Causal inference with misspecified exposure mappings

Fredrik Sävje*

July 11, 2019

Abstract

Exposure mappings facilitate investigations of complex causal effects when units interfere in experiments. Current methods assume that the exposures are correctly specified. The assumption can, however, not be verified, and it is questionable in many settings. This paper investigates whether inferences about exposure effects can be drawn when the exposures are misspecified. The main result is a law of large numbers under weak conditions on the errors introduced by the misspecification. In particular, the rate of convergence is determined by the dependence between units' specification errors, and consistency is achieved even when the errors are large as long as they are sufficiently independent. The limiting distribution of the estimator is also discussed. Asymptotic normality is achieved under stronger conditions than those needed for consistency. Similar conditions also facilitate conservative variance estimation.

Keywords: XXX, YYY, ZZZ.

*Department of Political Science and Department of Statistics & Data Science, Yale University.

1 Introduction

Investigators face two challenges when drawing causal inferences in the presence of interference. The first is definitional. Conventional treatment effects are not well-defined when units interfere, and the effects would in any case not capture causal aspects of the interference itself. The second challenge is inferential. Conventional estimators may be not be applicable, or they may not perform well. A common solution to both challenges is the introduction of exposure mappings.

An exposure mapping provides a terse representation of the nominal treatments assigned to the units in the sample. The exposures are defined to capture aspects of the treatment assignment deemed relevant or interesting for the question at hand. However, the way the exposures are later used requires that they also capture all structural causal information in the study. That is, the exposures are assumed to be correctly specified. The assumption is useful because it allows investigators to use standard causal inference techniques under no interference. The downside is that detailed knowledge of the structure of the interference is required to make the specification correct. We must for example know (or presume to know) whether the treatment assigned to one unit has the potential to affect the outcome of any other unit in the sample. Such insights are rare, and investigators have been forced to make unverified and often questionable assumptions to draw causal inference under interference.

This paper considers inference about exposure effects when the exposures are misspecified. It first provides a definition of an exposure effect that is robust to misspecification. The effect is defined as the average difference in expected outcomes for the two exposures under consideration. The definition has the advantage that it coincides with the conventional exposure effect when the exposures are correctly specified but remains well-defined when the exposures are misspecified.

The paper next considers whether inferences can be drawn about these misspecification-robust exposure effects. The focus is on conventional estimators of exposure effects. The main contribution is to show that the estimators are consistent for the misspecification-robust exposure effects given weak conditions on the specification errors. The critical

condition is that the dependence between unit's specification errors is sufficiently controlled. Like the assumptions currently used to investigate exposure effects, the weak dependence assumption is generally not testable. Its strength is instead that it is considerably weaker than prevailing assumptions. Assuming the exposures are correctly specified is equivalent to assuming that the specification errors are uniformly zero. A weak dependence allows for potentially grave misspecification as long as the units' exposures are not misspecified in the same way. The final contribution is a discussion about variance estimation. The tasks are less tractable than point estimation, but conditions allowing some progress are discussed.

2 Related work

The first glimpse of the idea of exposure mappings can be seen in Halloran & Struchiner (1995) who discuss causal inference under interference and provide some foundational definitions. This initial work was later extended by Sobel (2006) and Hudgens & Halloran (2008) who consider effects that we today would recognize as exposure effects. The authors consider exposures based on proportions of treated units in neighborhoods. They ask what the effect is when, say, 25% of a unit's neighbors are treated versus when 75% are treated. Two assumptions are used: *partial interference* and *stratified interference*. The first stipulates that only the treatment assignments of units in a unit's neighborhood (defined as disjoint groups) affect the unit's outcome. The second assumption stipulates that only the proportion of treated units affects the outcome. Taken together, the two assumptions amount to assuming that the neighborhood proportion of treated units together with a unit's own treatment assignment is a complete description of the causal structure, or in other words, that the specification the authors use is correct.

The results were later extended to other settings. For example, Toulis & Kao (2013) consider when interference is restricted to neighborhoods in known social networks, which may not be disjoint. The idea was finally taken to full generality by Manski (2013) and Aronow & Samii (2017). The authors point out that key methodological tool in all these approaches is a terse description of treatment assignment. The insight suggests a general-

ization of the approach to any setting where summaries of treatment assignments are the interest. In more detail, the authors consider a function that maps from the full set of treatments to some low-dimensional representation thereof. The elements of the codomain of this function are given labels, and contrasts between outcomes under different labels are interpreted as causal effects. Manski (2013) calls these labels "effective treatments," and Aronow & Samii (2017) call them "exposures." The latter term has become standard and will be used here. Independently of the terminology, both sets of authors impose the assumption that the exposures are correctly specified. That is, they assume that the exposures provide a complete description of the causal structure. Inference can then proceed as usual but with the exposures substituted for the nominal treatments.

The assumption that the exposures are correctly specified is critical to prove the theoretical results in Manski (2013) and Aronow & Samii (2017). The assumption has a direct parallel to the no-interference assumption often used to facilitate inference about average treatment effects. That is, the average effect of the units' own treatment on their own outcome. This may be seen an exposure effect based on a particularly simple exposure mapping with two level only depending on a unit's own treatment assignment. The necessity of the no-interference in this setting has recently been investigated by Sävje et al. (2018) and Chin (2018). The authors, respectively, provide a law of large number and a central limit theorem for marginal direct treatment effects under completely unmodeled interference. In other words, they show that inferences can be drawn about average treatment effects even if the units' treatments are misspecified. These results are the inspiration of the ones presented here. The paper essentially connects these ideas with the literature discussed above, generalizing the results to exposure mappings of arbitrary complexity. The proof strategies are, however, quite different, which provides additional insights.

Discussions about misspecification under interference are rare, but they are not limited to the two aforementioned contributions. One strand builds on Fisher (1935). The idea is that randomization tests can be constructed without the need for correctly specified exposures as long as the tested hypotheses is sufficiently precise (Aronow, 2012; Bowers et al.,

2013; Athey et al., 2018; Basse et al., 2018). The approach requires that the response of a subset of units are perfectly known under the null hypothesis for all possible treatment assignments for some subset of assignments. Rosenbaum (2007) uses similar ideas to show that certain test statistics can be inverted to form an estimate of the shift in the ranks of the outcomes for treated units relative to controls without any restrictions on the interference. Choi (2017) extends the approach to estimate the size, rather than rank shifts, of attributable effects of treatment under the assumption of non-negative effects.

The Fisherian approach does not easily accommodate estimation of exposure effects. Closer to the current investigation is a set contributions using more conventional estimation approaches. Eckles et al. (2017) discuss strategies to minimize bias introduced by violations to no-interference assumptions. Basse & Airoldi (2018) and Karwa & Airoldi (2018) provide impossibility results for inference about causal quantities when no assumptions are made about the interference structure. Egami (2018) studies estimation of spillover effects in partially unobserved interference networks, which is a way to formalize misspecification.

3 Misspecified exposures

3.1 Preliminaries

Consider a sample of n units indexed by $\mathcal{U} = \{1, 2, \ldots, n\}$ and a set of treatments indexed by $\mathcal{Z} \subseteq \mathbb{N}$. Each unit is assigned one of the treatments: $z_i \in \mathcal{Z}$. The assignments of all units are collected in $\mathbf{z} = (z_1, \ldots, z_n) \in \Omega = \mathcal{Z}^n$. Binary treatments are often used, in which case $\Omega = \{0, 1\}^n$, but the current discussion applies more generally.

Let the function $y_i \colon \Omega \to \mathbb{R}$ map to the observed outcome for unit i under a specific (potentially counterfactual) assignment of treatments (Neyman, 1923; Holland, 1986). That is, $y_i(\mathbf{z})$ is the response of i when the treatments are assigned as $\mathbf{z}$. The elements of the image of the function are *potential outcomes*. The potential outcomes are assumed to be well-defined throughout the paper. This requires that no hidden versions exist of the treatments in Ω and that the outcomes are not inherently random. Finally, the potential

outcomes are also assumed to be bounded. Boundedness is unnecessarily strong, but it eases the exposition. Lemma A3 in the appendix provides details on how it can be weaken.

Condition 1 (Bounded potential outcomes)**.** For some $k_1 < \infty$ and all $i \in \mathcal{U}$ and $\mathbf{z} \in \Omega$, $|y_i(\mathbf{z})| \leq k_1$.

The treatments to assign are selected at random according to some probability space. Let $\mathbf{Z}$ be a random variable denoting which treatment vector was randomly selected. The distribution of $\mathbf{Z}$ will be referred to as the *assignment mechanism* or the *design* of the study. The design is the sole source of randomness under consideration, and the sample of units is considered fixed. The observed outcome Y_i for unit i is defined as the potential outcome corresponding to the randomly selected intervention: $Y_i = y_i(\mathbf{Z})$.

3.2 Exposures

The potential outcomes contain all causal information about the sample, and any causal quantity of interest can be expressed solely based on them. Definitions of such causal quantities may, however, be complex, and it is often hard to formulate and interpret them. Exposures and exposure mappings are used to make the definitions more intuitive.

The idea is that sets of treatments, i.e., subsets of Ω, often share similar causal interpretations. The exposure mappings are used to encode this information. Two treatment vectors are mapped to the same exposure if they are deemed similar. The exposures are, in other words, labels on groups of treatments that share the same or a similar causal interpretation. For example, if a vaccine is the focus of the study as in Hudgens & Halloran (2008), one exposure could be that 75% of a unit's neighbors are vaccinated. Another exposure could be that only a 25% are vaccinated.

To state this formally, consider a set of exposure labels indexed by $\Delta \subseteq \mathbb{N}$. A function $d_i \colon \Omega \to \Delta$ exists for each unit that maps from all possible assignments to the exposures. That is, the exposure of unit i is $d_i(\mathbf{z})$ when the treatments are assigned according to $\mathbf{z}$. If $d_i(\mathbf{z}) = d_i(\mathbf{z}')$, then $\mathbf{z}$ has a similar causal interpretation as $\mathbf{z}'$ with respect to unit i. In the

example above, $\mathbf{z}$ and $\mathbf{z}'$ could be similar in that 25% of unit i's neighbors are vaccinated in both, but they might differ in which quarter is vaccinated. The realized exposure is a random variable because the treatments are randomly assigned. Let $D_i = d_i(\mathbf{Z})$ denote the realized exposure for unit i, and let $\pi_i(d) = \Pr(D_i = d)$ be its marginal distribution. A positivity assumption will initially be made on the distribution of D_i. This is later relaxed in Section 5.2.

Condition 2. An exposure $d \in \Delta$ satisfies *positivity* if $1/\pi_i(d) \leq k_2$ for some $k_2 < \infty$ and all $i \in \mathcal{U}$.

3.3 Conventional exposure effects

To facilitate a simple definition of exposure effects, the exposures are conventionally assumed to be correctly specified. That is to say that $d_i(\mathbf{z}) = d_i(\mathbf{z}')$ implies $y_i(\mathbf{z}) = y_i(\mathbf{z}')$ for all units and treatment assignments. Manski (2013) calls the assumption "constant treatment response," and Aronow & Samii (2017) call it "properly specified exposure mappings."

Correctly specified exposure mappings implies that each exposure corresponds to a unique and well-defined potential outcome for every unit. Under the assumption, a function $\tilde{y}_i \colon \Delta \to \mathbb{R}$ exists for each unit such that $\tilde{y}_i(d_i(\mathbf{z})) = y_i(\mathbf{z})$ for all $\mathbf{z} \in \Omega$. In other words, the exposures are assumed to accurately capture the causal structure in a sample. Since the full treatment vector provides no causal information in addition to what a unit's exposure already provides, we can use $\tilde{y}_i$ defined on Δ rather than the more cumbersome potential outcomes defined on the full Ω. The reduction in complexity can be considerable since $|\Omega|$ grows exponentially in n while $|\Delta|$ typically is fixed.

Causal effects can then be defined in the usual manner as contrasts between potential outcomes produced by the exposures. For example, the average causal effect of exposure $a \in \Delta$ relative to $b \in \Delta$ is:

$$\tilde{\tau}(a, b) = \frac{1}{n} \sum_{i=1}^{n} \big[\tilde{y}_i(a) - \tilde{y}_i(b)\big].$$

The interpretation of these effects is often straightforward because the exposures are chosen to have a natural causal interpretation.

3.4 Exposure effects under misspecification

The exposures are assumed to be correctly specified because the construction of $\tilde{y}_i$ requires it. In particular, no function $\tilde{y}_i : \Delta \to \mathbb{R}$ exists under misspecification such that $\tilde{y}_i(d_i(\mathbf{z})) = y_i(\mathbf{z})$ for all $\mathbf{z} \in \Omega$. Without such functions, the conventional exposure effect becomes ill-defined. A solution must provide analogues of exposure-based potential outcomes that remain well-defined even when the exposures are misspecified.

Let $\bar{y}_i : \Delta \to \mathbb{R}$ be a function such that $\bar{y}_i(d) = \mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d]$ where the expectation is taken over the design. The interpretation remains essentially the same as for $\tilde{y}_i$. The function captures the expected potential outcome under each exposure for each unit, so $\bar{y}_i(d)$ is the potential outcome we expect to be realized when unit i is assigned to exposure d. A definition of an exposure effect under misspecification is immediate.

Definition 1. The *misspecification-robust exposure effect* for exposures a and b is:

$$\tau(a, b) = \frac{1}{n} \sum_{i=1}^{n} \big[\bar{y}_i(a) - \bar{y}_i(b)\big].$$

Effects building on this idea has been discussed before in the literature. The earliest examples are the effects introduced by Hudgens & Halloran (2008). The authors derive their main results assuming that the exposures are correctly specified (i.e., under partial and stratified interference). They do, however, define the effect assuming only partial interference, which implicitly allows from some misspecification. The way they proceed is exactly as in Definition 1: they marginalize over all assignments that map to the same exposure.

The misspecification-robust exposure effect is also related to a discussion in Aronow & Samii (2017, Section 8). The authors derive the expectation of their estimator when their assumption that the exposures are correctly specified is relaxed. They show the expectation

is a particular weighted average of the potential outcomes defined on the full treatment vector, and this weighted average can be shown to coincide with Definition 1.

A more distant parallel can be draw with causal effects in the absence of interference. It used to be convention to assume that the causal effect of the units' treatment on their own outcome was the same for all units. Constant effects allowed focus to be directed towards a single well-defined causal parameter applicable to each unit under study. In this framework, if the effects were suspected to differ between the units, a model was made of the heterogeneity, and the model was assumed to be correctly specified. Investigators grew skeptical of these approaches because of the strong assumptions they involved, and focus shifted to unconditional or conditional average causal effects. These definitions do not presume that the effects are constant or can be captured by a model. Instead, the effects marginalize of any heterogeneity that may exist. If effect heterogeneity is the interest, the inferential targets are defined to be average effect for different types of units. This captures aspects of the heterogeneity of interest while marginalizing over all irrelevant aspects of the heterogeneity, bypassing the need to assume that the heterogeneity is perfectly modelled. Such heterogeneity-robust causal effects are analogous to the misspecification-robust exposure effects defined above.

3.5 Specification errors

Misspecification introduces specification errors. The errors can be formalized as differences between the actual outcomes and the outcomes predicted by the exposures. Or, equivalently, as differences between the potential outcomes based on the full treatment vector and the potential outcomes based on the exposures.

Definition 2 (Specification error)**.** $\varepsilon_i = Y_i - \bar{y}_i(D_i)$.

Assuming that the exposures are correctly specified is the same as assuming that the specification errors are zero with probability one. This insight suggests a way to weaken the assumption. Rather than assuming that the specification errors are zero, it may be

sufficient to ensure that the errors are sufficiently controlled. This is the idea explored in this paper.

The pair-wise dependence between errors is the critical factor to control. In particular, consistent estimation is possible if the dependence averaged over all pairs of units diminishes as the sample grows. The relevant pair-wise dependence is captured by $\mathrm{E}[\varepsilon_i\varepsilon_j \,|\, D_i, D_j]$. This quantity may, however, not provide much intuition about the sources of the dependence. The dependence between errors comes from two sources. The first is the conditioning event itself, capturing the fact that knowledge about j's exposure may provide information about i's outcome when the exposures are misspecified, and conversely with i's exposure and j's outcome. An example of this is when unit j interferes with unit i in a way that is not captured in i's exposure The second source is dependence in excess of what can be explained by the conditioning event. This capture the fact two units' errors can be dependent if misspecified in the same way even if exposures provide no information.

A better understanding about the two parts may be gained when we note that the specification errors are to some degree in our control because we decide how to define the exposures. Consider the case above when j's exposure provide information about i's outcome in excess of the information provided by its own exposure. A simple way to eliminate this misspecification error is to redefine i's exposure to include also the exposure of j. If i's redefined exposure is (D_i, D_j), no part of i's specification error can be explained by j's exposure. The idea of redefined exposures is straightforward, but its application is not. If applied to all units in the sample, the redefined exposure will be the interaction of all units' nominal exposures, and much of the reduction in complexity is lost. The idea does, however, suggest a decomposition of the specification error that will prove useful.

Let $\bar{y}_{ij} : \Delta \times \Delta \to \mathbb{R}$ be a function such that $\bar{y}_{ij}(d, q) = \mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d, D_j = q]$.[1] That is, $\bar{y}_{ij}(d, q)$ is the potential outcome of unit i when defined over the exposures for both i and j. The potential outcome based on the redefined exposures will be a more ac-

[1]The function may not be unambiguously defined if $D_i = d$ and $D_j = q$ is a measure zero event. The concern is valid but technical, so it is ignored for the moment. See Section A in the appendix for rigorous definitions.

curate representation of unit i's outcome than the potential outcome based on the nominal exposures because the former includes more information about the treatment vector. The difference between $\bar{y}_{ij}(d,q)$ and $\bar{y}_i(d)$, thus, captures the part of the specification error for unit i explainable by j's exposure.

Definition 3 (Explainable specification error)**.** $e_{ij}(d,q) = \bar{y}_{ij}(d,q) - \bar{y}_i(d)$.

While $\bar{y}_{ij}(d,q)$ is more accurate than $\bar{y}_i(d)$, it will generally not be correctly specified. The remaining error is the part that cannot be explained by j's exposure. This part is strictly speaking not unexplainable because the full treatment vector will always perfectly explain the potential outcomes in current setting, but it is unexplainable with respect to pairwise refinements of the exposures. Similar to Definition 2, the error not explainable by j's exposure is the difference between the actual outcome and the outcome predicted by the redefined exposures.

Definition 4 (Unexplainable specification error)**.** $u_{ij} = Y_i - \bar{y}_{ij}(D_i, D_j)$.

The overall specification error can now be decomposed using the explainable and unexplainable specification error. We have, in particular, $\varepsilon_i = e_{ij} + u_{ij}$ for any pair of units i and j, where $e_{ij} = e_{ij}(D_i, D_j)$ is the realized explainable specification error. The relevant quantities for the proposition soon to be presented is based on the components of the decomposed error.

Definition 5. The *average explainable error dependence* for exposure $d \in \Delta$ is:

$$E_d = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} e_{ij}(d,d) e_{ji}(d,d),$$

and the *average unexplainable error dependence* for the same exposure is:

$$U_d = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d).$$

Definition 5 captures pair-wise dependencies between errors of units. As implied by the discussion above, if $e_{ij}(d,d) = 0$, the knowledge that $D_j = d$ provides no insights about

Y_i in excess to the knowledge that $D_i = d$. Thus, $e_{ij}(d,d)e_{ji}(d,d)$ is non-zero only when the exposures of i and j both provide information about the other unit's outcome. The magnitude of the explainable errors matters only insofar that the dependence of the errors ensure that they are of a large magnitude simultaneously. Consider a vaccination trial as an example. Unit j in this trial is an asymptomatic (immune) potential carrier of the disease under study, while unit i will show symptoms, which are the outcomes of interest, when infected. The exposure assigned to j may provide information about i's outcome not contained in i's exposure in this case, because it may provide additional information about whether unit i is infected. Part of i's error is thus explainable by j's exposure, and $e_{ij}(d,d)$ is non-zero. However, i's exposure contains no information about j's symptoms, so $e_{ji}(d,d) = 0$. The lack of symmetry means that there is no dependence, according to this measure, between the errors.

Investigators may suspect that the explainable parts of the errors are fairly symmetric in many application. If they are perfectly symmetric so that $e_{ij}(d,d) = e_{ji}(d,d)$ for all units, E_d would collapse to a measure of magnitude. The point here is that perfectly symmetry is needed for this results. The definition makes clear that either low magnitude or asymmetry in the errors are sufficient. The insight is perhaps made clearer with the following inequality:

$$E_d \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[e_{ij}(d,d)]^2.$$

In fact, E_d will be small, or even negative, if the pair-wise explainable errors tend to have different signs.

Turning now to the unexplainable error. The fact that U_d captures dependence is immediate by the use of a covariance in its definition. To build intuition, consider the vaccine trial again. Consider when the exposure of a unit is defined to capture the nominal treatments assigned to units close, in some sense, to the unit in question (e.g., in their household, or in a neighbor in a social network). For illustration, assume that the experiment is so all encompassing in the studied community so the vaccinations have the potential to induce herd immunity. The exposures of any pair of units will in this case provide little informa-

tion about whether treatment assignment seen as a whole has induced herd immunity. If there is variation in whether head immunity is induced over the assignment mechanism, units' errors will exhibit great dependence for certain exposures even in cases where the explainable errors are small or zero, because pair-wise exposure cannot capture the global behavior here. In particular, if two units are assigned to exposures under which they would be infected by the disease without herd immunity, their errors would be highly dependent because infected at the same time if and only if experiment does not induced herd immunity. Generalizing from this example, the unexplainable error dependence captures whether the exposures are misspecified in the same way.

4 Point estimation

Commonly-used estimators for exposure effects build on ideas originally introduced in the survey sampling literature. Aronow & Samii (2017) focus on a version of the Horvitz-Thompson estimator (Horvitz & Thompson, 1952). They also discuss extensions to the Hájek estimator (Hájek, 1971) and various estimations facilitating covariate adjustments. Karwa & Airoldi (2018) show that the Horvitz-Thompson estimator generally is inadmissible when exposures are correctly specified and provide a large set of alternative estimators exploiting covariates and other auxiliary information.

The initial focus here is the Horvitz-Thompson estimator. Investigators will, however, likely benefit from the usual refinements also when exposures are misspecified, and such extensions are discussed in the following section.

Definition 6. The *Horvitz-Thompson estimator* for exposure effect $\tau(a, b)$ is:

$$\hat{\tau}(a, b) = \frac{1}{n} \sum_{i=1}^{n} \frac{D_{ia} Y_i}{\pi_i(a)} - \frac{1}{n} \sum_{i=1}^{n} \frac{D_{ib} Y_i}{\pi_i(b)},$$

where $D_{id} = \mathbb{1}[D_i = d]$ is an indicator denoting whether unit i's exposure is $d \in \Delta$.

The use of the Horvitz-Thompson estimator in this context builds on the idea that the observed potential outcomes can be seen as a sample from a finite population consist-

ing of the potential outcomes of all units in the study. An assignment mechanism that disproportionately assigns certain exposure will, seen through a survey sampling lens, be oversampled, and must be given lower weight than other potential outcomes when observed. For the Horvitz-Thompson estimator, these weights are the reciprocal of the probability of observing the outcomes.

When exposures are correctly specified, each realized outcome is equal to the potential outcome defined on the corresponding realized exposure. The positivity assumption stated in Condition 2 thus ensures that the reweighed outcomes for each term will be equal to the corresponding potential outcome in expectation. Under misspecification, the realized outcomes will vary even if the realized exposure is fixed, so the same logic does not apply. In expectation, each term gives a linear combination of all potential outcomes under the same exposure label. Two insights about the estimator makes this linear combination interpretable. First, $\pi_i(d)$ is fixed for all potential outcomes under the same exposure label, so each coefficient in the linear combination is proportional to the probability that the corresponding potential outcome is realized. Second, the expectation of D_{id} is $\pi_i(d)$, so the coefficients sum to one. The resulting convex combination is thus a conditional expectation. More precisely, the combination is equal to $\bar{y}_i(d) = \mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d]$, and the following result is immediate.

Proposition 1 (Unbiasedness)**.** *If Condition 2 holds for a and b, then:* $\mathrm{E}[\hat{\tau}(a, b)] = \tau(a, b)$.

The proposition is essentially a rephrasing of Proposition 8.1 in Aronow & Samii (2017). The implications of the result are, however, clearer when connected with an explicit target parameter as here. The proposition shows that the Horvitz-Thompson estimator is unbiased for the misspecification-robust exposure effect no matter how severe the misspecification is; no restrictions on the quantities in Definition 5 are needed. Of course, control over the location of the sampling distribution is not enough for inference. It is, however, comforting first step. The Horvitz-Thompson estimator is known for emphasizing unbiasedness at the cost of mean square error, so if it was shown not to control the bias, we would suspect behavior more generally also was poor.

4.1 Controlling design dependence

Definition 5 provides control of the dependence introduced by the specification errors. Another channel through which dependence can be introduced is the exposures themselves. For example, if the exposures are defined as the proportion of treated unit in the whole sample, it would follow that $D_1 = D_2 = \cdots = D_n$, and the Horvitz-Thompson estimator would exhibit considerable variation in large samples even if the exposures were correctly specified. To proceed, we must control the dependence between the exposures introduced by the design.

Definition 7. The *average design dependence* for exposure $d \in \Delta$ is:

$$C_d = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i \neq j} \big| \mathrm{Cov}\big(D_{id}, D_{jd}\big) \big|.$$

The average design dependence captures how strong the dependence on average is between two units' assigned exposures. It tells us how much information on average a unit's exposure provides about other units' exposures. Contrast this with the error dependence discussed in the previous section, which tells us how much information on average a unit's exposure provides about another unit's error.

The definition may appear unfamiliar, but the concept is capture is not. It can be seen a measure of effective sample size with respect to the assignment mechanism. If C_d diminishes in n, the effective sample size grows with the nominal size. The definition has a direct parallel in Aronow & Samii (2017) where C_d is, implicitly, defined as:

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{i \neq j} \mathbb{1}\big[D_{id} \perp\!\!\!\perp D_{jd}\big],$$

where $\mathbb{1}[D_{id} \perp\!\!\!\perp D_{jd}]$ is an indicator taking the value one when D_{id} and D_{jd} are independent. Clearly, this is a stronger dependence concept than the one captured by Definition 7.

It is rare that quantities like these are defined when interference is assumed to be absent. Instead, particular designs, such as complete randomization, are directly considered. This is generally not possible in our case because the complexity of the exposure mappings give

rise to intricate distributions of the exposures even if the design on the nominal treatments is simple.[2] The example with $D_1 = \cdots = D_n$ that opened this section is one such case. Intuition can, however, be built about C_d by considering these conventional designs on the exposures. If the exposure mappings were to induce a Bernoulli distribution on the exposures, the covariances are zero, so $C_d = 0$. If instead a pair-matched design is induced on the exposures, then $C_d = n^{-1}$, and if a complete randomization design with $\pi_i(d) = 0.5$ is induced, then $C_d = 0.25n^{-1}$.

4.2 Variance bound

We now have the components needed to characterize the behavior of the estimator beyond its expectation.

Proposition 2 (Bound on variance)**.** *If Conditions 1 and 2 hold for a and b, then:*

$$\mathrm{Var}\big(\hat{\tau}(a,b)\big) \leq 8k_1^2k_2n^{-1} + 20k_1^2k_2^2\big[C_a + C_b\big] + 4\big[E_a + E_b + U_a + U_b\big],$$

where C_d, E_d and U_d are given by Definitions 5 and 7.

The bound demonstrates that three aspects are relevant for the variance of the estimator. The first term captures the variation induced by the fact that the exposures are randomly assigned. That is, even when the exposures are independent and correctly specified, the estimator would still vary over assignments because different exposures (and thus potential outcomes) are realized. The second term captures the variation induced by dependence between exposures. That is, even when the exposures are correctly specified, the estimator tends to vary more when exposures are highly dependent. In some cases, such dependence can reduce the variance of the estimator, but that requires either additional restrictions on the potential outcomes (e.g., that they have the same sign) or additional restrictions on the design (e.g., that the number of units assigned to each exposure is fixed).

[2]The design and the exposure mappings are, however, known, so C_d can be calculated, although it might not be straightforward to do so computationally.

The final term in the bound captures variance stemming from misspecification. Recall that Definition 5 captures the dependence between the explainable and unexplainable specification errors. If the specification errors are strongly positively correlated, the estimator may be less stable. The bound makes clear that the magnitude of the specification errors is less of a concern. Large specification errors will affect the variance, but their effect is absorbed by the first term, so the large sample behavior is only affected by the dependence captured in the third term.

4.3 Large sample behavior

4.3.1 Asymptotic regime

The asymptotic regime used for the large sample investigation considers a sequence of fixed samples indexed by n. All quantities pertaining to the samples, such as the potential outcomes and designs, will thus have their own sequence also indexed by n. The index is, however, suppressed when no confusion ensues.

The regime differs from the conventional setup in that the sample is fixed and no population exists in the usual sense. The asymptotic properties discussed below are therefore not with respect to some sampling distribution, but they apply uniformly to all sequences of samples that satisfy the stated conditions. A consequence is that the samples need not be related in any specific way, and no assumptions about IID sampling or other stabilizing mechanisms are needed. This is particularly useful when interference is the focus, because units tend to neither be independent nor identically distributed in such cases.

This type of regime has been used extensively in the literature on design-based sampling (e.g., Isaki & Fuller, 1982). It has more recently seen increasing use in the design-based causal inference literature (e.g., Freedman, 2008; Lin, 2013).

4.3.2 Limiting behavior

Because the estimator is unbiased, the bound on the variance directly describes the estimator's asymptotic behavior in the L^2-norm. Control over the terms in the bound thus

provides consistency through Markov's inequality. The following two conditions provide the control.

Condition 3. An exposure $d \in \Delta$ satisfies *limited design dependence* if $C_d = o(1)$.

Condition 4. An exposure $d \in \Delta$ satisfies *limited specification error dependence* if $E_d \leq A_n$ and $U_d \leq A_n$ for some sequence $A_n = o(1)$.

Proposition 3 (Consistency)**.** *If Conditions 1 and 2 hold for a and b, then:*

$$\hat{\tau}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

If Conditions 3 and 4 also hold, then: $\hat{\tau}(a,b) - \tau(a,b) = o_p(1)$.

The limiting distribution of the estimator is less tractable than its limit. A situation where progress can be made is when the specification errors are very small relative to the sample size. Recall Definition 2 and consider the following decomposition of the estimator:

$$\hat{\tau}(a,b) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{D_{ia}\bar{y}_i(a)}{\pi_i(a)} - \frac{1}{n}\sum_{i=1}^{n}\frac{D_{ib}\bar{y}_i(b)}{\pi_i(b)}\right] + \frac{1}{n}\sum_{i=1}^{n}\frac{(D_{ia} - D_{ib})\varepsilon_i}{\pi_i(D_i)}.$$

The first term of the expression is the Horvitz-Thompson estimator if we could directly observe the misspecification-robust potential outcomes defined in Section 3.4. This part is, thus, unaffected by the misspecification, and any properties that would hold for the estimator when exposures are correctly specified would also hold for this term. In particular, if the rate of convergence of the second term is faster than the first, Slutsky's theorem tells us that the first term determines the limiting distribution. For example, if the estimator is known to convergence to a normal distribution at a root-n rate under correctly specified exposures, then a sufficient condition for the same to be true under misspecification is that $A_n = o(n^{-1})$ in Condition 4. The assumption is considerably stronger than Condition 4, bordering to assuming correctly specified exposures. I conjecture that limiting distribution can be characterized under considerably weaker conditions.

A last resort if no asymptotic approximation is available as a sound basis for hypothesis testing and interval estimation is Chebyshev's inequality. The inequality guarantees 95%

coverage rates for confidence intervals constructed as 4.47 standard errors wide windows on either side of the point estimate. Naturally, this interval estimator would be tremendously conservative in most situations, and it still requires a reasonable variance estimator, but it may be the only option in some settings.

5 Extensions

5.1 Improved estimators

As noted above, the Horvitz-Thompson estimator is rarely a good choice for actual empirical work, but it provides theoretical insights that can be built on. This section provides results for common refinements of the Horvitz-Thompson estimator.

The first refinement accounts for the realized number of units assigned to the exposures of interest. The Hájek estimator (Hájek, 1971) does this by dividing each term in the estimator with the sum of the reciprocals of the assignment probabilities for the units assigned to the exposure, rather than dividing by n. The change can absorb some of the variability in the estimator introduced by randomness in the number of units assigned to each exposure. The ratio structure introduces bias, but it is generally small enough to still grant improvements in mean square error. The denominator can generally be shown to be well-behaved, so the estimator's limited behavior can be linked to the Horvitz-Thompson estimator through linearization.

Definition 8. The *Hájek estimator* for exposure effect $\tau(a, b)$ is:

$$\hat{\tau}_{\text{HÁ}}(a,b) = \left(\sum_{i=1}^{n} \frac{D_{ia}Y_i}{\pi_i(a)} \Big/ \sum_{i=1}^{n} \frac{D_{ia}}{\pi_i(a)}\right) - \left(\sum_{i=1}^{n} \frac{D_{ib}Y_i}{\pi_i(b)} \Big/ \sum_{i=1}^{n} \frac{D_{ib}}{\pi_i(b)}\right).$$

Proposition 4 (Consistency of the Hájek estimator)**.** *If Conditions 1, 2, 3 and 4 hold for a and b, then $\hat{\tau}_{\text{HÁ}}(a, b)$ is consistent for $\tau(a, b)$ and converges at the following rate:*

$$\hat{\tau}_{\text{HÁ}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

Investigators commonly use estimators that do not explicitly adjust for the assignment probabilities. One such example is the difference-in-means estimator. This estimator can be shown to coincide with the Hájek estimator whenever the assignment probabilities are the same for all units: $\pi_i(d) = \pi_j(d)$ for all $i, j \in \mathcal{U}$. Proposition 4 thus implies that the difference-in-means estimator can be used in similar situation also under misspecification. Investigators should, however, not blindly use the difference-in-means estimator for exposure effects because the exposure mappings may not induce equal assignment probabilities on the exposures even if they are equal for the nominal treatments. Another estimator coinciding with the Hájek estimator is the ordinary least squares (OLS) estimator. The unweighted version requires equal assignment probabilities just like the difference-in-means estimator, but a weighted OLS estimator is equivalent to the Hájek estimator in the general case.

A disadvantage of all estimators discussed so far is their inability to exploit auxiliary information. A simple modification of the Horvitz-Thompson estimator allows us to incorporate such information. The idea is that information beside the observed potential outcomes themselves might allow us to predict the potential outcomes we do not observe. If this prediction is sufficiently good, the predicted outcomes can be used to offset chance imbalances introduced by the randomization. Särndal et al. (1992) call it the difference estimator in a sampling setting, and the name will be used here as well.

Definition 9. The *difference estimator* for exposure effect $\tau(a, b)$ is:

$$\hat{\tau}_{\text{DE}}(a, b) = \frac{1}{n} \sum_{i=1}^{n} \big[\hat{y}_i(a) - \hat{y}_i(b)\big] + \frac{1}{n} \sum_{i=1}^{n} \frac{(D_{ia} - D_{ib})\big[Y_i - \hat{y}_i(D_i)\big]}{\pi_i(D_i)},$$

where $\hat{y}_i(d)$ is a prediction of unit i's potential outcome when assigned to $d \in \Delta$.

The definition of the estimator reveals the idea that motivate its use. The first term is simply the average difference in predicted potential outcomes. If the predictions are of high quality, this term will be an accurate estimator of the exposure effect. The issue is that the predictions may have systematic errors. The second term is included to ensure robustness. If the predictions are of low quality, this term will compensate for the errors in the first

term, and it ensures that the estimator performs well in expectation. The estimator bears a resemblance in this regard to the class of doubly robust estimators used in observational studies when the assignment mechanism is unknown (see, e.g., Robins & Rotnitzky, 2001).

The properties of the difference estimator depend on the way the predictions are constructed. In particular, the estimator can be shown to retain the advantageous properties of the Horvitz-Thompson estimator if the predictions are external to the study. External here means that they do not depend on the treatment assignment. As the only randomness under consideration here stems from the assignment mechanism, independence between $\hat{y}_i(d)$ and $\mathbf{Z}$ implies that the predictions are non-random. The probability space can extended to accommodate random predictions if one wants to account for the consequences of external variability. Such variability could affect the rate of convergence if units' predictions are sufficiently dependent, but it is otherwise inconsequential to the results. The results presented here presume that the predictions are fixed to ease exposition; random predictions are considered in the appendix.

Proposition 5 (Unbiasedness of the difference estimator)**.** *If Condition 2 holds for a and b, and the predictions are non-random, then:* $\mathrm{E}[\hat{\tau}_{\text{DE}}(a,b)] = \tau(a,b)$.

Proposition 6 (Consistency of the difference estimator)**.** *If Conditions 1, 2, 3 and 4 hold for a and b, and the predictions are non-random, then $\hat{\tau}_{\text{DE}}(a,b)$ is consistent for $\tau(a,b)$ and converges at the following rate:*

$$\hat{\tau}_{\text{DE}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

The difference estimator seemingly provides advantages at no cost. Good predictions of the potential outcomes confer improvements in finite samples, but the estimator has the same robustness guarantees as the Horvitz-Thompson estimator both in finite and large samples. The no-cost advantages are superficial. The mean square error may increase when the predictions are poor, so investigators should use the difference estimator only when the predictions are expected to be of reasonably high quality.

The quality of the predictions is, however, less of a concern than their construction. Covariate information can be used to make the predictions, but the assigned exposures and

the observed outcomes can generally not be used because it would induce dependence between the predictions and $\mathbf{Z}$. More precisely, if $\mathbf{x}_i$ denotes a vector of covariates describing characteristics of unit i, we can form the predictions as $\hat{y}_i(d) = f(d, \mathbf{x}_i)$ for some function f. The function f can, however, not be constructed using $(Y_1, Y_2, \ldots, Y_n)$ or $(D_1, D_2, \ldots, D_n)$. This illustrates that the construction of f truly needs to be external to treatment assignment when used for the predictions in the difference estimator. This severely limits its applicability. Split-sample or leave-one-out approaches (see, e.g., Williams, 1961) that often are used to solve the issue cannot be used here because the misspecification may induce dependence between subsamples that otherwise appear isolated.

An estimator facilitating dependence between the predictions of the potential outcomes and the treatment assignments is inspired by the generalized regression estimator commonly used in the sampling literature. The estimator has received recent attention in the causal inference literature as well (see, e.g., Lin, 2013; Middleton, 2018).

The estimator uses a linear working model for the relationship between the potential outcomes and the covariates. The working model is used to construct the predictions: $\hat{y}_i(d) = \mathbf{x}_i\boldsymbol{\beta}(d)$ for some vector of coefficients $\boldsymbol{\beta}(d)$ indexed by $d \in \Delta$, so different coefficients are used for different exposures. No assumptions are made about the validity of the model, but the quality of the predictions are related to how well the model can approximate the potential outcomes. It remains to pick the coefficients $\boldsymbol{\beta}(d)$. The generalized regression estimator allows of dependence between the coefficients and the treatment assignments, so the coefficients can be estimated in the sample. For example, we may pick them as the minimizing solution to $\sum_{i=1}^{n} D_{id}[Y_i - \mathbf{x}_i\boldsymbol{\beta}(d)]^2$ as is often done in applications. But other choices exist, and the estimator is largely agnostic about how the coefficients were constructed.

Definition 10. The *generalized regression estimator* for exposure effect $\tau(a, b)$ is:

$$\hat{\tau}_{\text{GR}}(a, b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\big[\hat{\boldsymbol{\beta}}(a) - \hat{\boldsymbol{\beta}}(b)\big] + \frac{1}{n}\sum_{i=1}^{n} \frac{(D_{ia} - D_{ib})\big[Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}(D_i)\big]}{\pi_i(D_i)},$$

where $\hat{\boldsymbol{\beta}}(a)$ and $\hat{\boldsymbol{\beta}}(b)$ are two random vectors of the same dimensions as $\mathbf{x}_i$.

The conventional approach to investigating the properties of the generalized regression estimator is to assume that the vector of coefficients constructed in the sample convergences to some fixed vector asymptotically. This ensures that the dependence between units' predictions is small in large samples, which provides consistency. The assumption can be weaken to only require that the length of the vector of coefficients is asymptotically bounded, thereby bypassing the need of assuming a well-defined limit.

Proposition 7 (Consistency of the generalized regression estimator)**.** *Assume* $\mathbf{x}_i \in \mathcal{X}$ *for some bounded* $\mathcal{X} \subset \mathbb{R}^p$ *and* $\mathrm{E}\big[\|\hat{\boldsymbol{\beta}}(d)\|\big] = \mathcal{O}(1)$*. If Conditions 1, 2, 3 and 4 hold for* a *and* b*, then* $\hat{\tau}_{\text{GR}}(a,b)$ *is consistent for* $\tau(a,b)$ *and converges at the following rate:*

$$\hat{\tau}_{\text{GR}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

5.2 Lack of positivity

Positivity conditions are often seen as innocuous in experiments because the investigator controls the design and can ensure that it holds. This may not be the case when the focus is effects of exposures. The exposure mappings are often complex, and it may not be feasible to construct a design that would induce the desired distribution over the exposures. Investigators will often settle for a heuristic choice, and this might induce violations of Condition 2.

The positivity condition can fail in two ways. The first is when it is fundamental impossible that a unit is assigned to a certain exposure. For example, a person living in a single-person household cannot be assigned to the exposure that at least two household members are vaccinated. This may be formalized such that no $\mathbf{z} \in \Omega$ exists so that $d_i(\mathbf{z}) = d$ for some $d \in \Delta$. The consequences are more than just statistical. If it is nonsensical to talk about a unit being assigned a certain exposure, it is nonsensical to consider exposure effects that include the unit in its average. Unless the investigator is comfortable stipulating a metaphysical model allowing extrapolation to unrealizable potential outcomes, the only solution is to exclude such units from the average. The result may be that the number of

units under study is fewer than the length of $\mathbf{z}$, but this is not an issue. In the following discussion, it will be assumed that such exclusions have been made if necessary. That is, if the effect of exposures a and b is the focus, then $\{a, b\} \subseteq \{d_i(\mathbf{z}) : \mathbf{z} \in \Omega\}$ for all units.

The second way the positivity condition can fail is through the design. That is, assignments $\mathbf{z} \in \Omega$ exist so that $d_i(\mathbf{z}) = d$, but the design is such that $\pi_i(d) = 0$. Statistical issues are the only sequelae in this case, which all have cures. Two situations must be considered. The first is when the assignment probability for some exposure is exactly zero, $\pi_i(d) = 0$. The second is when the probability approaches zero asymptotically. Both are problematic, but they have different solutions.

Superficially, the first situation appears most acute. The reciprocal of the assignment probability is used in the Horvitz-Thompson estimator, and a probability that is exactly zero would render the estimator ill-defined. A simple solution suggests itself once we realize that the denominator is zero only when nominator is zero with probability one. Defining $0/0$ as zero makes the estimator well-defined without positivity, and this is the approach that will be taken here.

It remains to describe the behavior of the estimator. We must here consider both the assignment probabilities that are exactly zero and those that approach zero. Consider the following quantities:

$$\bar{S}_d = \frac{1}{n}\sum_{i=1}^{n} S_i(d), \qquad \text{and} \qquad \Pi(d, p) = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{1 - S_i(d)}{[\pi_i(d)]^p + S_i(d)}\right]^{1/p},$$

where $S_i(d) = \mathbb{1}[\pi_i(d) = 0]$. The first quantity counts the number of assignment probabilities that are exactly equal to zero, and the second is proportional the pth moment of the reciprocals of the remaining probabilities. The quantities allow us to weaken the positivity assumption in a controllable way. In particular, Condition 2 is the same as $\bar{S}_d = 0$ and $\Pi(d, p) \leq k < \infty$ as $p \to \infty$. The following proposition shows that neither part is necessary for consistency. However, the weaken positivity condition comes at the cost of potentially slower convergence rates. This is captured by a strengthening of the definition of the design

dependence, namely:

$$C_d(s) = \left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \left| \mathrm{Cov}\big(D_{id}, D_{jd}\big) \right|^s \right]^{1/s}.$$

The extended definition collapses to Definition 7 when $s = 1$, but generally $C_d = \mathcal{O}\big(C_d(s)\big)$ when $s > 1$. Thus, Condition 3 using $C_d(s)$ when $s > 1$ is stronger than the original version. The additional machinery admits consistency without positivity. The following proportion only considers the Horvitz-Thompson estimator, but a similar result should hold the estimators in the previous section.

Proposition 8 (Consistency without positivity)**.** *Assume* $\Pi(d, p) \leq k < \infty$ *for* $d \in \{a, b\}$ *and some* $p > 2$*. Also assume* $\bar{S}_d = o(1)$ *and* $C_d\big(p/(p-2)\big) = o(1)$ *for* $d \in \{a, b\}$*. If Conditions 1 and 4 hold, then the Horvitz-Thompson estimator is consistent for the misspecification-robust exposure effect and converges at the following rate:*

$$\hat{\tau}(a, b) - \tau(a, b) = \mathcal{O}_p\big(n^{-0.5} + \bar{S}_a + \bar{S}_b + \tilde{C}_{ap}^{0.5} + \tilde{C}_{bp}^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big),$$

where $\tilde{C}_{dp}$ *is short-hand for* $C_d\big(p/(p-2)\big)$*.*

6 Variance estimation

6.1 Current results

Variance estimation for exposure effect estimators is challenging because the variance consists of pair-wise products of potential outcomes, and some of theaw outcomes are not observable simultaneously. The issue is not unique to exposure effects, but exposure mappings tend to induce complex distributions on the exposures, which tend to exacerbate the issue.

The solution suggested by Aronow & Samii (2017) is to use Young's inequality to bound the unobservable parts of the variance expression, which gives the following estimator:

$$\widehat{\mathrm{Var}}_{\mathrm{AS}}\big(\hat{\tau}(a,b)\big) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ia}-D_{ib})(D_{ja}-D_{jb})P_{ij}(D_i,D_j)Y_iY_j + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{D_{ia}}{\pi_i(a)}+\frac{D_{ib}}{\pi_i(b)}\right]\big[S_{ij}(D_i,a)+S_{ij}(D_i,b)\big]Y_i^2,$$

where:

$$P_{ij}(d,q) = \frac{\pi_{ij}(d,q)-\pi_i(d)\pi_j(q)}{\pi_{ij}(d,q)\pi_i(d)\pi_j(q)+S_{ij}(d,q)}, \qquad S_{ij}(d,q) = \mathbb{1}[\pi_{ij}(d,q)=0],$$

and $\pi_{ij}(d,q) = \Pr(D_i = d, D_j = q)$ is the joint probability of unit i and j's exposures. The authors show that the estimator is conservative when exposures are correctly specified. That is, they show that the estimator is greater or equal to true variance in expectation.

What does not appear to be fully appreciated in the literature is that this fix may make the estimator excessively conservative. In fact, unless the assumption of correctly specified exposures is complemented with:

$$\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(a,a)+S_{ij}(b,b)+S_{ij}(a,b)\big] = \mathcal{O}(n),$$

the normalized variance $n\widehat{\mathrm{Var}}_{\mathrm{AS}}\big(\hat{\tau}(a,b)\big)$ generally diverges to infinity. I will not offer a solution to this problem. The remark instead serves as an illustration of the difficulty of variance estimation for complex exposure effects. In also provides insights about the mechanics of the estimator, which will aid our understanding of its behavior under misspecification.

6.2 Variance estimation and misspecification

Variance estimation could for this reason be seen as an open question even when exposures are correctly specified.

The same does not hold under misspecification.

Proposition 9 (Expectation of variance estimator)**.** *If Conditions 1 and 2 hold, then:*

$$\mathrm{E}\Big[\widehat{\mathrm{Var}}_{\mathrm{AS}}\big(\hat{\tau}(a,b)\big)\Big] = \mathrm{Var}\big(\hat{\tau}(a,b)\big) + B_1 + B_2(a,b) + B_2(b,a) + B_3(a) + B_3(b)$$

$$+ 2B_4(a,b) - B_4(a,a) - B_4(b,b),$$

where:

$$B_1 = \frac{1}{n^2}\sum_{i=1}^{n}\big[\bar{y}_i(a) - \bar{y}_i(b)\big]^2,$$

$$B_2(d,q) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j\neq i}\Big(S_{ij}(d,d)\big[\bar{y}_i(d) + \bar{y}_j(d)\big]^2 + S_{ij}(d,q)\big[\bar{y}_i(d) - \bar{y}_j(q)\big]^2\Big),$$

$$B_3(d) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(d,a) + S_{ij}(d,b)\big]\,\mathrm{Var}(\varepsilon_i \mid D_i = d),$$

$$B_4(d,q) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(d,q)\big]\Big[\bar{y}_i(d)e_{ji}(q,d) + \bar{y}_j(q)e_{ij}(d,q) + e_{ij}(d,q)e_{ji}(q,d) + \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = d, D_j = q)\Big].$$

The terms after the variance on the right hand side in the proposition captures factor biasing the variance estimator. These biases help us understand when variance estimation is possible under misspecification. The term B_1 stems from what Holland (1986) describes as the fundamental problem of causal inference, namely that a unit cannot simultaneously be assign to two different treatments. The joint distribution of the potential outcomes affects the variance, but the distribution can only be estimated if both potential outcomes are observed simultaneously. This would, however, require simultaneous assignment of two different treatments to the same unit. The first term captures the bias arising from our inability to estimate this aspect of the potential outcomes. The issue is not unique to variance estimation for exposure effects, and similar bias terms arise for most causal inference problems in finite samples.

The distribution of the exposures may, as noted above, be complex, and the joint exposure probabilities may be zero for a considerable number of pairs of units. The issue is similar to the first source of bias, but it is now induced by the design. The bound used to solve the issue introduces bias, and that bias is captured in the terms $B_2(a,b)$ and $B_2(b,a)$. These biases arise also when the exposures are correctly specified.

The remaining terms capture the bias stemming from misspecification; if the exposures are correctly specified, these terms are zero. The bias comes from two sources. The first arises from the use of Young's inequality in the construction of the estimator. When a pair of units cannot be assigned to a certain set of exposures simultaneously, their corresponding specification errors cannot interact and do not affect the variance. However, these are exactly the terms that need to be bounded to ensure conservativeness under correctly specified exposures. The bound is on the observed outcomes, and this will include the specification errors. The consequence is that the variance estimator is affected by the magnitude of the corresponding errors. The terms $B_3(a)$ and $B_3(b)$ capture this part of the bias. These terms are non-negative by construction, like the previous terms.

The terms of real concern are the last three: $2B_4(a,b)$, $B_4(a,a)$, and $B_4(b,b)$. These capture the bias introduced by our inability to estimate the dependence in the specification errors. Unlike the previous terms, the signs of the terms are unknown, so they may introduce negative bias. The consequence is that we may systematically underestimate the variance when the specification errors are large, and our inferences would then be anti-conservative.

The problem has no immediate solution, but some progress can be made. Similar to the discussion in Section 4.3, if the specification errors can be assumed to be negligible relative to the sample size, the terms given by $B_4(d,q)$ will be negligible relative to the other terms, and the variance estimator is ensured to be conservative asymptotically.

An alternative approach is to incorporate more information about the structure of the interference in the variance estimator. In particular, the anti-conservative behavior of the estimator stems from negative interactions of errors in $B_4(d,q)$. One may remove such interactions by setting $S_{ij}(d,q) = 1$ for the corresponding pairs of units, even if $\pi_{ij}(d,q) > 0$ holds. This will move the corresponding terms from $B_4(d,q)$, where negative interactions are possible, to $B_3(d)$, where no interactions exist. It may be hard to discern whether the interaction between two specific units' errors is negative or positive; a conservative approach is to set $S_{ij}(d,q) = 1$ for all pairs of units where an interaction of any type is

suspected.

For example, if it is assumed that units only interfere with each other within known disjoint groups (i.e., partial interference), one may set $S_{ij}(d,q) = 1$ if either $\pi_{ij}(d,q)$ is zero or if unit i and j belong to the same group. A redefinition of $S_{ij}(d,q)$ along these lines would ensure that $B_4(d,q) = 0$, so the variance estimator remains conservative. Of course, such knowledge about the interference would allow for the definition of exposure mappings that are correctly specified, which would obviate all concerns about misspecification. Investigators could, however, be interested keeping the main exposure mapping simple to facilitate interpretation. They can then proceed with misspecified exposures for point estimation, and use the more intricate information about the interference structure only when estimating variance.

A third approach is a combination of the previous two. One may set $S_{ij}(d,q) = 1$ for pairs of units where negative interactions are suspected to be particularly large. Unless one presumes to have caught all interacting terms, one must entertain the possibility that $B_4(d,q)$ is negative. However, setting $S_{ij}(d,q) = 1$ for the terms deemed most problematic may make the assumption that the remaining errors are small more reasonable. It should also be noted that the other bias terms tend to be large and possible, and they generally provide considerably leeway with respect control over $B_4(d,q)$.

The focus on the expectation of the variance estimator should be seen as an analogy for its more general behavior. Specifically, the precision of the variance estimator will be poor if joint exposure probabilities are small even if they are never exactly zero. Investigators should be aware that the variance estimator may be very imprecise, and particularly when the number of exposures is large. The concern is, however, not specific to misspecification.

Middleton (2018) introduces an improved version of the variance estimator that admits less conservative estimation in expectation under correctly specified exposures. The improvement only requires information about the design, so it can be used also under misspecification. Middleton's approach is a way to mitigate the excessive conservativeness discussed in the previous section. It will, however, not solve the issues discussed in this sec-

tion. The estimation can also be improved if more is known about the potential outcomes than what is stipulated by Condition 1. For example, if the potential outcomes are known to have the same sign, so that either $\bar{y}_i(d) \geq 0$ or $\bar{y}_i(d) \leq 0$ for all units and exposures, then the second term of the estimator can changed to:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{D_{ia}S_{ij}(a,b)}{\pi_i(a)} + \frac{D_{ib}S_{ij}(b,a)}{\pi_i(b)}\right]Y_i^2,$$

which admits a less conservative estimator when exposures are correctly specified. The modification can also be used under misspecification, but concerns about interactions between errors must still be tended to.

7 Concluding remarks

To be added...

References

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1), 3–16.

Aronow, P. M. & Samii, C. (2017). Estimating average causal effects under general interference. *Annals of Applied Statistics*, 11(4), 1912–1947.

Athey, S., Eckles, D., & Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521), 230–240.

Basse, G., Feller, A., & Toulis, P. (2018). Conditional randomization tests of causal effects with interference between units. arXiv:1709.08036v3.

Basse, G. W. & Airoldi, E. M. (2018). Limitations of design-based causal inference and A/B testing under arbitrary and network interference. *Sociological Methodology*, 48(1), 136–151.

Bohrnstedt, G. W. & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association*, 64(328), 1439–1442.

Bowers, J., Fredrickson, M. M., & Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. *Political Analysis*, 21(1), 97–124.

Chin, A. (2018). Central limit theorems via Stein's method for randomized experiments under interference. arXiv:1804.03105v1.

Choi, D. (2017). Estimation of monotone treatment effects in network experiments. *Journal of the American Statistical Association*, 112(519), 1147–1155.

Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1).

Egami, N. (2018). Unbiased estimation and sensitivity analysis for network-specific spillover effects: Application to an online network experiment. arXiv:1708.08171v3.

Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver & Boyd.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2), 180–193.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, part one". In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Halloran, M. E. & Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology*, 6(2), 142–151.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 832–842.

Isaki, C. T. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377), 89–96.

Karwa, V. & Airoldi, E. M. (2018). A systematic investigation of classical causal inference strategies under mis-specification due to network interference. arXiv:1810.08259v1.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals of Applied Statistics*, 7(1), 295–318.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1), 1–23.

Middleton, J. A. (2018). A unified theory of regression adjustment for design-based inference. arXiv:1803.06011v1.

Neyman, J. (1990/1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4), 465–472. (Original work published 1923).

Robins, J. M. & Rotnitzky, A. (2001). Comment on "inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4), 920–936.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477), 191–200.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Sävje, F., Aronow, P. M., & Hudgens, M. G. (2018). Average treatment effects in the presence of unknown interference. arXiv:1711.06399v3.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476), 1398–1407.

Toulis, P. & Kao, E. (2013). Estimation of causal peer influence effects. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research* (pp. 1489–1497). Atlanta.

Williams, W. H. (1961). Generating unbiased ratio and regression estimators. *Biometrics*, 17(2), 267.

A Rigorous definitions of the potential outcome

Condition 2 ensure that $\bar{y}_i$ is uniquely defined by $\bar{y}_i(d) = \mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d]$. This no longer holds when the condition is relaxed in Section 5.2. In particular, $D_i = d$ can then be a null set for some $i \in \mathcal{U}$, and the definition provided in the paper is ambiguous. The issue was ignored in the main text to expedite the exposition, but it will be addressed here. In particular, let the full definition be:

$$\bar{y}_i(d) = \begin{cases} \mathrm{Avg}\big(\{y_i(\mathbf{z}) : \mathbf{z} \in \Omega \text{ and } d_i(\mathbf{z}) = d\}\big) & \text{if } \pi_i(d) = 0, \\ \mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d] & \text{else,} \end{cases}$$

where $\mathrm{Avg}\big(A\big)$ gives the arithmetic mean of the elements in the set A.

A similar issue arises for the definition of $\bar{y}_{ij}$ in Section 3.5. In particular, the function is not uniquely defined if $\pi_{ij}(d, q) = 0$ for some pairs of units. Therefore, consider the following as the full definition:

$$\bar{y}_{ij}(d, q) = \begin{cases} \bar{y}_i(d) & \text{if } \pi_{ij}(d, q) = 0, \\ \mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d, D_j = q] & \text{else.} \end{cases}$$

It follows that $e_{ij}(d, q) = 0$ when $\pi_{ij}(d, q) = 0$. This captures the intuition that learning $D_j = q$ provides no information about $\bar{y}_i(d)$ when $D_i = d$ is not simultaneously possible with $D_j = q$. Similarly, set $\mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d) = 0$ when $\pi_{ij}(d, q) = 0$.

B Proofs

B.1 Miscellaneous lemmas

Lemma A1. *For any N random variables $X_1, X_2, \ldots, X_N$ defined on the same probability space:*

$$\operatorname{Var}\big(X_1 + X_2 + \cdots + X_N\big) \leq \Big(\sqrt{\operatorname{Var}(X_1)} + \sqrt{\operatorname{Var}(X_2)} + \cdots + \sqrt{\operatorname{Var}(X_N)}\Big)^2.$$

Proof. Write the variance of the sum as a double sum of covariances:

$$\operatorname{Var}(X_1 + \cdots + X_N) = \sum_{i=1}^{N}\sum_{j=1}^{N} \operatorname{Cov}(X_i, X_j).$$

Separate the covariances using the Cauchy–Schwarz inequality and reorder the summation:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \operatorname{Cov}(X_i, X_j) \leq \sum_{i=1}^{N}\sum_{j=1}^{N} \sqrt{\operatorname{Var}(X_i)\operatorname{Var}(X_j)} = \left(\sum_{i=1}^{N} \sqrt{\operatorname{Var}(X_i)}\right)^2. \qquad \square$$

Lemma A2. *For any N random variables $X_1, X_2, \ldots, X_N$ defined on the same probability space:*

$$\operatorname{Var}(X_1 + \cdots + X_N) \leq N \operatorname{Var}(X_1) + \cdots + N \operatorname{Var}(X_N).$$

Proof. Apply Lemma A1 to get:

$$\operatorname{Var}(X_1 + \cdots + X_N) \leq \left(\sum_{i=1}^{N} \sqrt{\operatorname{Var}(X_i)}\right)^2 = N^2\left(\frac{1}{N}\sum_{i=1}^{N} \sqrt{\operatorname{Var}(X_i)}\right)^2$$

The square is a convex function, so Jensen's inequality gives:

$$N^2\left(\frac{1}{N}\sum_{i=1}^{N} \sqrt{\operatorname{Var}(X_i)}\right)^2 \leq N^2\left(\frac{1}{N}\sum_{i=1}^{N} \operatorname{Var}(X_i)\right) \leq N\sum_{i=1}^{N} \operatorname{Var}(X_i). \qquad \square$$

Lemma A3. *If Condition 1 holds, then for all $i \in \mathcal{U}$ and $d \in \Delta$:*

1. $|\bar{y}_i(d)| \leq k_1$,
2. $\mathrm{E}[Y_i^2 \mid D_i = d] \leq k_1^2$,

3. $\big|\mathrm{E}[\varepsilon_i\varepsilon_j \mid D_i = D_j = d]\big| \leq 4k_1^2$.

Proof. Consider each statement in turn:

1. Recall the definition of $\bar{y}_i(d)$, and note:

$$|\bar{y}_i(d)| = \big|\mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d]\big| \leq \mathrm{E}\big[|y_i(\mathbf{Z})| \;\big|\; D_i = d\big] \leq \mathrm{E}\big[k_1 \;\big|\; D_i = d\big] = k_1.$$

2. Note that $|Y_i| = |y_i(\mathbf{Z})| \leq k_1$, so:

$$\mathrm{E}[Y_i^2 \mid D_i = d] = \mathrm{E}\big[|y_i(\mathbf{Z})|^2 \;\big|\; D_i = d\big] = k_1^2.$$

3. Using a similar logic as in the previous parts of the proof:

$$\begin{aligned}\big|\mathrm{E}[\varepsilon_i\varepsilon_j \mid D_i = D_j = d]\big| &\leq \mathrm{E}\big[\big(|Y_i| + |\bar{y}_i(d)|\big)\big(|Y_j| + |\bar{y}_j(d)|\big) \;\big|\; D_i = D_j = d\big] \\ &\leq \mathrm{E}\big[4k_1^2 \;\big|\; D_i = D_j = d\big] = 4k_1^2. \quad \square\end{aligned}$$

B.2 Proposition 1

Proposition 1. *If Condition 2 holds for a and b, then:* $\mathrm{E}[\hat{\tau}(a,b)] = \tau(a,b)$.

Proof. For a generic exposure $d \in \Delta$ satisfying Condition 2:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{E}[D_{id}Y_i]}{\pi_i(d)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_i(d)\,\mathrm{E}\big[Y_i \;\big|\; D_i = d\big]}{\pi_i(d)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_i(d)\bar{y}_i(d)}{\pi_i(d)} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i(d),$$

so:

$$\mathrm{E}[\hat{\tau}(a,b)] = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{E}[D_{ia}Y_i]}{\pi_i(a)} - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{E}[D_{ib}Y_i]}{\pi_i(b)} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i(a) - \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i(b) = \tau(a,b). \qquad \square$$

B.3 Proposition 2

Proposition 2. *If Conditions 1 and 2 hold for a and b, then:*

$$\mathrm{Var}\big(\hat{\tau}(a,b)\big) \leq 8k_1^2k_2n^{-1} + 20k_1^2k_2^2\big[C_a + C_b\big] + 4\big[E_a + E_b + U_a + U_b\big],$$

where C_d, E_d and U_d are given by Definitions 5 and 7.

Proof. Let $R_i(d) = D_{id}/\pi_i(d)$, so:

$$\hat{\tau}(a,b) = \frac{1}{n}\sum_{i=1}^{n} R_i(a)Y_i - \frac{1}{n}\sum_{i=1}^{n} R_i(b)Y_i$$

Apply Lemma A2 to get:

$$\operatorname{Var}\big(\hat{\tau}(a,b)\big) \leq 2\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} R_i(a)Y_i\right) + 2\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} R_i(b)Y_i\right)$$

Note that $Y_i = \bar{y}_i(D_i) + \varepsilon_i$, so for a generic exposure $d \in \Delta$:

$$\frac{1}{n}\sum_{i=1}^{n} R_i(d)Y_i = \frac{1}{n}\sum_{i=1}^{n} R_i(d)\bar{y}_i(d) + \frac{1}{n}\sum_{i=1}^{n} R_i(d)\varepsilon_i$$

Apply Lemma A2 again:

$$\begin{aligned}
2\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} R_i(d)Y_i\right) &\leq 4\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} R_i(d)\bar{y}_i(d)\right) + 4\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} R_i(d)\varepsilon_i\right) \\
&= \frac{4}{n^2}\sum_{i=1}^{n}\operatorname{Var}\big(R_i(d)\bar{y}_i(d)\big) + \frac{4}{n^2}\sum_{i=1}^{n}\operatorname{Var}\big(R_i(d)\varepsilon_i\big) \\
&\quad + \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\operatorname{Cov}\big(R_i(d)\bar{y}_i(d), R_j(d)\bar{y}_j(d)\big) \\
&\quad + \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\operatorname{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big)
\end{aligned}$$

Consider the first two terms of the expression. Note that:

$$\mathrm{E}[\varepsilon_i \mid D_i = d] = \mathrm{E}[Y_i - \bar{y}_i(d) \mid D_i = d] = \mathrm{E}[y_i(\mathbf{Z}) \mid D_i = d] - \bar{y}_i(d) = 0 \tag{1}$$

so

$$\begin{aligned}
\operatorname{Cov}\big(R_i(d)\bar{y}_i(d), R_i(d)\varepsilon_i\big) &= \mathrm{E}\big[[R_i(d)]^2\bar{y}_i(d)\varepsilon_i\big] - \mathrm{E}[R_i(d)\bar{y}_i(d)]\,\mathrm{E}[R_i(d)\varepsilon_i] \\
&= \frac{\bar{y}_i(d)}{\pi_i(d)}\,\mathrm{E}[\varepsilon_i \mid D_i = d] - \bar{y}_i(d)\,\mathrm{E}[\varepsilon_i \mid D_i = d] = 0
\end{aligned}$$

and in turn:

$$\operatorname{Var}\big(R_i(d)Y_i\big) = \operatorname{Var}\big(R_i(d)\bar{y}_i(d) + R_i(d)\varepsilon_i\big) = \operatorname{Var}\big(R_i(d)\bar{y}_i(d)\big) + \operatorname{Var}\big(R_i(d)\varepsilon_i\big)$$

Condition 2 and Lemma A3 give the following bound:

$$\operatorname{Var}\big(R_i(d)Y_i\big) \leq \operatorname{E}\big[[R_i(d)]^2 Y_i^2\big] = \frac{\operatorname{E}[Y_i^2 \mid D_i = d]}{\pi_i(d)} \leq k_1^2 k_2$$

so:

$$\frac{4}{n^2}\sum_{i=1}^{n} \operatorname{Var}\big(R_i(d)\bar{y}_i(d)\big) + \frac{4}{n^2}\sum_{i=1}^{n} \operatorname{Var}\big(R_i(d)\varepsilon_i\big) = \frac{4}{n^2}\sum_{i=1}^{n} \operatorname{Var}\big(R_i(d)Y_i\big) \leq \frac{4k_1^2 k_2}{n}$$

Recall $R_i(d) = D_{id}/\pi_i(d)$ and consider the third term of the variance expression:

$$\begin{aligned}
\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \operatorname{Cov}\big(R_i(d)\bar{y}_i(d), R_j(d)\bar{y}_j(d)\big) &= \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \frac{\bar{y}_i(d)\bar{y}_j(d)}{\pi_i(d)\pi_j(d)} \operatorname{Cov}\big(D_{id}, D_{jd}\big) \\
&\leq \frac{4k_1^2 k_2^2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \big|\operatorname{Cov}\big(D_{id}, D_{jd}\big)\big| \\
&= 4k_1^2 k_2^2 C_d
\end{aligned}$$

which again uses Condition 2 and Lemma A3.

Now consider the fourth and final term. Recall $S_{ij}(d,q) = \mathbb{1}[\pi_{ij}(d,q) = 0]$ and decompose the sum as such:

$$\begin{aligned}
\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \operatorname{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big) = \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,d) \operatorname{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big) \\
+ \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} \big[1 - S_{ij}(d,d)\big] \operatorname{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big)
\end{aligned}$$

Start with the terms for which $S_{ij}(d,d) = 0$, and note that the derivation in (1) implies:

$$\operatorname{E}\big[R_i(d)\varepsilon_i\big] = \operatorname{E}\big[\operatorname{E}\big[R_i(d)\varepsilon_i \;\big|\; D_i\big]\big] = \operatorname{E}\big[R_i(d)\operatorname{E}\big[\varepsilon_i \;\big|\; D_i\big]\big] = 0$$

which allows the covariances to be decomposed as:

$$\begin{aligned}
\operatorname{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big) &= \operatorname{E}\big[R_i(d)\varepsilon_i R_j(d)\varepsilon_j\big] - \operatorname{E}\big[R_i(d)\varepsilon_i\big]\operatorname{E}\big[R_j(d)\varepsilon_j\big] \\
&= \operatorname{E}\big[R_i(d)R_j(d)\varepsilon_i\varepsilon_j\big] \\
&= \frac{\Pr(D_i = D_j = d)}{\pi_i(d)\pi_j(d)} \operatorname{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\Pr(D_i = D_j = d) - \pi_i(d)\pi_j(d) + \pi_i(d)\pi_j(d)}{\pi_i(d)\pi_j(d)} \\
&\qquad \times \mathrm{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big] \\
&= \frac{\mathrm{Cov}\big(D_{id}, D_{jd}\big)}{\pi_i(d)\pi_j(d)}\,\mathrm{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big] \\
&\qquad + \mathrm{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big]
\end{aligned}
$$

where $S_{ij}(d,d) = 0$ ensures that $\mathrm{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big]$ is unambiguously defined.

Consider the first term through the lens of Condition 2 and Lemma A3:

$$
\frac{\mathrm{Cov}\big(D_{id}, D_{jd}\big)}{\pi_i(d)\pi_j(d)}\,\mathrm{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big] \leq 4k_1^2k_2^2\big|\mathrm{Cov}\big(D_{id}, D_{jd}\big)\big|
$$

For the second term, recall that $\varepsilon_i = e_{ij} + u_{ij}$, so:

$$
\begin{aligned}
\mathrm{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big] &= \mathrm{E}\Big[\big[e_{ij} + u_{ij}\big]\big[e_{ji} + u_{ji}\big] \;\Big|\; D_i = D_j = d\Big] \\
&= \mathrm{E}\big[e_{ij}e_{ji} + e_{ij}u_{ji} + u_{ij}e_{ji} + u_{ij}u_{ji} \;\big|\; D_i = D_j = d\big]
\end{aligned}
$$

Note that $e_{ij} = e_{ij}(d,d)$ and $e_{ji} = e_{ji}(d,d)$ are constant conditional on $D_i = D_j = d$, and:

$$
\mathrm{E}[u_{ij} \mid D_i, D_j] = \mathrm{E}\big[Y_i - \bar{y}_{ij}(D_i, D_j) \;\big|\; D_i, D_j\big] = \mathrm{E}[Y_i \mid D_i, D_j] - \bar{y}_{ij}(D_i, D_j) = 0
$$

so:

$$
\mathrm{E}\big[e_{ij}u_{ji} \;\big|\; D_i = D_j = d\big] = e_{ij}(d,d)\,\mathrm{E}\big[u_{ji} \;\big|\; D_i = D_j = d\big] = 0
$$

which gives:

$$
\begin{aligned}
\mathrm{E}\big[\varepsilon_i\varepsilon_j \;\big|\; D_i = D_j = d\big] &= e_{ij}(d,d)e_{ji}(d,d) + \mathrm{E}\big[u_{ij}u_{ji} \;\big|\; D_i = D_j = d\big] \\
&= e_{ij}(d,d)e_{ji}(d,d) + \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d)
\end{aligned}
$$

The two terms taken together then give for $S_{ij}(d,d) = 0$:

$$
\begin{aligned}
&\mathrm{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big) \\
&\qquad \leq 4k_1^2k_2^2\big|\mathrm{Cov}\big(D_{id}, D_{jd}\big)\big| + e_{ij}(d,d)e_{ji}(d,d) + \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d).
\end{aligned}
$$

For terms with $S_{ij}(d,d) = 1$, $R_i(d)R_j(d)$ is constant at zero, so:

$$\mathrm{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big) = \mathrm{E}\big[R_i(d)R_j(d)\varepsilon_i\varepsilon_j\big] = 0$$

Recall that $e_{ij}(d,d) = 0$ and $\mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d) = 0$ when $S_{ij}(d,d) = 1$, so:

$$\begin{aligned}&\mathrm{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big)\\ &\qquad\leq 4k_1^2k_2^2\big|\mathrm{Cov}\big(D_{id}, D_{jd}\big)\big| + e_{ij}(d,d)e_{ji}(d,d) + \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d).\end{aligned}$$

also for $S_{ij}(d,d) = 1$. It follows that:

$$\begin{aligned}\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}\big(R_i(d)\varepsilon_i, R_j(d)\varepsilon_j\big) &\leq \frac{16k_1^2k_2^2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big|\mathrm{Cov}\big(D_{id}, D_{jd}\big)\big|\\ &\quad + \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}e_{ij}(d,d)e_{ji}(d,d)\\ &\quad + \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d)\\ &= 16k_1^2k_2^2C_d + 4E_d + 4U_d\end{aligned}$$

The derivations for all four terms taken together yield:

$$2\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}R_i(d)Y_i\right) \leq \frac{4k_1^2k_2}{n} + 20k_1^2k_2^2C_d + 4E_d + 4U_d$$

so the variance is bounded as:

$$\begin{aligned}\mathrm{Var}\big(\hat{\tau}(a,b)\big) &\leq 2\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}R_i(a)Y_i\right) + 2\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}R_i(b)Y_i\right)\\ &\leq \frac{8k_1^2k_2}{n} + 20k_1^2k_2^2\big[C_a + C_b\big] + 4\big[E_a + E_b + U_a + U_b\big].\end{aligned}$$

□

B.4 Proposition 3

Proposition 3. *If Conditions 1 and 2 hold for a and b, then:*

$$\hat{\tau}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

If Conditions 3 and 4 also hold, then: $\hat{\tau}(a,b) - \tau(a,b) = o_p(1)$.

Proof. The proof presumes that E_a, E_b, U_a and U_b are non-negative. If they are not, the contravening quantities can be set to zero, and the proof would apply.

Consider the root mean square error:

$$\sqrt{\mathrm{E}\Big[\Big(\hat{\tau}(a,b) - \tau(a,b)\Big)^2\Big]} = \sqrt{\mathrm{Var}\big(\hat{\tau}(a,b)\big)}$$
$$\leq \sqrt{8k_1^2k_2n^{-1} + 20k_1^2k_2^2\big[C_a + C_b\big] + 4\big[E_a + E_b + U_a + U_b\big]}$$

where the first equality follows from Proposition 1. Concavity of the square root gives:

$$\sqrt{\mathrm{Var}\big(\hat{\tau}(a,b)\big)} \leq 3k_1k_2^{0.5}n^{-0.5} + 5k_1k_2\big[C_a^{0.5} + C_b^{0.5}\big] + 2\big[E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big]$$

which gives the rate of convergence in the L^2-norm:

$$\sqrt{\mathrm{E}\Big[\Big(\hat{\tau}(a,b) - \tau(a,b)\Big)^2\Big]} = \mathcal{O}\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big)$$

which in turn gives the rate of convergence in probability by Markov's inequality (see, e.g., Lemma A5 in the appendix of Sävje et al., 2018). Conditions 3 and 4 complete the proof. □

B.5 Proposition 4

The linearization used to prove consistency for the Hájek estimator requires an alternative representation of the estimator.

Definition A1 (Components for the Hájek estimator)**.**

$$\mu_d = \sum_{i=1}^{n} \bar{y}_i(d), \qquad \hat{\mu}_d = \sum_{i=1}^{n} \frac{D_{id}Y_i}{\pi_i(d)}, \qquad \text{and} \qquad \hat{n}_d = \sum_{i=1}^{n} \frac{D_{id}}{\pi_i(d)}.$$

Lemma A4. *Given Condition 1, $\mu_d = \mathcal{O}(n)$ for $d \in \Delta$.*

Proof. Consider the definition of μ_d through the lens of Lemma A3:

$$\mu_d = \sum_{i=1}^{n} \bar{y}_i(d) \leq \sum_{i=1}^{n} |\bar{y}_i(d)| \leq \sum_{i=1}^{n} k_1 = k_1 n.$$ □

Lemma A5. *Given Conditions 1 and 2,* $(\hat{\mu}_d - \mu_d)/n = \mathcal{O}_p\big(n^{-0.5} + C_d^{0.5} + E_d^{0.5} + U_d^{0.5}\big)$.

Proof. Note that $\hat{\tau}(a,b) = (\hat{\mu}_a - \hat{\mu}_b)/n$ and $\tau(a,b) = (\mu_a - \mu_b)/n$, so the proofs of Propositions 1 and 2 can be copied almost in verbatim to show:

$$\mathrm{E}\big[\hat{\mu}_d\big] = \mu_d, \qquad \text{and} \qquad \frac{\mathrm{Var}(\hat{\mu}_d)}{n^2} \leq \frac{2k_1^2k_2}{n} + 10k_1^2k_2^2C_d + 2E_d + 2U_d.$$

The logic of the proof of Proposition 3 then gives:

$$\sqrt{\mathrm{E}\big[(\hat{\mu}_d - \mu_d)^2/n^2\big]} = \mathcal{O}\big(n^{-0.5} + C_d^{0.5} + E_d^{0.5} + U_d^{0.5}\big).$$

Markov's inequality completes the proof as in the proof of Proposition 3. □

Lemma A6. *Given Condition 2,* $(\hat{n}_d - n)/n = \mathcal{O}_p\big(n^{-0.5} + C_d^{0.5}\big)$.

Proof. The first step is to show that $\mathrm{E}[\hat{n}_d] = n$ when d satisfies Condition 2:

$$\mathrm{E}[\hat{n}_d] = \sum_{i=1}^{n} \frac{\mathrm{E}[D_{id}]}{\pi_i(d)} = \sum_{i=1}^{n} \frac{\pi_i(d)}{\pi_i(d)} = n.$$

Next consider the variance:

$$\mathrm{Var}(\hat{n}_d) = \sum_{i=1}^{n} \frac{\mathrm{Var}(D_{id})}{[\pi_i(d)]^2} + \sum_{i=1}^{n}\sum_{j\neq i} \frac{\mathrm{Cov}(D_{id}, D_{jd})}{\pi_i(d)\pi_j(d)},$$

where by Condition 2:

$$\frac{\mathrm{Var}(D_{id})}{[\pi_i(d)]^2} = \frac{\pi_i(d)[1-\pi_i(d)]}{[\pi_i(d)]^2} \leq k_2, \qquad \text{and} \qquad \frac{\mathrm{Cov}(D_{id}, D_{jd})}{\pi_i(d)\pi_j(d)} \leq k_2^2\big|\mathrm{Cov}(D_{id}, D_{jd})\big|,$$

so $\mathrm{Var}(\hat{n}_d)/n^2 \leq k_2 n^{-1} + k_2^2 C_d$.

The logic of the proof of Proposition 3 then gives:

$$\sqrt{\mathrm{E}\big[(\hat{n}_d - n)^2/n^2\big]} = \mathcal{O}\big(n^{-0.5} + C_d^{0.5}\big),$$

and Markov's inequality completes the proof. □

Proposition 4. *If Conditions 1, 2, 3 and 4 hold for a and b, then $\hat{\tau}_{\text{HÁ}}(a,b)$ is consistent for $\tau(a,b)$ and converges at the following rate:*

$$\hat{\tau}_{\text{HÁ}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

Proof. Note that $\hat{\tau}_{\text{HÁ}}(a,b) = \hat{\mu}_a/\hat{n}_a - \hat{\mu}_b/\hat{n}_b$ and $\tau(a,b) = \mu_a/n - \mu_b/n$, so we can write:

$$\hat{\tau}_{\text{HÁ}}(a,b) - \tau(a,b) = \left(\frac{\hat{\mu}_a}{\hat{n}_a} - \frac{\mu_a}{n}\right) - \left(\frac{\hat{\mu}_b}{\hat{n}_b} - \frac{\mu_b}{n}\right)$$

For a generic exposure d consider:

$$\frac{\hat{\mu}_d}{\hat{n}_d} - \frac{\mu_d}{n} = \frac{\hat{\mu}_d/n}{\hat{n}_d/n} - \frac{(\mu_d/n)(\hat{n}_d/n)}{\hat{n}_d/n} = \frac{(\hat{\mu}_d - \mu_d)/n}{\hat{n}_d/n} - \frac{(\mu_d/n)(\hat{n}_d - n)/n}{\hat{n}_d/n}$$

where Lemma A6 ensures that we can ignore the event $\hat{n}_d = 0$.

Let $f(x,y) = x/y$ and consider a Taylor expansion of the two terms around $(0,1)$:

$$\begin{aligned}
\frac{(\hat{\mu}_d - \mu_d)/n}{\hat{n}_d/n} &= f\big((\hat{\mu}_d - \mu_d)/n, \hat{n}_d/n\big) = (\hat{\mu}_d - \mu_d)/n + r_1 \\
\frac{(\mu_d/n)(\hat{n}_d - n)/n}{\hat{n}_d/n} &= f\big((\mu_d/n)(\hat{n}_d - n)/n, \hat{n}_d/n\big) = (\mu_d/n)(\hat{n}_d - n)/n + r_2
\end{aligned}$$

where $r_1 = o_p\big((\hat{\mu}_d - \mu_d)/n + (\hat{n}_d - n)/n\big)$ and $r_2 = o_p\big((\mu_d/n)(\hat{n}_d - n)/n + (\hat{n}_d - n)/n\big)$ because Lemmas A4, A5 and A6 give convergence of $(\hat{\mu}_d - \mu_d)/n$ and $(\mu_d/n)(\hat{n}_d - n)/n$ to zero and of $\hat{n}_d/n$ to one. Lemma A4 gives $(\mu_d/n)(\hat{n}_d - n)/n = \mathcal{O}_p\big((\hat{n}_d - n)/n\big)$, so by Lemmas A5 and A6:

$$\hat{\tau}_{\text{HÁ}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big). \qquad \square$$

B.6 Proposition 5 and 6

The main text considers non-random predictions. The choice was made to expedite exposition, and the more general case is covered here. Throughout, it will be assumed that the predictions are sufficiently well-behaved asymptotically so their second moments exist: $\mathrm{E}\big[|\hat{y}_i(d)|^2\big] \leq k$ for all $i \in \mathcal{U}$ and $d \in \Delta$.

Condition A1. The predictions are said to be *external* for exposure $d \in \Delta$ if they are jointly independent of treatment assignment: $\big(\hat{y}_1(d), \hat{y}_2(d), \ldots, \hat{y}_n(d)\big) \perp\!\!\!\perp \mathbf{Z}$.

Definition A2. The *average prediction dependence* for exposure $d \in \Delta$ is:

$$P_d = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \big|\mathrm{Cov}\big(\hat{y}_i(d), \hat{y}_j(d)\big)\big|.$$

An alternative approach to focusing on the dependence between predictions is to consider their convergence. In particular, the approach used in the proof of Lemma A1 can be used to bound P_d as:

$$P_d = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left|\mathrm{Cov}\big(\hat{y}_i(d),\hat{y}_j(d)\big)\right| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathrm{Var}(\hat{y}_i(d))}\right)^2.$$

Lemma A7. *Assume the predictions are external. If Condition 2 holds for a and b, then:* $\mathrm{E}[\hat{\tau}_{\text{DE}}(a,b)] = \tau(a,b)$.

Proof. Rewrite the estimator as:

$$\begin{aligned}
\hat{\tau}_{\text{DE}}(a,b) &= \frac{1}{n}\sum_{i=1}^{n}\big[\hat{y}_i(a)-\hat{y}_i(b)\big] + \frac{1}{n}\sum_{i=1}^{n}\frac{(D_{ia}-D_{ib})\big[Y_i-\hat{y}_i(D_i)\big]}{\pi_i(D_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\big[\hat{y}_i(a)-\hat{y}_i(b)\big] + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_{ia}Y_i}{\pi_i(a)} - \frac{D_{ib}Y_i}{\pi_i(b)} - \frac{D_{ia}\hat{y}_i(a)}{\pi_i(a)} + \frac{D_{ib}\hat{y}_i(b)}{\pi_i(b)}\right] \\
&= \hat{\tau}(a,b) + \frac{1}{n}\sum_{i=1}^{n}\big[\hat{y}_i(a)-\hat{y}_i(b)\big] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_{ia}\hat{y}_i(a)}{\pi_i(a)} - \frac{D_{ib}\hat{y}_i(b)}{\pi_i(b)}\right].
\end{aligned}$$

Taking expectations yields:

$$\mathrm{E}\big[\hat{\tau}_{\text{DE}}(a,b)\big] = \tau(a,b) + \frac{1}{n}\sum_{i=1}^{n}\Big(\mathrm{E}\big[\hat{y}_i(a)\big]-\mathrm{E}\big[\hat{y}_i(b)\big]\Big) - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathrm{E}\big[D_{ia}\hat{y}_i(a)\big]}{\pi_i(a)} - \frac{\mathrm{E}\big[D_{ib}\hat{y}_i(b)\big]}{\pi_i(b)}\right].$$

We get $\mathrm{E}\big[D_{id}\hat{y}_i(d)\big] = \mathrm{E}\big[D_{ia}\big]\,\mathrm{E}\big[\hat{y}_i(a)\big] = \pi_i(d)\,\mathrm{E}\big[\hat{y}_i(a)\big]$ by independence between $\mathbf{Z}$ and the predictions, so by Condition 2:

$$\frac{\mathrm{E}\big[D_{id}\hat{y}_i(d)\big]}{\pi_i(d)} = \frac{\pi_i(d)\,\mathrm{E}\big[\hat{y}_i(a)\big]}{\pi_i(d)} = \mathrm{E}\big[\hat{y}_i(a)\big],$$

and the two last terms in the expression of $\mathrm{E}\big[\hat{\tau}_{\text{DE}}(a,b)\big]$ cancel. □

Lemma A8. *Assume the predictions are external and $P_d = o(1)$ for a and b. If Conditions 1, 2, 3 and 4 hold for a and b, then $\hat{\tau}_{\text{DE}}(a,b)$ is consistent for $\tau(a,b)$ and converges at the following rate:*

$$\hat{\tau}_{\text{DE}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5} + P_a^{0.5} + P_b^{0.5}\big).$$

Proof. Write the estimator as in the proof for Proposition 5:

$$\hat{\tau}_{\text{DE}}(a,b) = \hat{\tau}(a,b) + \frac{1}{n}\sum_{i=1}^{n}\big[\hat{y}_i(a) - \hat{y}_i(b)\big] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_{ia}\hat{y}_i(a)}{\pi_i(a)} - \frac{D_{ib}\hat{y}_i(b)}{\pi_i(b)}\right]$$

Apply Lemma A2 to get:

$$\begin{aligned}\operatorname{Var}(\hat{\tau}_{\text{DE}}(a,b)) \leq 5\operatorname{Var}\big(\hat{\tau}(a,b)\big) + \frac{5}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n}\hat{y}_i(a)\right) + \frac{5}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n}\hat{y}_i(b)\right) \\ + \frac{5}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n}\frac{D_{ia}\hat{y}_i(a)}{\pi_i(a)}\right) + \frac{5}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n}\frac{D_{ib}\hat{y}_i(b)}{\pi_i(b)}\right)\end{aligned}$$

The first term is bounded by Proposition 2. Consider the two subsequent terms:

$$\frac{5}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n}\hat{y}_i(d)\right) = \frac{5}{n^2}\sum_{i=1}^{n}\operatorname{Var}\big(\hat{y}_i(d)\big) + \frac{5}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\operatorname{Cov}\big(\hat{y}_i(d),\hat{y}_j(d)\big) \leq \frac{5k}{n} + 5P_d,$$

where $\operatorname{E}\big[|\hat{y}_i(d)|^2\big] \leq k$. Next consider the last two terms in the variance expression:

$$\frac{5}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n}\frac{D_{id}\hat{y}_i(d)}{\pi_i(d)}\right) = \frac{5}{n^2}\sum_{i=1}^{n}\frac{\operatorname{Var}\big(D_{id}\hat{y}_i(d)\big)}{[\pi_i(d)]^2} + \frac{5}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\operatorname{Cov}\big(D_{id}\hat{y}_i(d), D_{jd}\hat{y}_j(d)\big)}{\pi_i(d)\pi_j(d)}$$

Remembering that predictions are external and applying the covariance decomposition in Bohrnstedt & Goldberger (1969) yield:

$$\begin{aligned}\operatorname{Cov}\big(D_{id}\hat{y}_i(d), D_{jd}\hat{y}_j(d)\big) = \pi_i(d)\pi_j(d)\operatorname{Cov}\big(\hat{y}_i(d),\hat{y}_j(d)\big) + \operatorname{E}\big[\hat{y}_i(d)\big]\operatorname{E}\big[\hat{y}_j(d)\big]\operatorname{Cov}\big(D_{id}, D_{jd}\big) \\ + \operatorname{Cov}\big(D_{id}, D_{jd}\big)\operatorname{Cov}\big(\hat{y}_i(d),\hat{y}_j(d)\big)\end{aligned}$$

Note that $0 \leq \pi_i(d)\pi_j(d) + \operatorname{Cov}\big(D_{id}, D_{jd}\big) \leq 1$. Furthermore:

$$\operatorname{E}\big[\hat{y}_i(d)\big]\operatorname{E}\big[\hat{y}_j(d)\big]\operatorname{Cov}\big(D_{id}, D_{jd}\big) \leq k\big|\operatorname{Cov}(D_{id}, D_{jd})\big|,$$

where $\operatorname{E}\big[|\hat{y}_i(d)|^2\big] \leq k$. It follows that:

$$\frac{\operatorname{Cov}\big(D_{id}\hat{y}_i(d), D_{jd}\hat{y}_j(d)\big)}{\pi_i(d)\pi_j(d)} \leq k_2^2\big|\operatorname{Cov}\big(\hat{y}_i(d),\hat{y}_j(d)\big)\big| + kk_2^2\big|\operatorname{Cov}(D_{id}, D_{jd})\big|.$$

Recall once more the independence between the assignments and predictions:

$$\frac{\operatorname{Var}\big(D_{id}\hat{y}_i(d)\big)}{[\pi_i(d)]^2} = \frac{[\pi_i(d)]^2 \operatorname{Var}\big(\hat{y}_i(d)\big) + \operatorname{Var}\big(D_{id}\big)\big(\operatorname{E}\big[\hat{y}_i(d)\big]\big)^2 + \operatorname{Var}\big(D_{id}\big)\operatorname{Var}\big(\hat{y}_i(d)\big)}{[\pi_i(d)]^2}$$
$$\leq k + kk_2 + kk_2 \leq 3kk_2.$$

Taken together:

$$\frac{5}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n}\frac{D_{id}\hat{y}_i(d)}{\pi_i(d)}\right) \leq \frac{15kk_2}{n} + 5k_2^2 P_d + 5kk_2^2 C_d.$$

Combined with Proposition 2, the variance is bounded as:

$$\operatorname{Var}(\hat{\tau}_{\text{DE}}(a,b)) \leq \big(40k_1^2k_2 + 30kk_2 + 10k\big)n^{-1} + \big(100k_1^2k_2^2 + 5kk_2^2\big)\big[C_a + C_b\big]$$
$$+ 20\big[E_a + E_b + U_a + U_b\big] + \big(5k_2^2 + 5\big)\big[P_a + P_b\big]$$

Proposition 5 together with the logic of the proof of Proposition 3 provide convergence in the L^2-norm:

$$\sqrt{\operatorname{E}\Big[\big(\hat{\tau}_{\text{DE}}(a,b) - \hat{\tau}(a,b)\big)^2\Big]} = \mathcal{O}\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5} + P_a^{0.5} + P_b^{0.5}\big).$$

Markov's inequality completes the proof. □

Proposition 5. *If Condition 2 holds for a and b, and the predictions are non-random, then:* $\operatorname{E}[\hat{\tau}_{\text{DE}}(a,b)] = \tau(a,b)$.

Proof. Non-random predictions satisfy Condition A1. □

Proposition 6. *If Conditions 1, 2, 3 and 4 hold for a and b, and the predictions are fixed, then $\hat{\tau}_{\text{DE}}(a,b)$ is consistent for $\tau(a,b)$ and converges at the following rate:*

$$\hat{\tau}_{\text{DE}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

Proof. Non-random predictions satisfy Condition A1 and $P_d = 0$ for all $d \in \Delta$. □

B.7 Proposition 7

Proposition 7. *Assume* $\mathbf{x}_i \in \mathcal{X}$ *for some bounded* $\mathcal{X} \subset \mathbb{R}^p$ *and* $\mathrm{E}\big[\|\hat{\boldsymbol{\beta}}(d)\|\big] = \mathcal{O}(1)$. *If Conditions 1, 2, 3 and 4 hold for* a *and* b, *then* $\hat{\tau}_{\text{GR}}(a,b)$ *is consistent for* $\tau(a,b)$ *and converges at the following rate:*

$$\hat{\tau}_{\text{GR}}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + C_a^{0.5} + C_b^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big).$$

Proof. Rewrite the estimator as:

$$\hat{\tau}_{\text{GR}}(a,b) = \hat{\tau}(a,b) + \left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \frac{D_{ia}\mathbf{x}_i}{\pi_i(a)}\right)\right]\hat{\boldsymbol{\beta}}(a) - \left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \frac{D_{ib}\mathbf{x}_i}{\pi_i(b)}\right)\right]\hat{\boldsymbol{\beta}}(b).$$

Using the same approach as in the proofs of Propositions 1, 2 and 3, it can be shown that:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \frac{D_{id}\mathbf{x}_i}{\pi_i(d)}\right)\right] = \mathcal{O}_p\big(n^{-0.5} + C_d^{0.5}\big).$$

By Markov's inequality, $\mathrm{E}\big[\|\hat{\boldsymbol{\beta}}(d)\|\big] = \mathcal{O}(1)$ implies $\hat{\boldsymbol{\beta}}(d) = \mathcal{O}_p(1)$, so:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \frac{D_{id}\mathbf{x}_i}{\pi_i(d)}\right)\right]\hat{\boldsymbol{\beta}}(d) = \mathcal{O}_p\big(n^{-0.5} + C_d^{0.5}\big).$$

The proposition then follows from Proposition 3. □

B.8 Proposition 8

Proposition 8. *Assume* $\Pi(d,p) \leq k < \infty$ *for* $d \in \{a,b\}$ *and some* $p > 2$. *Also assume* $\bar{S}_d = o(1)$ *and* $C_d\big(p/(p-2)\big) = o(1)$ *for* $d \in \{a,b\}$. *If Conditions 1 and 4 hold, then the Horvitz-Thompson estimator is consistent for the misspecification-robust exposure effect and converges at the following rate:*

$$\hat{\tau}(a,b) - \tau(a,b) = \mathcal{O}_p\big(n^{-0.5} + \bar{S}_a + \bar{S}_b + \tilde{C}_{ap}^{0.5} + \tilde{C}_{bp}^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big),$$

where $\tilde{C}_{dp}$ *is short-hand for* $C_d\big(p/(p-2)\big)$.

Proof. First note that:

$$\hat{\tau}_{\text{ZHT}}(a,b) = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-S_i(a))D_{ia}Y_i}{\pi_i(a)+S_i(a)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-S_i(b))D_{ib}Y_i}{\pi_i(b)+S_i(b)},$$

with probability one since $D_{id}=0$ when $S_i(d)=1$. Thus, its expectation is:

$$\mathrm{E}[\hat{\tau}_{\text{ZHT}}(a,b)] = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i(a) - \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i(b) - \frac{1}{n}\sum_{i=1}^{n}S_i(a)\bar{y}_i(a) + \frac{1}{n}\sum_{i=1}^{n}S_i(b)\bar{y}_i(b),$$

and:

$$\left|\mathrm{E}[\hat{\tau}_{\text{ZHT}}(a,b)] - \tau(a,b)\right| \leq \frac{1}{n}\sum_{i=1}^{n}S_i(a)\left|\bar{y}_i(a)\right| + \frac{1}{n}\sum_{i=1}^{n}S_i(b)\left|\bar{y}_i(b)\right| \leq k_1[\bar{S}_a + \bar{S}_b].$$

Next consider the variance:

$$\mathrm{Var}\left(\hat{\tau}_{\text{ZHT}}(a,b)\right) \leq 2\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-S_i(a))D_{ia}Y_i}{\pi_i(a)+S_i(a)}\right) + 2\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-S_i(b))D_{ib}Y_i}{\pi_i(b)+S_i(b)}\right)$$

where Lemma A2 was used. Using a similar decomposition as in Proposition 2:

$$\begin{aligned}
&2\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-S_i(d))D_{id}Y_i}{\pi_i(d)+S_i(d)}\right) \\
&\qquad\leq \frac{4}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{(1-S_i(d))D_{id}Y_i}{\pi_i(d)+S_i(d)}\right) \\
&\qquad\quad + \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}\left(\frac{(1-S_i(d))D_{id}\bar{y}_i(d)}{\pi_i(d)+S_i(d)}, \frac{(1-S_j(d))D_{jd}\bar{y}_j(d)}{\pi_j(d)+S_j(d)}\right) \\
&\qquad\quad + \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}\left(\frac{(1-S_i(d))D_{id}\varepsilon_i}{\pi_i(d)+S_i(d)}, \frac{(1-S_j(d))D_{jd}\varepsilon_j}{\pi_j(d)+S_j(d)}\right).
\end{aligned}$$

Consider the first term:

$$\begin{aligned}
\frac{4}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{(1-S_i(d))D_{id}Y_i}{\pi_i(d)+S_i(d)}\right) &\leq \frac{4}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\frac{(1-S_i(d))D_{id}Y_i^2}{[\pi_i(d)]^2+S_i(d)}\right] \\
&= \frac{4}{n^2}\sum_{i=1}^{n}\frac{(1-S_i(d))\,\mathrm{E}[Y_i^2 \mid D_i=d]}{\pi_i(d)+S_i(d)} \leq \frac{4k_1^2}{n^2}\sum_{i=1}^{n}\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)} \leq \frac{4k_1^2\Pi(d,p)}{n},
\end{aligned}$$

where the last inequality follows from Jensen's inequality when $p \geq 1$.

Consider the second term next:

$$\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}\left(\frac{(1-S_i(d))D_{id}\bar{y}_i(d)}{\pi_i(d)+S_i(d)},\frac{(1-S_j(d))D_{jd}\bar{y}_j(d)}{\pi_j(d)+S_j(d)}\right)$$
$$\leq \frac{4k_1^2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]\left|\mathrm{Cov}\big(D_{id},D_{jd}\big)\right|$$

Apply Hölder's inequality with conjugates $p/2$ and $p/(p-2)$:

$$\frac{4k_1^2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]\left|\mathrm{Cov}\big(D_{id},D_{jd}\big)\right|$$
$$\leq \frac{4k_1^2}{n^2}\left[\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]^{p/2}\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]^{p/2}\right]^{2/p}$$
$$\times\left[\sum_{i=1}^{n}\sum_{j\neq i}\left|\mathrm{Cov}\big(D_{id},D_{jd}\big)\right|^{p/(p-2)}\right]^{(p-2)/p}$$
$$= 4k_1^2\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]^{p/2}\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]^{p/2}\right]^{2/p}$$
$$\times\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left|\mathrm{Cov}\big(D_{id},D_{jd}\big)\right|^{p/(p-2)}\right]^{(p-2)/p}$$

Then:

$$\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]^{p/2}\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]^{p/2}\right]^{2/p}$$
$$\leq\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]^{p/2}\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]^{p/2}\right]^{2/p}$$
$$=\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]^{p/2}\right)^2\right]^{2/p}$$
$$\leq\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]^{p}\right]^{2/p}=\big[\Pi(d,p)\big]^2$$

where the last inequality follows from Jensen's inequality. Taken together:

$$\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}\left(\frac{(1-S_i(d))D_{id}\bar{y}_i(d)}{\pi_i(d)+S_i(d)},\frac{(1-S_j(d))D_{jd}\bar{y}_j(d)}{\pi_j(d)+S_j(d)}\right)\leq 4k_1^2\big[\Pi(d,p)\big]^2C_d\big(p/(p-2)\big).$$

Consider the final term:

$$\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\operatorname{Cov}\left(\frac{(1-S_i(d))D_{id}\varepsilon_i}{\pi_i(d)+S_i(d)},\frac{(1-S_j(d))D_{jd}\varepsilon_j}{\pi_j(d)+S_j(d)}\right)$$
$$\leq \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]\operatorname{Cov}\big(D_{id}\varepsilon_i, D_{jd}\varepsilon_j\big)$$

If either $S_i(d)=1$ or $S_j(d)=1$, then:

$$\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]\operatorname{Cov}\big(D_{id}\varepsilon_i, D_{jd}\varepsilon_j\big)=0.$$

Furthermore, if $S_i(d)=0$ and $S_j(d)=0$ but $S_{ij}(d,d)=\mathbb{1}[\pi_{ij}(d,d)=0]=1$, then:

$$\operatorname{Cov}\big(D_{id}\varepsilon_i, D_{jd}\varepsilon_j\big)=\mathrm{E}\big[D_{id}D_{jd}\varepsilon_i\varepsilon_j\big]-\mathrm{E}\big[D_{id}\varepsilon_i\big]\,\mathrm{E}\big[D_{jd}\varepsilon_j\big]=0,$$

since $\mathrm{E}[D_{id}\varepsilon_i]=0$ when $S_i(d)=0$ as shown in the proof of Proposition 2. Note that $S_i(d)=0$ and $S_j(d)=0$ are implied by $S_{ij}(d,d)=0$, so:

$$\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{(1-S_i(d))}{\pi_i(d)+S_i(d)}\right]\left[\frac{(1-S_j(d))}{\pi_j(d)+S_j(d)}\right]\operatorname{Cov}\big(D_{id}\varepsilon_i, D_{jd}\varepsilon_j\big)$$
$$=\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\big[1-S_{ij}(d,d)\big]\operatorname{Cov}\big(D_{id}\varepsilon_i, D_{jd}\varepsilon_j\big)}{\pi_i(d)\pi_j(d)+S_i(d)+S_j(d)}$$

Using the same decomposition as in the proof of Proposition 2:

$$\frac{\operatorname{Cov}\big(D_{id}\varepsilon_i, D_{jd}\varepsilon_j\big)}{\pi_i(d)\pi_j(d)}=\frac{\operatorname{Cov}\big(D_{id}, D_{jd}\big)}{\pi_i(d)\pi_j(d)}\,\mathrm{E}\big[\varepsilon_i\varepsilon_j\;\big|\;D_i=D_j=d\big]+\mathrm{E}\big[\varepsilon_i\varepsilon_j\;\big|\;D_i=D_j=d\big],$$
$$\leq\frac{4k_1^2\big|\operatorname{Cov}\big(D_{id}, D_{jd}\big)\big|}{\pi_i(d)\pi_j(d)}+\mathrm{E}\big[\varepsilon_i\varepsilon_j\;\big|\;D_i=D_j=d\big]$$

when $S_{ij}(d,d)=0$ using Lemma A3, so:

$$\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\big[1-S_{ij}(d,d)\big]\operatorname{Cov}\big(D_{id}\varepsilon_i, D_{jd}\varepsilon_j\big)}{\pi_i(d)\pi_j(d)+S_i(d)+S_j(d)}$$
$$\leq\frac{16k_1^2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\big[1-S_{ij}(d,d)\big]\big|\operatorname{Cov}\big(D_{id}, D_{jd}\big)\big|}{\pi_i(d)\pi_j(d)+S_i(d)+S_j(d)}$$

$$+ \frac{4}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \big[1 - S_{ij}(d,d)\big] \, \mathrm{E}\big[\varepsilon_i \varepsilon_j \;\big|\; D_i = D_j = d\big]$$

Use Hölder's inequality as above to separate the factors in the first term:

$$\begin{aligned}
&\frac{16k_1^2}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \frac{\big[1 - S_{ij}(d,d)\big] \big|\mathrm{Cov}\big(D_{id}, D_{jd}\big)\big|}{\pi_i(d)\pi_j(d) + S_i(d) + S_j(d)} \\
&\qquad \leq \frac{16k_1^2}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \left[\frac{(1 - S_i(d))}{\pi_i(d) + S_i(d)}\right] \left[\frac{(1 - S_j(d))}{\pi_j(d) + S_j(d)}\right] \big|\mathrm{Cov}\big(D_{id}, D_{jd}\big)\big| \\
&\qquad \leq 16k_1^2 \big[\Pi(d,p)\big]^2 C_d\big(p/(p-2)\big).
\end{aligned}$$

Focusing on the second term, recall from the proof of Proposition 2:

$$\mathrm{E}\big[\varepsilon_i \varepsilon_j \;\big|\; D_i = D_j = d\big] = e_{ij}(d,d) e_{ji}(d,d) + \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d)$$

when $S_{ij}(d,d) = 0$. It follows:

$$\begin{aligned}
&\frac{4}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \big[1 - S_{ij}(d,d)\big] \, \mathrm{E}\big[\varepsilon_i \varepsilon_j \;\big|\; D_i = D_j = d\big] \\
&\quad = \frac{4}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} e_{ij}(d,d) e_{ji}(d,d) + \frac{4}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d) = 4E_d + 4U_d,
\end{aligned}$$

because $e_{ij}(d,d) = \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = D_j = d) = 0$ when $S_{ij}(d,d) = 1$. Taken together:

$$\mathrm{Var}\big(\hat{\tau}_{\text{ZHT}}(a,b)\big) \leq \frac{8k_1^2 k}{n} + 20k_1^2 k^2 \bar{C} + 4E_a + 4E_b + 4U_a + 4U_b.$$

where $\bar{C} = C_a\big(p/(p-2)\big) + C_b\big(p/(p-2)\big)$ and $\Pi(d,p) \leq k$ for $d \in \{a,b\}$ by the premise of the proposition.

Decompose the root mean square error with respect to the exposure effect into the estimator's bias and variance:

$$\sqrt{\mathrm{E}\Big[\big(\hat{\tau}(a,b) - \tau(a,b)\big)^2\Big]} \leq \big|\mathrm{E}[\hat{\tau}_{\text{ZHT}}(a,b)] - \tau(a,b)\big| + \sqrt{\mathrm{Var}\big(\hat{\tau}(a,b)\big)},$$

which gives:

$$\sqrt{\mathrm{E}\Big[\big(\hat{\tau}(a,b) - \tau(a,b)\big)^2\Big]} = \mathcal{O}\big(n^{-0.5} + \bar{S}_a + \bar{S}_b + \bar{C}^{0.5} + E_a^{0.5} + E_b^{0.5} + U_a^{0.5} + U_b^{0.5}\big),$$

and Markov's inequality gives the rate of convergence in probability. □

B.9 Proposition 9

Proposition 9. *If Conditions 1 and 2 hold, then:*

$$\mathrm{E}\Big[\widehat{\mathrm{Var}}_{\mathrm{AS}}\big(\hat{\tau}(a,b)\big)\Big] = \mathrm{Var}\big(\hat{\tau}(a,b)\big) + B_1 + B_2(a,b) + B_2(b,a) + B_3(a) + B_3(b) \\ + 2B_4(a,b) - B_4(a,a) - B_4(b,b),$$

where:

$$B_1 = \frac{1}{n^2}\sum_{i=1}^{n}\big[\bar{y}_i(a) - \bar{y}_i(b)\big]^2,$$

$$B_2(d,q) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j\neq i}\Big(S_{ij}(d,d)\big[\bar{y}_i(d) + \bar{y}_j(d)\big]^2 + S_{ij}(d,q)\big[\bar{y}_i(d) - \bar{y}_j(q)\big]^2\Big),$$

$$B_3(d) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(d,a) + S_{ij}(d,b)\big]\,\mathrm{Var}(\varepsilon_i \mid D_i = d),$$

$$B_4(d,q) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(d,q)\big]\Big[\bar{y}_i(d)e_{ji}(q,d) + \bar{y}_j(q)e_{ij}(d,q) + e_{ij}(d,q)e_{ji}(q,d) \\ + \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = d, D_j = q)\Big].$$

Proof. Consider:

$$P_{ij}(d,q) = \frac{\pi_{ij}(d,q) - \pi_i(d)\pi_j(q)}{\pi_{ij}(d,q)\pi_i(d)\pi_j(q) + S_{ij}(d,q)} \\ = \frac{1 - S_{ij}(d,q)}{\pi_i(d)\pi_j(q)} - \frac{1 - S_{ij}(d,q)}{\pi_{ij}(d,q) + S_{ij}(d,q)} - S_{ij}(d,q)\pi_i(d)\pi_j(q)$$

which allows the following decomposition of the first term of the variance estimator:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ia} - D_{ib})(D_{ja} - D_{jb})P_{ij}(D_i, D_j)Y_iY_j \\ = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{[1 - S_{ij}(D_i, D_j)](D_{ia} - D_{ib})(D_{ja} - D_{jb})}{\pi_i(D_i)\pi_j(D_j)}Y_iY_j \\ - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{[1 - S_{ij}(D_i, D_j)](D_{ia} - D_{ib})(D_{ja} - D_{jb})}{\pi_{ij}(D_i, D_j) + S_{ij}(D_i, D_j)}Y_iY_j$$

$$- \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ij}(D_i, D_j)(D_{ia} - D_{ib})(D_{ja} - D_{jb})\pi_i(D_i)\pi_j(D_j)Y_iY_j$$

Note that $S_{ij}(D_i, D_j) = 0$ with probability one, so:

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{[1 - S_{ij}(D_i, D_j)](D_{ia} - D_{ib})(D_{ja} - D_{jb})}{\pi_i(D_i)\pi_j(D_j)} Y_iY_j = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(D_{ia} - D_{ib})Y_i}{\pi_i(D_i)} \right)^2$$
$$= \left(\frac{1}{n} \sum_{i=1}^{n} \left[\frac{D_{ia}Y_i}{\pi_i(a)} - \frac{D_{ib}Y_i}{\pi_i(b)} \right] \right)^2 = \big(\hat{\tau}(a, b)\big)^2$$

and:

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ij}(D_i, D_j)(D_{ia} - D_{ib})(D_{ja} - D_{jb})\pi_i(D_i)\pi_j(D_j)Y_iY_j = 0$$

Consider the expectation of the second term of the decomposition:

$$\mathrm{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{[1 - S_{ij}(D_i, D_j)](D_{ia} - D_{ib})(D_{ja} - D_{jb})}{\pi_{ij}(D_i, D_j) + S_{ij}(D_i, D_j)} Y_iY_j \right]$$
$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \big[1 - S_{ij}(a, a)\big] \, \mathrm{E}\big[Y_iY_j \;\big|\; D_i = D_j = a\big]$$
$$+ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \big[1 - S_{ij}(b, b)\big] \, \mathrm{E}\big[Y_iY_j \;\big|\; D_i = D_j = b\big]$$
$$- \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \big[1 - S_{ij}(a, b)\big] \, \mathrm{E}\big[Y_iY_j \;\big|\; D_i = a, D_j = b\big]$$
$$+ \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{E}\big[Y_i^2 \;\big|\; D_i = a\big] + \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{E}\big[Y_i^2 \;\big|\; D_i = b\big]$$

which follows from $S_{ii}(a, a) = S_{ii}(b, b) = 0$ by Condition 2, and $S_{ii}(a, b) = 1$ by the fundamental problem of causal inference. Consider the conditional expectations:

$$\mathrm{E}\big[Y_iY_j \;\big|\; D_i = d, D_j = q\big] = \mathrm{E}\Big[\big(\bar{y}_i(D_i) + e_{ij} + u_{ij}\big)\big(\bar{y}_j(D_j) + e_{ji} + u_{ji}\big) \;\Big|\; D_i = d, D_j = q \Big]$$
$$= \bar{y}_i(d)\bar{y}_j(q) + \bar{y}_i(d)e_{ji}(q, d) + \bar{y}_j(q)e_{ij}(d, q) + e_{ij}(d, q)e_{ji}(q, d)$$
$$+ \mathrm{Cov}(u_{ij}, u_{ji} \mid D_i = d, D_j = q)$$

which follows from $\mathrm{E}[u_{ij} \mid D_i = d, D_j = q] = 0$. So:

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{[1 - S_{ij}(D_i, D_j)](D_{ia} - D_{ib})(D_{ja} - D_{jb})}{\pi_{ij}(D_i, D_j) + S_{ij}(D_i, D_j)}Y_iY_j\right] \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(a,a)\big]\bar{y}_i(a)\bar{y}_j(a) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(b,b)\big]\bar{y}_i(b)\bar{y}_j(b) \\
&\qquad\quad - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(a,b)\big]\bar{y}_i(a)\bar{y}_j(b) + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\big[Y_i^2 \;\big|\; D_i = a\big] \\
&\qquad\quad + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\big[Y_i^2 \;\big|\; D_i = b\big] + B_4(a,a) + B_4(b,b) - 2B_4(a,b)
\end{aligned}$$

Turning to the expectation of the second term in the variance estimator:

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{D_{ia}}{\pi_i(a)} + \frac{D_{ib}}{\pi_i(b)}\right]\big[S_{ij}(D_i, a) + S_{ij}(D_i, b)\big]Y_i^2\right] \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(a,a) + S_{ij}(a,b)\big]\,\mathrm{E}\big[Y_i^2 \;\big|\; D_i = a\big] \\
&\qquad\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(b,a) + S_{ij}(b,b)\big]\,\mathrm{E}\big[Y_i^2 \;\big|\; D_i = b\big] \\
&\qquad\quad + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\big[Y_i^2 \;\big|\; D_i = a\big] + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\big[Y_i^2 \;\big|\; D_i = b\big] \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(a,a) + S_{ij}(a,b)\big][\bar{y}_i(a)]^2 \\
&\qquad\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(b,a) + S_{ij}(b,b)\big][\bar{y}_i(b)]^2 \\
&\qquad\quad + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\big[Y_i^2 \;\big|\; D_i = a\big] + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}\big[Y_i^2 \;\big|\; D_i = b\big] \\
&\qquad\quad + B_3(a) + B_3(b)
\end{aligned}$$

because:

$$\mathrm{E}\big[Y_i^2 \;\big|\; D_i = d\big] = \mathrm{E}\big[\big(\bar{y}_i(D_i) + \varepsilon_i\big)^2 \;\big|\; D_i = d\big] = [\bar{y}_i(d)]^2 + \mathrm{Var}(\varepsilon_i \mid D_i = d)$$

since $\mathrm{E}[\varepsilon_i \mid D_i = d] = 0$.

Combined, this gives:

$$\begin{aligned}
\mathrm{E}\Big[\widehat{\mathrm{Var}}_{\mathrm{AS}}\big(\hat{\tau}(a,b)\big)\Big] = \mathrm{E}\Big[\big(\hat{\tau}(a,b)\big)^2\Big] &- \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(a,a)\big]\bar{y}_i(a)\bar{y}_j(a) \\
&- \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(b,b)\big]\bar{y}_i(b)\bar{y}_j(b) \\
&+ \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1 - S_{ij}(a,b)\big]\bar{y}_i(a)\bar{y}_j(b) \\
&+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(a,a) + S_{ij}(a,b)\big][\bar{y}_i(a)]^2 \\
&+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[S_{ij}(b,a) + S_{ij}(b,b)\big][\bar{y}_i(b)]^2 \\
&+ B_3(a) + B_3(b) + 2B_4(a,b) - B_4(a,a) - B_4(b,b)
\end{aligned}$$

Note that $S_{ij}(d,d) = S_{ji}(d,d)$, so:

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,d)[\bar{y}_i(d)]^2 &= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,d)\Big[[\bar{y}_i(d)]^2 + [\bar{y}_j(d)]^2\Big] \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,d)\Big[\big[\bar{y}_i(d) + \bar{y}_j(d)\big]^2 - 2\bar{y}_i(d)\bar{y}_j(d)\Big] \\
&= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,d)\big[\bar{y}_i(d) + \bar{y}_j(d)\big]^2 \\
&\qquad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,d)\bar{y}_i(d)\bar{y}_j(d)
\end{aligned}$$

Also $S_{ij}(d,q) = S_{ji}(q,d)$, so:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,q)[\bar{y}_i(d)]^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ji}(q,d)[\bar{y}_j(q)]^2 \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,q)\Big[[\bar{y}_i(d)]^2 + [\bar{y}_j(q)]^2\Big] \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,q)\Big[\big[\bar{y}_i(d) - \bar{y}_j(q)\big]^2 + 2\bar{y}_i(d)\bar{y}_j(q)\Big]
\end{aligned}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,q)\big[\bar{y}_i(d)-\bar{y}_j(q)\big]^2 + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(d,q)\bar{y}_i(d)\bar{y}_j(q)$$

This gives:

$$\begin{aligned}
\mathrm{E}\Big[\widehat{\mathrm{Var}}_{\mathrm{AS}}\big(\hat{\tau}(a,b)\big)\Big] &= \mathrm{E}\Big[\big(\hat{\tau}(a,b)\big)^2\Big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1-S_{ij}(a,a)\big]\bar{y}_i(a)\bar{y}_j(a) \\
&\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1-S_{ij}(b,b)\big]\bar{y}_i(b)\bar{y}_j(b) \\
&\quad + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\big[1-S_{ij}(a,b)\big]\bar{y}_i(a)\bar{y}_j(b) \\
&\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(a,a)\bar{y}_i(a)\bar{y}_j(a) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(b,b)\bar{y}_i(b)\bar{y}_j(b) \\
&\quad + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} S_{ij}(a,b)\bar{y}_i(a)\bar{y}_j(b) + B_2(a,b) + B_2(b,a) \\
&\quad + B_3(a) + B_3(b) + 2B_4(a,b) - B_4(a,a) - B_4(b,b) \\
&= \mathrm{E}\Big[\big(\hat{\tau}(a,b)\big)^2\Big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\bar{y}_i(a)\bar{y}_j(a) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\bar{y}_i(b)\bar{y}_j(b) \\
&\quad + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\bar{y}_i(a)\bar{y}_j(b) + B_2(a,b) + B_2(b,a) \\
&\quad + B_3(a) + B_3(b) + 2B_4(a,b) - B_4(a,a) - B_4(b,b)
\end{aligned}$$

Recall:

$$B_1 = \frac{1}{n^2}\sum_{i=1}^{n}\big[\bar{y}_i(a)-\bar{y}_i(b)\big]^2 = \frac{1}{n^2}\sum_{i=1}^{n}[\bar{y}_i(a)]^2 + \frac{1}{n^2}\sum_{i=1}^{n}[\bar{y}_i(b)]^2 - \frac{2}{n^2}\sum_{i=1}^{n}\bar{y}_i(a)\bar{y}_i(b)$$

so:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\bar{y}_i(a)\bar{y}_j(a) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\bar{y}_i(b)\bar{y}_j(b) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\bar{y}_i(a)\bar{y}_j(b) \\
&\qquad = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big[\bar{y}_i(a)\bar{y}_j(a) + \bar{y}_i(b)\bar{y}_j(b) - 2\bar{y}_i(a)\bar{y}_j(b)\big] - B_1
\end{aligned}$$

and:

$$\begin{aligned}\mathrm{E}\Big[\widehat{\mathrm{Var}}_{\mathrm{AS}}\big(\hat{\tau}(a,b)\big)\Big] = \mathrm{E}\Big[\big(\hat{\tau}(a,b)\big)^2\Big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big[\bar{y}_i(a)\bar{y}_j(a) + \bar{y}_i(b)\bar{y}_j(b) - 2\bar{y}_i(a)\bar{y}_j(b)\big] \\ + B_1 + B_2(a,b) + B_2(b,a) + B_3(a) + B_3(b) \\ + 2B_4(a,b) - B_4(a,a) - B_4(b,b)\end{aligned}$$

Finally:

$$\begin{aligned}&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big[\bar{y}_i(a)\bar{y}_j(a) + \bar{y}_i(b)\bar{y}_j(b) - 2\bar{y}_i(a)\bar{y}_j(b)\big] \\ &\quad = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big[\bar{y}_i(a) - \bar{y}_i(b)\big]\big[\bar{y}_j(a) - \bar{y}_j(b)\big] = \left(\frac{1}{n}\sum_{i=1}^{n}\big[\bar{y}_i(a) - \bar{y}_i(b)\big]\right)^2 = \big(\tau(a,b)\big)^2\end{aligned}$$

so:

$$\mathrm{E}\Big[\big(\hat{\tau}(a,b)\big)^2\Big] - \big(\tau(a,b)\big)^2 = \Big(\mathrm{E}\big[\hat{\tau}(a,b)\big]\Big)^2 = \mathrm{Var}\big(\hat{\tau}(a,b)\big) \qquad \square$$