 1.1 in [9] 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

\[
\mathbb{P}\left(\frac{\sqrt{n}(\hat{f}_{0n}, \hat{f}_{1n} - \tau(f_{0n}^*, f_{1n}^*))}{\sigma(f_{0n}, f_{1n})} \leq t\right) = \Phi(t) - \frac{(1 - 2q)\alpha_s + (1 - 2p)\alpha_r + 3(\kappa_s + \kappa_r)}{6n(N - n)/N} \Phi''(t) + \Delta_n
\]

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

\[
\alpha_s := \frac{1}{N} \sum_{i=1}^{N} \sigma^{-3} g_{1s}(V_i)^3
\]

\[
\alpha_r := \frac{1}{N} \sum_{i=1}^{N} \sigma^{-3} g_{1r}(V_i)^3
\]

\[
\kappa_s := \frac{2Ng(1 - q)}{N - 1} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma^{-3} g_{2s}(V_i, V_j)g_{1s}(V_i)g_{1s}(V_j)
\]

\[
\kappa_r := \frac{2Nqp(1 - p)}{N - 1} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma^{-3} g_{2r}(V_i, V_j)g_{1r}(V_i)g_{1r}(V_j)
\]

Under Assumption B.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 $\hat{F}_{0n} := C(\hat{F}_{0n}, \hat{F}_{1n})$ is consistent for the finite-population distribution $F_{01}^p := C(F_{01}^p, F_{11}^p)$, 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 (B.2), and Cramér’s condition (B.1) holds with the same constant $\theta$.

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 $\sigma^*$ in analogy to their sampling/randomization counterparts. We then have that for the bootstrap distribution,

\[
\mathbb{P}^*\left(\frac{\sqrt{n}(\hat{f}_{0n}, \hat{f}_{1n}^* - \tau(\hat{f}_{0n}^*, \hat{f}_{1n}^*))}{\sigma(f_{0n}^*, f_{1n}^*)} \leq t\right) = \Phi(t) - \frac{(1 - 2q)\alpha_s^* + (1 - 2p)\alpha_r^* + 3(\kappa_s^* + \kappa_r^*)}{6n(N - n)/N} \Phi''(t) + \Delta_n^*
\]

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

\[
\alpha_s^* := \frac{1}{N} \sum_{i=1}^{N} (\sigma^*)^{-3} g_{1s}(V_i^*)^3
\]

\[
\alpha_r^* := \frac{1}{N} \sum_{i=1}^{N} (\sigma^*)^{-3} g_{1r}(V_i^*)^3
\]

\[
\kappa_s^* := \frac{2Ng(1 - q)}{N - 1} \sum_{i=1}^{N} \sum_{j=1}^{N} (\sigma^*)^{-3} g_{2s}(V_i, V_j)g_{1s}(V_i)g_{1s}(V_j)
\]

\[
\kappa_r^* := \frac{2Nqp(1 - p)}{N - 1} \sum_{i=1}^{N} \sum_{j=1}^{N} (\sigma^*)^{-3} g_{2r}(V_i, V_j)g_{1r}(V_i)g_{1r}(V_j)
\]

**Convergence of bootstrap cumulants.** It remains to be shown that $\sigma^*$ and the cumulants $\alpha_s^*, \alpha_r^*, \kappa_s^*, \kappa_r^*$ converge in probability at a root-$n$ rate to their population counterparts. Under the assumption of six bounded moments for $g_{1s}, g_{1r}, g_{2s}, g_{2r}$ and noting that $0 < p < 1$, this can be established using a finite-population CLT. The conclusion then follows from Theorem 1.1 in [9].

\[\Box\]
C.1. Randomization Distribution for \( \hat{F}_0(y_0), \hat{F}_1(y_1) \). We first compute the randomization covariance \( \text{Cov}_{W,R}(\hat{F}_0(y_0), \hat{F}_1(y_1)) \) given the population distribution \( F_0^p(y_0, y_1) \), where

\[
\hat{F}_0(y_0) = \frac{1}{n(1-p)} \sum_{i=1}^N R_i (1 - W_i) \mathbb{1}\{Y_i(0) \leq y_0\}
\]

\[
\hat{F}_1(y_1) = \frac{1}{np} \sum_{i=1}^N R_i W_i \mathbb{1}\{Y_i(1) \leq y_1\}
\]

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 \((Y_i(0), Y_i(1))\) in the population. We then have

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

\[
= \frac{1}{n^2 p(1-p)} \sum_{i=1}^N \sum_{j \neq i}^N \mathbb{E}[R_i R_j (1 - W_i) W_j] A_{0i} A_{1j}
\]

\[
= \frac{1}{n^2 p(1-p)} \sum_{i=1}^N \sum_{j \neq i}^N \frac{n(n-1)}{N(N-1)} \frac{n^2 p(1-p)}{n(n-1)} A_{0i} A_{1j}
\]

\[
= \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i}^N A_{0i} A_{1j} = \frac{1}{N(N-1)} \left( \sum_{i=1}^N A_{0i} \left( \sum_{j=1}^N A_{1j} \right) - \sum_{i=1}^N A_{0i} A_{1i} \right)
\]

so that

\[
\text{Cov}(\hat{F}_0(y_0), \hat{F}_1(y_1)) = -\frac{1}{N-1} \left( F_0^p(y_0, y_1) - F_0^p(y_0) F_0^p(y_1) \right)
\]

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

\[
\text{Cov}(\hat{F}_0(y_0), \hat{F}_0(y_1)) = \frac{1}{n^2 (1-p)^2} \sum_{i=1}^N \sum_{j \neq i}^N \mathbb{E}[R_i R_j (1 - W_i)(1 - W_j)] B_{0i} B_{1j}
\]

\[
= \frac{1}{n^2 (1-p)^2} \left[ \sum_{i=1}^N \frac{n(n-1)}{n} B_{0i} B_{1i} + \sum_{i=1}^N \sum_{j \neq i}^N \frac{n(n-1)(n(1-p) - 1)}{n^2} \right]
\]

\[
= \frac{1}{n(N-1)} \sum_{i=1}^N B_{0i} B_{1i} + \frac{(n-1)(n(1-p) - 1)}{n^2 (1-p)} \left[ \frac{1}{N^2} \sum_{i=1}^N \sum_{j \neq i}^N B_{0i} B_{1j} \right]
\]

\[
= \left( \frac{1}{N} \sum_{i=1}^N B_{0i} B_{1i} \right) \left( \frac{1}{n(1-p)} - \frac{(n-1)(n(1-p) - 1)}{N n^2 (1-p)} \right)
\]

\[
= \left( \min\{F_0^p(y_0), F_0^p(y_1)\} - F_0^p(y_0) F_0^p(y_1) \right) \left( \frac{1}{n(1-p)} - \frac{1}{N} + O \left( \frac{1}{nN} \right) \right)
\]

Similarly,

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

Furthermore,

\[
\text{Cov}(\hat{F}_0(y_0), \hat{F}_0(y_1)) = \frac{1}{n} \min\{F_0^p(y_0), F_0^p(y_1)\} - F_0^p(y_0) F_0^p(y_1)
\]

\[
\text{Cov}(\hat{F}_1(y_0), \hat{F}_1(y_1)) = \frac{1}{n} \min\{F_1^p(y_0), F_1^p(y_1)\} - F_1^p(y_0) F_1^p(y_1)
\]
We let \( H \) denote the covariance kernel of the randomization process with elements
\[
H_{00}(y_0, y'_0) = \lim_n n \text{Cov}(\hat{F}_0(y_0), \hat{F}_0(y'_0)) = \left( \frac{1}{1-p} - \frac{n}{N} \right) \left( \min\{F_0^p(y_0), F_0^p(y'_0)\} - F_0^p(y_0)F_0^p(y'_0) \right)
\]
(C.1) \( H_{01}(y_0, y_1) = \lim_n n \text{Cov}(\hat{F}_0(y_0), \hat{F}_1(y_1)) = \lim_n \frac{n}{N} \left( F_0^p(y_0, y_1) - F_0^p(y_0)F_0^p(y_1) \right) \)
\( H_{11}(y_1, y'_1) = \lim_n n \text{Cov}(\hat{F}_1(y_1), \hat{F}_1(y'_1)) = \left( \frac{1}{p} - \frac{n}{N} \right) \left( \min\{F_1^p(y_1), F_1^p(y'_1)\} - F_1^p(y_1)F_1^p(y'_1) \right) \)
Note also that \( \frac{1}{1-p} \geq 1 \geq \frac{n}{N} \geq 0 \) and \( \frac{1}{p} \geq 1 \geq \frac{n}{N} \geq 0 \), so that \( H_{00}(\cdot, \cdot) \) and \( H_{11}(\cdot, \cdot) \) are nonnegative.

### C.2. Proofs of Section 2.5.

#### C.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 isotope 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 C.1. (Ordering of Distributions)** Let \( F_0, F_1 \) 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 isotope coupling. Then for any convex function, the randomization distribution for \( \tau_{ATE} \) satisfies
\[
E_{F_{01}^{iso}}[v(\hat{\tau}_{ATE})] \geq E_{F_{01}}[v(\hat{\tau}_{ATE})]
\]
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 isotope ( 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. [4], [15], and [41]). For illustrative purposes, we give a complete proof here.

**Proof:** In order to establish second-order stochastic dominance with respect to the isotope 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
\[
\hat{\tau}_{ATE} = \frac{1}{n} \left( B_{-ij} + R_i W_i (Y_i(0)/(1-p) + Y_i(1)/p) + R_j W_j (Y_j(0)/(1-p) + Y_j(1)/p) \right)
\]
where \( B_{-ij} := \sum_{k \neq i,j} R_k (W_k(Y_k(0)/(1-p) + Y_k(1)/p) - Y_k(0)/(1-p)) - (R_i Y_i(0) + R_j Y_j(0))/(1-p) \).

We can now consider the change in \( E[v(\hat{\tau}_{ATE})] \) 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 \( E[v(\hat{\tau}_{ATE})] \) if and only if
\[
0 \leq P(W_i = 1, W_j = 0) \left\{ \frac{v}{n} \left( B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p \right) - v \left( \frac{1}{n} \left( B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p \right) \right) \right\}
\]
\[
+ P(W_i = 0, W_j = 1) \left\{ v \left( \frac{1}{n} \left( B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p \right) \right) - v \left( \frac{1}{n} \left( B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p \right) \right) \right\}
\]
\[
= p(1-p) \left\{ v \left( \frac{1}{n} \left( B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p \right) \right) + v \left( \frac{1}{n} \left( B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p \right) \right) 
\]
\[
- v \left( \frac{1}{n} \left( B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p \right) \right) - v \left( \frac{1}{n} \left( B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p \right) \right) \right\}
\]
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 isotope assignment by pairwise substitutions of this form, the isotope assignment maximizes the expectation
\[
E[v(\hat{\tau}_{ATE})] = E \left[ v \left( \frac{1}{n} \sum_{i=1}^N R_i \left( W_i Y_i(1)/p - (1 - W_i) Y_i(0)/(1-p) \right) \right) \right]
\]
for all convex functions \( v(\cdot) \). Therefore the distribution of \( \hat{\tau}_{ATE} \) under the isotope assignment dominates that under any alternative coupling, as claimed above.
Proof of Proposition 2.1: The claim in the proposition follows immediately from Lemma C.1 and the observation that the function \( v(y) = y^2 \) is strictly convex

C.3. Proofs for Section 5.

C.3.1. Proof of Theorem 5.1. From standard results (see e.g. Example 19.6 in [42]), the class \( \mathcal{F} := \{(-\infty, y] : y \in \mathbb{R}\} \) is Glivenko-Cantelli, so that \( (\hat{F}_0 - F_0, \hat{F}_1 - F_1) \) 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 [42]).

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 [43]). Let \( \mathcal{F} := \{1[y \leq (-\infty, t]) : t \in \mathbb{R}\} \) 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 C.2. Suppose that \( (Y_1(0), Y_1(1)) \overset{iid}{\sim} F_{01} \). Then the randomization process

\[
\hat{G}_n := \sqrt{n} \left( \hat{F}_0 - F_0 - \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\} \right)
\]

converges in outer probability to \( G \) under the bounded Lipschitz metric,

\[
\sup_{h \in BL_1} |E_W h(\hat{G}_n) - Eh(G)| \rightarrow 0
\]

in outer probability, where \( G \) is a Gaussian process with covariance kernel \( H \).

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

\[
F_{01}^i(y_0, y_1) := \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}
\]

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

\[
F_{01}^c(y_0, y_1) := \frac{1}{np} \sum_{i=1}^{N} R_i W_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}
\]

Using this notation we can write

\[
\sqrt{n}(F_{01}^i(y_0, y_1) - F_{01}^p(y_0, y_1)) = \sqrt{n}(F_{01}^i(y_0, y_1) - F_{01}^c(y_0, y_1)) + \sqrt{n}(F_{01}^c(y_0, y_1) - F_{01}^p(y_0, y_1))
\]

Since \( R_i, W_i \) are drawn at random and without replacement, it follows from Theorem 3.1 in [6] that

\[
\sqrt{n}(F_{01}^c(y_0, y_1) - F_{01}^p(y_0, y_1)) \overset{d}{\rightarrow} G_{F_{01}^p}
\]

for Brownian bridges \( G_{F_{01}^p} \). 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,

\[
\sqrt{n}(F_{01}^c - F_0^p) \overset{d}{\rightarrow} G_{F_0^p} + G_{F_0^p}
\]

Finally, \( \hat{F}_1(y_1) \equiv F_1^c(y_1) \) and \( \hat{F}_0(y_0) \equiv \frac{1}{np-1}(F_0^c(y_0) - pF_0^c(y_0)) \), establishing the claim, where the structure of the covariance kernel follows from the point-wise calculations in the derivation of (C.1).
C.3.2. Proof of Theorem 5.2: From Assumption 2.1 it is immediate that \( \tau(F_0, F_1) \) is Hadamard-differentiable. Lemma C.2 and the functional delta method, see e.g. Theorem 20.8 in [42], 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

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

by randomizing from \( \hat{F}_0 \). We also define the asymptotic covariance kernel \( H^{iso} \) corresponding to the coupling \( C^{iso} \) in analogy to (C.1) where \( F_{01} \) is replaced with \( C^{iso}(F_0, F_1) \). We first show the two following Lemmas:

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

\[
\hat{G}_n^* := \sqrt{n} \left( \begin{array}{c} \hat{F}_0^* - \hat{F}_0 \\ \hat{F}_1^* - \hat{F}_1 \end{array} \right)
\]

converges in outer probability to \( G \) under the bounded Lipschitz metric, that is

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

in outer probability, where \( G \) is a Gaussian processes with covariance kernel \( H \).

**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 replicates \( \hat{F}_0, \hat{F}_1 \) are generated by randomization from the samples \( (\hat{Y}_i(1), \hat{Y}_i(1))_{\iota=1}^{\infty} \) corresponding to the joint distribution \( \hat{F}_{01} := C^{iso}(\hat{F}_0, \hat{F}_1) \).

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

\[
\|H^{iso} - H^{iso}\| \overset{a.s.}{\to} 0
\]

which completes the proof.

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

C.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 C.4 follows from Lemma C.4 and the functional Delta method (e.g. Theorem 20.8 in [42]).

C.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,

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

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 $C_X$ is the isotone coupling,

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

By the same line of reasoning, the least favorable coupling in the set $C_B$ 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 $\square$

C.4.2. Proof of Theorem 6.1. The proofs of Theorems 5.1-5.3 followed from Lemmas C.2-C.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 C.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 [6] holds for multivariate distributions of arbitrary (finite) dimension. Specifically, if we let $F_{01}^p(y_0, y_1, b, e)$ and $F_{01}^p(y_0, y_1, b, e)$ 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

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

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

Next, let $F_{01}^p(y_0, y_1, b, e)$ 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 [43] to obtain

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

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

It then follows that the randomization process

$$\hat{G}_n := \sqrt{n} \left( \hat{F}_0(y_0, b) - \hat{F}_0^p(y_0, b) \right)$$

converges in outer probability to $G$ under the bounded Lipschitz metric,

$$\sup_{h \in BL_1} |E_{W_i} h(\hat{G}_n) - Eh(G)| \to 0$$

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

We next show that we can extend Lemma C.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 C.3 then follows from the dominated convergence theorem for the integral over $F_B(b)$, noticing that the isotone copula $C_B^\text{iso}(F_0, F_1)$ is continuous in its arguments.

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

$$\hat{G}_n := \sqrt{n} \left( \frac{\hat{F}_0(y_0, b) - \hat{F}_0^p(y_0, b)}{\hat{F}_1(y_1, b) - \hat{F}_1^p(y_1, b)} \right)$$

converges in outer probability to $G^*$, a Brownian bridge with covariance kernel $H^*$, where $H^*$ is the covariance kernel derived in Appendix C.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 $\sigma^2(F_0, F_1)$ is conservative for $\sigma^2(F_{01})$ $\square$
REFERENCES


