A Design-Based Riesz Representation Framework for Randomized Experiments

Christopher Harshaw[1], Fredrik Sävje[2], and Yitan Wang[2]

[1] *University of California, Berkeley*
[2] *Yale University*

October 25, 2022

Abstract

We describe a new design-based framework for drawing causal inference in randomized experiments. Causal effects in the framework are defined as linear functionals evaluated at potential outcome functions. Knowledge and assumptions about the potential outcome functions are encoded as function spaces. This makes the framework expressive, allowing experimenters to formulate and investigate a wide range of causal questions. We describe a class of estimators for estimands defined using the framework and investigate their properties. The construction of the estimators is based on the Riesz representation theorem. We provide necessary and sufficient conditions for unbiasedness and consistency. Finally, we provide conditions under which the estimators are asymptotically normal, and describe a conservative variance estimator to facilitate the construction of confidence intervals for the estimands.

We thank P. M. Aronow, Xiaohong Chen, Jon Erickson, Avi Feller, Kenneth Hung, Brian Karrer, Laurent Lessard, Joel Middleton, Benjamin Recht, Jamie Robins, Liang Shi, Rahul Singh, Daniel A. Spielman, Vasilis Syrgkanis, and Angela Zhou for helpful comments and discussions. Part of this work was done while Christopher Harshaw was visiting the Simons Institute for the Theory of Computing. Christopher Harshaw gratefully acknowledges support from Foundations of Data Science (FODSI), NSF grant DMS-2023505, and Yitan Wang gratefully acknowledges support from ONR Award N00014-20-1-2335.

Contents

1 Introduction **3**

2 Illustrations **4**

2.1 Policing Experiments 5

2.2 Cash Transfer Programs 6

3 The Framework **8**

3.1 Treatments and Potential Outcomes 8

3.2 Effect Functionals 9

3.3 Examples of Effect Functionals 10

3.4 Model Spaces 13

3.5 Positivity 15

3.6 Examples of Model Spaces 16

4 The Riesz Estimator **22**

4.1 Overview and Definition 22

4.2 Existence and Unbiasedness 24

4.3 Constructing the Riesz Representors 27

4.4 Previous Uses of Riesz Representors 28

4.5 Examples of Riesz Estimators 30

5 Properties of the Riesz Estimator **33**

5.1 Asymptotic Regime and Regularity Conditions 33

5.2 Necessary and Sufficient Conditions for Consistency 35

5.3 Sufficient Conditions for Consistency 38

5.4 Asymptotic Normality 39

6 Variance Estimation and Inference **40**

6.1 Overview 40

6.2 Tensor Spaces and Tensorization of the Variance 41

6.3 Constructing Variance Riesz Representors 42

6.4 Variance Bound for the Riesz Estimator 45

6.5 Confidence Intervals and Consistency of Variance Estimator 48

7 Concluding remarks **49**

1 Introduction

Randomized experiments have been widely adopted in a diverse set of fields, including clinical trials, policy evaluations, and social science research. However, the use of experiments is limited, partly because conventional experimental methodologies cannot accommodate the range of settings empirical researchers are interested in. An important such limitation is the assumption of no interference, stating that a unit's outcome or behavior is unaffected by the treatment assigned to other units. Another limitation is the widespread focus on binary treatments, as opposed to, for example, continuous or set-valued treatments.

A growing body of work seeks to address these limitations, but both conventional and recently developed frameworks lack sufficient expressiveness to formulate the full range of questions empirical researchers are interested in. Furthermore, recently developed frameworks and methods still tend to require strong assumptions, and they are generally justified by an assumption of sampling from an imagined super-population, in contrast to a design-based justification in which all randomness under consideration stems from the experiment itself. This restricts the applicability and interpretability of current frameworks and methods.

In this paper, we describe a new design-based experimental framework for causal estimation under interference to address these limitations. The purpose of the framework is to be sufficiently expressive to allow experimenters to define and investigate a wide range of causal questions involving continuous or discrete treatments under rich and complex interference structures. At the same time, the framework is constructed to be sufficiently tractable to admit precise estimation and inference of the estimands defined with it. The framework unifies and generalizes all previously developed design-based frameworks that we are aware of.

There are three main contributions of this paper.

1. We describe a new design-based causal inference framework, in which causal effects are defined using linear functionals evaluated on potential outcome functions. This gives empirical researchers an expressive tool to formulate new types of estimands that are relevant for policy and substantive theory. In the framework, assumptions about the potential outcome functions are formalized as function spaces, giving empirical researchers flexibility also in formulating any knowledge or beliefs they might have about the environment under study.
2. We describe a new treatment effect estimator, which we call the Riesz estimator. The estimator can be used for any estimand defined in the framework as long as a type of positivity assumption holds. We provide both finite- and large-sample analyses of the estimator, including conditions for unbiasedness, consistency and asymptotic normality. These conditions are empirically non-trivial, in the sense that there exist many practically relevant situations in which they do not hold, but we believe the

conditions are sufficiently easy to understand, given their generality, that empirical researchers will find it feasible to assess their validity.

3. We describe a new conservative variance estimator for the Riesz estimator, and provide conditions under which the variance estimator is consistent. This facilitates the construction of asymptotically valid confidence intervals.

The statistical methodology developed in this paper uses several insights from functional analysis, including the Riesz representation theorem that has given the paper its title. We believe these insights shed light on the underlying principles that facilitate inference of causal effects more generally in the design-based paradigm, both with and without interference, and we believe the insights will be of independent interest to many researchers working in causal inference.

The Riesz representation theorem has previously been used in the semiparametric causal inference literature. To the best of our knowledge, the earliest uses were by Newey (1994) to characterize the variance of semiparametric estimators and independently by Robins, Rotnitzky, and Zhao (1994) to construct doubly robust estimators. More recently, the theorem has been prominently featured in work by Rotnitzky, Smucler, and Robins (2020), Hirshberg and Wager (2021) and Chernozhukov, Newey, and Singh (2022a), among others. We describe previous uses of the Riesz representation theorem in causal inference in Section 4.4. To the best of our knowledge, we are the first to use the Riesz representation theorem in a design-based setting, and the application of the theorem in this paper is different from the application in the semiparametric literature. Nevertheless, there are commonalities, and we hope that this paper will serve as starting point for building bridges between the design-based and semiparametric perspectives on causal inference.

Because the framework and methods we describe in this paper generalize and unify many approaches previously described in the design-based causal inference literature, we find it advantageous to review and highlight these connections throughout the paper, rather than describing the connections to prior work in a dedicated section.

2 Illustrations

We first describe a restricted version of our framework to illustrate its use. Consider an experiment with n units, where each unit is assigned a continuous treatment $Z_i \in [0,1]$ selected at random. We collect the treatments in a vector $\boldsymbol{Z} = (Z_1, \dots, Z_n)$. The distribution of $\boldsymbol{Z}$ can be complex and does not necessarily factor into independent and identical components. The units can potentially interfere with each other, meaning that the outcome of one unit could depend on the treatments of other units. Therefore, the potential outcome function $y_i(\boldsymbol{z})$ for each unit maps from $[0,1]^n$ to $\mathbb{R}$.

The conventional estimand in this setting is the so-called Global Average Treatment Effect, or the "all-or-nothing" effect, which is the difference in outcomes when all units are assigned to the extreme ends of the treatment variable interval: $n^{-1}\sum_{i=1}^{n}[y_i(\mathbf{1}) - y_i(\mathbf{0})]$. This estimand has been investigated by Eckles, Karrer, and Ugander (2017), Chin (2019), Harshaw, Sävje, Eisenstat, Mirrokni, and Pouget-Abadie (2021) and Leung (2022b), among others. While the all-or-nothing effect is useful in some settings, and an estimand that our framework can accommodate, it is a crude summary of the ways the treatments affect the outcomes. Empirical researchers often seek a deeper and more nuanced picture of how the treatments are causally related to the outcomes. We provide two examples of such estimands in this section.

2.1 Policing Experiments

A type of policy that has received recent attention in the social sciences is intensive policing in high-crime settings. The general idea is that there might be increasing returns to scale of policing. That is, intensive policing, such as around-the-clock patrolling, in a few selected locations might be more effective in reducing crime than moderate policing in many locations. An example of such a study is Blattman, Green, Ortega, and Tobón (2021), who investigate how intensive policing affects crime in the city of Bogotá, Colombia.

Here, the units are some type of geographical areas, such as neighborhoods, and Z_i denotes the amount of policing allocated to area i. Policing is arguably best seen as a continuous variable (e.g., the number of hours patrolled), but following the convention in this literature, Blattman et al. (2021) discretize the treatment variable into two levels, corresponding to high and low levels of patrolling. In our illustration, $Z_i \in [0, 1]$ is continuous, where $Z_i = 0$ denotes no policing and $Z_i = 1$ denotes the most policing the policy maker finds relevant or imaginable. The outcome $y_i(\boldsymbol{z})$ is some measure of crime activity, such as the number of reported crimes in the neighborhood.

It is common to focus on a version of the all-or-nothing effect when studying the effects of intensive policing, although the estimand is often defined only informally or implicitly. However, the focus on the all-or-nothing effect can be problematic in this setting, because the effect is typically not relevant to policy makers. This is because neither the "all" policy (e.g., patrolling all the neighborhoods all the time) nor the "nothing" policy (never patrol any neighborhoods) are viable or desirable policing policies. Instead, the relevant patrolling policies are arguably stochastic, because otherwise organized criminals could adapt their behavior to avoid encountering the patrols.

If we are interested evaluating the performance of stochastic policy F relative to some other policy G, where F and G are distribution functions over $[0, 1]^n$ for the assignment of

$\boldsymbol{Z}$, then a reasonable estimand is

$$\frac{1}{n}\sum_{i=1}^{n}\int_{[0,1]^n} y_i(\boldsymbol{z})\,\mathrm{d}\mu(\boldsymbol{z}),$$

where $\mu(\boldsymbol{z}) = F(\boldsymbol{z}) - G(\boldsymbol{z})$ is a signed measure. Note that F and G can implicitly depend on characteristics of the units through the unit indices, so they can be targeted policies.

2.2 Cash Transfer Programs

Another type of policy that has received attention in the social sciences in general, and in development economics in particular, is conditional or unconditional cash transfer programs, the latter of which is sometimes referred to as universal basic income (UBI) programs. Here, the units are people, and Z_i denotes the cash transfer that person i receives, where $Z_i = 0$ denotes no transfer and $Z_i = 1$ denotes the maximum possible transfer. An example of a study investigating such a program is Haushofer and Shapiro (2016), who investigate how different amounts of cash in an unconditional cash transfer program targeting poor households in Kenya affected consumption and well-being.

The typical estimand in a cash transfer experiment is a contrast between two levels of the transfer, resulting in a version of the all-or-nothing effect. A possible alternative estimand is the effect of a marginal increase of the transfer, which for example would be relevant to a policy maker who is considering making changes to an existing cash transfer program. We can use the derivative of the potential outcome function to capture such a marginal effect. For example, the overall effect of a marginal increase of the transfer is captured by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial y_i}{\partial z_j}(\boldsymbol{s}),$$

where $\boldsymbol{s} \in [0,1]^n$ is the current transfer level. It is also possible to consider the overall effect under some alternative, counterfactual policy $\boldsymbol{s}' \in [0,1]^n$ different from the current policy $\boldsymbol{s}$.

Note that this marginal effect takes into account all spillover effects of the increase. We could imagine that increasing the transfer could have a positive effect for the person who receives the transfer but be detrimental to other people because of crowding-out or inflationary effects. As highlighted by Egger, Haushofer, Miguel, Niehaus, and Walker (2022), among others, general equilibrium effects, which is a type of spillover effect, is an important consideration when evaluating cash transfer programs.

It is possible to define effects that provide a more nuanced picture of these spillover effects. One can for example decompose the overall effect into direct and indirect effects:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial y_i}{\partial z_i}(\boldsymbol{s}) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\partial y_i}{\partial z_j}(\boldsymbol{s}).$$

The first estimand captures the marginal effect of increasing the transfer on the person receiving the transfer, while the second estimand captures effect on other people. The first estimand is related to the Incremental Causal Effect, which is studied by Rothenhäusler and Yu (2019) in a super-population setting. The second estimand is informative of potential spillover and general equilibrium effects. If units are suspected to primarily interfere along the edges of a known graph, then it might aid interpretability and generalizability to restrict the indirect estimand to units in neighborhoods of the graph. Letting $\mathcal{N}_i$ denote the indices of the units adjacent to unit i in the graph, we can define a restricted indirect effect as

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_i}\frac{\partial y_i}{\partial z_j}(\boldsymbol{s}).$$

This type of restriction of the estimand might make sense even if units interfere in ways that are not captured by the graph. That is, the restricted indirect effect is well-defined and might be useful even if units interfere with units outside of their neighborhoods, $\mathcal{N}_i$.

The purpose of these examples is to illustrate questions that cannot be investigated using existing design-based frameworks and methods, but can be investigated using the framework we describe in this paper. It is sometimes possible to translate some of these questions into a form that can be accommodated by existing frameworks, but that translation must be done on a case-by-case basis, and the translation tends to be opaque, in our view. The causal inference literature based on a super-population justification has made progress on some of these types of estimands, but it is often difficult or impossible to express general forms of interference when working in a super-population framework, limiting its applicability. However, recent work in the super-population framework has make considerable progress in the direction of allowing for more general forms of interference (see, e.g., Li & Wager, 2022; Ogburn, Sofrygin, Diaz, & van der Laan, 2022; Tchetgen Tchetgen, Fulcher, & Shpitser, 2021). We see this paper as complementary to the ongoing work in the super-population literature.

3 The Framework

3.1 Treatments and Potential Outcomes

The type of experiment we consider in this paper consists of two primitives: a treatment variable and a set of outcome measurements. The treatment variable captures various interventions the experimenter can make. These interventions could potentially affect the world, including the outcomes, and our aim is to estimate these effects.

Let $\mathcal{Z}$ denote the set of all interventions that the experimenter potentially could do. We do not impose any restrictions on properties of the interventions, meaning that $\mathcal{Z}$ can have whatever structure most suitable for the application at hand. We will use $z \in \mathcal{Z}$ to denote a particular intervention, which we refer to as a treatment assignment. By virtue of being a randomized experiment, an element from $\mathcal{Z}$ is selected at random to be the realized intervention. Together with an appropriate σ-algebra and a probability measure, $\mathcal{Z}$ forms a probability space that describes how the intervention is selected. Let Z denote the randomly selected intervention. We refer to the distribution of Z as the *experimental design*, which we take to be known to the experimenter. The only randomness under consideration in this paper is that which is induced by the experimental design.

There are n outcome measurements in the experiment, which are indexed by integers $i \in [n] = \{1, \dots, n\}$. For convenience, we will refer to the measurements as *units*, but they do not need to be distinct units as conventionally understood. Each unit has an associated potential outcome function y_i that maps the treatment variable to an observed value of the measurement. The function describes what the outcome would have been under a particular, potentially counterfactual, intervention. It is assumed that all units have well-defined potential outcome functions, in the sense that an unambiguous outcome is produced by each intervention in $\mathcal{Z}$, and that this outcome can be observed. Let $Y_i = y_i(Z)$ denote the (random) realized outcome for unit i.

We define $\mathcal{Y}$ to be the function space of all functions $y : \mathcal{Z} \to \mathbb{R}$ that are measurable and square integrable with respect to the design:

$$\mathcal{Y} = \left\{\, y \in \mathbb{R}^{\mathcal{Z}} : \quad y \text{ is measurable} \quad \text{and} \quad \operatorname{E}\!\left[y(Z)^2\right] < \infty \,\right\}.$$

The space $\mathcal{Y}$ contains all potential outcome functions that we will consider in this paper. This space might be smaller than the complete set of all real-valued functions on $\mathcal{Z}$, but the purpose of $\mathcal{Y}$ is not to encode knowledge or beliefs the experimenter might have about how the interventions affect the outcomes. The purpose is instead to impose enough structure on $\mathcal{Y}$ to ensure that the objects we construct on the space are well-defined and sensible.

A special case of our framework is the conventional causal inference setting. Here, $\mathcal{Z} = \{0, 1\}^n$ is a set of all binary vectors, corresponding to active and control treatment. In examples involving binary treatment vectors, we will write the treatment variable as a vector for clarity: $\boldsymbol{Z} = (Z_1, \dots, Z_n)$. An example of a possible intervention is $\boldsymbol{z} = (0, 1, 1, 0, \dots, 1, 0) \in \mathcal{Z}$. In this setting, there is generally a correspondence between the ith unit and the ith coordinate of the treatment vector, Z_i, which is said to be the treatment assigned to the unit. Our framework does not require such a correspondence between the units and the treatment variable.

The framework also encompasses other settings, as illustrated by the following examples. Zigler and Papadogeorgou (2021) consider bipartite experiments, in which the units receiving treatment are different from the units for which we measure outcomes, meaning that $\mathcal{Z} = \{0, 1\}^m$ for some $m \neq n$, and there is no obvious mapping from the dimensions of the treatment variable to the units. Aronow, Samii, and Wang (2021), Papadogeorgou, Imai, Lyall, and Li (2020) and Pollmann (2020) consider spatial experiments, in which the possible interventions are geographical locations, meaning that $\mathcal{Z}$ is a set of vectors of geographical coordinates. VanderWeele and Hernan (2013) considers experiments with hidden versions of treatment, in which the nominal treatments are coarsened versions of $\mathcal{Z}$. Hirano and Imbens (2004), Galvao and Wang (2015), Kennedy, Ma, McHugh, and Small (2017), and Rothenhäusler and Yu (2019), among others, consider continuous treatments, $\mathcal{Z} \subseteq \mathbb{R}^n$, in a super-population framework. Kennedy (2019) and Hu, Li, and Wager (2022) consider the causal effect of changes to the experimental design, which for example could be an increase in the probability of receiving treatment. Basse, Ding, Feller, and Toulis (2019) consider causal effects of group formation, where the treatment is the assignment of units into groups. All of these examples fit in the current framework.

3.2 Effect Functionals

The starting point for the definition of a causal estimand in our framework is a set of individual treatment effects. Each individual effect is defined by a linear functional evaluated at the corresponding unit's potential outcome function. A linear functional is a function $\theta : \mathcal{Y} \to \mathbb{R}$ with the properties $\theta_i(f + g) = \theta_i(f) + \theta_i(g)$ and $\theta_i(\alpha f) = \alpha\theta_i(f)$ for $f, g \in \mathcal{Y}$ and $\alpha \in \mathbb{R}$. We refer to these functionals as *effect functionals*. This approach to effect definition is more expressive than conventional approaches, and it accommodates a wide range of types of treatments, meaning that we do not need to impose any particular structure on $\mathcal{Z}$ to define our effects.

The experimenter specifies an effect functional θ_i for each unit $i \in [n]$, and the individual treatment effect is given by $\tau_i = \theta_i(y_i)$. The estimand of interest is the average of the individual treatment effects:

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \tau_i = \frac{1}{n}\sum_{i=1}^{n} \theta_i(y_i).$$

Because this class of estimands is large and includes essentially all estimands previously considered in the design-based causal inference literature as special cases, we do not find it useful to give τ a particular name. For convenience, we will refer to τ as the aggregate treatment effect. The aggregate treatment effect is not the same as the average treatment effect (ATE), as conventionally defined, although we can reproduce the ATE estimand by using an appropriate set of effect functionals, as described in Section 3.6.1.

The effect functionals θ_i do not need to be the same for all units; indeed, they will typically be different. However, the functionals will often have the same structural interpretation, as illustrated by the examples below. The aggregate treatment effect can be extended to any linear combination of the individual treatment effects, as the coefficients of the combination can be absorbed into the functionals.

There are settings where the effect functionals of interest are not defined on the full function space $\mathcal{Y}$. For example, if a functional is the derivative of the potential outcome function at a certain point, then the potential outcome function must be differentiable at that point for the functional to take a well-defined value. In the remainder of the paper, we implicitly impose the assumption that the true potential outcome functions y_i are in the subset of $\mathcal{Y}$ on which the effect functionals of interest are well-defined. It is possible that this subset is different for different units. Note that this assumption is not related to estimation, and it is imposed only to ensure that the treatment effects exist and are sensible.

The use of functionals to define parameters of interest has a long history in econometrics and the super-population causal inference literature. The function that the functional is evaluated on in these literatures is typically a conditional expectation function with respect to some infinite population rather than individual potential outcome functions, as in this paper, meaning that the functionals capture population-level characteristics. To the best of our knowledge, we are the first to explicitly use functionals to define treatment effects in a design-based setting, and the first to consider functionals directly on potential outcome functions.

3.3 Examples of Effect Functionals

The use of effect functionals facilitates a large and expressive class of estimands. We believe it is useful to categorize these estimands into different types. We will review some of these types in this section along illustrative examples taken from the previous literature. The types of estimands we review here are partially overlapping and not exhaustive. Given its expressiveness, we anticipate that empirical researchers will use the framework to define estimands that we have not yet envisioned.

The first type is evaluation functionals, which capture the value of the potential outcome functions at particular treatment assignments. A slight extension of this type is contrastive functionals, which capture contrasts of potential outcome functions at different treatment assignments. The conventional definition of the average treatment effect is implicitly constructed using contrastive functionals. There are other examples.

- The Global Average Treatment Effect, or the all-or-nothing effect, discussed in Section 2 is constructed using contrastive functionals. In particular, the effect is the contrast in average outcomes between when all units are treated and when no unit is treated:
$$\frac{1}{n}\sum_{i=1}^{n}\big[y_i(\mathbf{1}) - y_i(\mathbf{0})\big].$$
The effect functional $\theta(y) = y(\mathbf{1}) - y(\mathbf{0})$ reproduces this estimand.
- We can define various direct and indirect interference effects by making the functional unique to each unit. For example, using the functional $\theta_i(y) = y(\boldsymbol{e}_i) - y(\mathbf{0})$ for unit i, where $\boldsymbol{e}_i$ is the standard basis vector of dimension n, produces a type of direct effect estimand. A related effect is given by $\theta_i(y) = y(\mathbf{1}) - y(\mathbf{1} - \boldsymbol{e}_i)$. The functional $\theta_i(y) = y(\mathbf{1} - \boldsymbol{e}_i) - y(\mathbf{0})$ produces a type of indirect or spillover effect estimand.

The second type is integration functionals, which capture the value of potential outcome functions over some set of interventions with respect to some measure. Integration functionals can for example be used to capture the average, expected or accumulation of an outcome under a treatment policy. We have already seen one such estimand in the intensive policing example in Section 2. There are other examples.

- Hudgens and Halloran (2008) define a direct effect estimand under interference as the average contrast in the expected outcome when a unit is treated and when it is not treated:
$$\frac{1}{n}\sum_{i=1}^{n}\big\{\mathrm{E}[y_i(\boldsymbol{Z}) \mid Z_i = 1] - \mathrm{E}[y_i(\boldsymbol{Z}) \mid Z_i = 0]\big\}.$$
Because the expectation operator is linear, this effect is captured by the effect functional $\theta_i(y) = \int_{\mathcal{Z}} y(\boldsymbol{z})\,\mathrm{d}\mu_i(\boldsymbol{z})$, where μ_i is a signed measure defined as $\mu_i(\boldsymbol{z}) = (2z_i - 1)\Pr(\boldsymbol{Z} = \boldsymbol{z} \mid Z_i = z_i)$. Sävje, Aronow, and Hudgens (2021) describe a related estimand that often is easier to interpret. Here, the signed measure is $\mu_i(\boldsymbol{z}) = (2z_i - 1)\Pr(\boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i})$, where $\boldsymbol{Z}_{-i}$ denotes the vector $\boldsymbol{Z}$ excluding its ith element. Both these estimands can be seen as combining contrastive and integration aspects, because the treatment assignments are partially but not fully pinned down.
- Aronow and Samii (2017) investigate causal inference when the treatment assignments are summarized using exposure mappings. An exposure mapping $d_i : \{0,1\}^n \to \Delta$ is a unit-specific mapping from the treatment vector to a low-dimensional, discrete space of labels, $\Delta \subset \mathbb{N}$. The effect of the exposures as considered by Aronow and

Samii (2017) requires additional assumptions to be defined, but it is possible to define a corresponding effect without those assumptions, and the original authors discuss the possibility of such an extension. This effect is referred to as the expected exposure effect and is studied by, for example, Sävje (2021) and Leung (2022a). The definition is similar to the definitions in the previous bullet point. The functional corresponding to the expected exposure effect for exposures $a, b \in \Delta$ is $\theta_i(y) = \int_{\mathcal{Z}} y(\boldsymbol{z}) \, \mathrm{d}\mu_i(\boldsymbol{z})$, where the measure is

$$\mu_i(\boldsymbol{z}) = \big\{ \mathbb{1}[d_i(\boldsymbol{z}) = a] - \mathbb{1}[d_i(\boldsymbol{z}) = b] \big\} \Pr\big(\boldsymbol{Z} = \boldsymbol{z} \bigm| d_i(\boldsymbol{Z}) = d_i(\boldsymbol{z})\big).$$

While exposure mappings have conventionally only been considered for binary treatments, the idea generalizes to arbitrary treatment variables. The expected exposure effect is related to the effect defined by VanderWeele and Hernan (2013) to study multiple version of treatments.

- Aronow et al. (2021), Papadogeorgou et al. (2020) and Pollmann (2020) consider causal inference for spatial treatments, where the treatments are localized to points in some two-dimensional plane. The estimands they study can be written using effect functionals in the current framework, but we will depart somewhat from their exact definitions to illustrate that the current framework can accommodate more general types of estimands. For example, unlike many papers in this literature, we do not impose the restriction that treatments only can take place in a small number of candidate locations.

  Let $S \subset \mathbb{R}^2$ be a region in the two-dimensional plane, representing geographical coordinates. The set of possible interventions is some measurable collection of subsets of S. For example, a possible treatment assignment is $\{(a, b), (c, d)\} \in \mathcal{Z}$, indicating that treatment is assigned to the two points (a, b) and (c, d). Note that $\mathcal{Z}$ generally is uncountable and cannot be represented by a binary vector; for example, a possible treatment assignment could be some region of S. A possible effect of interest in this setting is the comparison between being close to a treated location and being far from all treated locations. A way to formalize this effect is with the effect functional

$$\theta_i(y) = \int_{C_i} y(\{s\}) \, \mathrm{d}s - y(\{\}),$$

  where $C_i \subset S$ is the region of the plane that is considered to be close to unit i. This estimand compares the outcome when there is only one treated location compared to no treated locations, which might not be relevant for policy if there will always be some treated locations. A possible alternative estimand to address this uses the effect functional

$$\theta_i(y) = \int_{\mathcal{Z}} \int_{C_i} y\Big(\{s\} \cup (S \setminus C_i)\Big) \,\mathrm{d}s \,\mathrm{d}F(S) - \int_{\mathcal{Z}} y(S \setminus C_i) \,\mathrm{d}F(S),$$

where F is some distribution of the treatment assignments, perhaps representing the current policy. This estimand considers the effect of modifying the policy F in a way that forces there to be either one or no treated location close to unit i.

The third and final type of estimand we consider here is based on differentiation functionals, which capture how the potential outcome functions change in certain circumstances. These functionals can for example be used to describe marginal effects for continuous treatments, as in the cash transfer example presented in Section 2. There are other examples.

- Hu, Li, and Wager (2022) defines an infinitesimal policy effect as the average derivative of the expected outcome over the units with respect to the probability of being assigned active treatment. Let F_π denote the distribution of a binary treatment vector under a set of experimental designs parametrized by π. In the case of Hu et al. (2022), the set contains all Bernoulli designs and π is the marginal treatment probability, which for simplicity we set to be the same for all units. The expected outcome for a unit with potential outcome function y under design π is $\mathrm{E}_\pi[y(\boldsymbol{Z})] = \int_{\mathcal{Z}} y(\boldsymbol{z}) \,\mathrm{d}F_\pi(\boldsymbol{z})$, so the infinitesimal policy effect at the individual level is given by the effect functional

$$\theta(y) = \frac{\partial \,\mathrm{E}_\pi[y(\boldsymbol{Z})]}{\partial \pi}.$$

  It is possible to extend this to other designs and to multi-dimensional, partial parameterizations. A related estimand has been studied by Kennedy (2019) in a superpopulation framework.

3.4 Model Spaces

Experimenters use their knowledge and beliefs about the possible effects of the interventions to impose structure on the potential outcome functions. This structure facilitates estimation and inference. In this paper, we formalize the structure on the potential outcome functions using subspaces of the full function space $\mathcal{Y}$. We refer to these subspaces as *model spaces*.

Let $\mathcal{M}_i \subseteq \mathcal{Y}$ denote the model space of unit i. We do not require that $\mathcal{M}_i$ has any particular structure, other than being a subspace of $\mathcal{Y}$ and that the effect functional θ_i is well-defined on $\mathcal{M}_i$. The central results of this paper hold for both finite- or infinite-dimensional model spaces, but some of the computational procedures we describe apply only to model spaces with finite dimensions. The use of infinite-dimensional model spaces also involves a bias–variance trade-off that we do not investigate in this paper. We save the full extension to model spaces with infinite dimensions for future work.

The purpose of the model spaces is to restrict the possible potential outcome functions, as captured by the following assumption.

Assumption 1 (Correctly specified models)**.** Each unit's potential outcome function is in the model space specified for that unit: $y_i \in \mathcal{M}_i$ for all $i \in [n]$.

The assumption states that the models specified by the experimenter are correct, in the sense that there is no loss of information about the behavior of the potential outcome function y_i when we restrict our attention to the model space $\mathcal{M}_i$. As illustrated by the examples later in this section, this type of assumption is widely used in the causal inference literature. However, there is recent work that considers relaxations (e.g., Leung, 2022a; Sävje, 2021; Sävje et al., 2021). Assumption 1 is strong, and it is unlikely to hold exactly in most settings, but it can sometimes be seen as a useful approximation. We will maintain the assumption of correctly specified model spaces throughout this paper, and describe a relaxation in future work.

Experimenters are free to specify the model spaces however they wish. When the model spaces are finite-dimensional, we suspect experimenters often will find it convenient to specify them using a set of explicit functions. The model space is then the span of these functions. If the functions are linearly independent, they act as a basis for the model space. We provide concrete examples of model spaces created from basis functions in Section 3.6.

When working with model spaces created from explicit basis functions, some experimenters will find it convenient to define their estimands based on the coefficients for the basis functions corresponding to the true potential outcome function y_i, because this is reminiscent of how estimands often are defined in conventional linear regression studies. In Section S4.2 of the supplement, we show that any linear combination of the coefficients corresponding to y_i when written in a basis of the model space is an effect functional as defined in this section. This means this approach to defining estimands is accommodated by our framework. However, the disadvantage is that the estimand is well-defined only if the model space is correctly specified. The advantage of using an explicit effect functional is that we can separate the definition of the treatment effect from the assumptions we impose to estimate it.

The type of model encoded by the model spaces is not the same as a conventional statistical outcome model. A conventional statistical model in a causal inference setting is a parametrization of the conditional expectation function $\mathrm{E}[Y_i \mid Z_i, X_i]$ for some treatment variable Z_i and some vector of covariates X_i in a super-population, or a parametrization of the full conditional distribution. The models we consider here describe how the treatment variable relates to the outcomes for one particular unit; no restrictions are imposed on how the outcome relates to some possible set of covariates or on the heterogeneity in the potential outcomes between different units. That is, the model spaces limit how the interventions (e.g., the full treatment vector) affects the units, but they do not impose any regularity or homogeneity between units of those effects. To appreciate this difference, note that a d-dimensional conventional statistical model can be parametrized with d parameters, but if the model spaces $\mathcal{M}_i$ are all d-dimensional, the joint model is parametrized with dn parameters. In this sense, the framework we describe is nonparametric even when each individual model space is finite-dimensional.

3.5 Positivity

A model space does not ensure that we observe all aspects of the potential outcome function that are relevant for the estimand of interest. For example, if the experimental design is such that Z takes a certain value in $\mathcal{Z}$ with probability one, then we will only ever see one value of each potential outcome function, and will have no hope to estimate anything other than trivial estimands. The following assumption ensures that the design is such that all relevant aspects of the potential outcome functions are observable.

Assumption 2' (Simple positivity)**.** For all $i \in [n]$, if $u, v \in \mathcal{M}_i$ are equal almost surely on $\mathcal{Z}$, then the effect functional θ_i is the same on the two functions: $\theta_i(u) = \theta_i(v)$.

Simple positivity concerns the experimental design, the estimand, and the model spaces. It stipulates that the design provides information about all functions in the model spaces that are relevant for the estimand. Assumption 2' does not hold if there exists two functions $u, v \in \mathcal{M}_i$ such that $\theta_i(u) \neq \theta_i(v)$, but $u(Z) = v(Z)$ almost surely under the experimental design. By linearity of the effect functional, an equivalent definition of simple positivity is that all functions $u \in \mathcal{M}_i$ that are almost surely zero on $\mathcal{Z}$ must evaluate as zero on the effect functional: $\theta_i(u) = 0$.

A complication here is that the model space $\mathcal{M}_i$ might contain limit points, in a sense that will be made precise in Section 4.2, that are outside the model space, and positivity as defined in Assumption 2' does not pertain to such limit points. We address this complication by requiring that the effect functional is sufficiently well-behaved to ensure that observability holds also for these limit points, as formalized by the following assumption.

Assumption 2 (Positivity)**.** There exists some $C < \infty$ such that $|\theta_i(u)| \leq C\sqrt{\mathrm{E}[u(Z)^2]}$ for all $u \in \mathcal{M}_i$ and $i \in [n]$.

The positivity assumption stipulates that the effect functionals are bounded linear functionals. This can be interpreted as a slight strengthening of simple positivity; Assumption 2 implies Assumption 2', and if the model spaces are finite-dimensional, the two assumptions are equivalent. This is because all linear functionals are bounded on the non-zero functions in a finite-dimensional model space, so the only functions that must be regulated are those with $\mathrm{E}[(u(Z))^2] = 0$, which is what Assumption 2' does.

Assumption 2 is a direct generalization of conventional positivity assumptions made in the design-based causal inference literature. We will return to this point in the examples later in this section. Similar to the conventional positivity assumption, our two versions of the assumption are related to the existence of an unbiased estimator of an estimand, as described in the following proposition. The proof is provided in Section S1.1. The central idea behind the proposition is that if positivity does not hold, then it is possible to modify the potential outcome functions in a way that changes the estimand while holding the joint distribution of the observed outcomes constant.

Proposition 3.1. *Given correctly specified models (Assumption 1), simple positivity (Assumption 2') is a necessary condition for the existence of an unbiased estimator for the estimand $\tau = n^{-1}\sum_{i=1}^{n}\theta_i(y_i)$, and positivity (Assumption 2) is a sufficient condition.*

There is a gap between the necessary condition (Assumption 2') and the sufficient condition (Assumption 2) in Proposition 3.1, in the sense that it is unclear if an unbiased estimator exists if Assumption 2' holds but Assumption 2 does not. This gap is closed when the model spaces have finite dimensions, because the two assumptions then coincide, meaning that Assumption 2' is a necessary and sufficient condition for the existence of an unbiased estimator. We conjecture that Assumption 2 is necessary and sufficient for the existence of an unbiased estimator when the model spaces have infinite dimensions. However, we find no urgency in closing this gap, because estimation using infinite-dimensional model spaces will typically involve a bias–variance trade-off, and empirical researchers will typically opt for a biased estimator in this setting even if an unbiased estimator exists.

It is possible to investigate whether the positivity assumption holds before the experiment is run, because the assumption only depends on aspects of the experiment that are known to the experimenter. In Section S1.3 in the supplement, we describe a computationally efficient procedure for determining whether positivity holds when the model space has finite dimensions. If the experimenter aims to investigate several estimands in the same experiment, it could become tedious to check positivity with respect to each estimand separately. In Section S1.1 in the supplement, we describe a stronger notion of positivity. For finite-dimensional model spaces, the stronger notion of positivity can also be investigated before the experiment is run, and if it holds, positivity as defined in Assumption 2 is ensured for all possible estimands.

3.6 Examples of Model Spaces

3.6.1 Stable unit treatment value assumption

The stable unit treatment value assumption (SUTVA) states that there are no hidden versions of treatment and that a unit's outcome depends only on its own treatment assignment. In the case with binary treatments, $\mathcal{Z} = \{0,1\}^n$, the assumption implies that each unit effectively has only two potential outcomes. A way to express this is to write the potential outcome function as

$$y_i(\boldsymbol{z}) = \begin{cases} a_i & \text{if } z_i = 1, \\ c_i & \text{if } z_i = 0, \end{cases}$$

where, as above, we write the treatment variable as a vector $\boldsymbol{z} = (z_1, \dots, z_2)$ when considering binary treatments. Here, a_i is the outcome of unit i when assigned active treatment, and c_i is the outcome under the control treatment. The assumption allows us to interpret each potential outcome function as a mapping from $\{0, 1\}$, which is common in the literature, so we can write simply $y_i(1) = a_i$ and $y_i(0) = c_i$. However, to keep the notation consistent, we will always consider y_i as a mapping from $\mathcal{Z}$.

The model spaces corresponding to the stable unit treatment value assumption are finite-dimensional, so we can specify the model spaces using basis functions. A possible choice for these basis functions is to use the indicator functions $g_{i,1}(\boldsymbol{z}) = \mathbb{1}[z_i = 1]$ and $g_{i,2}(\boldsymbol{z}) = \mathbb{1}[z_i = 0]$, so that $g_{i,1}$ takes the value one if unit i is assigned active treatment, $Z_i = 1$, and $g_{i,0}$ takes the value one if it is assigned control, $Z_i = 0$. The corresponding model space is then

$$\mathcal{M}_i = \Big\{ u \in \mathcal{Y} : u = b_1 g_{i,1} + b_2 g_{i,2} \quad \text{for} \quad b_1, b_2 \in \mathbb{R} \Big\},$$

meaning that the true potential outcome function is $y_i = a_i g_{i,1} + c_i g_{i,2}$. The notation $u = \sum_{k=1}^{d} b_k g_{i,k}$ is meant to indicate that the function u is such that $u(z) = \sum_{k=1}^{m} b_k g_{i,k}(z)$ for all $z \in \mathcal{Z}$. Note that this potential outcome model does not impose any homogeneity of the potential outcome functions between the units; for example, a_i provides no information about a_j for $j \neq i$.

An alternative but equivalent choice for basis functions to capture the stable unit treatment value assumption is the constant function $g_{i,1}(\boldsymbol{z}) = 1$ and the unit's treatment variable $g_{i,2}(\boldsymbol{z}) = z_i$. The true potential outcome function is then given by $y_i = c_i g_{i,1} + (a_i - c_i) g_{i,2}$, which is the same function as above.

Because the basis functions $g_{i,1}(\boldsymbol{z}) = \mathbb{1}[z_i = 1]$ and $g_{i,2}(\boldsymbol{z}) = \mathbb{1}[z_i = 0]$ are disjoint, in the sense that at most one is non-zero for any treatment assignment, positivity for any effect functional is ensured by positivity for these basis functions alone. That is, Assumption 2 holds for any effect functional if $\Pr(g_{i,1}(\boldsymbol{Z}) = 1) > 0$ and $\Pr(g_{i,2}(\boldsymbol{Z}) = 1) > 0$. Note that this condition is the same as $0 < \Pr(Z_i = 1) < 1$, which is the conventional positivity assumption under the stable unit treatment value assumption.

3.6.2 Exposure mappings

In their study of causal inference using exposure mappings, Aronow and Samii (2017) assume the exposures are correctly specified. Using the unit-specific exposure mapping

$d_i : \mathcal{Z} \to \Delta$, the assumption states that as long as two treatment assignments $z, z' \in \mathcal{Z}$ map to the same exposure in Δ, the outcome for the corresponding unit is the same:

$$y_i(z) = y_i(z') \text{ for all } z, z' \in \mathcal{Z} \text{ such that } d_i(z) = d_i(z').$$

We can express this assumption using model spaces. Let $m = |\Delta|$ denote the number of exposures, and index the exposures in Δ with the integers $[m] = \{1, \dots, m\}$. Define m basis functions for $k \in [m]$ as $g_{i,k}(z) = \mathbb{1}[d_i(z) = k]$. The model space corresponding to the correctly specified exposure assumption is then given by

$$\mathcal{M}_i = \Big\{ u \in \mathcal{Y} : u = \textstyle\sum_{k=1}^{m} b_k g_{i,k} \quad \text{for} \quad b_1, \dots, b_m \in \mathbb{R} \Big\}. \tag{1}$$

Similar to the stable unit treatment value assumption, this model does not impose any homogeneity between units nor any relationship between the coefficients of different units corresponding to the true potential outcome functions when written in the bases of the model spaces.

Because the basis functions are disjoint also in this setting, positivity for any effect functional is ensured if $\Pr(d_i(Z) = k) > 0$ holds for all $i \in [n]$ and $k \in [m]$. However, unlike under the stable unit treatment value assumption, this might not be a necessary for positivity to hold for practically relevant estimands. If the number of exposures is three or more, $m \geq 3$, and the estimand only involve two of the exposures, say $a, b \in [m]$, then positivity only requires positivity with respect to those exposures. That is, we only need $\Pr(d_i(Z) = a) > 0$ and $\Pr(d_i(Z) = b) > 0$. If we have $\Pr(d_i(Z) = c) = 0$ for some $c \in [m] \setminus \{a, b\}$, the design provides no information about dimension c of the model space, but that is unproblematic as long as the effect functional does not depend on that dimension.

Auerbach and Tabord-Meehan (2021) use an approach that can be interpreted as a generalization of the exposure mapping approach, in which the exposures are local configurations of a rooted network. They consider estimation using a network sieve based on a distance pseudo-metric on these rooted networks.

3.6.3 Linear-in-means models

Exposure mappings can be used to model interference that takes place in a network, but it can be difficult or impossible to model more intricate structure because the exposures must be discrete. As an alternative to the exposure mapping approach, consider a design-based version of the linear-in-means model, which is used extensively in applied work (see, e.g., Cai, Janvry, & Sadoulet, 2015; Dupas, 2014; Oster & Thornton, 2012). Methodological investigations of the model are provided by Chin (2019), Leung (2020) and Hu et al. (2022), among others.

In this model, units are presumed to be affected by the share of treated units among their neighbors in a network. Let $\mathcal{N}_i$ be the indices of the units that are adjacent to unit i in the graph describing the network. The linear-in-means model then stipulates that

$$y_i(\boldsymbol{z}) = \beta_{i,1} + \beta_{i,2} z_i + \beta_{i,3} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} z_j,$$

where $|\mathcal{N}_i|$ is the number of units adjacent to unit i. That is, the potential outcome function depends on the unit's own treatment assignment and the share of treated units in its neighborhood.

It is common to add a random error term that is assumed to be mean zero and independent between units to capture heterogeneity between units in their responses. There is no need to add such an error term in this case, because no homogeneity is enforced between the units by the model spaces. That is, the potential outcome function for one unit is completely uninformative of the potential outcome functions of other units, as captured by the fact that the coefficients $\beta_{i,k}$ are indexed by i, so we already accommodate full unit heterogeneity. Empirical researchers sometimes add error terms for other reasons, for example to capture measurement errors. One can interpret the current setting as conditioning on those measurement errors. Alternatively, one could extend the framework by including an explicit measurement error, which means that the extended framework would blend components from the design-based and super-population approaches. We will not investigate such an extension in this paper.

It is not possible to express the linear-in-means model using exposure mappings. We could have an exposure for each possible level of $(z_i, \sum_{j \in \mathcal{N}_i} z_j)$, resulting in a model with $2|\mathcal{N}_i| + 2$ dimensions for unit i. This model is, however, much larger than the three-dimensional model above. More importantly, the exposure mapping model fails to encode the presumed linear relationship between a unit's outcome and the number of treated units in its neighborhood.

We can, however, express the linear-in-means model using model spaces. Let $g_{i,1}(\boldsymbol{z}) = 1$ be a constant function, $g_{i,2}(\boldsymbol{z}) = z_i$ be the unit's own treatment indicator, and $g_{i,3}(\boldsymbol{z}) = |\mathcal{N}_i|^{-1} \sum_{j \in \mathcal{N}_i} z_j$ be the number of treated units in unit i's neighborhood. The model space corresponding to the network interference model is then given by Equation (1) from the previous subsection using these functions as the basis functions. Of course, it is possible to make the model more elaborate by adding additional basis functions.

Positivity is somewhat more intricate to check in this setting, because the basis functions are neither disjoint nor orthogonal. To ensure positivity for all effect functionals, we need that $\mathrm{E}[g_{i,k}(\boldsymbol{Z})^2] > 0$ for all $k \in \{1, 2, 3\}$, and that $\mathrm{E}[g_{i,k}(\boldsymbol{Z}) g_{i,\ell}(\boldsymbol{Z})]$ is not equal to $\sqrt{\mathrm{E}[g_{i,k}(\boldsymbol{Z})^2] \, \mathrm{E}[g_{i,\ell}(\boldsymbol{Z})^2]}$ for all $k \neq \ell$. Essentially, this means that the design induces variation in all basis functions and that no two basis functions are collinear. We might be able to relax some of these conditions for certain effect functionals if they do not depend on all aspects of the potential outcome functions.

The linear-in-means model can be too restrictive to be tenable in practice, but the linearity assumption can be relaxed. For example, Munro, Wager, and Xu (2021) investigate a setting where the interference is known to be mediated through a set of market prices, but they impose essentially no restrictions on how the prices affect the units. This can be seen as a greatly generalized version of the linear-in-means model. The market price here plays the same role as the share of treated neighbors in linear-in-means model, corresponding to $g_{i,3}(\boldsymbol{z})$, but linearity is not imposed, so the model space would consist be various, possibly infinite, non-linear transformations of the market prices. Munro et al. (2021) consider when market price is convergent, meaning that $g_{i,3}(\boldsymbol{Z}) \xrightarrow{p} a \in \mathbb{R}$. This facilitates estimation of direct treatment effects, as a unit's own treatment assignment then only has a negligible on the market price. However, because the prices are convergent, and thus exhibit negligible variation in large samples, positivity might not hold for functionals capturing effects of changes in the prices. Munro et al. (2021) address this by considering a design that introduces unit-level perturbations of the prices.

3.6.4 Spatial potential outcome models

In the setting of spatial causal inference for a finite number of candidate treatment location, Pollmann (2020) considers several potential outcome models. One model stipulates that treatment locations close to each unit have additive effects on the unit. Another model stipulates that only the treated location closest to each unit has an effect on the unit. The treatments assigned to locations very far from each unit, as dictated by some threshold distance, are assumed to never have an effect.

To create model spaces for these models, let $\mathcal{N}_i$ be the indices of the candidate treatment locations that are sufficiently close to unit i to potentially affect its outcome, meaning that they are not beyond the threshold distance. The basis functions for unit i under the assumption of additive effects would then be $g_{i,k}(\boldsymbol{z}) = z_{J_i(k)}$ for all $k \in \{1, \dots, |\mathcal{N}_i|\}$, where $J_i(k)$ is an indexing of the elements in $\mathcal{N}_i$. We typically also want to include a constant function, $g_{i,0}(\boldsymbol{z}) = 1$, among the basis functions, as otherwise the unit cannot have a non-zero outcome unless there is a non-zero treatment effect.

For the assumption that only the closest treatment location matters, we again would have a basis function for each element in $\mathcal{N}_i$, plus a constant function. However, the function $g_{i,k}(\boldsymbol{z})$ would now be such that it takes the value one only when location $J_i(k) \in \mathcal{N}_i$ is treated and no location closer to i is treated. For example, if $J_i(1)$ is the closest location to i and $J_i(2)$ is the next closest, then $g_{i,2}(\boldsymbol{z}) = (1 - z_{J_i(1)}) z_{J_i(2)}$.

3.6.5 Treatment timing experiments

Consider a setting in which we want to investigate the effects of the timing of one particular intervention, rather than the effects of a set of several different interventions implemented at one particular time point. To make things concrete, consider an agricultural experiment in which we want to investigate when it is best to apply fertilizer for a particular type of crop. The units $i \in [n]$ are here plots of land used to grow the crop of interest, and the intervention is the time at which the fertilizer was applied to a particular plot. Let $z_i \in [-1, 1]$ denote the time at which plot i was fertilized, where $z_i = 0$ is the current recommended fertilization time, and $z_i = -1$ and $z_i = 1$ are the earliest and latest time points under consideration. For example, if $z_i = 0$ is May 23, then $z_i = -0.7$ might be May 4. The full intervention set is $\mathcal{Z} = [-1, 1]^n$, collecting all possible combinations of fertilizer timing for the n plots. The potential outcome functions $y_i(\boldsymbol{z})$ give the crop yield from plot i when the fertilizer timing of all plots is $\boldsymbol{z}$.

To investigate whether the crop yields are greater when the fertilizer is applied earlier or later than the current recommendation, we set out to estimate the sum of the gradients of the potential outcome functions at $\boldsymbol{z} = \mathbf{0}$. An effect functional that captures this effect is $\theta(y) = \mathbf{1}^{\mathsf{T}} \nabla y(\mathbf{0})$, where $\mathbf{1}$ and $\mathbf{0}$ denote the vectors of ones and zeros, and ∇ is the gradient operator. This is similar to the effect defined in Equation 2.2 for the cash transfer illustration in Section 2.2.

The question remains what model spaces to use. In the experiment, the plots are sufficiently separated geographically, so we are comfortable ruling out interference between plots, meaning that the fertilizer timing on plot i has no effect on the yield on plot j. This means that the potential outcome functions effectively map from $[-1, 1]$ to $\mathbb{R}$. It is reasonable to assume that the potential outcome functions are smooth, in the sense that small changes in fertilizer timing will lead to small changes in the crop yield, which also ensure that the effect functional is well-defined for all potential outcome functions. This yields the model spaces

$$\mathcal{M}_i = \big\{ \ y \in \mathcal{Y} : \quad y \text{ is continuously differentiable} \quad \text{ and } \quad y(\boldsymbol{z}) = y(\boldsymbol{z}') \ \text{ when } \ z_i = z_i' \ \big\}.$$

Note that this is a very large model space, and it might not be possible to construct a stable estimator with a model space that is this permissive. We will return to this point in Section 4.5.3, where we discuss estimation using this model space.

4 The Riesz Estimator

4.1 Overview and Definition

There are two components of the framework at this point. The first is an effect functional for each unit, θ_i, which defines the estimand: $\tau = n^{-1} \sum_{i=1}^{n} \theta_i(y_i)$. The second is a model space for each unit, $\mathcal{M}_i$, which encodes prior knowledge or beliefs about each unit's potential outcome function. The construction of the effect functionals and the model spaces could be complex, but the framework allows us to abstract away from this complexity, which is one of its benefits.

Being able to abstract away from the complexity allows us to take a unified approach to constructing estimators for estimands defined using the framework. The approach we describe in this section can be applied to any estimand as long as positivity holds. This is in contrast to the current convention in the design-based causal inference literature, where estimators often are constructed in an ad hoc way that is tailored uniquely to the estimand and model assumptions at hand. However, some estimators previously considered in the design-based literature are special cases of the estimators we describe here.

We give an overview of our estimation approach and the definition of the estimator in this subsection, postponing the discussion of some of the technical details of the definition to the next subsection. The construction of the estimator uses the Riesz representation theorem from functional analysis. The theorem states that any bounded linear functional defined on a function space can be represented by a member of the function space itself, which we call the *Riesz representor* and denote ψ_i. The Riesz representor represents the linear functional in the sense that an inner product of any function in the function space and representor equals the functional evaluated at that function.

In our setting, the relevant function spaces are the model spaces. The model spaces contain functions that map from the interventions in $\mathcal{Z}$, so the Riesz representors can be interpreted both as functions and as random variables. That is, in addition to interpreting $\psi_i \in \mathcal{M}_i$ as a function, we can also interpret it as the random variable $\psi_i(Z)$. The potential outcome functions y_i can be interpreted in the same way, yielding the random variables $Y_i = y_i(Z)$. We seek a Riesz representor that will represent the effect functional on the model space with respect to the expectation of the product of the two corresponding random variables.

Definition 1. A function $\psi_i \in \mathcal{M}_i$ is a *Riesz representor* for an effect functional θ_i on the model space $\mathcal{M}_i$ if $\theta_i(u) = \mathrm{E}[\psi_i(Z)u(Z)]$ for all $u \in \mathcal{M}_i$.

If the model space has infinite dimensions and is not closed, in a sense that will be made precise in the next subsection, then it is possible that the Riesz representor is not in the model space: $\psi_i \notin \mathcal{M}_i$. However, the Riesz representor is then on the boundary of the model space. To make Definition 1 fully rigorous, we need to consider a type of closure of

$\mathcal{M}_i$ discussed in the next subsection. For the purpose of building an initial understanding of the Riesz estimator, concerns about whether $\mathcal{M}_i$ is closed are not important.

The usefulness of the representation described in Definition 1 is that it essentially gives us a direct observation of the effect functional. The value of the effect functional when evaluated at the true potential outcome function $\theta_i(y_i)$ is unattainable because our observation of the potential outcome function $Y_i = y_i(Z)$ provides only a very limited picture of the whole function y_i. Generally, it is unclear how to relate Y_i to the individual treatment effect $\theta_i(y_i)$, which is the reason for the conventional ad hoc approach to estimation in this setting. The usefulness of the Riesz representor is that it tells us exactly how to relate Y_i to $\theta_i(y_i)$.

If the Riesz representor ψ_i exists and is known, we observe $\psi_i(Z)Y_i$. By construction, $\psi_i(Z)Y_i$ is a direct observation of a random variable whose expectation is the individual treatment effect, because $\mathrm{E}[\psi_i(Z)y_i(Z)] = \theta_i(y_i)$. That is, the Riesz representor provides a way to relate our observation of Y_i to the individual treatment effect. That is, we attain a direct but noisy observation of the effect functional evaluated at the true potential outcome function. Note, however, that this is true only when the model space is correctly specified, meaning that $y_i \in \mathcal{M}_i$, because Riesz representor is only ensured to represent the functional θ_i on this space.

Definition 2 (The Riesz estimator)**.** Let ψ_i be a Riesz representor of θ_i on $\mathcal{M}_i$ for each $i \in [n]$. The *Riesz estimator* of $\tau = n^{-1}\sum_{i=1}^{n} \theta_i(y_i)$ is

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \psi_i(Z)Y_i.$$

The Riesz estimator is a linear estimator in the sense that it is a (random) linear combination of the observed outcomes. Following the logic discussed earlier in this section, one can interpret the terms of the estimator $\psi_i(Z)Y_i$ as estimators of the individual treatment effects $\theta_i(y_i)$. Each of these unit-level estimators will be terribly imprecise, but they are unbiased by construction, so precision could potentially be achieved by averaging. The idea is similar to other design-based, linear estimators based on re-weighting; indeed, as we show in Section 4.5.1, the Riesz estimator is a generalization of the Horvitz–Thompson estimator. However, whereas the Horvitz–Thompson estimator requires a small number of discrete treatments, the Riesz estimator does not have any such limitations.

One can also interpret the Riesz estimator as a type of plug-in estimator. With this interpretation, we first construct an estimator $\widehat{y}_i$ of the whole potential outcome function y_i. An estimator of the unit-level treatment effect $\tau_i = \theta_i(y_i)$ is then constructed by plugging the estimated function into the effect functional: $\widehat{\tau}_i = \theta_i(\widehat{y}_i)$. These unit-level estimators are averaged to form an estimator of the aggregated treatment effect τ. For a particular choice of function estimators $\widehat{y}_i$, this approach coincides exactly with the Riesz estimator. We prefer the Riesz representation interpretation, as it is more general and less notationally involved, but some readers might prefer the plug-in interpretation as it may be more familiar. We describe this alternative interpretation in Section S4.1 of the supplement.

4.2 Existence and Unbiasedness

Showing that the Riesz representor exists in our setting is somewhat intricate. The challenge is that the experimental design might not provide information about the full model space, in the sense that the design might not induce variability in some dimensions of the space. This is the same concern discussed in Section 3.5, where we introduced our version of the positivity assumption. On a technical level, the problem is that the Riesz representation theorem concerns Hilbert spaces, but the model spaces as currently defined do not necessarily have associated inner products that are appropriate for our estimation problem.

To navigate this problem, we will consider equivalence classes of potential outcome functions that are observationally indistinguishable. We say that two functions $u, v \in \mathcal{Y}$ are *observationally equivalent* if $u(Z) = v(Z)$ almost surely with respect to the experimental design. If two distinct functions u and v are observationally equivalent, then the probability that the design produces an assignment that allows us to distinguish them is zero. We can use this equivalence relation to partition the space of all potential outcome functions $\mathcal{Y}$ into equivalence classes. We denote the equivalence class that contains all functions that are observationally equivalent to some function $u \in \mathcal{Y}$ as $[u]$. We call the quotient space that collects these equivalence classes the *unrestricted outcome space*, and it is denoted $\mathcal{L}(\mathcal{Y})$. Note that the unrestricted outcome space is itself a vector space built from the vector space $\mathcal{Y}$.

By viewing each equivalence class $[u] \in \mathcal{L}(\mathcal{Y})$ as being associated with a random variable $u(Z)$, we can interpret $\mathcal{L}(\mathcal{Y})$ as the collection of all distinct random variables that can be constructed from the functions in $\mathcal{Y}$. The random variables that are constructed in this way from functions in the same equivalence class are technically different random variables, but they are equal almost surely. The reason $\mathcal{L}(\mathcal{Y})$ is called an outcome space is because the random variables that represent the units' observed outcomes $Y_i = y_i(Z)$ can be seen as elements of this set.

We endow the unrestricted outcome space $\mathcal{L}(\mathcal{Y})$ with an inner product that is the expectation of the product of the two random variables associated with the corresponding two equivalence classes. That is, for equivalence classes $[u], [v] \in \mathcal{L}(\mathcal{Y})$, we define the inner product as $\langle [u], [v] \rangle = \mathrm{E}[u(Z)v(Z)]$. It follows that $\mathcal{L}(\mathcal{Y})$ together with this inner product is a Hilbert space, as the completeness property is satisfied. Indeed, we have just used the standard approach to constructing the L^2 function space on $\mathcal{Z}$ induced by the experimental design (see, e.g., Chapter 3 of Rudin, 1987).

We construct similar quotient spaces for each model space $\mathcal{M}_i$. We denote these spaces $\mathcal{L}(\mathcal{M}_i)$ and refer to them as *model outcome spaces*, or just outcome spaces for short. The outcome space for unit i contains the equivalence classes of all functions that are observationally equivalent to functions in $\mathcal{M}_i$. The space is formally defined as the closure of the subspace of all equivalence classes in $\mathcal{L}(\mathcal{Y})$ that contain at least one function from $\mathcal{M}_i$:

$$\mathcal{L}(\mathcal{M}_i) = \operatorname{cl}\big(\big\{[u] \in \mathcal{L}(\mathcal{Y}) : u \in \mathcal{M}_i\big\}\big),$$

where the closure is taken with respect to the metric induced by $\langle \cdot, \cdot \rangle$ on $\mathcal{L}(\mathcal{Y})$. Taking the closure ensures that $\mathcal{L}(\mathcal{M}_i)$ contains all its limit points. When the model space is finite-dimensional, the set $\{[u] \in \mathcal{L}(\mathcal{Y}) : u \in \mathcal{M}_i\}$ is a finite-dimensional subspace of $\mathcal{L}(\mathcal{Y})$ and therefore already closed, meaning that taking the closure is redundant. However, taking the closure is important when the model space is infinite-dimensional.

An equivalence class $[u] \in \mathcal{L}(\mathcal{M}_i)$ for some potential outcome function $u \in \mathcal{M}_i$ contains all potential outcome functions that are observationally indistinguishable from u under the current experimental design. In this sense, the most we can possibly learn about the true potential outcome function y_i from the experiment is which equivalence class it belongs to. As with the unrestricted outcome space $\mathcal{L}(\mathcal{Y})$, we can view each equivalence class $[u] \in \mathcal{L}(\mathcal{M}_i)$ as being associated with a random variable $u(Z)$, so $\mathcal{L}(\mathcal{M}_i)$ can be interpreted as the collection of all distinct random variables that can be constructed from the functions in the model space $\mathcal{M}_i$.

Proposition 4.1. *Together with the inner product $\langle [u], [v] \rangle = \operatorname{E}[u(Z)v(Z)]$, each model outcome space $\mathcal{L}(\mathcal{M}_i)$ is a Hilbert space.*

The proposition shows that the Riesz representation theorem is applicable to $\mathcal{L}(\mathcal{M}_i)$, which we use to construct Riesz representors on the model spaces $\mathcal{M}_i$. Our approach is to first construct a Riesz representor $\psi_i^\dagger$ on $\mathcal{L}(\mathcal{M}_i)$ for an extension of the effect functional θ_i to the outcome space, which we denote $\theta_i^\dagger$. We then use the correspondence between $\mathcal{L}(\mathcal{M}_i)$ and $\mathcal{M}_i$ to translate $\psi_i^\dagger$ to a representor ψ_i on the model space.

The central insight behind the extension of the effect functional is that positivity (Assumption 2) implies that the equivalence classes in the outcome space $\mathcal{L}(\mathcal{M}_i)$ contain the same information about the effect functional θ_i as the functions in the model space $\mathcal{M}_i$. Given positivity, $\theta_i(u) = \theta_i(v)$ for any two observationally equivalent functions $u, v \in \mathcal{M}_i$, so we only need to know which equivalence class the true potential outcome function y_i belongs to in order to know $\theta_i(y_i)$. This suggests using the definition $\theta_i^\dagger(E) = \theta_i(u)$ where $u \in E \cap \mathcal{M}_i$ for all equivalence classes $E \in \mathcal{L}(\mathcal{M}_i)$ that contain at least one function u from the model space. A complication here is that $\mathcal{L}(\mathcal{M}_i)$ could contain limit points that are completely outside the model space $\mathcal{M}_i$. Assumption 2 together with Hahn–Banach theorem allow us to extend the definition of $\theta_i^\dagger$ to also these limit points. This culminates in the following proposition. The proof of the proposition appear in Section S1.2 of the supplement.

Proposition 4.2. *Given positivity (Assumption 2), there exists an equivalence class $\psi_i^\dagger \in \mathcal{L}(\mathcal{M}_i)$ such that every function $\psi_i \in \psi_i^\dagger$ satisfies*

$$\theta_i(u) = \mathrm{E}[\psi_i(Z)u(Z)] \quad \textit{for all} \quad u \in \mathcal{M}_i.$$

We say that any function $\psi_i \in \psi_i^\dagger$ is a Riesz representor for the effect functional θ_i on the model space $\mathcal{M}_i$.

Proposition 4.2 simultaneously defines the Riesz representor rigorously and proves their existence. A consequence of the proposition is that the Riesz representors when seen as functions are not necessarily unique, because the equivalence class $\psi_i^\dagger$ could contain many functions. However, all Riesz representors in $\psi_i^\dagger$ are almost surely equal by construction. Therefore, the Riesz representor is unique, in an almost surely sense, when seen as a random variable.[1]

Proposition 4.2 allows us to better understand the caveat of Definition 1 discussed directly following the definition in the previous subsection. If the equivalence class $\psi_i^\dagger$ contains at least one function from $\mathcal{M}_i$, we can pick the Riesz representor be one of those functions, ensuring that the representor is in the model space as stated in the definition. However, if the equivalence class $\psi_i^\dagger$ is a limit point, it may not contain any functions from $\mathcal{M}_i$, in which case the Riesz representor function will not be in the model space. The Riesz representor can then be expressed as an infinite series of functions in the model space, meaning that the representor is on the boundary of the model space. If the model space has finite dimensions, there always exists a Riesz representor that is in the model space.

The existence of Riesz representors ensures the existence of the Riesz estimator, giving us the following theorem.

Theorem 4.3. *Given correctly specified model spaces and positivity (Assumptions 1 and 2), the Riesz estimator exists and is unbiased:* $\mathrm{E}[\widehat{\tau}] = \tau$.

Proof. Positivity ensures the existence of the Riesz representors and thus the Riesz estimator. By linearity of expectation, the definition of Riesz representors and correctly specified model spaces,

$$\mathrm{E}[\widehat{\tau}] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\psi_i(Z)y_i(Z)] = \frac{1}{n}\sum_{i=1}^{n} \theta_i(y_i) = \tau. \qquad \square$$

[1] There are functions outside $\mathcal{L}(\mathcal{M}_i)$ that also represent of the effect functionals on the model space, but that are not almost surely equal to the Riesz representor. We conjecture that estimators built on these alternative representors generally will have higher variance, but we will not explore that in this paper.

Unbiasedness is a direct consequence of the construction of the Riesz representors. While the theorem holds for any model space, the Riesz estimator might perform poorly when the model spaces have high or infinite dimensions. The reason is a bias–variance trade-off similar to the conventional bias–variance trade-off in nonparametric statistics. To achieve unbiasedness when the model spaces are complex, the estimator must be complex unless the estimand is sufficiently simple, so we must accept a high variance. The Riesz estimator as defined here achieves unbiasedness, so it will often be imprecise, especially when the dimensions of the model spaces are high. If the model spaces and effect functionals are sufficiently complex, the Riesz estimator will not be useful in practice because it would require unreasonably large samples to achieve acceptable precision. We sketch a possible solution to this problem at the end of the next subsection, but we leave a proper investigation of how to navigate this bias–variance trade-off for future work.

4.3 Constructing the Riesz Representors

The Riesz representors depend solely on the model spaces, the effect functionals, and the experimental design, all of which are known to the experimenter. This means that the Riesz estimator in principle can be constructed. However, constructing and evaluating the Riesz representors can be non-trivial computational tasks in practice. To facilitate the construction of the Riesz estimator, we require that the model spaces are of a form that make it tractable to work with them, as captured by the following assumption.

Assumption 3. The outcome spaces $\mathcal{L}(\mathcal{M}_i)$ are separable.

Separability means that there exists a dense subset of $\mathcal{L}(\mathcal{M}_i)$ that is countable. This implies that $\mathcal{L}(\mathcal{M}_i)$ has a countable orthonormal basis, so $\mathcal{L}(\mathcal{M}_i)$ has at most countable dimensions. If a model space has finite dimensions, $\mathcal{L}(\mathcal{M}_i)$ is always separable. We believe Assumption 3 will be unproblematic in essentially all practical applications, and it can be seen as a regularity condition.

Provided that $\mathcal{L}(\mathcal{M}_i)$ is separable, all Riesz representors of θ_i on $\mathcal{M}_i$ satisfy

$$\psi_i(Z) = \sum_{k=1}^{\infty} \theta_i(\phi_{i,k})\phi_{i,k}(Z) \quad \text{almost surely,} \tag{2}$$

where $[\phi_{i,1}], [\phi_{i,2}], \ldots$ is an orthonormal basis of $\mathcal{L}(\mathcal{M}_i)$. The construction of the Riesz estimator therefore boils down to whether we can construct a set of orthonormal basis functions such that the basis functions can be evaluated on the realized treatment variable, $\phi_{i,k}(Z)$, and the effect functional can be evaluated on the basis functions, $\theta_i(\phi_{i,k})$.

When the model space has finite dimensions, the number of basis functions is also finite, meaning that there are only finite terms in the sum in Equation (2). If the model space is specified using a set of explicit basis functions $g_{i,1}, \dots, g_{i,d}$, as is the case in many of the examples in this paper, an orthonormal basis can be constructed via a Gram–Schmidt orthogonalization procedure. The orthogonalization procedure requires that the experimenter knows or can compute the cross-moments $\mathrm{E}[g_{i,\ell}(Z)g_{i,k}(Z)]$, which can be done using the Monte Carlo method. However, in what follows, we assume that the orthonormal basis is known exactly, without any Monte Carlo error. In Section S1.3, we provide a description this procedure, and investigate its computational properties. If the cross-moments are known and all model spaces have d dimensions, the Riesz estimator can be computed using $\mathcal{O}(nd^3)$ arithmetic operations. Subsequent evaluations of the same estimator require only $\mathcal{O}(nd)$ arithmetic operations per evaluation.

When the model space has infinite dimensions, it will generally not be feasible to compute the Riesz estimator exactly. If the experimenter has access to a sequence of orthonormal basis functions, they can compute the Riesz estimator approximately by using the same procedure as for the finite-dimensional setting with a truncated sequence. For an appropriate basis and sufficiently generous truncation, the approximation error will be negligible for practical purposes. However, using a generous truncation will not resolve the bias–variance trade-off discussed at the end of the previous subsection. A way to navigate this trade-off is to carefully pick the sequence of basis functions so that the product of $\theta_i(\phi_{i,k})$ and $\mathrm{E}[\phi_{i,k}(Z)Y_i]$ quickly approaches zero, and then more aggressively truncate the sequence. This effectively creates a sieve version of the Riesz estimator. We save the investigation of the Riesz sieve estimator for future work.

4.4 Previous Uses of Riesz Representors

The Riesz representation theorem was developed in the early twentieth century independently by Fréchet (1907) and Riesz (1907). The theorem has played an important role in the semiparametric causal inference literature over the past 30 years. To the best of our knowledge, its earliest use was by Robins, Rotnitzky, and Zhao (1994) and independently by Newey (1994). Robins et al. (1994) developed doubly robust estimators using general representation theorems for efficient scores and influence functions. Newey (1994) described a general approach for estimating the asymptotic variance of semiparametric estimators, which features the L^2 inner product between the regression residual and the Riesz representor of the functional derivative. A common theme in this and following work is that a one-step bias correction, made possible through careful consideration of the influence function and Riesz representors, can often achieve the semiparametric efficiency bound.

The Riesz representation theorem has taken an increasingly prominent position in some recent work in the semiparametric literature. The focus has been to extend key insights from the existing literature to high-dimensional settings using machine learning methods (see, e.g., Robins, Li, Mukherjee, Tchetgen Tchetgen, & van der Vaart, 2017, and Chernozhukov et al., 2018). Part of this focus includes the development of automatic methods for Riesz representor estimation through various regularization schemes for a variety of types of local and global estimands (Chernozhukov, Escanciano, Ichimura, Newey, & Robins, 2022; Chernozhukov, Newey, & Singh, 2022a, 2022b; Hirshberg & Wager, 2021). The insights have also between extended to dynamic settings (Lewis & Syrgkanis, 2021) and the estimation of long term effects (Singh, 2021). Relatedly, Rotnitzky et al. (2020) describe the construction of doubly robust estimators in a semiparametric settings where a mixed-bias property holds. Finally, Riesz representors have been used to analyze sieve estimator in a time series setting and to analyze nonparametric instrumental variables estimation (Chen & Christensen, 2018; Chen, Liao, & Sun, 2014). To the best of our knowledge, the Riesz representation theorem has not previously been used in the design-based causal inference literature.

There are several similarities between how the Riesz representation theorem is used in the current paper and how it is used in the semiparametric literature. In both cases, the theorem is used to construct estimators with desirable statistical properties. The broad outline of the application of theorem is also the same: the underlying Hilbert space is an L^2 space in which the measure is a probability measure that captures the stochasticity from the data generating process. We find these connections exciting and hope that the current paper can act as a starting point for building bridges between design-based and super-population causal inference.

However, there are several important differences in our use of the Riesz representation theorem and its use in the semiparametric literature. In the semiparametric literature, parameters of interest are defined as functionals of a conditional expectation function of some outcome given treatment and covariates in a super-population. The source of stochasticity is sampling from the super-population. As a result, there is a single Riesz representor defined in the super-population, representing the functional with respect to the population distribution. This means that the Riesz representor must be estimated in the semiparametric literature, as the population distribution is unknown. In the current paper, parameters of interest are defined as functionals of individual potential outcome functions, thereby circumventing the need to consider super-population distributions. Furthermore, the source of stochasticity is random treatment assignment in a finite population. As a result, there is one Riesz representor for each unit, providing a noisy observation of the treatment effect of that specific unit, as described in Section 4.1. Because the Riesz representors in our setting are defined in the finite population with respect to a known experimental design, they can be computed, making estimation of the representors redundant.

Ultimately, the difference boils down to the fact that we are considering an experimental setting, with a finite population and known experimental design, whereas the semiparametric literature considers an observational setting, with a super-population and an unknown assignment mechanism that generally is assumed to be unconfounded.

4.5 Examples of Riesz Estimators

4.5.1 Average Treatment Effects

Our first example of a concrete Riesz estimator is for estimation of the average treatment effect under the stable unit treatment value assumption. As discussed in Section 3.6.1, a natural choice of basis functions for this model is $g_{i,1}(\boldsymbol{z}) = \mathbb{1}[z_i = 1]$ and $g_{i,2}(\boldsymbol{z}) = \mathbb{1}[z_i = 0]$, meaning that

$$\mathcal{M}_i = \Big\{ u \in \mathcal{Y} : u = b_1 g_{i,1} + b_2 g_{i,2} \quad \text{for} \quad b_1, b_2 \in \mathbb{R} \Big\}.$$

There are many functionals that capture the average treatment effect under this model. A common choice is the effect functional corresponding to the global average treatment effect: $\theta(y) = y(\mathbf{1}) - y(\mathbf{0})$. While we use this functional in this example for concreteness, the value of the estimand and the Riesz estimator are the same for any functional that captures the average treatment effect under correctly specified model spaces.

Using the approach described in the supplement, we can derive the Riesz representor by hand in only a few steps. The Riesz representor for unit i in this setting is

$$\psi_i(\boldsymbol{z}) = \frac{\mathbb{1}[z_i = 1]}{\Pr(Z_i = 1)} - \frac{\mathbb{1}[z_i = 0]}{\Pr(Z_i = 0)},$$

yielding the following Riesz estimator of the average treatment effect:

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\boldsymbol{Z}) Y_i = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}[Z_i = 1]}{\Pr(Z_i = 1)} Y_i - \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}[Z_i = 0]}{\Pr(Z_i = 0)} Y_i.$$

This is the conventional Horvitz–Thompson estimator for the average treatment effect under the stable unit treatment value assumption (Aronow & Middleton, 2013; Horvitz & Thompson, 1952; Narain, 1951).

The derivation of the Riesz estimator for exposure effects under the assumption of correctly specified exposure mappings, as discussed in Section 3.6.2, is almost identical. The resulting estimator is the Horvitz–Thompson-type estimator described by Aronow and Samii (2017). In particular, when estimating the exposure effect for exposures $a, b \in \Delta$ and given the exposure mapping $d_i : \mathcal{Z} \to \Delta$, the Riesz representor for unit i is

$$\psi_i(z) = \frac{\mathbb{1}[d_i(z) = a]}{\Pr(d_i(Z) = a)} - \frac{\mathbb{1}[d_i(z) = b]}{\Pr(d_i(Z) = b)}.$$

4.5.2 Network Effects

The next example considers the type of linear-in-means interference models discussed in Section 3.6.3. This model specified that the potential outcome functions were such that

$$y_i(\boldsymbol{z}) = \beta_{i,1} + \beta_{i,2} z_i + \beta_{i,3} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} z_j,$$

where $\mathcal{N}_i$ was the set of indices of the units adjacent to unit i in the graph describing the network. The natural basis functions for this model are

$$g_{i,1}(\boldsymbol{z}) = 1, \qquad g_{i,2}(\boldsymbol{z}) = z_i, \qquad \text{and} \qquad g_{i,3}(\boldsymbol{z}) = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} z_j.$$

The estimands we consider here are the global average treatment effect, $\theta_i^{\text{G}}(y) = y(\mathbf{1}) - y(\mathbf{0})$, and an indirect interference effect, $\theta_i^{\text{IN}}(y) = y(\mathbf{1} - \boldsymbol{e}_i) - y(\mathbf{0})$, where $\boldsymbol{e}_i$ is the ith standard basis function.

We need to know the experimental design to construct the Riesz representors in this setting. Consider a Bernoulli design, meaning that the treatments are assigned independently with some probability $\Pr(Z_i = 1) = p$, so that

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \prod_{i=1}^{n} p^{z_i} (1-p)^{1-z_i}.$$

Also in this setting can we derive the Riesz representors by hand. The Riesz representor for unit i and the indirect interference effect is

$$\psi_i^{\text{IN}}(\boldsymbol{z}) = \frac{|\mathcal{N}_i|^{-1} \sum_{j \in \mathcal{N}_i} (z_j - p)}{p(1-p)},$$

meaning that the Riesz estimator for the indirect interference effect is

$$\widehat{\tau}_{\text{IN}} = \frac{1}{np(1-p)} \sum_{i=1}^{n} \frac{Y_i}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} (z_j - p).$$

This estimator is related to the estimator described by Hu, Li, and Wager (2022). The difference is that Hu et al. (2022) implicitly use a larger model space. It is possible to reproduce the exact estimator of Hu et al. (2022) by extending the model spaces to allow each unit in the neighborhoods $\mathcal{N}_i$ to have a different effect on the potential outcome, effectively increasing the number of dimensions of each model space from 3 to $2 + |\mathcal{N}_i|$.

The Riesz representor of the global average treatment effect for unit i is

$$\psi_i^{\mathrm{G}}(\boldsymbol{z}) = \frac{\mathbb{1}[z_i = 1]}{p} - \frac{\mathbb{1}[z_i = 0]}{1 - p} + \frac{|\mathcal{N}_i|^{-1} \sum_{j \in \mathcal{N}_i}(z_j - p)}{p(1-p)}.$$

Thus, the Riesz estimator for the global average treatment effect is the sum of the estimator for the indirect effect and the Horvitz–Thompson estimator from the previous section, which can be seen as an estimator of a type of direct effect.

4.5.3 Treatment timing experiments

Finally, we return to the example of a fertilizer timing experiment discussed in Section 3.6.5. The full intervention set was there $\mathcal{Z} = [-1, 1]^n$, and we aimed to estimate the gradient of the potential outcome functions at zero: $\theta(y) = \mathbf{1}^{\mathsf{T}} \nabla y(\mathbf{0})$. The model spaces imposed no-interference, so that y_i only depends on the ith dimension of $\boldsymbol{z}$, and the spaces otherwise contained the set of all functions that are measurable, square integrable and continuously differentiable.

We will consider an experimental design for which $Z_1, \ldots, Z_n$ are jointly independent and each Z_i is distributed according to the Wigner semicircle distribution with parameter $R = 1$. This distribution is unimodal, bell-shaped, and centered at 0, which we believe would be appropriate in this application, but the primary motivation for using this design here is that the derivation of the Riesz estimator becomes straightforward. The estimator can be constructed for other designs, but one would then likely need to do it numerically, rather than analytically as in this example.

The model spaces have infinite dimensions, but they are separable (Assumption 3), so a countable orthonormal basis exists. One such basis is the set of Chebyshev polynomials of the second kind, which is what we will use here. The kth basis function for unit i is given explicitly as

$$\phi_{i,k}(\boldsymbol{z}) = \frac{\left(z_i + \sqrt{z_i^2 - 1}\right)^k - \left(z_i - \sqrt{z_i^2 - 1}\right)^k}{2\sqrt{z_i^2 - 1}}.$$

The effect functional when evaluated on the kth basis function is

$$\theta(\phi_{i,k}) = -k \cos(\pi k / 2),$$

where $\cos(\cdot)$ is the cosine function. This means that the Riesz representor for unit i is

$$\psi_i(\boldsymbol{z}) = \sum_{k=1}^{\infty} -k \cos(\pi k/2) \phi_{i,k}(\boldsymbol{z}) = \sum_{k=1}^{\infty} (-1)^k 2k \phi_{i,2k}(\boldsymbol{z}),$$

and the Riesz estimator for τ is

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{\infty} (-1)^k 2k \phi_{i,2k}(\boldsymbol{Z}) Y_i.$$

The Riesz estimator will be unbiased for the average gradient of the potential outcome functions at zero in this setting. However, the estimator does not try to navigate the bias–variance trade-off we discussed at the end of Section 4.2. Therefore, this estimator will not be useful in practice because of its high variance. As we already have remarked, a way to make the estimator practically useful is to truncate the series to include only a finite number of basis functions. We leave it to future work to investigate how to best conduct such a truncation.

5 Properties of the Riesz Estimator

5.1 Asymptotic Regime and Regularity Conditions

We consider an asymptotic regime that consists of a sequence of experiments with a growing number of observations. Each experiment in the sequence is associated with its own experimental design, set of model spaces, linear functionals, and so on. Therefore, nearly all variables and parameters are implicitly indexed by n to denote which experiment in the sequence they are associated with, but following the current convention, we leave this indexing implicit for notational tidiness. We investigate the asymptotic properties of the Riesz estimator subject to conditions on the sequence of experiments. This asymptotic regime, or minor variations of it, has been widely used in the design-based causal inference literature (see, e.g., Freedman, 2008; Leung, 2022a; Lin, 2013; Sävje et al., 2021).

To state our regularity conditions, it will prove useful to consider the product spaces of the model spaces and the outcome spaces. Let $\mathcal{M}_{(n)} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_n$ be the *product model space*, which contains all vector-valued potential outcome functions. That is, functions in this space are of the form $\boldsymbol{u} : \mathcal{Z} \to \mathbb{R}^n$ and describe the outcomes produced by an intervention $z \in \mathcal{Z}$ for all units. If the model spaces are correctly specified, then the true vector-valued potential outcome function $\boldsymbol{y} = (y_1, \ldots, y_n)$ is in the product model space: $\boldsymbol{y} \in \mathcal{M}_{(n)}$. Similarly, let $\mathcal{L}(\mathcal{M}_{(n)}) = \mathcal{L}(\mathcal{M}_1) \times \cdots \times \mathcal{L}(\mathcal{M}_n)$ be the *product outcome space*. This space contains equivalence classes of vector-valued functions from $\mathcal{M}_{(n)}$ that are observationally equivalent. Similar to the unit-level outcome spaces, each equivalence class $[\boldsymbol{u}] \in \mathcal{L}(\mathcal{M}_{(n)})$ can be interpreted as a multivariate random variable $\boldsymbol{u}(Z)$ of dimension n.

We will occasionally consider the behavior of the Riesz estimator when the vector-valued potential outcome function take some particular value. We use semicolon to extend common probability operators to clarify which function we consider in such settings. For example, $\mathrm{E}[\widehat{\tau}; \boldsymbol{u}]$ and $\mathrm{Var}(\widehat{\tau}; \boldsymbol{u})$ denote the expectation and variance of the Riesz estimator when the true vector-valued potential outcome function is equal to some $\boldsymbol{u} \in \mathcal{M}_{(n)}$.

Our regularity conditions concern the magnitude of functions in the product spaces. This magnitude is captured by norms $\|\cdot\| : \mathcal{L}(\mathcal{M}_{(n)}) \to \mathbb{R}_+$ on the product outcome space. We consider several types of norms in our analysis, as described below. For notational convenience, we also consider the norms on the product model space $\mathcal{M}_{(n)}$ that are induced by the norms on product outcome space $\mathcal{L}(\mathcal{M}_{(n)})$. That is, for any $\boldsymbol{u} \in \mathcal{M}_{(n)}$, we define its norm to be the norm of the equivalence class it belongs to in the product outcome space: $\|\boldsymbol{u}\| = \|[\boldsymbol{u}]\|$. This is technically a seminorm, because it is not positive definite on $\mathcal{M}_{(n)}$, but that distinction will not be important for our purposes. More generally, the distinction between a function $\boldsymbol{u} \in \mathcal{M}_{(n)}$ and its equivalence class $[\boldsymbol{u}] \in \mathcal{L}(\mathcal{M}_{(n)})$ will typically not be important in what follows, and we will often use the notation $\boldsymbol{u}$ to refer to both.

The first type of norm we will consider, which also has been implicitly used in the design-based causal inference, is what we refer to as the max-p norm:

$$\|\boldsymbol{u}\|_{\max,p} = \max_{i \in [n]} \mathrm{E}\big[|u_i(Z)|^p\big]^{1/p}.$$

For example, it is common to assume that the fourth moment of the outcome Y_i exists for all units, and this is equivalent to assuming that the max-4 norm of the true vector-valued potential outcome function $\boldsymbol{y}$ exists.

The other type of norm we will consider is the mean-square norm, which is similar to the Euclidean norm for real vector spaces:

$$\|\boldsymbol{u}\|_{\text{MS}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\big[u_i(Z)^2\big]}.$$

The reason the mean-square norm will play an important role in our analysis is that it is the norm induced by the Hilbert space when $\mathcal{L}(\mathcal{M}_{(n)})$ is endowed with the inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\text{MS}} = n^{-1} \sum_{i=1}^{n} \mathrm{E}[u_i(Z)v_i(Z)]$, which can be seen as the canonical inner product in this setting because of its correspondence with the inner product on $\mathcal{L}(\mathcal{Y})$.

The central regularity condition for our results in this section is that norms of functions in the product spaces are bounded. This condition corresponds directly to moment conditions conventionally made in the design-based causal inference literature. For example, if we impose the condition that $\|\boldsymbol{y}\|_{\text{MS}} \leq C$ for some $C < \infty$, we say that the vector-valued potential outcome function is in some ball given by the mean-square norm. This prevents the potential outcome functions from producing outcomes that are large outliers. Our regularity conditions do not require empirical researchers to pin down C. It suffices to know that such a ball exists for some finite radius.

5.2 Necessary and Sufficient Conditions for Consistency

We provide two analyses of the precision of the Riesz estimator in large samples. The first analysis, presented in the current subsection, yields necessary and sufficient conditions for consistency in mean square, and provides the rate of convergence in the most general setting. The analysis is based on the operator norm of a linear operator on the product outcome model space that exactly characterizes the variance. While this approach is somewhat involved, it highlights what drives consistency of design-based estimators, and these insights will prove useful both for empirical researchers and for future theoretical work. In the next subsection, we present an analysis of consistency using only sufficient conditions, which are more restrictive than the conditions in this subsection, but also simpler and more familiar.

Proposition 5.1. *Given correctly specified model spaces and positivity (Assumptions 1 and 2), either:*

(i) There exists a bounded linear operator $\mathcal{V} : \mathcal{L}(\mathcal{M}_{(n)}) \to \mathcal{L}(\mathcal{M}_{(n)})$ that exactly characterizes the variance of the Riesz estimator in finite samples:

$$\operatorname{Var}\big(\widehat{\tau}; \boldsymbol{u}\big) = \frac{1}{n}\|\mathcal{V}(\boldsymbol{u})\|_{\text{MS}}^2 \qquad \text{for all} \quad \boldsymbol{u} \in \mathcal{M}_{(n)},$$

(ii) Or the variance of the Riesz estimator cannot be characterized with respect to the mean-square norm:

$$\sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq C} \operatorname{Var}(\widehat{\tau}; \boldsymbol{u}) = \infty \qquad \text{for all} \quad C > 0.$$

The proposition describes two situations. In the first situation, the variance of the Riesz estimator is described by the mean-square norm of a transformation $\mathcal{V} : \mathcal{L}(\mathcal{M}_{(n)}) \to \mathcal{L}(\mathcal{M}_{(n)})$ of the vector-valued potential outcome function. That is, the variance of the Riesz estimator is exactly $n^{-1}\|\mathcal{V}(\boldsymbol{y})\|_{\text{MS}}^2$. We therefore refer to $\mathcal{V}$ as the *variance characterizing operator*. In the second situation, the variance of the Riesz estimator cannot be described using a regularity condition based on the mean-square norm. That is, if we impose the condition that $\|\boldsymbol{y}\|_{\text{MS}} \leq C$, then no matter how small we make C, we cannot rule out that the variance is infinite. In other words, we cannot rule out that the Riesz estimator is extremely poorly behaved in this setting even when the potential outcome functions are extremely well-behaved in a mean-square norm sense.

The variance characterizing operator depends solely on the experimental design, the effect functionals and the model spaces. This is powerful because these components are known to the experimenter before the experiment is run, so $\mathcal{V}$ is known without any knowledge of the true potential outcomes $\boldsymbol{y}$, other than what is encoded in the model spaces. However, it might be computationally challenging to construct the operator in practice.

In Section S2.1.1 of the supplement, we show that the operator can be represented as a closed-form matrix when the model spaces have finite dimensions, making the computation tractable. If the model spaces are infinite-dimensional and separable, then relevant aspects of the variance characterizing operator can be computed approximately under additional regularity conditions, but we will not explore that direction in this paper.

It can be challenging to test whether we are in the first or second situation, corresponding to when the variance characterizing operator does and does not exist. If the model spaces have finite dimensions, one can construct and inspect the closed form matrix mentioned in the previous paragraph; we are in the first situation if all entries of the matrix are finite. For infinite-dimensional model spaces, one must investigate whether a more elementary condition about whether a certain bilinear functional is bounded with respect to the mean-square norm, as described by Lemma S2.2 in Section S2.1.1 of the supplement. We do not have a general computational procedure for testing whether this condition holds, and one would need to proceed on a case by case basis.

While we cannot exactly delineate when the variance characterizing operator exists, there are many settings of interest where we can show that it does exist. One such setting is when the model spaces contain functions that are bounded with probability one on the experimental design. That is, if we know that $\|\boldsymbol{u}\|_{\max,\infty} < \infty$ by construction for all $\boldsymbol{u} \in \mathcal{L}(\mathcal{M}_{(n)})$, where $\|\cdot\|_{\max,p}$ is the max-p norm defined in the previous subsection, then the operator exists. This is the case, for example, when the cardinality of the intervention set $\mathscr{Z}$ is finite. Another setting where the operator exists is when the effect functionals are integration functionals with respect to measures that are sufficiently close to the experimental design. More generally, a sufficient (but not necessary) condition for the existence of the variance characterizing operator is that the essential supremum of each unit's Riesz representor is finite. That is, $\mathcal{V}$ exists if $\|\boldsymbol{\psi}\|_{\max,\infty} < \infty$, where $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)$ is the vector of all Riesz representors.

We cannot use the result of Proposition 5.1 to estimate the standard error of the Riesz estimator, because we can never evaluate the operator at the true potential outcome function. The potential outcome function $\boldsymbol{y}$ is unknown even after the experiment is run. Instead, the usefulness of the proposition is that it allows us to characterize the behavior of the estimator over a set of potential outcome functions.

Corollary 5.2. *Given correctly specified model spaces and positivity (Assumptions 1 and 2), the worst-case root mean square error of the Riesz estimator in any mean-square norm ball is the scaled product of the radius of the ball and the operator norm of $\mathcal{V}$:*

$$\sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq C} \sqrt{\mathrm{E}\big[(\widehat{\tau} - \tau)^2; \boldsymbol{u}\big]} = n^{-1/2} C \|\mathcal{V}\|_{op} \qquad \textit{for all} \quad C > 0.$$

If the variance characterizing operator does not exists, set $\|\mathcal{V}\|_{op} = \infty$.

A direct implication of the corollary is that root mean square error of the Riesz estimator is bounded from above by the product of the mean-square norm of the true potential outcome function $\boldsymbol{y}$ and the operator norm of the variance characterizing operator $\mathcal{V}$:

$$\sqrt{\mathrm{E}\big[(\widehat{\tau} - \tau)^2\big]} \leq n^{-1/2} \|\boldsymbol{y}\|_{\text{MS}} \|\mathcal{V}\|_{op}.$$

Furthermore, this inequality is sharp, also in finite samples, in the sense that for any $C > 0$, there exists some $\boldsymbol{u} \in \mathcal{M}_{(n)}$ with $\|\boldsymbol{u}\|_{\text{MS}} = c$ such that the root mean square error is $n^{-1/2} \|\boldsymbol{u}\|_{\text{MS}} \|\mathcal{V}\|_{op}$ when $\boldsymbol{y} = \boldsymbol{u}$.

The operator norm $\|\cdot\|_{op}$ captures how much the magnitude of the input can increase by the transformation described by the operator. That is, the operator norm of an operator f is how large the magnitude of the output $\|f(x)\|$ can possibly be relative to the magnitude of the input $\|x\|$. In our case, the operator norm for the linear transformation $\mathcal{V}$ is defined with respect to the mean-square norm:

$$\|\mathcal{V}\|_{op} = \sup_{\boldsymbol{u} \neq \mathbf{0}} \frac{\|\mathcal{V}(\boldsymbol{u})\|_{\text{MS}}}{\|\boldsymbol{u}\|_{\text{MS}}}.$$

The operator norm bound in Corollary 5.2 follows almost directly from this definition.

The strength of Corollary 5.2 is that it separates the transformation of $\boldsymbol{y}$ from the input $\boldsymbol{y}$ itself. This is useful because $\|\mathcal{V}\|_{op}$ only depends on the design, the effect functionals and the model spaces, so it is known, at least in principle, without any knowledge of the true potential outcomes. Furthermore, while $\boldsymbol{y}$ is unknown, experimenters often feel comfortable reasoning about its magnitude. The variance is therefore decomposed into one part, $\|\mathcal{V}\|_{op}$, that can be calculated without knowledge of the true potential outcomes, and one part, $\|\boldsymbol{y}\|_{\text{MS}}$, that empirical researchers often feel comfortable reasoning about, at least to the degree that they can judge whether it is asymptotically bounded.

Corollary 5.3. *If Assumptions 1 and 2 hold and the mean-square norm of the potential outcomes is asymptotically bounded, $\|\boldsymbol{y}\|_{\text{MS}} = \mathcal{O}(1)$, then a necessary and sufficient condition for consistency in mean square of the Riesz estimator is $\|\mathcal{V}\|_{op} = o(n^{1/2})$. Furthermore, the rate of convergence of the Riesz estimator is exactly the rate at which $n^{-1/2}\|\mathcal{V}\|_{op}$ approaches zero. A necessary and sufficient condition for root-n consistency of the Riesz estimator is $\|\mathcal{V}\|_{op} = \mathcal{O}(1)$.*

The corollary uses the operator norm bound to describe in what situations we can expect the Riesz estimator to be precise in large samples. The takeaway is that the precision of the estimator is determined by the operator norm $\|\mathcal{V}\|_{op}$, where a smaller operator norm means more precision. The estimator is root-n consistent if the operator norm is asymptotically bounded. The reason the corollary restricts the potential outcome function to to have asymptotically bounded mean-square norm is that the variance trivially becomes large if the magnitude of $\boldsymbol{y}$ approaches infinity. Conversely, if the potential outcome functions approach the zero function asymptotically, $\|\boldsymbol{y}\|_{\text{MS}} = o(1)$, then the Riesz estimator can be precise even when the operator norm is not well-controlled. However, both these edge cases are typically not relevant in practice, which is why Corollary 5.3 does not consider them.

We take the experimental design as given in this paper, but experimenters often have some control over what design is used in the experiment. Proposition 5.1 and its corollaries provide some guidance on how to select a well-performing design. Corollary 5.2 tells us that we should minimize the operator norm $\|\mathcal{V}\|_{op}$ in order to make the worst-case variance of the Riesz estimator as small as possible. This is, at least in principle, a feasible task because the operator norm can be calculated before running the experiment. The approach of using operator norms to guide the design of randomized experiments has been investigated in the setting of binary treatments under no interference by Efron (1971), Kapelner, Krieger, Sklar, Shalit, and Azriel (2020) and Harshaw, Sävje, Spielman, and Zhang (2021). Proposition 5.1 shows that this idea extends to a more general setting. We save the full investigation of how to design experiments in this setting for future work.

5.3 Sufficient Conditions for Consistency

The operator norm characterization in the previous subsection exhaustively describes when the Riesz estimator is precise in large samples, but some experimenters might find it somewhat opaque. To complement the previous analysis, we describe sufficient conditions for consistency of the Riesz estimator in this section that are more restrictive but perhaps easier to interpret. The alternative approach is based on dependency neighborhoods, which have been widely used in the recent literature on causal inference under interference. The dependency neighborhoods concern the model spaces in our setting.

Definition 3. The *model dependency neighborhood* for unit i is the smallest $\mathcal{D}_i \subset [n]$ such that $\mathcal{M}_i$ and $\bigcup_{j \in [n] \setminus \mathcal{D}_i} \mathcal{M}_j$ are independent with respect to the experimental design. Let $d_{\text{avg}} = n^{-1} \sum_{i=1}^{n} |\mathcal{D}_i|$ denote the average size of the dependence neighborhoods, and let $d_{\max} = \max_i |\mathcal{D}_i|$ denote the size of the largest dependence neighborhood.

We say that two sets of functions, such as $\mathcal{M}_i$ and $\bigcup_{j \in [n] \setminus \mathcal{D}_i} \mathcal{M}_j$, are independent if any finite-dimensional vector of functions from the first function space evaluated at Z is independent of any finite-dimensional vector of functions from the second function space also evaluated at Z. For example, if $u_1, \dots, u_k \in \mathcal{M}_i$ and $v_1, \dots, v_\ell \in \bigcup_{j \in [n] \setminus \mathcal{D}_i} \mathcal{M}_j$, then independence of $\mathcal{M}_i$ and $\bigcup_{j \in [n] \setminus \mathcal{D}_i} \mathcal{M}_j$ implies that $(u_1(Z), \dots, u_k(Z))$ and $(v_1(Z), \dots, v_\ell(Z))$ are independent. Intuitively, if unit j is in unit i's model dependency neighborhood, then unit j's outcome (possibly together with outcomes of other units) provides information about unit i's outcome under the experimental design. The size of a dependency neighborhood is one way to characterize the amount of dependence for a given unit, and the average d_{avg} and maximum $d_{\max}$ characterize the overall amount of dependence in the experiment.

Proposition 5.4. *Let p and q be values satisfying $1/p + 1/q = 1/2$. The root mean square error of the Riesz estimator is upper bounded by*

$$\sqrt{\mathrm{E}\big[(\widehat{\tau} - \tau)^2\big]} \leq n^{-1/2} d_{\mathrm{avg}}^{1/2} \|\boldsymbol{y}\|_{\max,p} \|\boldsymbol{\psi}\|_{\max,q},$$

where $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)$ is the vector of all Riesz representors and $\|\cdot\|_{\max,p}$ is the max-p norm discussed in Section 5.1.

The proposition has the same structure as Corollary 5.2, but with the operator norm substituted with $d_{\mathrm{avg}}^{1/2} \|\boldsymbol{\psi}\|_{\max,q}$. Because the Riesz representors are in the model spaces, we can consider the magnitude of the vector-valued Riesz representor $\boldsymbol{\psi}$ in the same way as we consider the magnitude of the vector-valued potential outcome function, here using the max-p norm.

The magnitude of $\boldsymbol{\psi}$ captures the degree to which the outcomes are weighted in the Riesz estimator, which can interpreted as a measure of how difficult it is to estimate the treatment effect. If $\|\boldsymbol{\psi}\|_{\max,q}$ is large, the outcomes are heavily weighted and the effect is difficult to estimate. If the estimand is the average expected outcome, $n^{-1} \sum_{i=1}^{n} \mathrm{E}[Y_i]$, which is easy to estimate, then the Riesz representors are constant at one, so $\|\boldsymbol{\psi}\|_{\max,q} = 1$. Positivity (Assumption 2) implies that the max-2 norm is finite for any fixed n, $\|\boldsymbol{\psi}\|_{\max,2} < \infty$, but $\|\boldsymbol{\psi}\|_{\max,q}$ could be infinite for $q > 2$ even when positivity holds, and it could be that $\|\boldsymbol{\psi}\|_{\max,2} \to \infty$ as n grows. Experimenters should try to ensure that $\|\boldsymbol{\psi}\|_{\max,q}$ is asymptotically bounded, which can be seen as a strengthening of the positivity assumption. When using exposure mappings, this corresponds to ensuring that the assignment probabilities of the relevant exposures are bounded away from zero by a constant. That is, if $d_i(z) = a$ is an exposure of interest for unit i, then positivity stipulates that $\Pr(d_i(Z) = a) > 0$, while $\|\boldsymbol{\psi}\|_{\max,q} = \mathcal{O}(1)$ stipulates that $\Pr(d_i(Z) = a) \geq c$ for some constant $c > 0$.

The following corollary of Proposition 5.4 shows that consistency is ensured for the Riesz estimator provided that the potential outcomes and Riesz representors are asymptotically bounded, and that there is not too much dependence between the model spaces.

Corollary 5.5. *Given correctly specified model spaces and positivity (Assumptions 1 and 2), limited average model dependence, $d_{\mathrm{avg}} = o(n)$, and $\|\boldsymbol{y}\|_{\max,p} = \mathcal{O}(1)$ and $\|\boldsymbol{\psi}\|_{\max,q} = \mathcal{O}(1)$ for some $1/p + 1/q = 1/2$, the Riesz estimator is consistent in mean square. If the condition on the average model dependence is strengthened to $d_{\mathrm{avg}} = \mathcal{O}(1)$, the Riesz estimator is root-n consistent.*

5.4 Asymptotic Normality

We conclude this section by investigating the limiting distribution of the Riesz estimator. We provide sufficient conditions under which the distribution approaches a normal distribution as the sample size grows. It is beyond the scope of this paper to provide necessary and sufficient condition for asymptotic normality.

The approach we take here is similar to the one we took in the previous subsection, using model dependency neighborhood to capture dependence between units. The conditions used to ensure asymptotic normality are stronger than those we used for consistency. We also impose an additional regularity condition to ensure that the limiting distribution of the Riesz estimator is not degenerate, which is implemented as an assumption that the convergence rate is not faster than root-n, which is the parametric rate. In theory, there are sequences in our asymptotic regime for which the rate is faster than this, but they are knife-edge situations where, for example, the dependence structure in the design perfectly aligns with the potential outcomes. We do not believe that these situations are practically relevant, so essentially nothing is lost by imposing the non-degeneracy assumption. This type of assumption is common in the design-based causal inference literature; examples include Condition 6 in Aronow and Samii (2017) and Assumption 5 in Leung (2022b).

Assumption 4 (Non-degeneracy)**.** $n \operatorname{Var}(\widehat{\tau}) \geq c$ for some $c > 0$ and all n.

Proposition 5.6. *Given correctly specified model spaces, positivity and non-degeneracy (Assumptions 1, 2 and 4), limited maximum model dependence, $d_{\max} = o(n^{1/4})$, and $\|\boldsymbol{y}\|_{\max,p} = \mathcal{O}(1)$ and $\|\boldsymbol{\psi}\|_{\max,q} = \mathcal{O}(1)$ for some $1/p + 1/q = 1/4$, the limiting distribution of the Riesz estimator is normal:*

$$\frac{\widehat{\tau} - \tau}{\sqrt{\operatorname{Var}(\widehat{\tau})}} \xrightarrow{d} \mathcal{N}(0, 1).$$

The proof of Proposition 5.6 appears in Section S2.3 of the supplement. The main technique of the proof is Stein's method adapted to dependency graphs, as described by Ross (2011).

6 Variance Estimation and Inference

6.1 Overview

To assess the precision of the Riesz estimator and construct confidence intervals, we aim to estimate its variance:

$$\operatorname{Var}(\widehat{\tau}) = \operatorname{Var}\Big(n^{-1} \textstyle\sum_{i=1}^{n} \psi_i(Z) Y_i\Big) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}\big(\psi_i(Z) Y_i, \psi_j(Z) Y_j\big).$$

The expression shows that we can reinterpret the variance as an average of n^2 bilinear functionals on the Cartesian products of the model spaces evaluated at the true potential outcome functions. This interpretation highlights a parallel to the estimation of τ, which is an average of linear functionals on the model spaces evaluated at the true potential outcome functions. We will approach the variance estimation problem in a similar way to how we approached the point estimation problem.

There are, however, two hurdles that prevent a direct application of the approach we used for point estimation. First, the Riesz representation theorem concerns linear functionals, but the variance consists of bilinear functionals. We navigate this issue by tensorizing the variance expression. For our purposes, the tensors are used to rewrite the bilinear functionals on the Cartesian products of the model spaces, $\mathcal{M}_i \times \mathcal{M}_j$, as linear functionals on the tensor product of the model spaces, $\mathcal{M}_i \otimes \mathcal{M}_j$. We describe tensor products and our use of them later in this section.

The second hurdle is that the structure of the variance estimation problem often introduces positivity violations, meaning that the equivalent to Assumption 2 generally does not hold in the variance estimation problem. That is, some aspects of the tensor products of the model spaces that are relevant for the variance are often inherently unobservable. Unlike positivity violations under the point estimation, this problem can typically not be addressed by changing the experimental design when estimating the variance. We navigate this issue by estimating alternative linear functionals on the tensor products, whose average is an upper bound for the true variance. This means that our variance estimator will be conservative in expectation. This approach is commonly used in the design-based causal inference literature (see, e.g., Aronow, Green, & Lee, 2014; Fogarty, 2018; Imbens & Menzel, 2021). The bound we use is inspired by the bound described by Aronow and Samii (2013, 2017) for discrete treatments. Our bound is a generalization of this bound to general tensor products of model spaces. It is possible to use the ideas described by Harshaw, Middleton, and Sävje (2021) to improve on the bound we use here, but we will not explore such an extension in this paper. Because we use the more straightforward approach, our variance bounds are often quite crude, yielding overly conservative variance estimators.

6.2 Tensor Spaces and Tensorization of the Variance

The tensor product $\mathcal{M}_i \otimes \mathcal{M}_j$ of two model spaces is the vector space for which a basis is the collection of all pairs of elements from a basis of $\mathcal{M}_i$ and a basis of $\mathcal{M}_j$. The tensor product contains an element associated with all elements in the Cartesian product of the model spaces: $(u, v) \in \mathcal{M}_i \times \mathcal{M}_j$. The element in $\mathcal{M}_i \otimes \mathcal{M}_j$ corresponding to the pair (u, v) is denoted $u \otimes v$, and such an element is called a simple tensor. The mapping from $\mathcal{M}_i \times \mathcal{M}_j$ into the set of simple tensors is bilinear, in the sense that if $u = \sum_{k=1}^{s} a_k u_k$ and $v = \sum_{\ell=1}^{r} b_\ell v_\ell$, then $u \otimes v = \sum_{k=1}^{s} \sum_{\ell=1}^{r} a_k b_\ell (u_k \otimes v_\ell)$. The tensor product contains all linear combinations of the simple tensors:

$$\mathcal{M}_i \otimes \mathcal{M}_j = \operatorname{span}\big(\{u \otimes v : u \in \mathcal{M}_i, v \in \mathcal{M}_j\}\big).$$

The tensor product of $\mathcal{M}_i$ and $\mathcal{M}_j$ is different from their Cartesian product. If both model spaces have dimension d, then $\mathcal{M}_i \times \mathcal{M}_j$ has $2d$ dimensions but $\mathcal{M}_i \otimes \mathcal{M}_j$ has d^2 dimensions.

We will use tensors to rewrite bilinear functionals on $\mathcal{M}_i \times \mathcal{M}_j$ as linear functionals on $\mathcal{M}_i \otimes \mathcal{M}_j$. This is done by first defining the linear functional to coincide with the bilinear functional for all simple tensors and then use linearity to extend the functional to the full tensor product. The relevant bilinear functionals in our case is the covariance terms in Equation 6.1 when interpreted as functionals of the potential outcome functions. That is, the bilinear functional $A_{i,j} : \mathcal{M}_i \times \mathcal{M}_j \to \mathbb{R}$ is defined as

$$A_{i,j}(u, v) = \operatorname{Cov}\big(\psi_i(Z)u(Z), \psi_j(Z)v(Z)\big),$$

for all $u \in \mathcal{M}_i$ and $v \in \mathcal{M}_j$. Under correctly specified model spaces, the variance can therefore be written as

$$\operatorname{Var}(\widehat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}(y_i, y_j).$$

Let $\mathcal{A}_{i,j} : \mathcal{M}_i \otimes \mathcal{M}_j \to \mathbb{R}$ denote the linear functional produced by tensorizing $A_{i,j}$. The tensorization starts by defining $\mathcal{A}_{i,j}$ to coincide with $A_{i,j}$ for all simple tensors, meaning that $\mathcal{A}_{i,j}(u \otimes v) = A_{i,j}(u, v)$ for all $u \in \mathcal{M}_i$ and $v \in \mathcal{M}_j$. This is followed by an extension of the definition to the full tensor product using linearity. That is, if $\rho = \sum_{k=1}^{m} a_k(u_k \otimes v_k)$ is a non-simple tensor written as a linear combination of m simple tensors, then $\mathcal{A}_{i,j}(\rho) = \sum_{k=1}^{m} a_k \mathcal{A}_{i,j}(u_k \otimes v_k)$. By definition, all tensors can be written as a linear combination of simple tensors, so the extension is complete.

We can now write the variance as an average of linear functionals on the tensor products of the model spaces:

$$\operatorname{Var}(\widehat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{A}_{i,j}(y_i \otimes y_j).$$

6.3 Constructing Variance Riesz Representors

Having written the variance as an average of linear functionals, we can now proceed in a similar fashion as in Section 4 to construct estimators for the variance functionals using Riesz representors. For point estimation, the relevant function spaces were the model spaces, so the Riesz representors were functions mapping from $\mathcal{Z}$ to $\mathbb{R}$. Here, the relevant spaces are the tensor products $\mathcal{M}_i \otimes \mathcal{M}_j$, meaning that the Riesz representors will be tensors. To construct these representors, we build Hilbert spaces from the tensor products, in a similar fashion as when we constructed the point estimator. Under a positivity assumption on the tensors, we can define linear functionals on the Hilbert spaces that capture the behavior of the variance functionals on the tensor products.

It will prove helpful to associate each tensor with a function mapping from $\mathcal{Z}$ to $\mathbb{R}$. We slightly overload the notation and use the same symbol for the tensor $\rho \in \mathcal{M}_i \otimes \mathcal{M}_j$ to also denote its associated function $\rho : \mathcal{Z} \to \mathbb{R}$. For a simple tensor $\rho = u \otimes v$, the corresponding function ρ is defined as the product $\rho(z) = u(z)v(z)$. Next, we use linearity to extend the definition to also non-simple tensors. That is, if $\rho = \sum_{k=1}^{m} a_k(u_k \otimes v_k)$ is a non-simple tensor, then the corresponding function is defined as $\rho(z) = \sum_{k=1}^{m} a_k u_k(z) v_k(z)$. This means that $\rho(Z)$ can be seen as a random variable associated with the tensor $\rho \in \mathcal{M}_i \otimes \mathcal{M}_j$.

A potential concern here is that some tensors $\rho \in \mathcal{M}_i \otimes \mathcal{M}_j$ may not be in $\mathcal{Y}$, because they may not be square integrable. This would prevent us from using the Riesz representation theorem. To address this, we impose a stronger moment condition on the model spaces than square integrability. For each unit, we define the *moment restricted model space* $\mathcal{M}_i^*$ to be the set of all functions in the model space $\mathcal{M}_i$ with bounded fourth moments:

$$\mathcal{M}_i^* = \{u \in \mathcal{M}_i : \mathrm{E}[u(Z)^4] < \infty\},$$

and we require the true potential outcome function to be in this set.

Assumption 5. The true potential outcome function for each unit has bounded fourth moment under the design: $y_i \in \mathcal{M}_i^*$.

Mirroring Section 4.2, we say that two tensors ρ and γ in $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$ are observationally equivalent if $\rho(Z) = \gamma(Z)$ almost surely under the experimental design. We define the *outcome tensor space* to be the closure of the equivalence classes that contain a function corresponding to a tensor in $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$:

$$\mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*) = \mathrm{cl}\big(\{[\rho] \in \mathcal{L}(\mathcal{Y}) : \rho \in \mathcal{M}_i^* \otimes \mathcal{M}_j^*\}\big).$$

The closure is taken with respect to the topology induced by the Hilbert space $\mathcal{L}(\mathcal{Y})$. By Hölder's inequality, the function associated with each tensor $\rho \in \mathcal{M}_i^* \otimes \mathcal{M}_j^*$ has bounded second moment, meaning that $\rho \in \mathcal{Y}$. This means that we can use the inner product from the Hilbert space $\mathcal{L}(\mathcal{Y})$ together with $\mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*)$ to form a new Hilbert space that can be used to create the Riesz representors for the variance functionals.[2]

Proposition 6.1. *Together with the inner product $\langle [\rho], [\gamma] \rangle = \mathrm{E}[\rho(Z)\gamma(Z)]$, each outcome tensor space $\mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*)$ is a Hilbert space.*

[2] An alternative approach to create a Hilbert space in this setting is to use the tensor products of the outcome spaces $\mathcal{L}(\mathcal{M}_i) \otimes \mathcal{L}(\mathcal{M}_j)$ with the inner product $\langle u_i \otimes u_j, v_i \otimes v_j \rangle = \mathrm{E}[u_i(Z)v_i(Z)]\,\mathrm{E}[u_j(Z)v_j(Z)]$. While this is a Hilbert space, it is not an useful for the construction of variance estimators.

To extend the variance functionals to the outcome tensor spaces, we must ensure that the outcome tensor spaces contains the same information about the functionals as the tensor spaces. This is done by the same type of positivity condition as we used for point estimation.

Definition 4. A linear functional $\mathcal{A}_{i,j} : \mathcal{M}_i^* \otimes \mathcal{M}_j^* \to \mathbb{R}$ satisfies *second-order positivity* if there exists $C < \infty$ such that $|\mathcal{A}_{i,j}(\rho)| \leq C\sqrt{\mathrm{E}[\rho(Z)^2]}$ for all $\rho \in \mathcal{M}_i^* \otimes \mathcal{M}_j^*$.

Second-order positivity states that if two tensors are observationally equivalent under the design, then they must have the same value when evaluated on the corresponding variance functional. While this positivity condition is essentially the same as Assumption 2, we refer to it as second-order positivity for clarity. In Section S1.3.4 of the supplement, we demonstrate how to computationally verify whether second-order positivity holds when the model spaces are finite dimensional. If second-order positivity holds, we can extend the variance functional $\mathcal{A}_{i,j}$ to the Hilbert space $\mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*)$, on which the Riesz representation theorem can be applied. The construction is almost identical to the one that appears in Section 4.2. We describe the extension in detail in Section S1.3.4 of the supplement.

Proposition 6.2. *If a linear functional $\mathcal{A}_{i,j}$ on the tensor space $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$ satisfies second-order positivity, then there exists an equivalence class $\Psi_{i,j}^\dagger \in \mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*)$ such that every tensor $\Psi_{i,j} \in \Psi_{i,j}^\dagger$ satisfies*

$$\mathcal{A}_{i,j}(\rho) = \mathrm{E}[\Psi_{i,j}(Z)\rho(Z)] \quad \text{for all} \quad \rho \in \mathcal{M}_i^* \otimes \mathcal{M}_j^*.$$

We say that any tensor $\Psi_{i,j}$ in this equivalence class is a second-order Riesz representor for the functional $\mathcal{A}_{i,j}$.

Proposition 6.2 establishes the existence of Riesz representors for the variance functional under second-order positivity. The procedure for constructing Riesz representors for the variance functionals is the same as constructing Riesz representors for the effect functionals, as discussed in Section 4.3, provided that the outcome spaces $\mathcal{L}(\mathcal{M}_i)$ are replaced with the outcome tensor spaces $\mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*)$.

We can use the Riesz representors for the variance functionals to construct a variance estimator. Provided that second-order positivity holds for all variance functionals, Riesz representors $\Psi_{i,j}$ exist for all those functionals. This means that the expectation of the random variable $\Psi_{i,j}(Z)Y_iY_j$, which is observed, is equal to $\mathcal{A}_{i,j}(y_i \otimes y_j)$, which in turn is the i,jth term of the variance expression. The average of $\Psi_{i,j}(Z)Y_iY_j$ over $i,j \in [n]$ would therefore be an unbiased estimator of the variance of the Riesz point estimator. Note, however, that this only holds if all variance functionals satisfy second-order positivity.

6.4 Variance Bound for the Riesz Estimator

Second-order positivity will generally not hold for all variance functionals $\mathcal{A}_{i,j}$. The reason is what Holland (1986) calls the fundamental problem of causal inference, which is the fact that we can never simultaneously observe the potential outcome functions evaluated at two different treatment assignments. This means that the diagonal tensor products $\mathcal{M}_i^* \otimes \mathcal{M}_i^*$ tend to violate positivity, because they typically contain observationally equivalent tensors for which the variance functional has different values. This is true even for simple model spaces and experimental designs, such as binary treatments under no interference and a Bernoulli design. Positivity violations tend to exist also for non-diagonal tensor products $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$ even for moderately complex model spaces and experimental designs. However, while second-order positivity violations are common, there situations where all variance functionals satisfy positivity, in which case unbiased variance estimation is possible. One such setting is studied by Harshaw, Sävje, Eisenstat, Mirrokni, and Pouget-Abadie (2021).

We will follow the current convention in the design-based causal inference literature and address these positivity violations by deriving an upper bound for the variance that satisfies second-order positivity. We construct the variance bound by defining a new linear functional $\mathcal{B}_{i,j} : \mathcal{M}_i^* \otimes \mathcal{M}_j^* \to \mathbb{R}$ for all $i, j \in [n]$ such that $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$ satisfies second-order positivity with respect to $\mathcal{B}_{i,j}$ and

$$\mathrm{VB}(\widehat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{B}_{i,j}(y_i \otimes y_j)$$

is an upper bound for the variance $\mathrm{Var}(\widehat{\tau})$. This means that we can construct Riesz-type estimators for all terms $\mathcal{B}_{i,j}(y_i \otimes y_j)$, and thereby estimate the variance bound, which acts as a conservative variance estimator.

At a high level, our approach to construct a variance bound is to decompose each variance functional $\mathcal{A}_{i,j} : \mathcal{M}_i^* \otimes \mathcal{M}_j^* \to \mathbb{R}$ into a sum of more elementary linear functionals based on basis representations of the Riesz representors ψ_i and ψ_j. If the variance functional violates second-order positivity, then at least one of these elementary linear functionals must also violate positivity. For all elementary linear functionals that do not satisfy positivity, we apply the bound described by Aronow and Samii (2013, 2017) to bound the functionals by two other functionals that satisfy positivity. After all positivity violations have been addressed for the elementary functionals, we re-collect terms to construct our variance bound functionals $\mathcal{B}_{i,j}$.

Under Assumption 3, each outcome model space $\mathcal{L}(\mathcal{M}_i)$ is separable and therefore has a countable orthonormal basis $\{[\phi_{i,\ell}]\}_{\ell=1}^{\infty}$. Each $[u] \in \mathcal{L}(\mathcal{M}_i)$ can be written uniquely as $[u] = \sum_{k=1}^{\infty} a_k [\phi_{i,k}]$. Using these bases, we can decompose the equivalence class of each Riesz representor used in the point estimator as $[\psi_i] = \sum_{k=1}^{\infty} \beta_{i,k} [\phi_{i,k}]$ for some set of coefficients

$\{\beta_{i,k}\}_{k=1}^{\infty}$. We can then decompose the variance functional as

$$\begin{aligned}\mathcal{A}_{i,j}(u \otimes v) &= \operatorname{Cov}\Big(\psi_i(Z)u(Z), \psi_j(Z)v(Z)\Big) \\ &= \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty} \beta_{j,k}\beta_{i,\ell} \operatorname{Cov}\Big(\phi_{i,k}(Z)u(Z), \phi_{j,\ell}(Z)v(Z)\Big),\end{aligned}$$

where $\{\phi_{i,\ell}\}_{\ell=1}^{\infty}$ is a sequence of arbitrary functions selected from each basis equivalence class. Define a new set of linear functionals $\mathcal{A}_{i,j,k,\ell} : \mathcal{M}_i^* \otimes \mathcal{M}_j^* \to \mathbb{R}$ on the simple tensors as

$$\mathcal{A}_{i,j,k,\ell}(u \otimes v) = \beta_{j,k}\beta_{i,\ell} \operatorname{Cov}\Big(\phi_{i,k}(Z)u(Z), \phi_{j,\ell}(Z)v(Z)\Big),$$

and use linearity to extend the definition to the full tensor product. The variance functional can thereby be reinterpreted as the sum of these elementary functionals: $\mathcal{A}_{i,j} = \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty} \mathcal{A}_{i,j,k,\ell}$.

If $\mathcal{A}_{i,j}$ does not satisfy positivity, then at least one of elementary functionals $\mathcal{A}_{i,j,k,\ell}$ does not satisfy positivity. When either $i \neq j$ or $k \neq \ell$, we bound these offending functionals using the Cauchy–Schwarz and AM–GM inequalities:

$$\begin{aligned}\mathcal{A}_{i,j,k,\ell}(u \otimes v) &= \beta_{i,k}\beta_{j,\ell} \operatorname{Cov}\Big(\phi_{i,k}(Z)u(Z), \phi_{j,\ell}(Z)v(Z)\Big) \\ &\leq \frac{1}{2}\Big[\beta_{i,k}^2 \operatorname{Var}\big(\phi_{i,k}(Z)u(Z)\big) + \beta_{j,\ell}^2 \operatorname{Var}\big(\phi_{j,\ell}(Z)v(Z)\big)\Big].\end{aligned}$$

Note that $\beta_{i,k}^2 \operatorname{Var}\big(\phi_{i,k}(Z)u(Z)\big)$ is exactly the functional $\mathcal{A}_{i,i,k,k}$ evaluated on the simple tensor $u \otimes u$, resulting in the bound $2\mathcal{A}_{i,j,k,\ell}(u \otimes v) \leq \mathcal{A}_{i,i,k,k}(u \otimes u) + \mathcal{A}_{j,j,\ell,\ell}(v \otimes v)$. It is computationally tractable to test for second-order positivity violations of a linear functional on $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$, meaning that we can discover which functionals need to be bounded. We describe this test in Section S1.3.4 in the appendix.

If one of the elementary functionals $\mathcal{A}_{i,j,k,\ell}$ that does not satisfy positivity is on the diagonal, meaning that $i = j$ and $k = \ell$, then we need an alternative bound. We cannot apply the Cauchy–Schwarz and AM–GM inequalities in this case because they would return the same the functional as we started with. Here, we use the fact that the second raw moment is a bound on the variance for any random variable:

$$\mathcal{A}_{i,i,k,k}(u \otimes u) = \beta_{i,k}^2 \operatorname{Var}\big(\phi_{i,k}(Z)u(Z)\big) \leq \beta_{i,k}^2 \operatorname{E}\big[(\phi_{i,k}(Z)u(Z))^2\big].$$

We introduce a new family of linear functionals $\mathcal{D}_{i,k} : \mathcal{M}_i^* \otimes \mathcal{M}_i^* \to \mathbb{R}$ to capture these terms. For simple tensors, let $\mathcal{D}_{i,k}(u \otimes v) = \operatorname{E}[\phi_{i,k}(Z)^2 u(Z)v(Z)]$, and use linearity to extend the definition to the full tensor product. This means that $\mathcal{A}_{i,i,k,k}(u \otimes u) \leq \mathcal{D}_{i,k}(u \otimes u)$. Lemma S1.8 in Section S1.3.4 of the supplement shows that $\mathcal{D}_{i,k}$ satisfies second-order positivity. The intuition behind this is that the random variable $(\phi_{i,k}(Z)u(Z))^2$ is directly informative of its expectation, so the design always provides sufficient information of $\mathcal{D}_{i,k}$.

After applying these bounds on all offending elementary functionals $\mathcal{A}_{i,j,k,\ell}$, we are left with a set of functionals that all satisfy second-order positivity. The final step is a bookkeeping exercise, collecting all functionals defined on the same tensor product into a single functional, which gives our variance bound functional $\mathcal{B}_{i,j}$. To this end, let $F_{i,j} \subseteq \mathbb{N}^2$ be the set of pairs of indices (k,ℓ) of elementary functionals $\mathcal{A}_{i,j,\ell,k}$ for i and j that satisfy second-order positivity. Let $\rho_{i,k} \in \{0,1\}$ be an indicator of whether $\mathcal{A}_{i,i,k,k}$ satisfies second-order positivity: $\rho_{i,k} = \mathbb{1}[(k,k) \in F_{i,i}]$. The variance bound functional is then defined as

$$\mathcal{B}_{i,j} = \begin{cases} \sum_{(k,\ell)\in F_{i,j}} \mathcal{A}_{i,j,k,\ell} & \text{if } i \neq j, \\ \sum_{(k,\ell)\in F_{i,j}} \mathcal{A}_{i,i,k,\ell} + \sum_{k=1}^{\infty} Q_{i,k}\big(\rho_{i,k}\mathcal{A}_{i,i,k,k} + (1-\rho_{i,k})\mathcal{D}_{i,k}\big) & \text{if } i = j, \end{cases}$$

where $Q_{i,k} = |\{(j,\ell) \in [n] \times [d] : (k,\ell) \notin F_{i,j}\}|$ is the number of elementary functionals $\mathcal{A}_{i,j,\ell,k}$ involving unit i and basis k that do not satisfy positivity. If the model spaces have infinite dimensions, then $[d] = \mathbb{N}$ in the definition of $Q_{i,k}$. For this bound to be useful, one must ensure that $Q_{i,k}$ is finite for all $(i,k) \in [n] \times [d]$.

The intuition behind the inequalities we have used to derive this variance bound is that they transfer mass, in lack of a better term, from the off-diagonal terms, $i \neq j$ or $k \neq \ell$, that do not satisfy positivity to diagonal terms. The counting variable $Q_{i,k}$ captures the amount of mass transferred to the diagonal term involving unit i and basis function k, corresponding to functionals $\mathcal{A}_{i,i,k,k}$ or $\mathcal{D}_{i,k}$, and the sum of $Q_{i,k}$ gives a rough indication of the overall conservativeness of the variance bound.

Proposition 6.3. *The variance bound* $\mathrm{VB}(\widehat{\tau})$ *based on functionals* $\mathcal{B}_{i,j}$ *is an upper bound on the variance of the Riesz estimator:*

$$\mathrm{Var}(\widehat{\tau}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{A}_{i,j}(y_i \otimes y_j) \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{B}_{i,j}(y_i \otimes y_j) = \mathrm{VB}(\widehat{\tau}).$$

The proposition states that $\mathrm{VB}(\widehat{\tau})$ indeed is a variance bound, and therefore can be used to construct a conservative variance estimator. Note that the inequality does not necessarily hold term by term, meaning that we generally will have some terms for which $\mathcal{A}_{i,j}(y_i \otimes y_j) \geq \mathcal{B}_{i,j}(y_i \otimes y_j)$.

To construct the variance estimator, first construct Riesz representors for the variance bound functionals, as described in the previous subsection. Let $\Psi_{i,j}^{\mathrm{VB}}$ denote the Riesz representors for $\mathcal{B}_{i,j}$ on $\mathcal{M}_i \otimes \mathcal{M}_j$. The estimator for the variance bound, which acts as our variance estimator, is then

$$\widehat{\mathrm{VB}}(\widehat{\tau}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Psi_{i,j}^{\mathrm{VB}}(Z) Y_i Y_j.$$

Some model spaces and designs make it easy to deduce that $\mathcal{A}_{i,j}(y_i \otimes y_j) = 0$ for certain pairs $i, j \in [n]$. In that case, the corresponding Riesz representor will capture this, in the sense that $\Psi_{i,j}^{\mathrm{VB}}(Z) = 0$ almost surely, meaning that nothing needs to changed to accommodate those situations. However, to speed up the computation of the variance estimator and improve numerical stability, it is advisable to not derive the Riesz representor for the terms that are known to be zero.

Proposition 6.4. *For each pair $i, j \in [n]$, the variance functional $\mathcal{B}_{i,j}$ satisfies second-order positivity with respect to the moment restricted tensor product $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$.*

Corollary 6.5. *Given correctly specified model spaces, first-order positivity, separability and bounded fourth moments (Assumptions 1, 2, 3, and 5), the variance estimator is unbiased for the variance bound and conservative in expectation for the true variance:*

$$\mathrm{E}\big[\widehat{\mathrm{VB}}(\widehat{\tau})\big] = \mathrm{VB}(\widehat{\tau}) \geq \mathrm{Var}(\widehat{\tau}).$$

6.5 Confidence Intervals and Consistency of Variance Estimator

With the variance estimator in hand, we have all the components needed to construct Wald-type confidence intervals. This type of confidence interval takes advantage of the fact that the estimator is asymptotically normal and approximates the sampling distribution with a normal distribution, yielding an interval for τ at the $1 - \alpha$ nominal confidence level that is centered at $\widehat{\tau}$ with a radius that is the product of the estimated standard error and the $1 - \alpha/2$ percentile of the standard normal distribution:

$$\widehat{\tau} \pm \Phi^{-1}(1 - \alpha/2)\sqrt{\widehat{\mathrm{VB}}(\widehat{\tau})},$$

where $\Phi^{-1} : [0, 1] \to \mathbb{R}$ is the quantile function of the standard normal deviate.

To prove that this interval is a valid confidence interval, we must show that the variance estimator is not only conservative in expectation, but that it is conservative with a probability approaching one. We will do this by showing that it is consistent for the variance bound. Because the variance estimator is a type of Riesz estimator, we could take the same approach as in Section 5.2 to show consistency. That is, we would construct a linear operator on the direct product of the outcome tensor spaces that characterizes the variance of the variance estimator, in a manner analogous to Proposition 5.1. However, unlike for the point estimator, for which this approach gave necessary and sufficient conditions, the approach would only give sufficient conditions for the variance estimator. The reason for this is that we know that the estimand corresponds to the variance bound functionals evaluated at simple tensors, because the true potential outcome tensors are simple by construction, but the operator norm bound considers all tensors. A possible way forward would be to use a restricted concept of an operator norm that only considers simple tensors, but that would be of limited practical value as such a restricted operator norm is hard to reason about and compute. For this reason, we will provide sufficient conditions for consistency of the variance estimator using an approach similar to the one taken in Section 5.3.

Definition 5. The *pairwise second-order dependency neighborhood* for unit pair (i, j) is the smallest $\mathcal{S}_{i,j} \subset [n]^2$ such that $\mathcal{M}_i^* \cup \mathcal{M}_j^*$ and $\mathcal{M}_r^* \cup \mathcal{M}_s^*$ are independent with respect to the experimental design for all pairs $(r, s) \in [n]^2 \setminus \mathcal{S}_{i,j}$. Let $s_{\text{avg}} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} |\mathcal{S}_{i,j}|$ denote the average size of the second-order dependence neighborhoods.

Proposition 6.6. *Let* $\boldsymbol{\Psi}^{\text{VB}} = (\Psi_{1,1}^{\text{VB}}, \dots, \Psi_{n,n}^{\text{VB}})$ *be a vector collecting all Riesz representors for the variance bound functionals* $\mathcal{B}_{i,j}$*. Given correctly specified model spaces, first-order positivity, separability and bounded fourth moments (Assumptions 1, 2, 3, and 5), that* $\|\boldsymbol{y}\|_{\max,p}$ *and* $\|\boldsymbol{\Psi}^{\text{VB}}\|_{\max,q}$ *are asymptotically bounded for* $p \geq 4$ *and* $q \geq 2$ *such that* $1/p + 1/2q = 1/4$*, and that* $s_{\text{avg}} = o(1)$*, the normalized variance estimator is consistent of the normalized variance bound:* $\mathrm{E}[(n\widehat{\mathrm{VB}}(\widehat{\tau}) - n\mathrm{VB}(\widehat{\tau}))^2] = o(1)$*.*

Corollary 6.7. *Given the conditions of Propositions 5.6 and 6.6 (asymptotic normality of the point estimator and consistency of the variance estimator), the Wald-type confidence intervals are asymptotically valid:*

$$\liminf_{n \to \infty} \Pr\bigl(\widehat{\tau} - R_\alpha \leq \tau \leq \widehat{\tau} + R_\alpha\bigr) \geq 1 - \alpha,$$

where $R_\alpha = \Phi^{-1}(1 - \alpha/2)\sqrt{\widehat{\mathrm{VB}}(\widehat{\tau})}$ *is the radius of the interval.*

7 Concluding remarks

The framework we have described in this paper and the associated Riesz estimator allow empirical researchers to answer a wide range of causal question in the design-based paradigm. The results also provide insights about what we believe are some of the foundations of design-based causal inference, as evident from the fact that the framework unifies and generalizes existing design-based frameworks. Even in situations where the intricacies of the current framework are superfluous for practical purposes, such as when one is estimating the conventional average treatment effect for binary treatments under no-interference, the framework provides an explanation of why conventional estimators work, such as the

Horvitz–Thompson estimator. We find that to be valuable on its own, and we hope those insights will prompt new investigations and discoveries.

One such more general insight is about what type of statistical problem causal inference is in the design-based paradigm. It is common to interpret the causal inference problem as a discrete missing data problem, in the sense that the cells in our data set corresponding to the counterfactual potential outcomes are missing. We interpret the framework and results in this paper to suggest that the missing data perspective is too limited to capture the full range of causal questions empirical researchers might want to ask. A conclusion we draw from this work is that the causal inference problem is instead best seen as a functional estimation problem.

There are many open questions related to the framework and the Riesz estimator, some of which we have already mentioned. The most pressing open question is estimation when the model spaces have infinite dimensions. Our theory allows us to show that the Riesz estimator exists in the infinite-dimensional setting, given positivity, and we can compute the Riesz estimator to arbitrary numerical precision, given that the model spaces are separable. However, it is an open question when the Riesz estimator based on infinite-dimensional model spaces is useful in practice. The concern is that the variance of the estimator might be very high, possibly infinite. We know of examples where the Riesz estimator based on infinite-dimensional model spaces is consistent, and we know of examples where it is not. It remains to better delineate these situations, and describe alternatives in the situations where the Riesz estimator, as described here, is not useful. We believe the best candidate for such an alternative is the sieve version of the Riesz estimator discussed at the end of Section 4.3.

Another pressing open question is the behavior of the Riesz estimator when the assumption of correctly specified model spaces does not hold. The fact that the model spaces can be large, possibly infinite-dimensional, means that the assumption of correct specification might be less problematic here than in the conventional setting. But it is nevertheless a strong assumption. It remains to be investigated what properties the Riesz estimator has when the model spaces are misspecified. We conjecture that this investigation will reveal connections to the sieve version of the Riesz estimator. A potential way to address misspecification is to consider a sequence of model spaces, indexed by n, with increasing complexity such that a model space for any finite n might be misspecified, but the sequence is correctly specified in the limit. The sequence of Riesz estimators corresponding to this sequence of model spaces could be interpreted as a type of sieve estimator, suggesting that infinite-dimensional model spaces and misspecification might be two perspectives of the same underlying question.

References

Aronow, P. M., Green, D. P., & Lee, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Annals of Statistics*, *42*(3), 850–871.

Aronow, P. M., & Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, *1*(1).

Aronow, P. M., & Samii, C. (2013). Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology*, *39*(1), 231–241.

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

Aronow, P. M., Samii, C., & Wang, Y. (2021). *Design-based inference for spatial experiments with interference.* (arXiv:2010.13599)

Auerbach, E., & Tabord-Meehan, M. (2021). *The local approach to causal inference under network interference.* (arXiv:2105.03810)

Basse, G., Ding, P., Feller, A., & Toulis, P. (2019). *Randomization tests for peer effects in group formation experiments.* (arXiv:1904.02308)

Blattman, C., Green, D. P., Ortega, D., & Tobón, S. (2021). Place-based interventions at scale: The direct and spillover effects of policing and city services on crime. *Journal of the European Economic Association*, *19*(4), 2022–2051.

Cai, J., Janvry, A. D., & Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, *7*(2), 81–108.

Chen, X., & Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, *9*(1), 39-84.

Chen, X., Liao, Z., & Sun, Y. (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, *178*, 639–658.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, *90*(4), 1501–1535.

Chernozhukov, V., Newey, W. K., & Singh, R. (2022a). Automatic debiased machine learning of causal and structural effects. *Econometrica*, *90*(3), 967–1027.

Chernozhukov, V., Newey, W. K., & Singh, R. (2022b). Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, *25*(3), 576–601.

Chin, A. (2019). Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal of Causal Inference*, *7*(2).

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, *82*(1), 197–228.

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

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, *58*(3), 403–417.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., & Walker, M. (2022). General equilibrium effects of unconditional cash transfers: Experimental evidence from Kenya. *Econometrica*, *in press*.

Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 1035-1056.

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

Fréchet, R. M. (1907). Sur les ensembles de fonctions et les opérations linéaires. *Comptes rendus de l'Académie des Sciences*, *144*, 1414–1416.

Galvao, A. F., & Wang, L. (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, *110*(512), 1528–1542.

Harshaw, C., Middleton, J. A., & Sävje, F. (2021). *Optimized variance estimation under interference and complex experimental designs.* (arXiv:2112.01709)

Harshaw, C., Sävje, F., Spielman, D., & Zhang, P. (2021). *Balancing covariates in randomized experiments with the gram-schmidt walk design.* (arXiv:1911.03071)

Harshaw, C., Sävje, F., Eisenstat, D., Mirrokni, V., & Pouget-Abadie, J. (2021). *Design and analysis of bipartite experiments under a linear exposure-response model.* (arXiv:2103.06392)

Haushofer, J., & Shapiro, J. (2016). The short-term impact of unconditional cash transfers to the poor: Experimental evidence from Kenya. *The Quarterly Journal of Economics*, *131*(4), 1973–2042.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin's statistical family* (pp. 73–84). Chichester: John Wiley & Sons.

Hirshberg, D. A., & Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, *49*(6), 3206–3227.

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.

Hu, Y., Li, S., & Wager, S. (2022). Average direct and indirect causal effects under interference. *Biometrika*, *in print*.

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

Imbens, G., & Menzel, K. (2021). A causal bootstrap. *The Annals of Statistics*, *49*(3), 1460–1488.

Kapelner, A., Krieger, A. M., Sklar, M., Shalit, U., & Azriel, D. (2020). Harmonizing optimized designs with classic randomization in experiments. *American Statistician*, *in print*, 1–12.

Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, *114*(526), 645–656.

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *79*(4), 1229–1245.

Leung, M. P. (2020). Treatment and spillover effects under network interference. *Review of Economics and Statistics*, *102*(2), 368–380.

Leung, M. P. (2022a). Causal inference under approximate neighborhood interference. *Econometrica*, *90*(1), 267–293.

Leung, M. P. (2022b). *Rate-optimal cluster-randomized designs for spatial interference.* (arXiv:2111.04219)

Lewis, G., & Syrgkanis, V. (2021). Double/debiased machine learning for dynamic treatment effects. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 22695–22707). Curran Associates, Inc.

Li, S., & Wager, S. (2022). Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, *50*(4), 2334–2358.

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

Munro, E., Wager, S., & Xu, K. (2021). *Treatment effects in market equilibrium.* (arXiv:2109.11647)

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, *3*, 169–175.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, *62*(6), 1349–1382.

Ogburn, E. L., Sofrygin, O., Diaz, I., & van der Laan, M. J. (2022). Causal inference for social network data. *Journal of the American Statistical Association*, *in print*.

Oster, E., & Thornton, R. (2012). Determinants of technology adoption: Peer effects in menstrual cup take-up. *Journal of the European Economic Association*, *10*(6), 1263–1293.

Papadogeorgou, G., Imai, K., Lyall, J., & Li, F. (2020). *Causal inference with spatio-temporal data: Estimating the effects of airstrikes on insurgent violence in Iraq.* (arXiv:2003.13555)

Pollmann, M. (2020). *Causal inference for spatial treatments.* (arXiv:2011.00373)

Riesz, F. (1907). Sur une espèce de géométrie analytique des systèmes de fonctions sommables. *Comptes rendus de l'Académie des Sciences*, *144*, 1409–1411.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen Tchetgen, E., & van der Vaart, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, *45*(5), 1951–1987.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Ross, N. (2011). Fundamentals of stein's method. *Probability Surveys*, *8*, 210–293.

Rothenhäusler, D., & Yu, B. (2019). *Incremental causal effects.* (arXiv:1907.13258)

Rotnitzky, A., Smucler, E., & Robins, J. M. (2020). Characterization of parameters with a mixed bias property. *Biometrika*, *108*(1), 231–238.

Rudin, W. (1987). *Real and complex analysis.* New York: McGraw-Hill.

Sävje, F. (2021). *Causal inference with misspecified exposure mappings.* (arXiv:2103.06471)

Sävje, F., Aronow, P. M., & Hudgens, M. G. (2021). Average treatment effects in the presence of unknown interference. *Annals of Statistics*, *49*(2), 673–701.

Singh, R. (2021). *A finite sample theorem for longitudinal causal inference with machine learning: Long term, dynamic, and mediated effects.* (arXiv:2112.14249)

Tchetgen Tchetgen, E. J., Fulcher, I. R., & Shpitser, I. (2021). Auto-g-computation of causal effects on a network. *Journal of the American Statistical Association*, *116*(534), 833–844.

VanderWeele, T. J., & Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, *1*(1), 1–20.

Zigler, C. M., & Papadogeorgou, G. (2021). Bipartite causal inference with interference. *Statistical Science*, *36*(1), 109–123.

Supplement

Contents

S1 Positivity and Riesz Representors **54**

S1.1 Positivity and Effect Estimation 54

S1.2 Existence of Riesz Representors 56

S1.3 Constructing Riesz Representors and Verifying Positivity 58

S1.3.1 An Equivalent Definition of Positivity 59

S1.3.2 Constructing an Orthonormal Basis 60

S1.3.3 Test for Positivity and Construction of Riesz Representors 61

S1.3.4 Second Order Positivity and Riesz Representors 62

S2 Analysis of the Riesz Estimator **66**

S2.1 Necessary and Sufficient Conditions for Consistency 66

S2.1.1 A Characterizing Linear Operator 66

S2.1.2 Mean Squared Error Rates thru the Operator Norm 71

S2.2 Sufficient Conditions for Consistency 72

S2.3 Asymptotic Normality of the Riesz Estimator 74

S3 Analysis of the Variance Estimation **77**

S3.1 Consistency of the Variance Estimation 77

S3.2 Asymptotic Validity of the Confidence Intervals 79

S4 Miscellaneous Results **80**

S4.1 Plug-In Interpretation of the Riesz Estimator 80

S4.2 Estimands from Coefficients 82

S1 Positivity and Riesz Representors

S1.1 Positivity and Effect Estimation

We begin by showing that, under the correctly specified model assumption, simple positivity is a necessary condition for unbiased treatment effect estimation while positivity is a sufficient condition.

Proposition 3.1. *Given correctly specified models (Assumption 1), simple positivity (Assumption 2') is a necessary condition for the existence of an unbiased estimator for the estimand* $\tau = n^{-1} \sum_{i=1}^{n} \theta_i(y_i)$*, and positivity (Assumption 2) is a sufficient condition.*

Proof. The fact that Assumption 2 is sufficient condition for unbiased estimation under Assumption 1 follows directly from Theorem 4.3, which shows that the Riesz estimator exists and is unbiased under the two conditions.

Now let us show that under Assumption 1, Assumption 2' is a necessary condition for existence of an unbiased estimator. An effect estimator is a function $T : \mathcal{Z} \times \mathbb{R}^n \to \mathbb{R}$ which is evaluated at the treatment Z and observed outcomes $Y_1, \dots Y_n$. The effect estimator is unbiased if $\mathrm{E}[T(Z, Y_1, \dots Y_n)] = (1/n) \sum_{i=1}^n \theta_i(y_i)$ for all potential outcome functions $y_i \in \mathcal{M}_i$.

We proceed by way of contradiction. At a high level, we will construct two sets of individual potential outcome functions under which the observational distributions are identical, but the estimands are different. This will establish that no unbiased estimator exists.

Suppose that Assumption 2' does not hold so that there exists a unit $k \in [n]$ such that the model space $\mathcal{M}_k$ contains functions u_k and v_k which are observationally equivalent but have different values under the effect functional θ_i. Unpacking this definition, this means that $u_k(Z) = v_k(Z)$ with probability 1 but $\theta_k(u_k) \neq \theta_k(v_k)$.

Consider two collections of individual potential outcome functions $\boldsymbol{y} = (y_1, \dots y_n)$ and $\boldsymbol{y}' = (y_1' \dots y_n')$. We set $y_k = u_k$ and $y_k' = v_k$ and for all $i \neq k$ we set $y_j = y_j'$. Define the aggregated effects

$$\tau = \frac{1}{n} \sum_{i=1}^n \theta_i(y_i) \quad \text{and} \quad \tau' = \frac{1}{n} \sum_{i=1}^n \theta_i(y_i') \enspace .$$

By construction, we have that $\tau \neq \tau'$. Let T be an unbiased estimator for the aggregate treatment effect. Using unbiasedness of the estimator in addition to the equivalence of the observational distributions under $\boldsymbol{y}$ and $\boldsymbol{y}'$, we have that

$$\tau = \mathrm{E}[T(Z, Y_1, \dots Y_n) \mid \boldsymbol{y}] = \mathrm{E}[T(Z, Y_1, \dots Y_n) \mid \boldsymbol{y}'] = \tau' \enspace .$$

We have arrived at a contradiction, as $\tau \neq \tau'$. □

We now present a stronger notion of positivity which is applicable primarily to finite-dimensional model spaces. The definition of *strong positivity* is given below.

Definition S1. A model space $\mathcal{M}_i$ satisfies *strong positivity* if for $u \in \mathcal{M}_i$, $\mathrm{E}[u(Z)^2] = 0$ implies that u is the zero function.

Unlike positivity (Assumption 2) which depends on the linear effect functional, strong positivity as defined above depends only on the model space and the underlying experimental design. If strong positivity holds, then no two distinct functions in the model space are observationally equivalent. To see this, let $u, v \in \mathcal{M}_i$ and observe that if $u \neq v$, then $u - v \neq 0$ so that strong positivity implies that $\mathrm{E}[(u(Z) - v(Z))^2] > 0$ and thus u and v

are not observationally equivalent. Thus, strong positivity implies that positivity holds for every linear functional on the model space. In other words, we have shown the following claim.

Proposition S1.1. *If a finite-dimensional model space $\mathcal{M}_i$ satisfies strong positivity with respect to the experimental design, then every linear functional $\theta_i : \mathcal{M}_i \to \mathbb{R}$ satisfies positivity on the model space $\mathcal{M}_i$ with respect to the design.*

Proof. As discussed by Section 3.5, Assumption 2' is equivalent to saying that all functions $u \in \mathcal{M}_i$ that are almost surely zero on $\mathcal{Z}$ must evaluate as zero on the effect functional: $\theta_i(u) = 0$. If strong positivity holds, then the only function in $\mathcal{M}_i$ that is almost surely zero on $\mathcal{Z}$ is the zero function. By linearity of the the effect functional, $\theta_i(u) = 0$ if u is the zero function. □

To test whether strong positivity holds, it suffices to compute the d-by-d matrix $\boldsymbol{H}$ whose entries are $\boldsymbol{H}_{k,\ell} = \mathrm{E}[g_{i,\ell}(Z)g_{i,k}(Z)]$ and compute whether its smallest eigenvalue is nonzero.

S1.2 Existence of Riesz Representors

In this section, we prove Proposition 4.2, which establishes existence of the Riesz estimators for the effect functionals, under the positivity condition (Assumption 2). This argument works for both finite and infinite dimensional model spaces, although it is non-constructive. In later sections, we will provide methods for explicitly constructing a Riesz representor when the model space is finite dimensional. At the end of this section, we will discuss how a similar technique may be used to prove Proposition 6.2, which establishes existence of second order Riesz representors under the assumption of second-order positivity, which is required for the variance estimator.

Fix a unit $i \in [n]$ with model space $\mathcal{M}_i \subseteq \mathcal{Y}$. We begin by proving Proposition 4.1, which states that $\mathcal{L}(\mathcal{M}_i)$ is a Hilbert space.

Proposition 4.1. *Together with the inner product $\langle [u], [v] \rangle = \mathrm{E}[u(Z)v(Z)]$, each model outcome space $\mathcal{L}(\mathcal{M}_i)$ is a Hilbert space.*

Proof. Recall that $\mathcal{L}(\mathcal{Y})$ is a Hilbert space. By construction, $\mathcal{L}(\mathcal{M}_i)$ is a closed subspace of the Hilbert space $\mathcal{L}(\mathcal{Y})$. Thus, it contains all its limit points and is itself a Hilbert space. □

Our next goal is to extend the linear effect functional θ_i defined on the model space $\mathcal{M}_i$ to a linear function $\theta_i^\dagger$ defined on the outcome model space $\mathcal{L}(\mathcal{M}_i)$. The first step towards this goal is to first show that the effect functional is identical for two observationally equivalent functions.

Lemma S1.2. *Suppose that $u, v \in \mathcal{M}_i$ are observationally equivalent. Under Assumption 2, we have that $\theta_i(u) = \theta_i(v)$.*

Proof. Suppose that $u, v \in \mathcal{M}_i$ are observationally equivalent. This means that $\mathrm{E}[(u(Z) - v(Z))^2] = 0$. The linearity of the effect functional together with the positivity condition and observational equivalence imply that

$$|\theta_i(u) - \theta_i(v)| = |\theta_i(u - v)| \leq C \cdot \mathrm{E}[(u(Z) - v(Z))^2] = 0 \ . \qquad \square$$

Now, we can use the fact that the effect functional is constant on equivalence classes to extend it to a bounded linear functional on the entire model outcome space $\mathcal{L}(\mathcal{M}_i)$.

Lemma S1.3. *Under Assumption 2, there exists a bounded linear functional $\theta_i^\dagger : \mathcal{L}(\mathcal{M}_i) \to \mathbb{R}$ such that $\theta_i^\dagger([u]) = \theta_i(u)$ for all $u \in \mathcal{M}_i$.*

Proof. Recall that $\mathcal{L}(\mathcal{M}_i) = \mathrm{cl}(\mathcal{X}_i)$, where $\mathcal{X}_i = \{[u] : u \in \mathcal{M}_i\}$. Observe that $\mathcal{X}_i$ is a subspace of $\mathcal{L}(\mathcal{Y})$.

We begin by constructing a functional $\theta_i^\dagger : \mathcal{X}_i \to \mathbb{R}$. For each equivalence class $E \in \mathcal{X}_i$, there exists at least one function $u_E \in E \cap \mathcal{M}_i$. Invoking the axiom of choice, let u_E be an arbitrary representative function from $E \cap \mathcal{M}_i$ and define $\theta_i^\dagger(E) = \theta_i(u_E)$. It follows from Lemma S1.2 that $\theta_i^\dagger([u]) = \theta_i(u)$ for all $u \in \mathcal{M}_i$.

We now seek to show that $\theta_i^\dagger$ is a bounded linear functional on $\mathcal{X}_i$. Let E_1 and E_1 be equivalence classes in $\mathcal{L}(\mathcal{M}_i)$, and consider their sum $E_1 + E_2$. Let the representative elements be denoted u_{E_1}, u_{E_2}, and $u_{E_1+E_2}$, respectively. By definition of the quotient space, $u_{E_1} + u_{E_2} \in E_1 + E2$ so that $u_{E_1} + u_{E_2}$ and $u_{E_1+E_2}$ are two equivalent functions in the model space $\mathcal{M}_i$. Thus, Lemma S1.2 together with linearity of the effect functional θ_i implies that

$$\theta_i^\dagger(E_1 + E_2) = \theta_i(u_{E_1+E_2}) = \theta_i(u_{E_1} + u_{E_2}) = \theta_i(u_{E_1}) + \theta_i(u_{E_2}) = \theta_i^\dagger(E_1) + \theta_i^\dagger(E_2) \ .$$

A similar argument shows that for $\theta_i^\dagger(\alpha \cdot E) = \alpha \cdot \theta_i^\dagger(E)$. This establishes that $\theta_i^\dagger$ is a linear functional on $\mathcal{X}_i$. Now, to show that $\theta_i^\dagger$ is bounded, we may apply the positivity condition again to obtain

$$|\theta_i^\dagger(E)| = |\theta_i(u_E)| \leq C\sqrt{\mathrm{E}[u_E(Z)^2]} = C\|E\| \ .$$

Finally, we seek to extend $\theta_i^\dagger$ from a bounded linear functional on the subspace $\mathcal{X}_i$ to a bounded linear functional on $\mathcal{L}(\mathcal{M}_i)$. The Hahn-Banach theorem guarantees that there exists such an bounded extension of $\theta_i^\dagger$ to the entirety of $\mathcal{M}_i$. $\square$

Finally, we are ready to prove Proposition 4.2, which establishes the existence of a Riesz representor $\psi_i \in \mathcal{M}_i$ for a linear functional $\theta_i : \mathcal{M}_i \to \mathbb{R}$ satisfying positivity. For completeness, we re-state the proposition below.

Proposition 4.2. *Given positivity (Assumption 2), there exists an equivalence class $\psi_i^\dagger \in \mathcal{L}(\mathcal{M}_i)$ such that every function $\psi_i \in \psi_i^\dagger$ satisfies*

$$\theta_i(u) = \mathrm{E}[\psi_i(Z)u(Z)] \quad \text{for all} \quad u \in \mathcal{M}_i.$$

We say that any function $\psi_i \in \psi_i^\dagger$ is a Riesz representor for the effect functional θ_i on the model space $\mathcal{M}_i$.

Proof. Let θ_i be the effect functional on $\mathcal{M}_i$. By Lemma S1.3, there exists a bounded linear functional $\theta_i^\dagger : \mathcal{L}(\mathcal{M}_i) \to \mathbb{R}$ such that $\theta_i^\dagger([u]) = \theta_i(u)$ for all $u \in \mathcal{M}_i$. By the Riesz representation theorem, there exists an equivalence class $\psi_i^\dagger \in \mathcal{L}(\mathcal{M}_i)$ such that $\theta_i^\dagger(E) = \langle E, \psi_i^\dagger \rangle$ for all equivalence classes $E \in \mathcal{L}(\mathcal{M}_i)$. We attain the desired result by combining these facts as

$$\theta_i(u) = \theta_i^\dagger([u]) = \langle [u], \psi_i^\dagger \rangle = \mathrm{E}[u(Z)\psi_i(Z)] \ . \qquad \square$$

A nearly identical argument may be used to prove Proposition 6.2. To avoid repeating the argument again, we sketch the high level ideas here. First, one may establish that $\mathcal{L}(\mathcal{M}_i \otimes \mathcal{M}_j)$ is a Hilbert space, which is listed in the main body as Proposition 6.1. When a linear functional $\mathcal{A}_{i,j} : \mathcal{M}_i \otimes \mathcal{M}_j \to \mathbb{R}$ satisfies second order positvity (Definition 4), then it is possible to extend this to a bounded linear functional defined on the Hilbert space $\mathcal{L}(\mathcal{M}_i \otimes \mathcal{M}_j)$. Finally, the Riesz representation theorem is applied, which proves Proposition 6.2.

S1.3 Constructing Riesz Representors and Verifying Positivity

In the previous section, we demonstrated that for an arbitrary model space $\mathcal{M}_i$ which is subspace of measurable functions with bounded second moment $\mathcal{Y}$, there always exists a Riesz representor for a linear functional provided that a positivity condition is satisfied. However, the proof was non-constructive, which means that although the Riesz representor exists, it is not obvious how to evaluate it. Moreover, it is not clear how to computationally or analytically verify that the positivity condition holds in this general setting. Similar concerns arise for the second-order Riesz representors as well.

In this section, we present algorithms for explicitly constructing the Riesz representors and computationally verifying whether the positivity condition holds in the more restricted setting when the model space is finite dimensional. Although this setting is more restricted, we believe that it will be most relevant for the majority of practitioners. To this end, we study finite dimensional vector quotient spaces and demonstrate how to explicitly and concretely verify positivity and construct Riesz representors in an algorithmically efficient manner. At the end of the section, we sketch how these computational methods may be extended to dealing with the second-order Riesz representors used in the the variance estimator.

We assume that each model space $\mathcal{M}_i$ is represented as a collection of d basis functions, $g_{i,1}, \dots g_{i,d}$. That is, each function $u \in \mathcal{M}_i$ admits a unique decomposition $u = \sum_{\ell=1}^{d} \alpha_\ell g_{i,\ell}$. We assume that the experimenter can perform the following three computational primitives:

1. Evaluate the basis function $g_{i,\ell}(z)$ for all $z \in \mathcal{Z}$.
2. Compute cross moments $\mathrm{E}[g_{i,\ell}(Z) g_{i,k}(Z)]$.
3. Evaluate effect functionals $\theta_i(g_{i,\ell})$.

We will measure the computational efficiency of our algorithms by the number of calls to these computational primitives. Using these primitives, we show that the Riesz representor may be explicitly represented as $\psi_i = \sum_{\ell=1}^{d} \beta_{i,\ell} g_{i,\ell}$ and thus evaluated. Additionally, we will show that verifying positivity may be done by checking whether a few carefully chosen linear combinations $c_k = \sum_{\ell=1}^{d} \alpha_{\ell,k} \theta_i(g_{i,\ell})$ are equal to zero.

S1.3.1 An Equivalent Definition of Positivity

In order to better test positivity, we derive a condition which is equivalent, provided that the model spaces are finite dimensional. For a fixed unit $i \in [n]$ with modelspace $\mathcal{M}_i$, we define the *design-null subspace* $\mathcal{S}_i$ as those functions in $\mathcal{M}_i$ that are observationally equivalent to the zero function:

$$\mathcal{S}_i = \{ u \in \mathcal{M}_i : \mathrm{E}[u(Z)^2] = 0 \} \ .$$

The following proposition demonstrates that when the model space is finite dimensional, positivity of a linear effect functional θ_i is equivalent to requiring that $\mathcal{S}_i$ is in the null space (i.e., kernel) of θ_i.

Proposition S1.4. *Suppose that the model space $\mathcal{M}_i$ is finite dimensional. An effect functional $\theta_i : \mathcal{M}_i \to \mathbb{R}$ satisfies positivity if and only if $\theta_i(u) = 0$ for all $u \in \mathcal{S}_i$.*

The usefulness of Proposition S1.4 is that it will admit simpler computational tests than the original definition of positivity. Indeed, the original definition of positivity would seem to require that we iterate over all pairs of observationally equivalent functions in $\mathcal{M}_i$, which seems like a daunting computational task. Instead, Proposition S1.4 states that it is enough to check that all vectors in the subspace $\mathcal{S}_i$ evaluate to zero under the linear functional. In particular, if we can find a finite set of vectors which span $\mathcal{S}_i$, then by linearity of θ_i, it suffices to check only their values under the linear functional. A more complete description of this test is given in the following sections.

S1.3.2 Constructing an Orthonormal Basis

We saw in Section 4.3 that constructing an orthonormal basis for the outcome model space $\mathcal{L}(\mathcal{M}_i)$ was the main computational challenge behind constructing a Riesz representor. Likewise, we saw in Section S1.3.1 that verifying positivity may be reduced to finding a set of vectors which span $\mathcal{S}_i$. In this Section, we show that a modified Gram–Schmidt Orthogonalization procedure may be used to accomplish both tasks at the same time.

The modified Gram–Schmidt Orthogonalization procedure is similar to the typical Gram–Schmidt Orthogonalization procedure, except that it creates two sets of vectors: one will form an orthonormal basis for $\mathcal{L}(\mathcal{M}_i)$ and the other will form a spanning set of $\mathcal{S}_i$. Like the typical Gram–Schmidt procedure, a sequence of functions $\phi_{i,1}, \dots \phi_{i,d}$ is constructed in an iterative fashion from the basis by projecting out previously seen iterates. The main difference in the procedure is that if a function $\phi_{i,\ell}$ has zero "norm", as measured by $\sqrt{\mathrm{E}[u(Z)^2]}$, then it is placed into the spanning set of $\mathcal{S}_i$. The procedure is given formally below as Algorithm S1.

Algorithm S1: Modified Gram–Schmidt Orthogonalization

```
Input  : Basis vectors $g_{i,1}, \dots, g_{i,d}$ for $\mathcal{M}_i$
Output: Functions $\{\phi_{i,\ell}\}_{\ell \in B^{(o)}}$ and $\{\phi_{i,\ell}\}_{\ell \in B^{(z)}}$ for disjoint index sets $B^{(o)}$ and $B^{(z)}$.
for $t = 1, \dots d$ do
    $u_{i,t} \leftarrow g_{i,t} - \sum_{s<t} \mathrm{E}[g_{i,t}(Z)\phi_{i,s}(Z)]\phi_{i,s}$
    if $\mathrm{E}[u_{i,t}(Z)^2] > 0$ then
        $\phi_{i,t} \leftarrow u_{i,t} / \sqrt{\mathrm{E}[u_{i,t}(Z)^2]}$
        $B^{(o)} \leftarrow B^{(o)} \cup \{t\}$
    else
        $\phi_{i,t} \leftarrow u_{i,t}$
        $B^{(z)} \leftarrow B^{(z)} \cup \{t\}$
end
return Functions $\{\phi_{i,\ell}\}_{\ell \in B^{(o)}}$ and $\{\phi_{i,\ell}\}_{\ell \in B^{(z)}}$ for disjoint index sets $B^{(o)}$ and $B^{(z)}$.
```

The following proposition guarantees that the output of the modified Gram–Schmidt procedure produces the desired results.

Proposition S1.5. *The modified Gram–Schmidt Orthogonalization procedure, as described by Algorithm S1, returns two sets of functions the following properties:*

1. *The functions $\{\phi_{i,\ell}\}_{\ell \in B^{(o)}}$ form an orthonormal basis for $\mathcal{L}(\mathcal{M}_i)$.*
2. *The functions $\{\phi_{i,\ell}\}_{\ell \in B^{(z)}}$ span the design-null subspace $\mathcal{S}_i$.*

In the Gram–Schmidt procedure, each of the functions $\phi_{i,1}, \dots \phi_{i,d}$ is expressed as a linear combination of the basis functions:

$$\phi_{i,k} = \sum_{s \leq k} \alpha_{s,k} g_{i,s} \ \ ,$$

where the coefficients $\alpha_{s,k}$ are computed in an iterative way. Using this decomposition, expectations can be re-expressed as

$$\mathrm{E}[g_{i,t}(Z)\phi_{i,s}(Z)] = \sum_{r \leq s} \alpha_{r,k} \, \mathrm{E}[g_{i,t}(Z) g_{i,r}(Z)] \ \ ,$$

so that only the computation of the cross moments $\mathrm{E}[g_{i,t}(Z)g_{i,s}(Z)]$ on the basis functions are required. Running the Gram–Schmidt orthogonalization in this way for a single unit requires pre-computing all $\mathcal{O}(d^2)$ cross moments $\mathrm{E}[g_{i,\ell}(Z)g_{i,k}(Z)]$ as well as $\mathcal{O}(d^3)$ arithmetic operations. Running Gram–Schmidt orthogonalization on the model spaces for all n units requires $\mathcal{O}(nd^3)$ arithmetic operations and computing all $\mathcal{O}(nd^2)$ within-unit cross moments. This will be the dominating cost of the tests for positivity and construction of the individual Riesz representors.

S1.3.3 Test for Positivity and Construction of Riesz Representors

Once the Gram–Schmidt procedure has been run, the test for positivity is relatively easy. It simply requires us to check that the effect functional is zero on the spanning set of $\mathcal{S}_i$. This test is presented formally below as Algorithm S2.

Algorithm S2: Positivity Test

```
Input  : Spanning set $\{\phi_{i,\ell}\}_{\ell \in B^{(z)}}$ of $\mathcal{S}_i$.
Output: Boolean, which tests whether positivity holds
for $k \in B^{(z)}$ do
    if $\theta_i(\phi_{i,k}) \neq 0$ then
        return False
end
return True
```

Recall that each function $\phi_{i,k}$ is represented as the linear combination of the basis functions, i.e. $\phi_{i,k} = \sum_{s \leq k} \alpha_{s,k} g_{i,s}$. By linearity, an evaluation $\theta_i(\phi_{i,k}) = \sum_{s \leq k} \alpha_{s,k} \theta_i(g_{i,s})$ is possible using the evaluation of the effect functional on the basis functions and $\mathcal{O}(d)$ arithmetic operations. Thus, in addition to the computational requirements of the Gram–Schmidt procedure, the test for positivity requires $\mathcal{O}(d^2)$ arithmetic operations and evaluating the effect functional at all $\mathcal{O}(d)$ basis functions. Testing positivity for all n effect functionals requires

$\mathcal{O}(nd^2)$ arithmetic operations and $\mathcal{O}(nd)$ evaluations of all effect functionals on the their basis functions.

Once the Gram–Schmidt procedure has been run, the construction of the Riesz representors is similarly straightforward. It simply requires us to take the weighted sum of the orthogonal basis functions, weighted by their evaluation under the effect functional. This test is presented formally below as Algorithm S2.

```
Algorithm S3: Construction of Riesz Representors
  Input  : Orthonormal basis $\{\phi_{i,\ell}\}_{\ell \in B^{(o)}}$ of $\mathcal{L}(\mathcal{M}_i)$.
  Output: Riesz representor $\psi_i$
  Set $\psi_i \leftarrow \sum_{\ell \in B^{(o)}} \theta_i(\phi_{i,\ell})\phi_{i,\ell}$
  return $\psi_i$
```

Recall that each function $\phi_{i,k}$ is represented as the linear combination of the basis functions, i.e. $\phi_{i,k} = \sum_{s \leq k} \alpha_{s,k} g_{i,s}$. By linearity, an evaluation $\theta_i(\phi_{i,k}) = \sum_{s \leq k} \alpha_{s,k} \theta_i(g_{i,s})$ is possible using the evaluation of the effect functional on the basis functions and $\mathcal{O}(d)$ arithmetic operations. In this way, the result of Algorithm S3 are coefficients $\beta_{i,\ell}$ such that the Riesz representor may be expressed as $\psi_i = \sum_{\ell=1}^{d} \beta_{i,\ell} g_{i,\ell}$.

In addition to the computational requirements of the Gram–Schmidt procedure, the construction of an individual Riesz representor requires $\mathcal{O}(d^2)$ arithmetic operations and evaluating the effect functional at all $\mathcal{O}(d)$ basis functions. Constructing all of the individual Riesz representors requires $\mathcal{O}(nd^2)$ arithmetic operations and $\mathcal{O}(nd)$ evaluations of all effect functionals on the their basis functions. Evaluating all n individual Riesz representors requires $\mathcal{O}(nd)$ evaluations of all basis functions and $\mathcal{O}(nd)$ arithmetic operations.

S1.3.4 Second Order Positivity and Riesz Representors

So far, this section has been concerned with testing positivity and computing individual Riesz representors, which relates primarily to the point estimation. We now discuss how these same computational techniques can be used for testing second-order positivity of linear functionals on the tensor space and constructing second-order Riesz representors. The overall techniques are more or less the same, so we go through the material more quickly.

The key difference here is that instead of working with the model space $\mathcal{M}_i$, we are working with the moment restricted tensor space $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$. The basis for the tensor space will be the simple tensors formed from the bases of the restricted model spaces $\mathcal{M}_i^*$ and $\mathcal{M}_j^*$ respectively, i.e. $g_{i,\ell} \otimes g_{j,k}$ for $\ell, k \in [d]$. In this way, each tensor $\rho \in \mathcal{M}_i^* \otimes \mathcal{M}_j^*$ may be represented uniquely as a linear combination of $g_{i,\ell} \otimes g_{j,k}$ for $\ell, k \in [d]$.

We can define the *second-order design null-space* $\mathcal{S}_{i,j} \subseteq \mathcal{M}_i^* \otimes \mathcal{M}_j^*$ as

$$\mathcal{S}_{i,j} = \{\rho \in \mathcal{M}_i^* \otimes \mathcal{M}_j^* : \mathrm{E}[\rho(Z)^2] = 0\} \ .$$

The benefit of this definition is that it allows for an alternative characterization of second-order positivity, which admits for efficient tests.

Proposition S1.6. *Suppose that the model spaces $\mathcal{M}_i$ and $\mathcal{M}_j$ are finite dimensional. A linear functional $\mathcal{A}_{i,j} : \mathcal{M}_i^* \otimes \mathcal{M}_j^* \to \mathbb{R}$ satisfies second order positivity if and only if $\mathcal{A}_{i,j}(\rho) = 0$ for all $\rho \in \mathcal{S}_{i,j}$.*

Next, we present a modified Gram–Schmidt Orthogonalization procedure applied to the moment restricted tensor space $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$. To distinguish this procedure from the previously introduced Gram–Schmidt orthogonalization procedure (Algorithm S1), we refer to this one as being “second order”, to reflect the fact that it involves pairs of units. Like its first-order counterpart, the second-order orthogonalization procedure will iteratively construct two sets a functions: functions that form an orthonormal basis for $\mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*)$ and a set of functions that will form a spanning set for $\mathcal{S}_{i,j}$. There are two main differences between the first and second order orthogonalization procedures: first, the second order procedure operates on the tensor space, which has d^2 basis vectors. This increases the computational resources required by the algorithm. Second, the functions are expressed as products of basis functions, and so the inner product now requires 4-way expectations to be computed.

The algorithm is given formally below as Algorithm S4.

Algorithm S4: Modified Second Order Gram–Schmidt Orthogonalization

```
Input  : Basis vectors $\{g_{i,\ell}\}_{\ell\in[d]}$ for $\mathcal{M}_i^*$ and $\{g_{j,k}\}_{k\in[d]}$ for $\mathcal{M}_j^*$
Output: Functions $\{\Gamma_{i,j,r}\}_{r\in B^{(o)}}$ and $\{\Gamma_{i,j,r}\}_{r\in B^{(z)}}$ for disjoint index sets $B^{(o)}$ and
        $B^{(z)}$.
Set $r \leftarrow 1$
for $\ell = 1, \ldots, d$ do
    for $k = 1, \ldots, d$ do
        $u_{i,j,r} \leftarrow g_{i,\ell} g_{j,k} - \sum_{s<r} \mathrm{E}[g_{i,\ell}(Z) g_{i,k}(Z) \Gamma_{i,j,s}(Z)] \Gamma_{i,j,s}$
        if $\mathrm{E}[u_{i,j,r}(Z)^2] > 0$ then
            $\Gamma_{i,j,r} \leftarrow u_{i,j,r} / \sqrt{\mathrm{E}[u_{i,j,r}(Z)^2]}$
            $B^{(o)} \leftarrow B^{(o)} \cup \{r\}$
        else
            $\Gamma_{i,j,r} \leftarrow u_{i,j,r}$
            $B^{(z)} \leftarrow B^{(z)} \cup \{r\}$
        $r \leftarrow r + 1$
    end
end
return Functions $\{\Gamma_{i,j,r}\}_{r\in B^{(o)}}$ and $\{\Gamma_{i,j,r}\}_{r\in B^{(z)}}$ for disjoint index sets $B^{(o)}$ and
    $B^{(z)}$.
```

Proposition S1.7. *The modified Second-Order Gram–Schmidt Orthogonalization procedure (Algorithm S4) returns two sets of functions the following properties:*

1. *The functions $\{\Gamma_{i,j,r}\}_{r\in B^{(o)}}$ form an orthonormal basis for $\mathcal{L}(\mathcal{M}_i^* \otimes \mathcal{M}_j^*)$.*
2. *The functions $\{\Gamma_{i,j,r}\}_{r\in B^{(z)}}$ span the design-null subspace $\mathcal{S}_{i,j}$.*

In the second order Gram–Schmidt procedure, each of the functions $\Gamma_{i,j,r}$ is expressed as a linear combination of the product of basis functions:

$$\Gamma_{i,j,r} = \sum_{\ell=1}^{d} \sum_{k=1}^{d} \alpha_{r,\ell,k} g_{i,\ell} g_{j,k} \ ,$$

where the coefficients $\alpha_{r,\ell,k}$ are computed in an iterative way. Using this decomposition, the expectations can be re-expressed as

$$\mathrm{E}[g_{i,\ell}(Z) g_{i,k}(Z) \Gamma_{i,j,s}(Z)] = \sum_{p=1}^{d} \sum_{q=1}^{d} \alpha_{s,p,q} \, \mathrm{E}[g_{i,\ell}(Z) g_{i,k}(Z) g_{i,p}(Z) g_{j,q}(Z)] \ ,$$

so that only the computation of the 4-way expectations of the basis functions is required.

For a single pair of units $i, j \in [n]$, there are $\mathcal{O}(d^4)$ such expectations and they may be precomputed at the beginning of the second order Gram–Schmidt procedure. The procedure itself requires $\mathcal{O}(d^6)$ arithmetic operations. Thus, running the procedure for all pairs of units incurs a computational requirement of $\mathcal{O}(n^2 d^4)$ 4-way expectation computations and $\mathcal{O}(n^2 d^6)$ arithmetic operations.

We now briefly discuss how to use the output of the second order Gram–Schmidt orthogonalization procedure to test for second-order positivity and construct second-order Riesz representors. By Propositions S1.6 and S1.7, a linear functional $\mathcal{A}_{i,j}$ on the moment restricted tensor space $\mathcal{M}_i^* \otimes \mathcal{M}_j^*$ satisfies second order positivity if and only if

$$\mathcal{A}_{i,j}(\Gamma_{i,j,r}) = 0 \quad \text{for all } r \in B^{(z)} \ ,$$

which admits a computationally efficient test, the details of which are nearly identical to Algorithm S2. The linear functional may be evaluated using linearity by decomposing each tensor into its basis decomposition and writing it as

$$\mathcal{A}_{i,j}(\Gamma_{i,j,r}) = \mathcal{A}_{i,j}\Big(\sum_{\ell=1}^{d}\sum_{k=1}^{d} \alpha_{r,\ell,k} g_{i,\ell} \otimes g_{j,k}\Big) = \sum_{\ell=1}^{d}\sum_{k=1}^{d} \alpha_{r,\ell,k} \mathcal{A}_{i,j}(g_{i,\ell} \otimes g_{j,k}) \ .$$

In order to construct the Riesz representor, we can use the explicit construction:

$$\Psi_{i,j}(Z) = \sum_{r \in B^{(o)}} \mathcal{A}_{i,j}(\Gamma_{i,j,r}) \Gamma_{i,j,r}(Z) \ ,$$

which can be further broken down into the basis functions for the tensor space. The details are nearly identical to the Riesz representor construction of Algorithm S3, so we omit them here.

Finally, we demonstrate that the following linear functional on the outcome tensor space always satisfies second order positivity, provided that a moment condition holds.

Lemma S1.8. *Let $\phi_{i,k}$ be a basis function for the moment restricted tensor space $\mathcal{M}_i^*$. The corresponding linear functional $\mathcal{D}_{i,k} : \mathcal{M}_i^* \otimes \mathcal{M}_i^* \to \mathbb{R}$ defined as*

$$\mathcal{D}_{i,k}(\rho) = \mathrm{E}[\phi_{i,k}(Z)^2 \rho(Z)]$$

satisfies second order positivity with respect to the experimental design.

Proof. By Hölder's inequality, we have that

$$\mathrm{E}[\phi_{i,k}(Z)^2 \rho(Z)] \leq \mathrm{E}[\phi_{i,k}(Z)^4]^{1/2} \cdot \mathrm{E}[\rho(Z)^2]^{1/2} \ .$$

Observe that $\phi_{i,k} \in \mathcal{M}_i^*$ so that $\mathrm{E}[\phi_{i,k}(Z)^4]$ is finite. □

S2 Analysis of the Riesz Estimator

S2.1 Necessary and Sufficient Conditions for Consistency

S2.1.1 A Characterizing Linear Operator

In this section, we show that the variance can be characterized by a linear operator $\mathcal{V} : \mathcal{L}(\mathcal{M}_{(n)}) \to \mathcal{L}(\mathcal{M}_{(n)})$ on the product outcome space. We first show a general representation theorem for bilinear forms on a Hilbert space. Next, we show that the variance of the Riesz estimator is a bilinear form on the product outcome space and thus the representation theorem can be applied.

The proposition below gives the representation theorem for symmetric and positive semidefinite bilinear forms on a general Hilbert space.

Proposition S2.1. *Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space and let $V : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be a bilinear form satisfying the following properties:*

- ***Symmetric**: $V(x, y) = V(y, x)$ for all $x, y \in \mathcal{H}$.*
- ***Positive Semi-definite**: $V(x, x) \geq 0$ for all $x \in \mathcal{H}$.*
- ***Bounded:** There exists $C < \infty$ such that $|V(x, y)| \leq C\|x\|\|y\|$ for all $x, y \in \mathcal{H}$*

Then, there exists a bounded linear operator $L : \mathcal{H} \to \mathcal{H}$ such that

$$V(x, x) = \|Lx\|^2 \quad \text{for all } x \in \mathcal{H} \enspace .$$

Proof. Observe that for a fixed $x \in \mathcal{H}$, we have that $y \mapsto V(x, y)$ is a linear functional. Moreover, this linear functional is bounded in the sense that

$$|V(x, y)| \leq C\|x\|\|y\| = C_x\|y\| \enspace ,$$

where we have used the boundedness of the bilinear form V. Thus, by the Riesz representation theorem, there exists a unique vector $f_x \in \mathcal{H}$ such that $V(x, y) = \langle f_x, y \rangle$. Define the operator $A : \mathcal{H} \to \mathcal{H}$ to be the mapping $x \mapsto f_x$, which is easily verified to be linear due to bilinearity of V. Thus, we have the representation

$$V(x, y) = \langle Ax, y \rangle \enspace .$$

We will now show that A is positive semi-definite, symmetric, and bounded. The fact that A is positive semidefinite follows directly from the fact that V is a positive semidefinite form, as $\langle Ax, x \rangle = V(x, x) \geq 0$. The fact that A is symmetric follows from the symmetry of V as

$$\langle Ax, y \rangle = V(x, y) = V(y, x) = \langle Ay, x \rangle \enspace .$$

Finally, the boundedness of A follows from boundedness of V as

$$\|Ax\| = \max_{\|y\|=1} \langle Ax, y \rangle = \max_{\|y\|=1} V(x, y) \leq C\|x\| \ .$$

Because A is a positive semi-definite, self-adjoint, and bounded operator, there exists a bounded linear operator $L : \mathcal{H} \to \mathcal{H}$ such that $A = L^{\mathsf{T}}L$. Now, the claim follows as

$$V(x, x) = \langle Ax, x \rangle = \langle L^{\mathsf{T}}Lx, x \rangle = \langle Lx, Lx \rangle = \|Lx\|^2 \ . \qquad \square$$

In order to apply Proposition S2.1, we need to show that the variance of the Riesz representor may be identified with a positive semi-definite and symmetric bilinear form on the product outcome space satisfying the appropriate conditions. The next lemma demonstrates that a certain bilinear form satisfies these properties.

Lemma S2.2. *Let $\boldsymbol{\psi} \in \mathcal{L}(\mathcal{M}_{(n)})$ be a fixed element of the product outcome space. and define the function $V : \mathcal{L}(\mathcal{M}_{(n)}) \times \mathcal{L}(\mathcal{M}_{(n)}) \to \mathbb{R}$ by*

$$V(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}\Big(u_i(Z)\psi_i(Z), v_j(Z)\psi_j(Z) \Big) \ .$$

Then, V is a bilinear form which is symmetric and positive semi-definite. Moreover, if the supremum of the variance over the $\|\cdot\|_{\text{MS}}$ unit ball is finite

$$\sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq 1} \operatorname{Var}(\widehat{\tau}; \boldsymbol{u}) = C \ ,$$

then the bilinear form V is bounded in the sense that there exists a value $C' < \infty$ such that

$$V(\boldsymbol{u}, \boldsymbol{v}) \leq C' \|\boldsymbol{u}\|_{\text{MS}} \|\boldsymbol{v}\|_{\text{MS}} \quad \textit{for all } \boldsymbol{u}, \boldsymbol{v} \in \mathcal{L}(\mathcal{M}_{(n)}) \ .$$

Proof. Both the bilinearity and the symmetry of V follow from symmetry and bilinearity of the covariance. The positive semi-definiteness of V follows from the non-negativity of variance as

$$V(\boldsymbol{u}, \boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}\Big(u_i(Z)\psi_i(Z), u_j(Z)\psi_j(Z) \Big) = \frac{1}{n} \operatorname{Var}\Big(\sum_{i=1}^{n} u_i(Z)\psi_i(Z) \Big) \geq 0 \ .$$

Now, we show the boundedness condition. By bilinearity, the boundedness condition is immediate if either $\|\boldsymbol{u}\| = 0$ or $\|\boldsymbol{u}\| = 0$, as both sides of the inequality would be zero. Suppose then that $\boldsymbol{u}$ and $\boldsymbol{v}$ are functions in $\mathcal{L}(\mathcal{M}_{(n)})$ with positive norm. Define $\tilde{\boldsymbol{u}} = \boldsymbol{u}/\|\boldsymbol{u}\|_{\text{MS}}$ and $\tilde{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|_{\text{MS}}$, so that $\|\tilde{\boldsymbol{u}}\|_{\text{MS}} = \|\tilde{\boldsymbol{v}}\|_{\text{MS}} = 1$. Using bilinearity and the

Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
V(\boldsymbol{u}, \boldsymbol{v}) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}\Big(u_i(Z)\psi_i(Z), u_j(Z)\psi_j(Z) \Big) \\
&= \frac{1}{n} \operatorname{Cov}\Big(\sum_{i=1}^{n} u_i(Z)\psi_i(Z), \sum_{j=1}^{n} v_j(Z)\psi_j(Z) \Big) \\
&= \frac{1}{n} \|\boldsymbol{u}\|_{\text{MS}} \|\boldsymbol{v}\|_{\text{MS}} \operatorname{Cov}\Big(\sum_{i=1}^{n} \tilde{u}_i(Z)\psi_i(Z), \sum_{j=1}^{n} \tilde{v}_j(Z)\psi_j(Z) \Big) \\
&\leq \frac{1}{n} \|\boldsymbol{u}\|_{\text{MS}} \|\boldsymbol{v}\|_{\text{MS}} \sqrt{ \operatorname{Var}\Big(\sum_{i=1}^{n} \tilde{u}_i(Z)\psi_i(Z) \Big), \operatorname{Var}\Big(\sum_{j=1}^{n} \tilde{v}_j(Z)\psi_j(Z) \Big) } \\
&\leq \frac{1}{n} \|\boldsymbol{u}\|_{\text{MS}} \|\boldsymbol{v}\|_{\text{MS}} \max \left[\operatorname{Var}\Big(\sum_{i=1}^{n} \tilde{u}_i(Z)\psi_i(Z) \Big), \operatorname{Var}\Big(\sum_{j=1}^{n} \tilde{v}_j(Z)\psi_j(Z) \Big) \right] \\
&\leq \frac{C}{n} \cdot \|\boldsymbol{u}\|_{\text{MS}} \|\boldsymbol{v}\|_{\text{MS}} \\
&= C' \cdot \|\boldsymbol{u}\|_{\text{MS}} \|\boldsymbol{v}\|_{\text{MS}} \ ,
\end{aligned}$$

where the final inequality follows from the assumption that the supremum of the variance over the $\|\cdot\|_{\text{MS}}$ unit ball is C. □

We are now ready to prove the existence of the variance characterizing linear operator.

Proposition 5.1. *Given correctly specified model spaces and positivity (Assumptions 1 and 2), either:*

(i) There exists a bounded linear operator $\mathcal{V} : \mathcal{L}(\mathcal{M}_{(n)}) \to \mathcal{L}(\mathcal{M}_{(n)})$ that exactly characterizes the variance of the Riesz estimator in finite samples:

$$\operatorname{Var}\big(\widehat{\tau}; \boldsymbol{u}\big) = \frac{1}{n} \|\mathcal{V}(\boldsymbol{u})\|_{\text{MS}}^2 \qquad \text{for all} \quad \boldsymbol{u} \in \mathcal{M}_{(n)},$$

(ii) Or the variance of the Riesz estimator cannot be characterized with respect to the mean-square norm:

$$\sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq C} \operatorname{Var}(\widehat{\tau}; \boldsymbol{u}) = \infty \qquad \text{for all} \quad C > 0.$$

Proof. Without loss of generality, set $C = 1$. Suppose that the supremum of the variance over the $\|\cdot\|_{\text{MS}}$ unit ball is finite.

Under the correctly specified assumption $\boldsymbol{y} \in \mathcal{L}(\mathcal{M}_{(n)})$, so that

$$\begin{aligned} n \cdot \mathrm{Var}(\widehat{\tau}) &= n \cdot \mathrm{Var}\Big(\frac{1}{n} \sum_{i=1}^{n} y_i(Z)\psi_i(Z)\Big) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(y_i(Z)\psi_i(Z), y_j(Z)\psi_j(Z)) \\ &= V(\boldsymbol{y}, \boldsymbol{y}) \ . \end{aligned}$$

Lemma S2.2 demonstrates that $V : \mathcal{L}(\mathcal{M}_{(n)}) \times \mathcal{L}(\mathcal{M}_{(n)})$ is a bilinear form which is symmetric, positive semidefinite, and bounded in terms of the $\|\cdot\|_{\text{MS}}$ norm in the sense that

$$|V(\boldsymbol{y}, \boldsymbol{y})| \leq C \|\boldsymbol{u}\|_{\text{MS}} \|\boldsymbol{v}\|_{\text{MS}} \ .$$

Thus, we may apply Proposition S2.1, which guarantees the existence of a bounded linear operator $\mathcal{V} : \mathcal{L}(\mathcal{M}_{(n)}) \times \mathcal{L}(\mathcal{M}_{(n)}) \to \mathbb{R}$ for which

$$n \cdot \mathrm{Var}(\widehat{\tau}) = V(\boldsymbol{y}, \boldsymbol{y}) = \|\mathcal{V}(y)\|_{\text{MS}}^2 \ . \qquad \square$$

We now describe how to numerically compute a matrix representation of this operator when the model spaces are finite dimensional. For each outcome model space $\mathcal{L}(\mathcal{M}_i)$, let $\{\phi_{i,\ell}\}_{\ell=1}^{d}$ be an orthonormal basis with respect to the inner product $\langle u, v \rangle = \mathrm{E}[u(Z)v(Z)]$. We can write each of the potential outcome functions in this basis as $y_i = \sum_{\ell=1}^{d} \alpha_{i,\ell}\phi_{i,\ell}$. The following lemma shows how to explicitly construct a matrix whose quadratic form represents the variance.

Proposition S2.3. *Let $\boldsymbol{H}$ be the $(d \cdot n) \times (d \cdot n)$ dimensional symmetric matrix whose entries are given as*

$$\boldsymbol{H}_{(i,\ell),(j,k)} = \mathrm{Cov}(\phi_{i,\ell}(Z)\psi_i(Z), \phi_{j,k}(Z)\psi_j(Z)) \ .$$

Then, the matrix $\boldsymbol{H}$ represents the variance of the Riesz representor in the following sense:

$$n \, \mathrm{Var}(\widehat{\tau}) = \boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{H} \boldsymbol{\alpha} \ .$$

where $\boldsymbol{\alpha} = (\alpha_{1,1}/\sqrt{n}, \dots, \alpha_{1,d}/\sqrt{n}, \dots, \alpha_{n,1}/\sqrt{n}, \dots, \alpha_{n,n}/\sqrt{n})$ is the $n \cdot d$-dimensional vector whose entries are scaled coefficients of the orthogonal basis functions. Furthermore, the following two conditions are satisfied:

1. *The matrix $\boldsymbol{H}$ is positive semidefinite and so admits the decomposition $\boldsymbol{H} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}$. Any such matrix $\boldsymbol{V}$ is a representation of the the operator $\mathcal{V}$ in the orthonormal basis.*

2. The largest eigenvalue of the matrix $\boldsymbol{H}$ *is equal to* $\|\mathcal{V}\|_{op}^2$.

Proof. By decomposing the variance into individual covariance terms and writing the potential outcome functions in their basis, we have that

$$\begin{aligned} n \cdot \mathrm{Var}(\widehat{\tau}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(y_i(Z)\psi_i(Z), y_j(Z)\psi_j(Z)) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\Big(\Big(\sum_{\ell=1}^{d} \alpha_{i,\ell}\phi_{i,\ell}(Z)\Big)\psi_i(Z), \Big(\sum_{k=1}^{d} \alpha_{j,k}\phi_{i,k}(Z)\Big)\psi_j(Z)\Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\ell=1}^{d}\sum_{k=1}^{d} \frac{\alpha_{i,\ell}}{\sqrt{n}}\frac{\alpha_{j,k}}{\sqrt{n}} \mathrm{Cov}(\phi_{i,\ell}(Z)\psi_i(Z), \phi_{i,k}(Z)\psi_j(Z)) \\ &= \boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{H}\boldsymbol{\alpha} \ , \end{aligned}$$

which establishes the first part of the proposition.

Next, observe that the mean squared norm $\|\boldsymbol{y}\|_{\mathrm{MS}}$ is equal to the ℓ_2 norm of the vector $\|\boldsymbol{\alpha}\|_2$.

$$\begin{aligned} \|\boldsymbol{y}\|_{\mathrm{MS}}^2 &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[y_i(Z)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\Big[\Big(\sum_{\ell=1}^{d} \alpha_{i,\ell}\phi_{i,\ell}(Z)\Big)^2\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{d}\sum_{k=1}^{d} \alpha_{i,\ell}\alpha_{i,k}\, \mathrm{E}[\phi_{i,\ell}(Z)\phi_{i,k}(Z)] \\ &= \sum_{i=1}^{n}\sum_{\ell=1}^{d}\Big(\frac{\alpha_{i,\ell}}{\sqrt{n}}\Big)^2 \\ &= \|\boldsymbol{\alpha}\|^2 \ . \end{aligned}$$

Let's show the first bulleted point of the proposition. For any Cholesky factorization $\boldsymbol{H} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}$, we have that $\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{H}\boldsymbol{\alpha} = \|\boldsymbol{V}\boldsymbol{\alpha}\|_2^2$. Because the ℓ_2 norm of the coefficients can be identified with the mean squared norm of the vector-valued potential outcome function, we have that $\|\boldsymbol{V}\boldsymbol{\alpha}\|_2^2 = \|\mathcal{V}(\boldsymbol{y})\|_{\mathrm{MS}}$, so that $\boldsymbol{V}$ is a matrix representation of the operator $\mathcal{V}$.

Next, we show the second bulleted point of the proposition. Recall that the largest eigenvalue is $\lambda_{\max}(\boldsymbol{H}) = \lambda_{\max}(\boldsymbol{L}^{\mathsf{T}}\boldsymbol{L}) = \|\boldsymbol{L}\|_2^2$, where $\|\cdot\|_2$ is the matrix norm induced by the ℓ_2 norm. Because the ℓ_2 norm on the coefficients $\boldsymbol{\alpha}$ can be identified with the mean squared norm of the vector-valued potential outcome function,$\|\boldsymbol{L}\|_2 = \|\mathcal{V}\|_{op}$. This establishes that

$\lambda_{\max}(\boldsymbol{H}) = \|\mathcal{V}\|_{op}$. □

S2.1.2 Mean Squared Error Rates thru the Operator Norm

Corollary 5.2. *Given correctly specified model spaces and positivity (Assumptions 1 and 2), the worst-case root mean square error of the Riesz estimator in any mean-square norm ball is the scaled product of the radius of the ball and the operator norm of $\mathcal{V}$:*

$$\sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq C} \sqrt{\mathrm{E}\big[(\widehat{\tau} - \tau)^2; \boldsymbol{u}\big]} = n^{-1/2} C \|\mathcal{V}\|_{op} \qquad \textit{for all} \quad C > 0.$$

If the variance characterizing operator does not exists, set $\|\mathcal{V}\|_{op} = \infty$.

Proof. The proof follows from direct applications of Theorem 4.3 which guarantees unbiasedness of the Riesz representor, Proposition 5.1 which proves existence of the variance characterizing operator, and the definition of the operator norm:

$$\sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq C} \sqrt{n \cdot \mathrm{E}[(\widehat{\tau} - \tau)^2; \boldsymbol{u}]} = \sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq C} \sqrt{n \cdot \mathrm{Var}(\widehat{\tau}; \boldsymbol{u})} = \sup_{\|\boldsymbol{u}\|_{\text{MS}} \leq C} \|\mathcal{V}(\boldsymbol{u})\|_{\text{MS}} = C \cdot \|\mathcal{V}\|_{op} \ . \ \square$$

Corollary 5.3. *If Assumptions 1 and 2 hold and the mean-square norm of the potential outcomes is asymptotically bounded, $\|\boldsymbol{y}\|_{\text{MS}} = \mathcal{O}(1)$, then a necessary and sufficient condition for consistency in mean square of the Riesz estimator is $\|\mathcal{V}\|_{op} = o(n^{1/2})$. Furthermore, the rate of convergence of the Riesz estimator is exactly the rate at which $n^{-1/2}\|\mathcal{V}\|_{op}$ approaches zero. A necessary and sufficient condition for root-n consistency of the Riesz estimator is $\|\mathcal{V}\|_{op} = \mathcal{O}(1)$.*

Proof. Corollary 5.2 demonstrates that the root mean squared error of the Riesz estimator may be bounded as

$$\sqrt{\mathrm{E}[(\widehat{\tau} - \tau)^2]} \leq \frac{1}{\sqrt{n}} \|\mathcal{V}\|_{op} \cdot \|\boldsymbol{y}\|_{\text{MS}} \ .$$

Because the sequence of potential outcome functions is asymptotically bounded in the $\|\cdot\|_{\text{MS}}$, we have that $\|\boldsymbol{y}\|_{\text{MS}} = \mathcal{O}(1)$ so that the standardized mean squared error is on the order of

$$\sqrt{\mathrm{E}[(\widehat{\tau} - \tau)^2]} = \mathcal{O}(n^{-1/2}\|\mathcal{V}\|_{op}) \ .$$

Thus, $\|\mathcal{V}\|_{op} = o(n^{1/2})$ is sufficient for convergence in mean squared error for all sequences which are asymptotically bounded with respect to the $\|\cdot\|_{\text{MS}}$ norm. Moreover, for some such sequences, the inequalities above hold with equality, and so the condition $\|\mathcal{V}\|_{op} = o(n^{1/2})$ is also necessary for consistency of sequences in this class.

A similar calculation shows that $\|\mathcal{V}\|_{op} = \mathcal{O}(1)$ is both necessary and sufficient for $\sqrt{n}$-convergence in mean squared error. □

S2.2 Sufficient Conditions for Consistency

In this section, we prove the sufficient conditions for consistency of the Riesz estimator, which appear in Section 5.3. We begin by deriving an upper bound on the variance of an individual Riesz estimator in terms of the moments of the potential outcome function and the individual Riesz representor.

Lemma S2.4. *Let p and q be values satisfying $1/p + 1/q = 1/2$. The variance of an individual treatment estimator is bounded as*

$$\mathrm{Var}(\widehat{\tau}_i) \leq \Big(\mathrm{E}[|y_i(Z)|^p]^{1/p} \cdot \mathrm{E}[|\psi_i(Z)|^q]^{1/q} \Big)^2 \ .$$

Proof. We can upper bound the variance by the raw second moment:

$$\mathrm{Var}(\widehat{\tau}_i) = \mathrm{E}[\widehat{\tau}_i^2] - \mathrm{E}[\widehat{\tau}_i]^2 \leq \mathrm{E}[\widehat{\tau}_i^2] = \mathrm{E}[|y_i(Z)\psi_i(Z)|^2] \ .$$

Define $p' = p/2$ and $q' = q/2$ and observe that p' and q' are conjugate pairs as

$$\frac{1}{p'} + \frac{1}{q'} = \frac{2}{p} + \frac{2}{q} = 2 \cdot \Big(\frac{1}{p} + \frac{1}{q} \Big) = 2 \cdot \frac{1}{2} = 1 \ .$$

Thus, we may use Hölder's inequality to obtain that

$$\begin{aligned} \mathrm{Var}(\widehat{\tau}_i) &\leq \mathrm{E}[|y_i(Z)\psi_i(Z)|^2] \\ &\leq \mathrm{E}[|y_i(Z)|^{2p'}]^{1/p'} \cdot \mathrm{E}[|\psi_i(Z)|^{2q'}]^{1/q'} \\ &= \Big(\mathrm{E}[|y_i(Z)|^{2p'}]^{1/2p'} \cdot \mathrm{E}[|\psi_i(Z)|^{2q'}]^{1/2q'} \Big)^2 \\ &= \Big(\mathrm{E}[|y_i(Z)|^p]^{1/p} \cdot \mathrm{E}[|\psi_i(Z)|^q]^{1/q} \Big)^2 \ . \end{aligned}$$

□

Next, we use the bounds on the variance of the individual Riesz estimator together with the dependency neighborhood conditions to obtain a bound on the variance of the Riesz estimator.

Lemma S2.5. *Let p and q be values satisfying $1/p + 1/q = 1/2$. The variance of the Riesz estimator is bounded above by*

$$\mathrm{Var}(\widehat{\tau}) \leq \frac{1}{n^2} \sum_{i=1}^{n} |\mathcal{D}_i| \Big(\mathrm{E}[|y_i(Z)|^p]^{1/p} \cdot \mathrm{E}[|\psi_i(Z)|^q]^{1/q} \Big)^2 \ .$$

Proof. We begin by decomposing the variance as

$$\operatorname{Var}(\widehat{\tau}) = \operatorname{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}(\widehat{\tau}_i, \widehat{\tau}_j) \ ,$$

By definition of the dependency neighborhoods, if $i \notin \mathcal{D}_j$ or $j \notin \mathcal{D}_i$, then τ_i and τ_j are independent, so that $\operatorname{Cov}(\widehat{\tau}_i, \widehat{\tau}_j) = 0$. Thus, we have the identity $\operatorname{Cov}(\widehat{\tau}_i, \widehat{\tau}_j) = \mathbb{1}[i \in \mathcal{D}_j]\mathbb{1}[j \in \mathcal{D}_i]\operatorname{Cov}(\widehat{\tau}_i, \widehat{\tau}_j)$. Substituting this identity into the variance calculation and using Cauchy-Schwarz inequality together with Lemma S2.4, we obtain that

$$\begin{aligned}
\operatorname{Var}(\widehat{\tau}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}[i \in \mathcal{D}_j]\mathbb{1}[j \in \mathcal{D}_i]\operatorname{Cov}(\widehat{\tau}_i, \widehat{\tau}_j) && \text{(indicator identity)}\\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}[i \in \mathcal{D}_j]\mathbb{1}[j \in \mathcal{D}_i]\sqrt{\operatorname{Var}(\widehat{\tau}_i)\operatorname{Var}(\widehat{\tau}_j)} && \text{(Cauchy-Schwarz)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{\mathbb{1}[j \in \mathcal{D}_i]\operatorname{Var}(\widehat{\tau}_i)\cdot\mathbb{1}[i \in \mathcal{D}_j]\operatorname{Var}(\widehat{\tau}_j)} \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\Big(\mathbb{1}[j \in \mathcal{D}_i]\operatorname{Var}(\widehat{\tau}_i) + \mathbb{1}[i \in \mathcal{D}_j]\operatorname{Var}(\widehat{\tau}_j)\Big) && \text{(AM-GM inequality)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}|\mathcal{D}_i|\operatorname{Var}(\widehat{\tau}_i) && \text{(collecting terms)}\\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}|\mathcal{D}_i|\Big(\mathrm{E}[|y_i(Z)|^p]^{1/p}\cdot\mathrm{E}[|\psi_i(Z)|^q]^{1/q}\Big)^2 && \text{(Lemma S2.4)} \ .\square
\end{aligned}$$

Finally, we are ready to use the previous lemma to derive a finite sample bound on the variance of the Riesz estimator.

Proposition 5.4. *Let p and q be values satisfying $1/p + 1/q = 1/2$. The root mean square error of the Riesz estimator is upper bounded by*

$$\sqrt{\mathrm{E}\big[(\widehat{\tau}-\tau)^2\big]} \leq n^{-1/2} d_{\mathrm{avg}}^{1/2}\|\boldsymbol{y}\|_{\max,p}\|\boldsymbol{\psi}\|_{\max,q},$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ is the vector of all Riesz representors and $\|\cdot\|_{\max,p}$ is the max-p norm discussed in Section 5.1.

Proof. Using Lemma S2.5 and the definition of the max-p norms, we have that the variance of the Riesz estimator may be bounded as

$$\begin{aligned}
\mathrm{Var}(\widehat{\tau}) &\leq \frac{1}{n^2}\sum_{i=1}^{n}|\mathcal{D}_i|\Big(\mathrm{E}[|y_i(Z)|^p]^{1/p}\cdot\mathrm{E}[|\psi_i(Z)|^q]^{1/q}\Big)^2 \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}|\mathcal{D}_i|\big(\|\boldsymbol{y}\|_{\max,p}\|\boldsymbol{\psi}\|_{\max,q}\big)^2 \\
&= \frac{1}{n}\Big(\frac{1}{n}\sum_{i=1}^{n}|\mathcal{D}_i|\Big)\Big(\|\boldsymbol{y}\|_{\max,p}\|\boldsymbol{\psi}\|_{\max,q}\Big)^2 \\
&= \frac{d_{\mathrm{avg}}}{n}\Big(\|\boldsymbol{y}\|_{\max,p}\|\boldsymbol{\psi}\|_{\max,q}\Big)^2 \enspace .
\end{aligned}$$

The proof follows by rearranging terms. □

Finally, the following corollary re-interprets the finite sample bound above in an asymptotic setting.

Corollary 5.5. *Given correctly specified model spaces and positivity (Assumptions 1 and 2), limited average model dependence, $d_{\mathrm{avg}} = o(n)$, and $\|\boldsymbol{y}\|_{\max,p} = \mathcal{O}(1)$ and $\|\boldsymbol{\psi}\|_{\max,q} = \mathcal{O}(1)$ for some $1/p+1/q = 1/2$, the Riesz estimator is consistent in mean square. If the condition on the average model dependence is strengthened to $d_{\mathrm{avg}} = \mathcal{O}(1)$, the Riesz estimator is root-n consistent.*

Proof. Observe that under Assumptions 1 and 2, the Riesz estimator is unbiased so that its mean squared error is equal to its variance. In other words, we have that

$$\sqrt{\mathrm{E}[(\widehat{\tau}-\tau)^2]} = \sqrt{\mathrm{Var}(\widehat{\tau})} \leq \sqrt{\frac{d_{\mathrm{avg}}}{n}}\cdot\|\boldsymbol{y}\|_{\max,p}\|\boldsymbol{\psi}\|_{\max,q} \enspace .$$

By assumption, $\|\boldsymbol{y}\|_{\max,p}$ and $\|\boldsymbol{\psi}\|_{\max,q}$ are asymptotically bounded as $\mathcal{O}(1)$ and so the normalized mean squared error is on the order of $\sqrt{\mathrm{E}[(\widehat{\tau}-\tau)^2]} = \mathcal{O}(\sqrt{d_{\mathrm{avg}}/n})$. Thus, the estimator is consistent in mean square if $d_{\mathrm{avg}} = o(n)$.

A similar calculation shows that $d_{\mathrm{avg}} = \mathcal{O}(1)$ guarantees $\sqrt{n}$-convergence in mean squared error, i.e. $\sqrt{n\cdot\mathrm{E}[(\widehat{\tau}-\tau)^2]} = \mathcal{O}(1)$. □

S2.3 Asymptotic Normality of the Riesz Estimator

In this section, we prove that the Riesz estimator is asymptotically normal under moment conditions on the outcomes and Riesz representors and a bound on the dependency between model spaces. Our technique will be the dependency graph version of Stein's method.

We begin by re-defining the notion of dependency neighborhoods used in this literature on Stein's method (Ross, 2011). Let $A_1, \ldots A_n$ be random variables indexed by integers

$[n]$. For each index $i \in [n]$, we define the *dependency neighborhood* to be the smallest set $\mathcal{D}_i \subset [n]$ such that

$$A_i \text{ is jointly independent of the variables } \{A_j : j \in [n] \setminus \mathcal{D}_i\} \ .$$

In the main paper, we introduce the notion of dependency neighborhoods at the level of the model spaces, but they induce dependency neighborhoods of random variables of the form $u_i(Z)$ for $u_i \in \mathcal{M}_i$

The following lemma is a finite sample bound on the Wassterstein distance between the normalized sum of random variables $A_1, \dots A_n$ and a normal distribution.

Lemma S2.6 (Lemma 3.6 of Ross (2011))**.** *Let $A_1, A_2, \dots A_n$ be random variables such that $\mathrm{E}[A_i^4] < \infty$, $\mathrm{E}[A_i] = 0$. Define $S = \frac{1}{n}\sum_{i=1}^n A_i$ and define $\sigma^2 = \mathrm{Var}(S)$, and define $X = S/\sigma$. Then for a standard normal $B \sim \mathcal{N}(0,1)$, we have*

$$d_W(X, B) \leq \frac{d_{\max}^2}{\sigma^3 n^3} \sum_{i=1}^n \mathrm{E}[|A_i|^3] + \sqrt{\frac{28}{\pi}} \cdot \frac{d_{\max}^{3/2}}{n^2 \sigma^2} \sqrt{\sum_{i=1}^n \mathrm{E}[A_i^4]} \ ,$$

where $d_{\max} = \max_{i \in [n]} |\mathcal{D}_i|$ is the maximum dependency degree of the random variables and $d_W(\cdot, \cdot)$ is the Wasserstein distance.

We will be considering the random variables $A_i = \widehat{\tau}_i - \tau_i$, which are the errors of the individual treatment effect estimates. To this end, we need to bound the pth moments of the absolute error of the individual treatment effect estimates. We begin with the following lemma which holds for general random variables.

Lemma S2.7. *For $1 \leq p \leq \infty$ and a random variable X, we have that*

$$\mathrm{E}\Big[\big|X - \mathrm{E}[X]\big|^p\Big] \leq 2^p \, \mathrm{E}\Big[|X|^p\Big] \ .$$

Proof. We may use Minkowski's inequality together with Jensen's inequality applied to $x \mapsto |x|^p$, we have that

$$\begin{aligned} \mathrm{E}\Big[\big|X - \mathrm{E}[X]\big|^p\Big]^{1/p} &\leq \mathrm{E}[|X|^p]^{1/p} + (|\mathrm{E}[X]|^p)^{1/p} && \text{(Minkowski's inequality)} \\ &\leq \mathrm{E}[|X|^p]^{1/p} + \mathrm{E}[|X|^p]^{1/p} && \text{(Jensen's inequality)} \\ &= 2\,\mathrm{E}[|X|^p]^{1/p} \ . && \square \end{aligned}$$

Next, we use this lemma together with Hölder's inequality to obtain a bound on the pth moments of absolute error of the individual treatment effect estimates.

Lemma S2.8. *Fix $1 \leq r \leq \infty$ and let p, q be value satisfying $1/p + 1/q = 1/r$. Under Assumption 1, the rth moment of the absolute error of the individual treatment effect estimate is*

$$\mathrm{E}\Big[|\widehat{\tau}_i - \tau_i|^r\Big] \leq \Big(2\,\mathrm{E}\Big[|y_i(Z)|^p\Big]^{1/p} \cdot \mathrm{E}\Big[|\psi_i(Z)|^q\Big]^{1/q}\Big)^r \ .$$

Proof. Define $p' = p/r$ and $q' = q/r$ and observe that p' and q' are conjugate pairs as

$$\frac{1}{p'} + \frac{1}{q'} = \frac{r}{p} + \frac{r}{q} = r \cdot \Big(\frac{1}{p} + \frac{1}{q}\Big) = r \cdot \frac{1}{r} = 1 \ .$$

Under Assumption 1, we have that $\mathrm{E}[\widehat{\tau}_i] = \tau_i$. Thus, we may apply Lemma S2.7 and obtain

$$\begin{aligned}
\mathrm{E}\Big[|\widehat{\tau}_i - \tau_i|^r\Big] &\leq 2^r\,\mathrm{E}\Big[|\widehat{\tau}_i|^r\Big] && \text{(Lemma S2.7)}\\
&= 2^r\,\mathrm{E}\Big[|y_i(Z)\psi_i(Z)|^r\Big] && \text{(definition of } \widehat{\tau}_i\text{)}\\
&\leq 2^r\,\mathrm{E}\Big[|y_i(Z)|^{rp'}\Big]^{1/p'} \cdot \mathrm{E}\Big[|\psi_i(Z)|^{rq'}\Big]^{1/q'} && \text{(Hölder's Inequality)}\\
&= \Big(2\,\mathrm{E}\Big[|y_i(Z)|^{rp'}\Big]^{1/rp'} \cdot \mathrm{E}\Big[|\psi_i(Z)|^{rq'}\Big]^{1/rq'}\Big)^r\\
&= \Big(2\,\mathrm{E}\Big[|y_i(Z)|^{p}\Big]^{1/p} \cdot \mathrm{E}\Big[|\psi_i(Z)|^{q}\Big]^{1/q}\Big)^r \ . \qquad \Box
\end{aligned}$$

Finally, we are ready to prove Proposition 5.6 asymptotic normality of the Riesz estimator under moment conditions and the assumption of limited dependence. We restate the proposition below for completeness.

Proposition 5.6. *Given correctly specified model spaces, positivity and non-degeneracy (Assumptions 1, 2 and 4), limited maximum model dependence, $d_{\max} = o(n^{1/4})$, and $\|\boldsymbol{y}\|_{\max,p} = \mathcal{O}(1)$ and $\|\boldsymbol{\psi}\|_{\max,q} = \mathcal{O}(1)$ for some $1/p + 1/q = 1/4$, the limiting distribution of the Riesz estimator is normal:*

$$\frac{\widehat{\tau} - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau})}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof. We seek to use Lemma S2.6 on the random variables $A_i = \widehat{\tau}_i - \tau_i$, which are the errors of the individual treatment effect estimates. Note that in this case, $S = (1/n)\sum_{i=1}^n A_i = \widehat{\tau} - \tau$, $\sigma = \sqrt{\mathrm{Var}(\widehat{\tau})}$, and $X = \frac{\widehat{\tau} - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau})}}$.

First, let us verify that condition required for Lemma S2.6 hold. By Assumption 1, the individual treatment effect estimators are unbiased so that $\mathrm{E}[A_i] = \mathrm{E}[\widehat{\tau}_i - \tau_i] = 0$. By Lemma S2.8 together with asymptotic boundedness of $\|\boldsymbol{y}\|_{\max,p}$ and $\|\boldsymbol{\psi}\|_{\max,q}$ yields

$$\mathrm{E}[A_i^4]^{1/4} = \mathrm{E}[(\widehat{\tau}_i - \tau_i)^4]^{1/4} \leq 2\,\mathrm{E}\Big[|y_i(Z)|^p\Big]^{1/p} \cdot \mathrm{E}\Big[|\psi_i(Z)|^q\Big]^{1/q} \leq 2\|\boldsymbol{y}\|_{\max,p}\|\boldsymbol{\psi}\|_{\max,q} < \infty \ .$$

Using Lemma S2.6, we have that for $B \sim \mathcal{N}(0,1)$,

$$d_W(X,B) \leq \frac{d_{\max}^2}{\sigma^3 n^3} \sum_{i=1}^n \mathrm{E}[|\widehat{\tau}_i - \tau_i|^3] + \sqrt{\frac{28}{\pi}} \cdot \frac{d_{\max}^{3/2}}{n^2 \sigma^2} \sqrt{\sum_{i=1}^n \mathrm{E}[|\widehat{\tau}_i - \tau_i|^4]} \enspace .$$

Using Lemma S2.8 for $r = 3$, the first sum may be bounded as

$$\sum_{i=1}^n \mathrm{E}[|\widehat{\tau}_i - \tau_i|^3] \leq \sum_{i=1}^n \Big(2\, \mathrm{E}\Big[|y_i(Z)|^p\Big]^{1/p} \cdot \mathrm{E}\Big[|\psi_i(Z)|^q\Big]^{1/q} \Big)^3 \leq n(2\|\boldsymbol{y}\|_{\max,p} \|\boldsymbol{\psi}\|_{\max,q})^3 \enspace .$$

Similarly, using Lemma S2.8 with $r = 4$, the second sum may be bounded as

$$\sum_{i=1}^n \mathrm{E}[|\widehat{\tau}_i - \tau_i|^4] \leq \sum_{i=1}^n \Big(2\, \mathrm{E}\Big[|y_i(Z)|^p\Big]^{1/p} \cdot \mathrm{E}\Big[|\psi_i(Z)|^q\Big]^{1/q} \Big)^4 \leq n(2\|\boldsymbol{y}\|_{\max,p} \|\boldsymbol{\psi}\|_{\max,q})^4 \enspace .$$

By assumption, the moment quantities $\|\boldsymbol{y}\|_{\max,p}$ and $\|\boldsymbol{\psi}\|_{\max,q}$ are asymptotically bounded so the Wasserstein distance is bounded by

$$d_W(X,B) = \mathcal{O}\Big(\frac{d_{\max}^2}{\sigma^3 n^2} + \frac{d_{\max}^{3/2}}{\sigma^2 n^{3/2}} \Big) \enspace .$$

By Assumption 4, the variance is bounded below as $\mathrm{Var}(\widehat{\tau}) = \sigma^2 \geq \Omega(1/n)$ and so the Wasserstein distance is bounded by

$$d_W(X,B) = \mathcal{O}\Big(\frac{d_{\max}^2}{n^{1/2}} + \frac{d_{\max}^{3/2}}{n^{1/2}} \Big) = \mathcal{O}\Big(\frac{d_{\max}^2}{n^{1/2}} \Big) \enspace .$$

By assumption, $d_{\max} = o(n^{1/4})$, so that the Wasserstein distance goes to zero. □

S3 Analysis of the Variance Estimation

S3.1 Consistency of the Variance Estimation

In this section, we prove Proposition 6.6, which establishes that the normalized variance estimator is mean squared consistent for the normalized variance bound.

Proposition 6.6. *Let $\boldsymbol{\Psi}^{\mathrm{VB}} = (\Psi_{1,1}^{\mathrm{VB}}, \dots, \Psi_{n,n}^{\mathrm{VB}})$ be a vector collecting all Riesz representors for the variance bound functionals $\mathcal{B}_{i,j}$. Given correctly specified model spaces, first-order positivity, separability and bounded fourth moments (Assumptions 1, 2, 3, and 5), that $\|\boldsymbol{y}\|_{\max,p}$ and $\|\boldsymbol{\Psi}^{\mathrm{VB}}\|_{\max,q}$ are asymptotically bounded for $p \geq 4$ and $q \geq 2$ such that $1/p +$*

$1/2q = 1/4$, and that $s_{\text{avg}} = o(1)$, the normalized variance estimator is consistent of the normalized variance bound: $\mathrm{E}[(n\widehat{\mathrm{VB}}(\widehat{\tau}) - n\mathrm{VB}(\widehat{\tau}))^2] = o(1)$.

Proof. Write the variance of the n-normalized variance estimator as

$$\mathrm{Var}(n\widehat{\mathrm{VB}}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\sum_{s=1}^{n}\mathrm{Cov}\big(Y_iY_j\Psi_{i,j}^{\text{VB}}(Z), Y_rY_s\Psi_{r,s}^{\text{VB}}(Z)\big).$$

Note that if $(r,s) \notin \mathcal{S}_{i,j}$, then the corresponding covariance term is zero, meaning that

$$\mathrm{Var}(n\widehat{\mathrm{VB}}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{(r,s)\in\mathcal{S}_{i,j}}\mathrm{Cov}\big(Y_iY_j\Psi_{i,j}^{\text{VB}}(Z), Y_rY_s\Psi_{r,s}^{\text{VB}}(Z)\big).$$

Using the Cauchy–Schwarz inequality, followed by the AM–GM inequality, and finally the fact that the second raw moment bounds the variance, we have

$$\begin{aligned}\mathrm{Cov}\big(Y_iY_j\Psi_{i,j}^{\text{VB}}(Z), Y_rY_s\Psi_{r,s}^{\text{VB}}(Z)\big) &\leq \sqrt{\mathrm{Var}\big(Y_iY_j\Psi_{i,j}^{\text{VB}}(Z)\big)\,\mathrm{Var}\big(Y_rY_s\Psi_{r,s}^{\text{VB}}(Z)\big)} \\ &\leq \frac{1}{2}\,\mathrm{Var}\big(Y_iY_j\Psi_{i,j}^{\text{VB}}(Z)\big) + \frac{1}{2}\,\mathrm{Var}\big(Y_rY_s\Psi_{r,s}^{\text{VB}}(Z)\big) \\ &\leq \frac{1}{2}\,\mathrm{E}\big[Y_i^2Y_j^2(\Psi_{i,j}^{\text{VB}}(Z))^2\big] + \frac{1}{2}\,\mathrm{E}\big[Y_r^2Y_s^2(\Psi_{r,s}^{\text{VB}}(Z))^2\big].\end{aligned}$$

For some $p > 1$ and $q > 1$ such that $2/p + 1/q = 1$, use Hölder's inequality to write

$$\mathrm{E}\big[Y_i^2Y_j^2(\Psi_{i,j}^{\text{VB}}(Z))^2\big] \leq \mathrm{E}\big[Y_i^{2p}\big]^{2/2p}\,\mathrm{E}\big[Y_j^{2p}\big]^{2/2p}\,\mathrm{E}\big[(\Psi_{i,j}^{\text{VB}}(Z))^{2q}\big]^{2/2q} \leq \|\boldsymbol{y}\|_{\max,2p}^4\|\boldsymbol{\Psi}^{\text{VB}}\|_{\max,2q}^2$$

and similarly for indices r and s. It follows that

$$\mathrm{Cov}\big(Y_iY_j\Psi_{i,j}^{\text{VB}}(Z), Y_rY_s\Psi_{r,s}^{\text{VB}}(Z)\big) \leq \|\boldsymbol{y}\|_{\max,2p}^4\|\boldsymbol{\Psi}^{\text{VB}}\|_{\max,2q}^2,$$

and

$$\begin{aligned}\mathrm{Var}(n\widehat{\mathrm{VB}}) &\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{(r,s)\in\mathcal{S}_{i,j}}\|\boldsymbol{y}\|_{\max,2p}^4\|\boldsymbol{\Psi}^{\text{VB}}\|_{\max,2q}^2 \\ &= \|\boldsymbol{y}\|_{\max,2p}^4\|\boldsymbol{\Psi}^{\text{VB}}\|_{\max,2q}^2\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\mathcal{S}_{i,j}| = s_{\text{avg}}\|\boldsymbol{y}\|_{\max,2p}^4\|\boldsymbol{\Psi}^{\text{VB}}\|_{\max,2q}^2.\end{aligned}$$

By assumption, $\|\boldsymbol{y}\|_{\max,2p}$ and $\|\boldsymbol{\Psi}^{\text{VB}}\|_{\max,2q}$ are asymptotically bounded, meaning that the variance of the normalized variance estimator is $\mathrm{Var}(n\widehat{\mathrm{VB}}) = \mathcal{O}(s_{\text{avg}})$. From $s_{\text{avg}} = o(1)$, we have $\mathrm{Var}(n\widehat{\mathrm{VB}}) = o(1)$. The proposition then follows from the fact that $\mathrm{E}[n\widehat{\mathrm{VB}}] = n\mathrm{VB}$. □

Using the non-degeneracy assumption, we can show that the ratio of the variance estimator and the variance bound goes to 1 in probability. This is an important lemma for the asymptotic validity of the normal-based confidence intervals.

Lemma S3.1. *Under the conditions of Proposition 6.6 and the non-degeneracy assumption (Assumption 4), we have that the ratio of the variance estimator and the variance bound converge to 1 in probability:* $(\widehat{\mathrm{VB}}/\mathrm{VB}) \xrightarrow{p} 1$.

Proof. Fix $\epsilon > 0$. Note that $\mathrm{E}[\widehat{\mathrm{VB}}] = \mathrm{VB}$ so by Chebyshev's inequality, we have

$$\Pr\Big(\Big|1 - \frac{\widehat{\mathrm{VB}}}{\mathrm{VB}}\Big| > \epsilon\Big) \leq \frac{1}{\epsilon^2}\operatorname{Var}\Big(\frac{\widehat{\mathrm{VB}}}{\mathrm{VB}}\Big) = \frac{1}{\epsilon^2}\operatorname{Var}\Big(\frac{n \cdot \widehat{\mathrm{VB}}}{n \cdot \mathrm{VB}}\Big) = \frac{1}{\epsilon^2}\frac{\operatorname{Var}(n \cdot \widehat{\mathrm{VB}})}{(n \cdot \mathrm{VB})^2} \ .$$

By Proposition 6.3, we have that $\operatorname{Var}(\widehat{\tau}) \leq \mathrm{VB}$. Additionally, the non-degeneracy assumption (Assumption 4), states that $n \cdot \operatorname{Var}(\widehat{\tau}) \geq c$ for some constant $c > 0$. Putting these together we have that $n \cdot \mathrm{VB} \geq c$. Using this, we have that

$$\Pr\Big(\Big|1 - \frac{\widehat{\mathrm{VB}}}{\mathrm{VB}}\Big| > \epsilon\Big) \leq \frac{1}{\epsilon^2} \cdot \frac{1}{c^2} \cdot \operatorname{Var}(n \cdot \widehat{\mathrm{VB}}) \to 0 \ ,$$

where the last line follows from Proposition 6.6. □

S3.2 Asymptotic Validity of the Confidence Intervals

We are now ready to establish asymptotic validity of the normal-based confidence intervals. This is demonstrated by Corollary 6.7.

Corollary 6.7. *Given the conditions of Propositions 5.6 and 6.6 (asymptotic normality of the point estimator and consistency of the variance estimator), the Wald-type confidence intervals are asymptotically valid:*

$$\liminf_{n\to\infty} \Pr\big(\widehat{\tau} - R_\alpha \leq \tau \leq \widehat{\tau} + R_\alpha\big) \geq 1 - \alpha,$$

where $R_\alpha = \Phi^{-1}(1 - \alpha/2)\sqrt{\widehat{\mathrm{VB}}(\widehat{\tau})}$ *is the radius of the interval.*

Proof. Write the probability as

$$\Pr\big(\widehat{\tau} - R_\alpha \leq \tau \leq \widehat{\tau} + R_\alpha\big) = \Pr\big(S_n\Phi^{-1}(\alpha/2) \leq T_n U_n \leq S_n \Phi^{-1}(1 - \alpha/2)\big),$$

where

$$S_n = \sqrt{\frac{\mathrm{VB}(\widehat{\tau})}{\operatorname{Var}(\widehat{\tau})}}, \qquad \text{and} \qquad T_n = \frac{\widehat{\tau} - \tau}{\sqrt{\operatorname{Var}(\widehat{\tau})}}, \qquad \text{and} \qquad U_n = \sqrt{\frac{\mathrm{VB}(\widehat{\tau})}{\widehat{\mathrm{VB}}(\widehat{\tau})}},$$

and symmetry of the standard normal deviate was used for $\Phi^{-1}(\alpha/2) = -\Phi^{-1}(1-\alpha/2)$.

Note that Proposition 6.3 (validity of variance bound) ensures that $S_n \geq 1$ for all n. This extends the interval in the probability expression for $T_n U_n$, meaning that

$$\Pr\big(S_n \Phi^{-1}(\alpha/2) \leq T_n U_n \leq S_n \Phi^{-1}(1-\alpha/2)\big) \geq \Pr\big(\Phi^{-1}(\alpha/2) \leq T_n U_n \leq \Phi^{-1}(1-\alpha/2)\big).$$

Let $F_n : \mathbb{R} \to [0,1]$ be the cumulative distribution function of the random variable $T_n U_n$. We then have

$$\Pr\big(\Phi^{-1}(\alpha/2) \leq T_n U_n \leq \Phi^{-1}(1-\alpha/2)\big) = F_n\big(\Phi^{-1}(1-\alpha/2)\big) - F_n\big(\Phi^{-1}(\alpha/2)\big).$$

Lemma S3.1 ensures that U_n^2 goes to 1 in probability. By the continuous mapping theorem, this implies that U_n goes to 1 in probability. This implies, together with Proposition 5.6 (asymptotic normality) and Slutsky's theorem, that the limiting distribution of $T_n U_n$ is standard normal: $\lim_{n\to\infty}[F_n(x) - \Phi(x)] = 0$ pointwise for all $x \in \mathbb{R}$. As a result,

$$\lim_{n\to\infty} \Big[F_n\big(\Phi^{-1}(1-\alpha/2)\big) - F_n\big(\Phi^{-1}(\alpha/2)\big)\Big] = 1 - \alpha/2 - \alpha/2 = 1 - \alpha. \qquad \square$$

S4 Miscellaneous Results

S4.1 Plug-In Interpretation of the Riesz Estimator

One can interpret the Riesz representation estimator as being motivated by a type plug-in principle for the design-based inference framework. We believe this interpretation will be more familiar to some people.

The idea is that the experimenter constructs an estimator for the potential outcome function y_i for each unit, denoted $\widehat{y}_i$. Note that the whole function is estimated, meaning that the estimated potential outcome function $\widehat{y}_i(z)$ can be evaluated at any $z \in \mathcal{Z}$. By plugging in the estimated potential outcome function into the linear functional, we arrive at an estimate of each unit's treatment effect $\widehat{\tau}_i = \theta_i(\widehat{y}_i)$. The average of these individual treatment effects is an estimate of the overall treatment effect τ:

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \theta_i(\widehat{y}_i).$$

Any individual $\widehat{y}_i$ or $\widehat{\tau}_i$ will typical be too imprecise to be useful by themselves, even in large samples. Indeed, these individual estimators will generally not concentrate in any meaningful sense. Still, the aggregation will result in lower variance, provided that the individual estimators are sufficiently uncorrelated.

Formally speaking, we will estimate the equivalence class in the outcome model space

$\mathcal{L}(\mathcal{M}_i)$. Although the basic idea of the plug-in principle holds whenever $\mathcal{L}(\mathcal{M}_i)$ is separable, suppose for sake of presentation that each model space $\mathcal{M}_i$ is finite-dimensional so that there exists an orthonormal basis $\{\phi_{i,\ell}\}_{\ell=1}^{d}$ for each outcome model space $\mathcal{L}(\mathcal{M}_i)$.

Under the correctly specified assumption, the equivalence class of the true potential outcome function admits the unique decomposition $y_i(Z) = \sum_{\ell=1}^{d} \alpha_{i,\ell}\phi_{i,\ell}(Z)$. Thus, the problem of estimating the equivalence class containing y_i can be recast as the problem of estimating the coefficients $\{\alpha_{i,\ell}\}_{\ell=1}^{d}$. One choice of estimated coefficients is $\hat{\alpha}_{i,\ell} = y_i(Z)\phi_{i,\ell}(Z)$, which yields the estimated potential outcome function

$$\widehat{y}_i = \sum_{\ell=1}^{d} \hat{\alpha}_{i,\ell}\phi_{i,\ell} = \sum_{\ell=1}^{d} \big(y_i(Z)\phi_{i,\ell}(Z)\big)\phi_{i,\ell}$$

This will lead to an unbiased estimator of the equivalence class containing the potential outcome function.

Proposition S4.1. *Under Assumptions 1 and 2, the plug-in estimator for the potential outcome function described above is unbiased:* $\mathrm{E}[\widehat{y}_i] = y_i$.

Proof. By linearity of expectation, the expectation of the estimated potential outcome function can be decomposed in the following way:

$$\mathrm{E}[\widehat{y}_i] = \mathrm{E}\Big[\sum_{\ell=1}^{d} \hat{\alpha}_{i,\ell}\phi_{i,\ell}\Big] = \sum_{\ell=1}^{d} \mathrm{E}[\hat{\alpha}_{i,\ell}]\phi_{i,\ell} \quad .$$

Under the correctly specified assumption, the equivalence class containing the true observed potential outcome admits the decomposition $y_i(Z) = \sum_{\ell=1}^{d} \alpha_{i,\ell}\phi_{i,\ell}(Z)$. Thus, the expected value of the estimated coefficients may be calculated as

$$\mathrm{E}[\hat{\alpha}_{i,\ell}] = \mathrm{E}[y_i(Z)\phi_{i,\ell}(Z)] = \mathrm{E}\Big[\Big(\sum_{k=1}^{d} \alpha_{i,k}\phi_{i,k}(Z)\Big)\phi_{i,\ell}(Z)\Big] = \sum_{k=1}^{d} \alpha_{i,k}\,\mathrm{E}[\phi_{i,k}(Z)\phi_{i,\ell}(Z)] = \alpha_{i,\ell}.$$

This completes the proof, as

$$\mathrm{E}[\widehat{y}_i] = \mathrm{E}\Big[\sum_{\ell=1}^{d} \hat{\alpha}_{i,\ell}\phi_{i,\ell}\Big] = \sum_{\ell=1}^{d} \mathrm{E}[\hat{\alpha}_{i,\ell}]\phi_{i,\ell} = \sum_{\ell=1}^{d} \alpha_{i,\ell}\phi_{i,\ell} = y_i \quad . \qquad \square$$

Strictly speaking, the plug-in estimator has only unbiasedly estimated an equivalence class in the outcome model space $\mathcal{L}(\mathcal{M}_i)$. In order to make sense of the plug-in $\theta_i(\widehat{y}_i)$, we have to select a representative function $u \in \mathcal{M}_i$ from the estimated equivalence class $[\widehat{y}_i]$ and plug it into the effect functional θ_i. Under the positivity assumption (Assumption 2), this yields the same result no matter the representative function chosen. This completes the unbiased estimation method based on the plug-in principle.

The following proposition shows that this plug-in estimator is equivalent to the Riesz estimator, due to a simple changing of terms.

Proposition S4.2. *The Riesz estimator and the plug-in estimator described above coincide.*

Proof. It suffices to show that the individual effect estimators will coincide. Let us start with the individual effect estimator derived using the plug-in principle and let us show that it recovers the individual effect estimator derived using Riesz representation.

$$\begin{aligned}
\widehat{\tau}_i &= \theta_i(\widehat{y}_i) \\
&= \theta_i\Big(\sum_{\ell=1}^{d} \hat{\alpha}_{i,\ell}\phi_{i,\ell}\Big) \\
&= \sum_{\ell=1}^{d} \hat{\alpha}_{i,\ell}\theta_i(\phi_{i,\ell}) \\
&= \sum_{\ell=1}^{d} y_i(Z)\phi_{i,\ell}(Z)\theta_i(\phi_{i,\ell}) \\
&= y_i(Z)\Big(\sum_{\ell=1}^{d} \theta_i(\phi_{i,\ell})\phi_{i,\ell}(Z)\Big) \\
&= y_i(Z)\psi_i(Z) \ .
\end{aligned}$$

□

S4.2 Estimands from Coefficients

In this section, we discuss how coefficients of a basis representation can be used to define estimands. For simplicity, we focus on finite-dimensional model spaces. Let $i \in [n]$ be a unit with model space $\mathcal{M}_i$. Suppose that $\mathcal{M}_i$ admits a basis $\{g_{i,\ell}\}_{\ell=1}^{d}$. By definition of a basis, every function $u \in \mathcal{M}_i$ admits a unique decomposition in terms of the basis, i.e. $u = \sum_{\ell=1}^{d} \alpha_\ell g_{i,\ell}$.

For a fixed coefficient $k \in [d]$, The mapping $\theta(u) = \alpha_k$ is a well-defined functional because the coefficient α_k is uniquely defined given the basis. Additionally, one can verify that this functional is linear. In particular, let $u, v \in \mathcal{M}_i$ have the basis expansions $u = \sum_{\ell=1}^{d} \alpha_\ell g_{i,\ell}$ and $v = \sum_{\ell=1}^{d} \beta_\ell g_{i,\ell}$. Note that $\theta_i(u) = \alpha_k$ and $\theta_i(v) = \beta_k$. Then, the sum $u+v$ as the basis expansion

$$u + v = \sum_{\ell=1}^{d} \alpha_\ell g_{i,\ell} + \sum_{\ell=1}^{d} \beta_\ell g_{i,\ell} = \sum_{\ell=1}^{d} (\alpha_\ell + \beta_\ell) g_{i,\ell}$$

So that $\theta_i(u+v) = \alpha_k+\beta_k = \theta_i(u)+\theta_i(v)$, which establishes that the functional θ_i is additive. A similar argument shows that the mapping is homogeneous, i.e. $\theta_i(c \cdot u) = c \cdot \theta_i(u)$.

Thus, functionals of the form $\theta_i(u) = \alpha_{i,\ell}$ are linear functionals on $\mathcal{M}_i$. Additionally, any linear combination of these functionals are linear functionals on $\mathcal{M}_i$. That means that $\theta(u) = \sum_{\ell=1}^{d} c_\ell \alpha_\ell$ is a linear functional on $\mathcal{M}_i$ for any real coefficients $\{c_\ell\}_{\ell=1}^{d}$.